Estimation Methods for Cross-Validation Prediction Accuracy: A Comparison of Proportional Bias

David A. Walker
Northern Illinois University

Using empirical data, the performance of the predictive effectiveness of four algorithms and a bootstrap method for cross-validation of a multiple regression equation were examined. Results indicated that the Browne algorithm was the most accurate in 8 of the 9 data situations. The Rozeboom algorithm, in a majority of conditions, had the second least amount of proportional bias. The Nicholson and Lord and Stein-Darlington formulas demonstrated a consistent pattern of low relative accuracy in many situations, with the most amount of proportional bias in 4 of 9 and in 6 of 9 data sets, respectively. The bootstrap method showed no discernable pattern of relative accuracy with results ranging from the most accurate in a situation to the least accurate in three different data situations.

For prediction studies derived via multiple regression that estimate how well a sample equation generalizes to other samples, or as Rozeboom (1978, p. 1350) called this "a sample regression's generalized validity," the use of empirical methods such as cross-validation or double cross-validation have been studied in the past (Lord & Novick, 1968; Mosier, 1951). However, there are cautions with these techniques. For example, when applied to small sample sizes, where the sample is split into two sub-samples, this can lead to less precise estimates of prediction (Browne, 1975; Cattin, 1980a; Cotter & Raju, 1982). To resolve this major limitation, the use of estimation algorithms has been noted in the scholarly literature, where the sample size is not split, but left whole, exacting a more accurate estimate of the criterion score estimation for cross-validation of a multiple regression equation (Allen, 1971; Claudy, 1978; Cotter & Raju, 1982; Gollob, 1967; Huberty & Mourad, 1980; Morris, 1984; 1986).

From a review of the literature in the area of proposed cross-validation techniques, the vast majority of the research conducted on the predictive effectiveness of multiple regression equations derived via an algorithm has been conducted as Monte Carlo studies. That is, very few studies reported results from cross-validation algorithms when empirical data were used. Of the small number of studies that did implement empirical data, most used data from large samples derived from business, government, or educational institutions (cf. Cotter & Raju, 1982; Huberty & Mourad, 1980; Kromrey & Hines, 1995). A few exceptions found in the literature that applied empirical data sets from smaller samples were Krus and Fuller (1982) who used cross-validation algorithms with data from a textbook, and Morris (1986) who employed Allen's (1971) PRESS (Predicted Error Sum of Square) technique with data from journal articles, textbooks, and professional conference papers.

Various formulas have been proposed for use as estimation algorithms for cross-validating a regression equation. From the literature, four formulas emerge as viable estimators. Many of the subsequent formulas are decidedly related algebraically and/or are hybrids of one another. Formulas 1, 2, and 4 are found in Huberty & Mourad (1980) and formula 3 is from Cattin (1980b).

Nicholson (1960) and Lord (1950) proposed R^2_{NL}, where:

$$R^2_{NL} = 1 - \left[\frac{N+p+1}{N-p-1}\right]\left[\frac{N-1}{N}\right]\left[1-R^2\right] \quad (1)$$

Stein (1960) and Darlington (1968) proposed R^2_{SD}, where:

$$R^2_{SD} = 1 - \left[\frac{N-1}{N-p-1}\right]\left[\frac{N-2}{N-p-2}\right]\left[\frac{N+1}{N}\right]\left[1-R^2\right] \quad (2)$$

Browne (1975) proposed R^2_B, rearranged by Cattin (1980b), where:

$$R^2_B = \frac{[(N-p-3)(R^2)^2]+R^2}{[(N-2p-2)R^2]+p} \quad (3)$$

Rozeboom (1978) restructured Browne's (1975) algorithm and proposed a simpler version, R^2_R, where:

$$R^2_R = 1-\left[\frac{N+p}{N-p}\right]\left[1-R^2\right] \quad (4)$$

where, N = Sample size, p = Number of X variables, and R^2 = Squared multiple correlation coefficient.

Finally, in prediction studies derived via multiple regression that have the intention of closely approximating a sample prediction equation, the bootstrap method, as well as the leave-one-out method and the jackknife method, have been found to be similar to cross-validation techniques (Efron, 1983; Gong, 2003; Huberty & Mourad, 1980; Kromrey & Hines, 1995; Lachenbruch, 1967). The bootstrap is a resampling method where the sampling properties of a statistic, in this instance R^2, are derived by recomputing its value for artificial samples. Thus, the sample data from this study will serve as pseudo-populations and 1,000 random samples with replacement will be drawn from these full samples. One thousand iterations will be used as an established threshold where all of the nine empirical data sets will have had convergence. Once the bootstrap method is repeated 1,000 times on each empirical data set, a distribution of bootstrapped estimates for R^2 will emerge, where the mean value (i.e., R^2_{BOOT}) of each bootstrapped distribution is the estimate for R^2.

Purpose

Using empirical data, the intention of the current research is to determine the stability of predictive effectiveness of the criterion score estimation of four algorithms and the bootstrap method for cross-validation of a multiple regression equation. The stability of predicative effectiveness is defined as the performance of the techniques in terms of relative accuracy as determined from bias and proportional bias (cf. Aaron, Kromrey, & Ferron, 1998; Morris, 1986). As bias multiples, the distance between the R^2_{CV} value and the R^2 value increases, which leads to diminished stability and a lesser proportion of the criterion score variance accounted for in the predicted Y value than in the original sample's Y value. The measures of bias are:

$$\text{Bias} = R^2 - R^2_{CV} \quad (5)$$

$$\text{Proportional Bias} = \text{Bias} / R^2 = 1 - (R^2_{CV} / R^2) \quad (6)$$

where, R^2_{CV} is defined by either R^2_{NL}, R^2_{SD}, R^2_R, R^2_B, or R^2_{BOOT}.

Thus, this study will compare the performance of four algorithms and the bootstrap method in a three-tier situation: (1) in the first set of empirical data, each will contain two regressor variates (p=2), variable sample sizes (N) = 12, 20, 30, and variable R^2 values = .799, .255, .358; (2) in the second set of data, each will have p =3, N = 20, 30, 93, and R^2 = .943, .617, .261; (3) in the third set of data, each will have p = 4, N = 13, 30, 36, and R^2 = .982, .640, .355.

Methods

The data sets for this research came from Agresti and Finlay (1986), Cohen and Cohen (1983), Hald (1965), Kerlinger and Pedhazur (1973), Rulon, Tiedeman, Tatsuoka, and Langmuir (1967), Sprinthall (2000), and Thurstone (1947). These data are well-known and found in textbooks utilized in graduate-level research design and statistics courses. Also, they exemplify the types of data often applied in social science research, with varying distributional characteristics, multicollinearity, sample sizes, regressor variates, and criterion variables such as predicting grade point average, psychiatric impairment, or job success.

The previously listed N, p, and R^2 values from the data sets will be entered into the four formulas for R^2_{NL}, R^2_{SD}, R^2_R, and R^2_B, which are part of a program written in SPSS (Statistical Package for the Social

Sciences) v. 14.0 by the author (see Appendix A). The data sets also will be bootstrapped in the program AMOS v. 5.0 (Analysis of Moment Structures) for R^2_{BOOT}.

Results and Discussion

Table 1 shows that in terms of the relative accuracy of prediction, overall, the Browne algorithm, R^2_B, was superior in every data situation, except one. The R^2_B showed a distinct pattern of low bias and high stability and appeared to have the most relative accuracy for predictive effectiveness of criterion score estimation. Furthermore, following the standard set by Kromrey and Hines (1995), estimates within .01 of the sample R^2 value can be thought of as statistically unbiased, which occurred in two situations with R^2_B (i.e., the Thurstone and Hald data sets).

For the other three algorithm estimation techniques, none were noticeably superior in all nine of the data sets. That is, based on empirical data with varying sample sizes, regressor variates, and R^2 values, no generalizable rules can be constructed concerning which of the remaining three algorithm-based cross-validation methods were the "best" to use in a particular condition. However, patterns from the results do emerge to allow for some suggestions. For example, in the majority of data sets (i.e., 6 of 9), the Rozeboom algorithm (R^2_R) had the second least amount of proportional bias of the remaining formulas. In the other three data situations, this formula's prediction accuracy was the next most precise. Though research by Huberty and Mourad (1980) and Cotter and Raju (1982) studied the same three cross-validation formulas (e.g., R^2_{NL}; R^2_{SD}; R^2_R) in different ways, and came to some differing conclusions pertaining to predictive accuracy, they both concluded that $R_R{}^2$ was a precise estimator in most empirical data circumstances. Another apparent pattern from the current study's results was that after R^2_B and R^2_R, the Nicholson and Lord (R^2_{NL}) and the Stein-Darlington (R^2_{SD}) formulas demonstrated consistent patterns of low relative accuracy in many situations, with the most amount of proportional bias in 4 of 9 and in 6 of 9 data sets, respectively.

For the bootstrap method, R^2_{BOOT} showed no discernable pattern of relative accuracy with results ranging from the most accurate in a situation to the least accurate in three different data situations. Of interest is that the R^2_{BOOT} method had the least amount of proportional bias, in fact it was less than 0.01, when used with the Hald data set, which was the only data set with multicollinearity (e.g., variance inflation factor > 38 and tolerance < 0.03). This situation was checked with a data set independent from the others used in this study, which also manifested multicollinearity (e.g., variance inflation factor > 30 and tolerance < .02). In the scholarly literature, results from a study conducted by Ayabe (1985) using a technique similar to the bootstrap method, the jackknife procedure, found inferior estimates as well. Kromrey and Hines (1995) found mixed results, similar to the current study's findings, with use of the bootstrap method with small sample sizes. However, when sample sizes were $N \geq 100$, they found more unbiased estimates when using the bootstrap.

Conclusion

Given the very unique characteristics of each data set in this study in the areas of dissimilar N, p, and R^2 values, the R^2_B algorithm was the most accurate in 8 of the 9 data situations. None of the remaining three proposed cross-validation algorithms, or the bootstrap method, were exceedingly superior or inferior to each other when compared based on proportional bias. Although it may be convenient to run all four cross-validation methods from the program in Appendix A to determine which one has the least amount of bias given a specific data situation, the definite preference is toward R^2_B in nearly every

Table 1. Bias Affiliated with Cross-Validation Estimation Methods

Data Set	R^2	Biases	R^2_{NL}	R^2_{SD}	R^2_R	R^2_B	R^2_{BOOT}
Kerlinger (1973) $N = 12$ $p = 2$ DV = Attitude Score	0.799	R^2_{CV} Bias Proportional Bias	0.693 0.106 **0.133**	0.667 0.132 **0.165**	0.719 0.080 **0.100**	0.775 0.024 **0.030**	0.675 0.124 **0.155**
Sprinthall (2000) $N = 20$ $p = 2$ DV = Anger Score	0.255	R^2_{CV} Bias Proportional Bias	0.042 0.213 **0.835**	0.016 0.239 **0.937**	0.089 0.166 **0.651**	0.221 0.034 **0.133**	0.073 0.182 **0.714**
Agresti (1986) $N = 30$ $p = 2$ DV = Psychiatric Impairment	0.358	R^2_{CV} Bias Proportional Bias	0.241 0.117 **0.327**	0.233 0.125 **0.349**	0.266 0.092 **0.257**	0.336 0.022 **0.061**	0.215 0.143 **0.399**
Thurstone (1947) $N = 20$ $p = 3$ DV = Volume of a Box	0.943	R^2_{CV} Bias Proportional Bias	0.919 0.024 **0.025**	0.915 0.028 **0.03**	0.923 0.02 **0.021**	0.935 0.008 **0.008**	0.929 0.014 **0.015**
Kerlinger (1973) $N = 30$ $p = 3$ DV = GPA	0.617	R^2_{CV} Bias Proportional Bias	0.516 0.101 **0.164**	0.506 0.111 **0.18**	0.532 0.085 **0.138**	0.588 0.029 **0.047**	0.478 0.139 **0.225**
Rulon (1967) $N = 93$ $p = 3$ DV = Success Score	0.261	R^2_{CV} Bias Proportional Bias	0.203 0.058 **0.222**	0.202 0.059 **0.226**	0.212 0.049 **0.188**	0.246 0.015 **0.057**	0.167 0.094 **0.360**
Hald (1965) $N = 13$ $p = 4$ DV = Heat Evolved (Cement)	0.982	R^2_{CV} Bias Proportional Bias	0.963 0.019 **0.019**	0.954 0.028 **0.029**	0.966 0.016 **0.016**	0.974 0.008 **0.008**	0.975 0.007 **0.007**
Kerlinger (1973) $N = 30$ $p = 4$ DV = GPA	0.640	R^2_{CV} Bias Proportional Bias	0.513 0.127 **0.198**	0.497 0.143 **0.223**	0.529 0.111 **0.173**	0.599 0.041 **0.064**	0.508 0.132 **0.206**
Cohen (1983) $N = 36$ $p = 4$ DV = Religious Attitude	0.355	R^2_{CV} Bias Proportional Bias	0.171 0.184 **0.518**	0.152 0.203 **0.572**	0.194 0.161 **0.454**	0.303 0.052 **0.146**	0.222 0.133 **0.375**

empirical data cross-validation circumstance. Thus, it is probably prudent to apply R^2_B first while regarding the proportional bias derived from R^2_R as a comparison. The remaining two algorithms, R^2_{NL} and R^2_{SD}, did not perform well in virtually any data situation. Though use of the AMOS bootstrap technique it not difficult (cf. Fan, 2003 for application instructions), the mixed results derived from R^2_{BOOT} should afford caution when used with small sample sizes (i.e., $N < 100$), except in situations of multicollinearity where the R^2_{BOOT} method showed the least amount of proportional bias.

Appendix A. Cross-Validation Algorithms Program

```
**************************************************************************
Copyright David A. Walker, 2006
Contact dawalker@niu.edu
Northern Illinois University, 101J Gabel, DeKalb, IL 60115
        **APA 5th Edition Citation**
Walker, D. A. (2006). Four estimators for sample cross-validation [Computer
    program]. DeKalb, IL: Author.
**************************************************************************.


**************************************************************************
NOTE: Between BEGIN DATA and END DATA, insert the multiple correlation
coefficient (R2), the sample size (N), and the number of regressor variates (p)
derived from your data
**************************************************************************.

DATA LIST LIST / R2(F9.3) p(F8.0) N(F8.0).
BEGIN DATA
.799 2 12
.255 2 20
.358 2 30
.943 3 20
.617 3 30
.261 3 93
.982 4 13
.640 4 30
.355 4 36
END DATA.
COMPUTE RNICHOL = (N+p+1)/ (N-p-1).
COMPUTE RLORD = (N-1)/(N).
COMPUTE RNICLORD = (1-(RNICHOL*RLORD)* (1-R2)).
COMPUTE RSTEIN1 = (N-1)/ (N-p-1).
COMPUTE RSTEIN2 = (N-2)/ (N-p-2).
COMPUTE RDARLING = (N+1)/(N).
COMPUTE RSTDARL = (1-(RSTEIN1*RSTEIN2*RDARLING)* (1-R2)).
COMPUTE RROZE = (1- (N+p)/ (N-p) * (1-R2)).
COMPUTE RBROWNE1 = ((N-p-3) * (R2)**2)+R2.
COMPUTE RBROWNE2 = ((N-2*p-2) * R2)+p.
COMPUTE RBROWNE = RBROWNE1 / RBROWNE2.
EXECUTE.
FORMAT RNICHOL TO RBROWNE (F9.3).
VARIABLE LABELS R2 'Multiple Correlation Coefficient'/p 'Number of Predictor Variables'/ N 'Sample
Size'/RNICLORD 'Nicholson-Lord'/ RBROWNE 'Browne'/RSTDARL 'Stein-Darlington'/ RROZE
'Rozeboom'/.
REPORT FORMAT=LIST AUTOMATIC ALIGN (CENTER)
MARGINS (*,110)
  /VARIABLES=N p R2 RNICLORD RSTDARL RROZE RBROWNE
  /TITLE "Estimation of the Sample Cross-Validity Expectancy".
```

References

Aaron, B., Kromrey, J. D., & Ferron, J. M. (1998, November). *Equating r-based and d-based effect size indices: Problems with a commonly recommended formula*. Paper presented at the annual meeting of the Florida Educational Research Association, Orlando, FL.

Agresti, A., & Finlay, B. (1986). *Statistical methods for the social sciences* (2nd ed.). SanFrancisco: Dellan Publishing Company.

Allen, D. M. (1971). *The prediction sum of squares as a criterion for selecting predictor variables* (Tech. Rep. No. 23). Lexington, KY: University of Kentucky, Department of Statistics.

Ayabe, C. R. (1985). Multicrossvalidation and the jackknife in the estimation of shrinkage of the multiple coefficient of correlation. *Educational and Psychological Measurement, 45*, 445-451.

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 the predictive power 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 the regression model. *Psychological Bulletin, 87*, 63-65.

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

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cotter, K. L., & Raju, N. S. (1982). An evaluation of formula-based population squared cross-validity estimates and factor score estimates in prediction. *Educational and Psychological Measurement, 42*, 493-519.

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

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvements on cross-validation. *Journal of the American Statistical Association, 78*, 316-331.

Fan, X. (2003). Using commonly available software for bootstrapping in both substantive and measurement analyses. *Educational and Psychological Measurement, 63*, 24-50.

Gollob, H. F. (1967, September). *Cross-validation using samples of size one*. Paper presented at the meeting of the American Psychological Association. Washington, D.C.

Gong, G. (2003). Cross-validation, the jackknife, and the bootstrap: Excess error estimation in forward logistic regression. In P. I. Good & J. W. Hardin (Eds.), *Common errors in statistics (and how to avoid them)* (pp. 173-186). Hoboken, NJ: John Wiley & Sons, Inc.

Hald, A. (1965). *Statisitcal theory with engineering applications* (6th ed.). New York: John Wiley & Sons, Inc.

Huberty, C. J., & Mourad, S. A. (1980). Estimation in multiple correlation/prediction. *Educational and Psychological Measurement, 40*, 101-112.

Kerlinger, F. N., & Pedhazur, E. J. (1973). *Multiple regression in behavioral research*. New York: Holt, Rinehart and Winston, Inc.

Kromrey, J. D., & Hines, C. V. (1995). Use of empirical estimates of shrinkage in multiple regression: A caution. *Educational and Psychological Measurement, 55*, 901-925.

Krus, D. J., & Fuller, E. A. (1982). Computer assisted multicrossvalidation in regression analysis. *Educational and Psychological Measurement, 42*, 187-193.

Lachenbruch, P. A. (1967). An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. *Biometrics, 23*, 639-645.

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.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental tests*. Reading, MA: Addison-Wesley.

Morris, J. D. (1984). Cross-validation with Gollob's estimator: A computational simplification. *Educational and Psychological Measurement, 44*, 151-154.

Morris, J. D. (1986). Microcomputer selection of a predictor weighting algorithm. *Multiple Linear Regression Viewpoints, 1*, 53-68.

Mosier, C. I. (1951). Problems and designs of crossvalidation. *Educational and Psychological Measurement, 11*, 5-11.

Nicholson, G. E. (1960). Prediction in future samples. In I. Olkin et al. (Eds.), *Contributions to probability and statistics*. Stanford, CA: Stanford University Press.

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

Rulon, P. J., Tiedeman, D. V., Tatsuoka, M. M., & Langmuir, C. R. (1967). *Multivariate statistics for personnel classification*. New York: John Wiley & Sons, Inc.

Sprinthall, R. C. (2000). *Basic statistical analysis* (6th ed.). Needham Heights, MA: Allyn and Bacon.

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

Thurstone, L. L. (1947). *Multiple-factor analysis*. Chicago: The University of Chicago Press.

Send correspondence to: David A. Walker
ETRA
Northern Illinois University
DeKab, IL 60115
Email: dawalker@niu.edu