Estimation of the Squared Cross-Validity Coefficient in the Context of Best Subset Regression

Eugene Kennedy
South Carolina Department of Education

A monte carlo study was conducted to examine the performance of several strategies for estimating the squared cross-validity coefficient of a sample regression equation in the context of best subset regression. Data were simulated for populations and experimental designs likely to be encountered in practice. The results indicated that a formula presented by Stein (1960) could be expected to yield estimates as good as or better than cross-validation, or several other formula estimators, for the populations considered. Further, the results suggest that sample size may play a much greater role in validity estimation in subset selection than is true in situations where selection has not occurred. *Index terms: Best subset regression, Cross-validity coefficient, Multiple regression, Predictive validity, Variable selection.*

The predictive power of a sample regression equation in the population and in future samples is often of primary importance to researchers. A measure widely used for this purpose is the squared cross-validity coefficient, R_c^2. This index is defined as the squared correlation of actual criterion values with those predicted from the sample equation for the population of interest. A natural choice as an estimator of this parameter is the sample squared multiple correlation, EST(R^2). This statistic, however, is known to exaggerate the predictive power of the sample equation and is widely deemed inappropriate in its original form (see Herzberg, 1969). This fact has given rise to a large body of empirical and theoretical work concerning the relative performance of various correction or shrinkage procedures (Cattin, 1980; Schmitt, Coyle, & Rauschenberger, 1977). Currently, the consensus among investigators appears to be that a formula proposed by Browne (1975) is preferable to other strategies for a variety of experimental situations (see Drasgow, Dorans, & Tucker, 1979).

Although in many respects the literature on estimation of R_c^2 is comprehensive, it is almost always assumed in these studies that the analyst has predetermined which predictors will constitute the model. Typically, however, this is not the case. In the majority of practical situations, the analyst will be confronted with a large array of potential predictors and few substantive guidelines as to which to include in the model. In this context, an algorithm for empirically identifying a ''best'' subset of predictors offers a solution. Among the many possibilities are stepwise regression, backward elimination, and all possible regressions (see Pedhazur, 1982).

These procedures afford a degree of convenience and empirical support for a model, but they also tend to introduce certain complications. In particular, once a model has been formulated using a best subset selection algorithm, the usual procedures for drawing inferences in regression analysis are not strictly applicable (for a thorough review see Hockings, 1976). Rencher and Pun (1980), for

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 12, No. 3, September 1988, pp. 231–237
© Copyright 1988 Applied Psychological Measurement Inc.
0146-6216/88/030231-07$1.60

example, reported that the expected value of EST(R^2) under subset selection can be twice that without selection. Lerner and Games (1981), working with social science datasets, have similarly noted substantial inflation of validity estimates after empirical model selection. Pope and Webster (1972) considered an order statistic approach to the ordered set of dependent F values in stepwise regression, on the grounds that fewer unrealistic assumptions are required than in the usual procedure. Hockings (1976) proposed that biased estimators may offer improvements over the usual least-squares estimator when selection has occurred. Finally, Diehr and Hoflin (1974) obtained approximate percentile points to the unknown distribution of the sample squared multiple correlation in best subset regression.

These results suggest that the subset regression strategy may warrant a unique set of guidelines for estimating the squared cross-validity coefficient of a sample regression equation. Unfortunately, there is little research on this issue. Most authors suggest splitting sample data, then using one portion for identification of the model and the other portion (not necessarily of equal size) for estimation of parameters. But cross-validation, as this is called, is known to have significant restrictions. In particular, a significant loss of information can be expected when all available data are not used for purposes of parameter estimation. When sample size is large (i.e., several thousand people), this loss is most likely minor. But for moderate size datasets, which typify many social applications, splitting data can yield seriously unstable parameter estimates (see Picard & Cook, 1984).

The present study compared several formula estimators of R^2_c. An extensive search of the statistical and psychological literature yielded several alternative estimation strategies for both the regular and the subset regression procedures. A first objective of the study was thus to determine which combinations of procedures and sets of experimental conditions would yield the best estimates (previous recommendations have provided few experimental or theoretical foundations). A second objective was to investigate the impact of different experimental designs on validity shrinkage when the sample equation is applied in the population.

A persistent problem in predictive validity studies is confusion concerning appropriate formula estimators of R^2_c. Researchers commonly apply the Wherry (1931) shrinkage formula in this context, but in fact this formula was designed to estimate the validity of a regression model, not the predictive validity of a sample regression equation. To illustrate the consequences of this error, the Wherry formula has been included in the simulation results. (Wherry, 1975, presented an excellent discussion of this problem.)

Method

One of the primary guidelines of the current study was to keep the experiments as relevant as possible for the applied social researcher. The first step in achieving this was to select a model frequently studied in the social psychology of education. The model posits that the occupation a student expects to achieve is a function of the student's (1) self-concept of academic ability, (2) curriculum enrollment, (3) academic performance, (4) achievement orientation, (5) academic aptitude, (6) peer influence, (7) teacher influence, (8) parent influence, (9) social class background, and (10) an error component.

The next step was to obtain a nationally representative sample of high school students. The source for this was the National Longitudinal Study of the High School Class of 1972 conducted by the Research Triangle Institute. The model was then estimated on black and white males. Unfortunately, missing data for these groups necessitated deleting several thousand students from the analysis. Because the size of the resulting sample would have presented problems for sampling procedures, it was decided to generate a population of hypothetical individuals from the covariance matrices of the sample data. The populations were generated with an algorithm based on the IMSL routine GGNML.

The populations were each composed of 2,000 simulated persons. The squared multiple correlations with all 9 predictors covered a span which encompasses many models in the social sciences. For black males R^2 was .12 and for white males R^2 was .20. The intercorrelations and standard deviations for black males (Population 1) and white males (Population 2) are presented in Table 1.

Previous research has shown that sample size and the ratio of predictors selected to the total in the set will affect validity estimation in the subset context (see Rencher & Pun, 1980). Again, to cover a broad span of experimental situations, sample size was set to 30, 70, and 150. To vary the number of predictors selected to the total in the set, the best 7, 6, and 5 predictors were selected from among the 9 possible. The sample size and predictor ratio factors generated 3 × 3 = 9 sampling conditions for each of the two populations, for a total of 18 experimental conditions.

The simulation strategy was as follows: For each cell of the design, 100 random samples (with replacement) were drawn. For each sample, the best K ($K = 7, 6, 5$) predictors were identified and a sample regression equation was computed. Then various strategies for estimating validity were applied, and these estimates were compared to the validity obtained when the sample equation was applied across the entire population. The accuracy of these procedures was assessed by the root mean square error of estimation (RMSE):

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

Validity Estimators

The estimators of the squared cross-validity coefficient considered in this study are listed in Table 2. The Browne (1975), Claudy (1978), Lord (1950), Stein (1960), and Rozeboom (1978) formulas and double cross-validation are all estimators of R_c^2 for unselected models. The Wherry formula is not an estimator of R_c^2 but is included here, as noted earlier, because it is frequently misused in this context. The Cohen and Cohen (1975) formula was proposed specifically for estimating R_c^2 in the subset case. However, little empirical or theoretical support was provided.

Subset Selection

These analyses employed backward elimination as the technique for identifying the best K predictors in sample data (see Klienbaum & Kuper, 1978). This procedure first estimates a regression equation with all possible predictors. The predictor with the smallest beta weight is then eliminated and the equation is reestimated with one fewer predictor. This continues until a prescribed number of predictors remains.

Results

Subset Selection and Validity Shrinkage

Because so much has been written on the need to consider sample size and number of predictors when validity is at issue in a regression problem, mean error of estimation was aggregated over all estimators of R_c^2, over all experimental situations for sample size and number of predictors. These results are presented in Table 3. As is obvious from the margins, when the number of predictors changes only small changes occur in mean error. For sample size, on the other hand, the increase from 30 to 150 observations brings a significant drop in mean error.

Table 4 presents the mean error of estimation for the sample validity estimate, EST(R^2), as well as the eight estimators considered in this study. The results for EST(R^2) provide some insight into factors affecting validity shrinkage when selection has occurred. First, these values are always positive, a pattern which reflects the known tendency of EST(R^2) to yield overly optimistic validity estimates. Second, as sample size increases and the number of predictors decreases, there is a reduction in the mean difference between EST(R^2) and R_c^2.

Upon closer examination of these patterns, it appears that the number of predictors in the model is not nearly as potent a factor as sample size. At 30 observations the average error for any number

Table 1
Zero-Order Correlations and Standard Deviations for
Population 1 (Above the Diagonal) and Population 2 (Below the Diagonal)

	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	SD
y	---	.280	.153	.019	.179	−.036	−.093	−.124	.154	.226	62.750
x_1	.301	---	.354	.310	.462	−.078	−.184	−.109	.318	.277	.619
x_2	.180	.222	---	.193	.289	.096	.021	−.080	.309	.277	3.694
x_3	.212	.502	.249	---	.319	.023	.038	.003	.324	.286	1.313
x_4	.293	.522	.267	.474	---	−.008	−.029	−.113	.285	.390	.479
x_5	.143	.107	.079	.171	.149	---	.244	.264	.043	−.003	.652
x_6	.000	.050	.073	.123	.063	.085	---	.299	.106	.129	.732
x_7	.006	−.014	−.032	−.067	−.072	.098	.246	---	−.089	−.085	.721
x_8	.245	.471	.290	.472	.443	.105	.114	−.017	---	.295	.860
x_9	.404	.482	.213	.510	.588	.230	.138	−.056	.571	---	1.311
SD	69.234	.673	3.202	1.466	.484	.621	.658	.659	.966	1.290	---

Table 2
Estimators of The Squared Cross-Validity Coefficient

Author	Estimator
Wherry (1931)	$\text{Est} = 1 - (N-1)\dfrac{(1-R^2)}{(N-p-1)}$
Browne (1975)	$\text{Est} = \dfrac{(N-p-3)R^4+w}{(N-2p-2)w+p}$
	Where $R^4 = (w)^2 - \dfrac{2p(1-w)^2}{(N-1)(N-p+1)}$
	$w =$ Wherry Validity Estimate
Claudy (1978)	$\text{Est} = 1 - \dfrac{(N-1)(N-2)(N-1)(1-R^2)}{(N-p-1)(N-p-2)N}$
Lord (1950)	$\text{Est} = 1 - \dfrac{(N-1)(N+p+1)(1-R^2)}{(N-p-1)N}$
Stein (1960)	$\text{Est} = 1 - \dfrac{(N-1)(N-2)(N+1)(1-R^2)}{(N-p-1)(N-p-2)N}$
Rozeboom (1978)	$\text{Est} = 1 - \dfrac{(N+p)(1-R^2)}{(N-p)}$
Cohen & Cohen (1975)	$\text{Est} = 1 - \dfrac{(N-1)(1-R^2)}{(N-m-1)}$
	Where $m =$ total predictors in set
Moiser (1951)	Est = Double Cross-Validation

Note. Claudy attributes his formula to Darlington (1968), but in fact the Darlington formula is an algebraic rearrangement of the Stein formula presented above.

Table 3
Aggregated Mean Error of Estimation
Based on Samples From Population 2

Number of Predictors	Sample Size 30	70	150	Mean
5	.0937	.0202	.0152	.0430
6	.0578	.0240	.0181	.0333
7	.0776	.0231	.0105	.0370
Mean	.0764	.0224	.0146	

of predictors examined is in the .30 range. As the sample size increases to 70 and then 150 the error drops, first to around .15 and then to .07. This suggests that the researcher should seek to obtain as many observations as possible. This result has not been reported in previous investigations and it highlights a significant difference between best subset and regular regression: Although validity estimates for the latter are highly sensitive to the number of predictors in the model, the validity estimates for the former are less sensitive.

Bias In Estimators of R_c^2

The mean error for the eight estimators of R_c^2 in Table 1 can serve as a basis for comparing the various strategies. First, in almost every experiment the Wherry formula gave the most biased estimate. This is cause for concern for the user of "canned" statistics packages which include this formula as part of their subset routines. Second, the prominence of the Browne (1975) formula is lost in these data. While in some instances its estimates were within a hundredth of a point, it was never as accurate as the Stein (1960) formula. Indeed, in 12 of the 18 experiments conducted, the Stein formula was superior to all others. Further, when it failed to yield the smallest error the differences were usually less than .001. The formula appears to be especially potent when sample size is small. Again, this result points to the need for special considerations in the context of subset selection.

Another means of examining the performance of various strategies for estimating R_c^2 is by examining the root mean squared error of estimation. Table 5 presents these values for each of the eight procedures discussed in this study. Unlike the results in Table 4, these values do not show any clear pattern. The Stein (1960) formula again yielded superior estimates, but there does not appear to be

Table 4
Mean Differences Between Estimated and True Squared
Cross-Validity Coefficients (Sample Estimate Minus Population Value)

N	P	EST(R^2)	Wherry	Browne	Claudy	Lord	Stein	Rozeboom	Cohen	Double Cross-Validation
Population 1 (R^2=.12)										
30	7	.3290	.1388	.0469	.0131	.0177	−.0026	.0297	.0785	.0685
	6	.3005	.1391	.0600	.0300	.0289	.0043	.0446	.0596	.0768
	5	.3032	.1716	.0962	.0711	.0606	.0383	.0801	.0606	.0922
70	7	.1725	.0872	.0319	.0122	.0020	−.0081	.0126	.0603	.0188
	6	.1550	.0803	.0318	.0173	.0051	−.0017	.0152	.0392	.0246
	5	.1523	.0913	.0488	.0392	.0236	.0176	.0345	.0372	.0385
150	7	.0810	.0401	.0105	.0031	−.0053	−.0078	.0001	.0277	.0105
	6	.0743	.0394	.0137	.0089	−.0005	−.0026	.0051	.0209	.0031
	5	.0892	.0607	.0393	.0369	.0268	.0252	.0326	.0365	.0229
Population 2 (R^2=.20)										
30	7	.3836	.2119	.0993	.0528	.0613	.0245	.0805	.1462	.0786
	6	.3294	.1769	.0833	.0488	.0475	.0173	.0668	.0858	.0555
	5	.3262	.2034	.1217	.0989	.0878	.0611	.1077	.0878	.0908
70	7	.1745	.0952	.0368	.0193	.0084	−.0025	.0198	.0692	.0105
	6	.1503	.0821	.0344	.0248	.0122	.0040	.0227	.0456	.0246
	5	.1325	.0755	.0337	.0256	.0096	.0033	.0209	.0236	.0251
150	7	.0835	.0459	.0164	.0118	.0035	.0009	.0089	.0345	−.0024
	6	.0811	.0493	.0242	.0214	.0124	.0104	.0179	.0323	.0080
	5	.0702	.0433	.0226	.0209	.0113	.0098	.0168	.0205	.0043

Table 5
Root Mean Squared Errors of Estimators of the Squared Cross-Validity Coefficient

N	P	EST(R^2)	Wherry	Browne	Claudy	Lord	Stein	Rozeboom	Cohen	Double Cross-Validation
Population 1 (R^2=.12)										
30	7	.3511	.2049	.1337	.1174	.1221	.1013	.1323	.1649	.1246
	6	.3260	.1999	.1287	.1096	.1087	.0905	.1216	.1350	.1367
	5	.3265	.2201	.1521	.1368	.1280	.1073	.1441	.1280	.1429
70	7	.1830	.1099	.0692	.0650	.0624	.0611	.0651	.0902	.0697
	6	.1701	.1107	.0793	.0770	.0733	.0708	.0764	.0866	.0745
	5	.1727	.1264	.1007	.0987	.0925	.0905	.0967	.0978	.0950
150	7	.0965	.0680	.0552	.0572	.0567	.0566	.0570	.0622	.0607
	6	.0910	.0673	.0561	.0571	.0565	.0565	.0567	.0594	.0561
	5	.1051	.0835	.0700	.0694	.0652	.0646	.0675	.0692	.0631
Population 2 (R^2=.20)										
30	7	.4028	.2626	.1754	.1572	.1627	.1365	.1756	.2169	.1525
	6	.3554	.2420	.1760	.1661	.1654	.1492	.1760	.1870	.1252
	5	.3535	.2592	.1988	.1882	.1807	.1646	.1946	.1807	.1375
70	7	.1913	.1283	.0934	.0936	.0914	.0899	.0936	.1124	.0967
	6	.1754	.1284	.1012	.0999	.0966	.0952	.0992	.1089	.0969
	5	.1646	.1293	.1121	.1136	.1118	.1113	.1130	.1133	.1081
150	7	.1009	.0752	.0622	.0631	.0623	.0623	.0627	.0695	.0708
	6	.1028	.0821	.0710	.0712	.0698	.0696	.0705	.0744	.0703
	5	.0900	.0726	.0633	.0634	.0616	.0614	.0624	.0633	.0641

a clear relationship to sample size or number of predictors in the model. The results for Populations 1 and 2 do not differ systematically.

Discussion

In the majority of the experiments considered above, the Stein (1960) formula yielded superior estimates of R_c^2. The data reported in Table 4 indicate that as sample size increases cross-validation becomes an acceptable estimator, but in general, the Stein formula could be expected to perform as well as or better than cross-validation, and to outperform the other formula estimators considered. This result has not been noted in previous studies.

Further, the present results indicate that for models generated in this manner for populations similar to the ones used in this study, sample size is a primary factor in shrinkage. The number of predictors in the model appears to have a much less prominent role here than is reported in studies where subset selection has not occurred (Drasgow et al., 1979). This result explains in part the superiority of the Stein (1960) formula. Of all shrinkage formulas considered, it is the most sensitive to sample size and among the least sensitive to changes in the number of predictors in the model.

Conclusions

Simulation is a useful research tool for dealing with difficult theoretical problems. But its use entails finding a balance between the limitless number of parameters that could be manipulated, and that group of manipulations relevant to practical applications. The former often leads to theoretical insight whereas the latter, though more limited, can suggest practical guidelines. In this project, obviously, the latter course was selected. The total number of predictors in the pool was not manipulated; the population multiple correlations were both low and did not differ by .1; the number of predictors selected did not cover a broad range, and only one subset selection strategy was considered. The reader is advised to bear these limitations in mind when considering the relevance of the present results to a particular application. Future efforts along these lines could provide useful insights.

References

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

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

Claudy, J. G. (1978). Multiple regression and validity estimation in one sample. *Applied Psychological Measurement*, *2*, 595–607.

Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale NJ: Erlbaum.

Darlington, R. B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin*, *69*, 161–182.

Diehr, G., & Hoflin, D. R. (1974). Approximating the distribution of the sample R^2 in best subset regressions. *Technometrics*, *16*, 317–320.

Drasgow, F., Dorans, N. J., & Tucker, L. R. (1979). Estimators of the squared cross-validity coefficient: A monte carlo investigation. *Applied Psychological Measurement*, *3*, 387–399.

Herzberg, P. (1969). The parameters of cross-validation. *Psychometrika Monograph*, *34* (2, Pt. 2, No. 16).

Hockings, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics*, *32*, 1–49.

Klienbaum, D. G., & Kuper, L. L. (1978). *Applied regression analysis and other multivariable methods*. North Scituate MA: Duxbury Press.

Lerner, J. V., & Games, P. A. (1981). Maximum R^2 improvement and stepwise multiple regression as related to over-fitting. *Psychological Reports*, *48*, 979–983.

Lord, F. M. (1950). *Efficiency of prediction when a regression equation from one sample is used in a new sample* (Research Bulletin No. 50-40). Princeton NJ: Educational Testing Service.

Moiser, C. I. (1951). Problems and designs of cross-validation. *Educational and Psychological Measurement*, *11*, 5–11.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and prediction*. New York: CBS College Publishing.

Picard, R. R., & Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association*, *79*, 575–583.

Pope, P. T., & Webster, J. T. (1972). The use of an *F*-statistic in stepwise regression procedures. *Technometrics*, *14*, 327–340.

Rencher, A. C., & Pun, F. C. (1980). Inflation of R^2 in best subset regression. *Technometrics*, *22*, 49–53.

Rozeboom, W. W. (1978). Estimation of cross-validated multiple correlations: A clarification. *Psychological Bulletin*, *85*, 1348–1351.

Schmitt, N., Coyle, B. W., & Rauschenberger, J. (1977). A Monte Carlo evaluation of three formula estimators of cross-validated multiple correlation. *Psychological Bulletin*, *84*, 751–758.

Stein, C. (1960). Multiple regression. In I. Olkin et al. (Eds.), *Contributions to probability and statistics*. Stanford CA: Stanford University Press.

Wherry, R. J. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. *Annals of Mathematical Statistics*, *2*, 440–457.

Wherry, R. J. (1975). Underprediction from overfitting: 45 years of shrinkage. *Personnel Psychology*, *28*, 1–18.

Author's Address

Send requests for reprints or further information to Eugene Kennedy, Office of Research, Department of Education, 1429 Senate Street, Columbia SC 29201, U.S.A.