DOCUMENT RESUME

ED 223 701 TM 820 848

AUTHOR Murphy, Kevin R.
TITLE Cost-Benefit Considerations in Choosing among Cross-Validation Methods.
PUB DATE Aug 82
NOTE 16p.; Paper presented at the Annual Meeting of the American Psychological Association (Washington, DC, August 23-27, 1982).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cost Effectiveness; *Estimation (Mathematics); Mathematical Formulas; Psychometrics; *Research Design; Sampling; Statistical Data; *Validity
IDENTIFIERS *Cross Validation

ABSTRACT

There are two general methods of cross-validation: empirical estimation, and formula estimation. In choosing a specific cross-validation procedure, one should consider both costs (e.g., inefficient use of available data in estimating regression parameters) and benefits (e.g., accuracy in estimating population cross-validity). Empirical cross-validation methods involve significant costs, since they are typically laborious and wasteful of data, but under conditions represented in Monte Carlo studies, they are generally not more accurate than formula estimates. Consideration of costs and benefits suggests that empirical estimation methods are typically not worth the cost, except in a limited number of cases in which Monte Carlo sampling assumptions are not met in the derivation sample. Designs which use multiple samples to estimate the cross-validity of a single regression equation are clearly preferable to single-sample designs; the latter are never expected to be more accurate than formula estimates and thus are never worth the cost. Multi-equation designs are more accurate than single equation designs, but they appear to estimate the wrong parameter, and thus are difficult to interpret. (Author)

Cost-Benefit Considerations in Choosing Among Cross-Validation Methods

Kevin R. Murphy

New York University

Running Head: Costs and Benefits in Cross-Validation

Mailing Address
Kevin R. Murphy
Dept. of Psychology
New York University
6 Washington Place
New York, NY 10003
(212) 598-2339

Abstract

There are two general methods of cross-validation: (a) empirical estimation, and (b) formula estimation. In choosing a specific cross-validation procedure, one should consider both costs (eg. inefficient use of available data in estimating regression parameters) and benefits (eg. accuracy in estimating population cross-validity). Empirical cross-validation methods involve significant costs, since they are typically laborious and wasteful of data, but under conditions represented in Monte Carlo studies, they are generally not more accurate than formula estimates. Consideration of costs and benefits suggests that empirical estimation methods are typically not worth the cost, except in a limited number of cases in which Monte Carlo sampling assumptions are not met in the derivation sample. Designs which use multiple samples to estimate the cross-validity of a single regression equation are clearly preferable to single-sample designs; the latter are never expected to be more accurate than formula estimates and thus are never worth the cost. Multi-equation designs are more accurate than single equation designs, but they appear to estimate the wrong parameter, and thus are difficult to interpret.

Cost-Benefit Considerations in Choosing among Cross-Validation Methods

The sample multiple correlation coefficient, R, is an upwardly biased estimator of the corresponding population parameter ρ and of the population cross-validation ρ_c (Darlington, 1968; Herzberg, 1969).[1] A number of cross-validation strategies have been proposed to counter this bias, but there appear to be no clear guidelines for choosing one strategy over another. Cross-validation methods differ in terms of their ease of application and in terms of the amount of data which they consume (costs). They may also differ in terms of their accuracy (benefits). The choice of a strategy which is laborious or which is wasteful of data implies some benefit which offsets these costs; the choice of a cumbermsome cross-validation strategy is unwarranted unless the method chosen is more accurate than simpler alternatives. The purpose of this paper is to outline the costs and benefits associated with different cross-validation strategies; in particular, this paper discusses the way in which the design of a cross-validation study affects the costs and benefits of different types of cross-validation.[2]

There are two general strategies for cross-validating a sample regression equation: (a) formula estimation, and (b) empirical estimation. Formula estimation involves adjusting the sample R by a function of R, N (the number of cases), and p (the number of variables), and using the adjusted R to estimate ρ or ρ_c, depending on the specific formula employed. Empirical methods of cross-validation involve three steps. First, one must collect two or more independent samples from the same population. Next, regression weights must be obtained in one sample (the derivation sample) and applied in another sample (the validation

sample). The correlation between this weighted linear combination of predictiors and the criterion is then used to estimate ρ_c (Mosier, 1951). Formula estimation appears to possess a number of advantages over empirical estimates. First, empirical methods do not allow researchers to use all available data in estimating regression weights; a sizable portion of the available data must be held out for use in a validation sample (Horst, 1966; Schmitt, Coyle & Rauschenberger, 1977). Since the stability of regression weights is highly dependent on the ratio of cases to predictors, the requirement that only some of the available data be used to estimate regression parameters is a serious liability, one which is not incurred when formula estimates are used. Second, formula estimates are very easily computed, whereas the computation of empirical estimates is somewhat laborious.[3] Third, and most important, formula estimates appear to be highly accurate (Cattin, 1980; Rozeboom, 1978). In fact, Monte Carlo comparisons between formula estimates and empirical estimates show that empirical estimates are generally not more accurate, and may in some cases be less accurate than formula estimates (Claudy, 1978; Schmitt, Coyle & Rauschenberger, 1977).

Monte Carlo studies differ from field studies in that the former typically involve truly random sampling from populations in which distributions assume reasonably simple and regular (eg. normal) forms, and in which population parameters are known. When field studies feature sampling procedures and distributions similar to those which occur in Monte Carlo studies, empirical estimates are decidedly inferior to formula estimates; the labor and waste of data inherent in empirical methods leads to no clear-cut gain in accuracy. The choice to conduct an empirical cross-validation strategy is therefore justified only when

important assumptions of the Monte Carlo model are not met. Given the robust nature of multiple regression, departures from normality or linearity assumptions are not likely to seriously affect the conclusions of Monte Carlo studies (Schmitt, Coyle & Rauschenberger, 1977). Serious violations of sampling assumptions may, however, have different effects on the accuracy of formula estimates and the accuracy of empirical estimates of cross-validity. The choice to employ an empirical estimation strategy rather than relying upon formula estimates may therefore be justified when the sampling assumptions of Monte Carlo studies are seriously violated.

Formula estimation procedures are completely insensitive to violations of random sampling assumptions; for any given N, p, and R, the estimate of ρ_c is the same regardless of the nature of the sample from which the regression equation was obtained. It follows that formula estimates may be seriously in error when the derivation sample is not representative of the population. The accuracy of empirical methods, on the other hand, depends entirely on the representativeness of the validation sample, since the validation sample R is used to estimate ρ_c. Thus, empirical methods may be used to accurately estimate the population cross-validity of a regression equation which is obtained from a biased (non-representative) derivation sample. In this situation, empirical estimates may be more accurate than formula estimates, and may therefore be worth the cost. The possible advantage of empirical estimation methods depends in part on the design of the cross-validation study.

The Design of the Empirical Estimation Study

A variety of empirical estimation methods have been described in Mosier (1951), Norman (1965), Gollob (Note 2), and Darlington (1968). These can be classified as either single-sample or multiple-sample

designs and can be further classified as either single-equation or multiple-equation designs. Each cross-validation design has its own strengths and weaknesses; some designs are more accurate across a variety of conditions, while other designs allow researchers to partially offset the costs inherent in empirical approaches.

Single-Sample Designs. The most common empirical cross-validation design is one in which the researcher collects a single sample and randomly partitions that sample into derivation and validation subsamples. A review of studies published in Personnel Psychology and Journal of Applied Psychology between 1976 and 1981 showed that 29 studies employed empirical cross-validation; of these, 24 employed single-sample designs. The single-sample design has previously been criticized as inefficient (Murphy, In Press); the present analysis suggests that there is never any justification for choosing this method of cross-validation over formula estimates.

The conceptual problem with a single-sample design is implied by the name. Although investigators employing this design speak of derivation samples and validation samples, individuals are in fact sampled from a broader population only once. When a sample is randomly partitioned into two subsamples, A and B, any result obtained in subsample A is likely to cross-validate well in subsample B; the similarity of subsamples A and B is a necessary consequence of random partitioning, and has nothing whatsoever to do with the generalizability of sample regression parameters. As N increases, the statistical similarity of subsamples A and B must also increase to the point that the $R_A = R_B = R_{CV}$, where R_{CV} represents a single-sample cross-validated R. This is true regardless of the nature of the original sample. Both single-sample

cross-validation estimates and formula estimates therefore invariably suggest that, when N is large, the sample R is a very accurate estimator of (Murphy, In Press). When the original sample is not representative of the population of interest, R is not an accurate cross-validity estimate, regardless of the sample size.

Empirical methods of cross-validation are likely to be more accurate than formula estimates only if the sampling assumptions of shrinkage formulas are clearly not met; in this case, empirical methods may allow you to adjust for the effects of both random and systematic sampling error. Single-sample cross-validation methods are an exception, since single-sample methods allow one to adjust for random sampling error only. There do not appear to be any plausible circumstances under which single-sample methods would be expected to yield cross-validity estimates which are systematically more accurate than formula estimates. It is likely, then, that single-sample estimates never yield benefits which justify their relative costs.

Multi-Sample Designs. Mosier (1951) clearly called for multiple, independent samples in estimating cross-validity. Monte Carlo studies have shown that multi-sample designs are highly accurate when the validation sample is representative of the population of interest (Claudy, 1978). Multi-sample cross-validity estimates may be significantly more accurate than formula estimates in the somewhat unusual case in which the validation sample is representative of the population of interest but the derivation sample is not. Even in this restricted case, however, empirical estimation procedures would be sensible only if the derivation sample was fairly large and the validation sample was fairly small. If a large, representative sample

were available for validation purposes, it would surely be simpler to apply multiple regression in that sample, and use a formula to estimate ρ_c

Overall, empirical estimation methods appear to be of very limited utility. Single-sample methods are never significantly more accurate than formula estimates. Multi-sample methods appear to be justified only when thevalidation sample is representative of the population of interest, the derivation sample is not, and the validation sample is considerably smaller than the derivation sample. This represents the only plausible case where the gain in accuracy could possibly offset the costs in terms of the labor involved and in terms of inefficient use of data in estimating regression parameters.

Single vs. Multiple Equations. The researcher who collects two independent samples has two options for empirical cross-validation. First, he or she can compute a single regression equation in one sample and validate that equation in another. Second, the researcher can compute a number of regression equations, and can use a pooled cross-validity to estimate ρ_c . For example, Mosier (1951) advocated double cross-validation, in which a regression equation is computed in each of two independent samples and is validated in the cross sample. Norman, (1965) described a more elaborate double-split technique which combines features of single-sample and multi-sample cross-validation. Gollob (Note 1) described a jacknife-like method which involved computing N separate regression equations.

Claudy (1978) has shown that pooled cross-validation estimates are more accurate than single-equation estimates. Accuracy, however, is purchased at the price of conceptual clarity. Cross-validation is generally defined as a method for estimating the population correlation

between a set of predictors, which are combined using a specific sample regression equation, and the criterion. When a number of different regression equations are used in a cross-validation study to estimate ρ_c, it is no longer clear precisely what is being cross-validated. Consider, for example, the situation depicted in Figure 1.

Insert figure 1 about here

Figure 1 depicts a double cross-validation study in which the final estimate of cross-validity is .57. The regression equations in samples 1 and 2 are different, and have different validities in the population. The double cross-validity of .57 is not a valid estimate of the population cross-validity of the sample 1 equation, nor is it a valid estimate of the population cross-validity of the sample 2 equation. It may provide an estimate of the average ρ_c when the equations are used interchangeably, but this paramenter is not likely to be of interest, since sample equations are not likely to be used in this way. In general, it appears that multi-equation cross-validity studies provide accurate estimates, but that they estimate the wrong parameter.

Summary

The choice between empirical estimation methods and formula estimates invariably involves a trade-off between the accuracy of the cross-validity estimate and the simplicity and efficiency of the estimation procedure. Empirical methods of cross-validation are justified only if they are more accurate than formula estimates and if

the gain in accuracy offsets the costs inherent in empirical methods. Recent Monte Carlo studies suggest that empirical estimates are generally less accuate than formula estimates (Claudy, 1978; Schmitt, Coyle & Rauschenberger, 1977). Thus, in a wide variety of situation, empirical estimates are clearly inferior to formula estimates.

Empirical cross-validity estimates may be more accurate than formula estimates in the limited set of cases where the validation sample is representative of the population of interest, the derivation sample is not representative, and the validation sample is considerably smaller than the derivation sample. The utility of empirical estimation methods is further restricted by the design of the cross-validation study. Single-sample designs cannot offer benefits which offset their costs. Multi-equation designs are more accurate than single-equation designs, but they estimate a parameter which is of limited interest. Altogether, it appears that empirical cross-validation techniques are rarely worth the time and effort.

Reference Notes

1. Gollob, H. F. Cross-validation using samples of size one. Presented at annual convention of the American Psychological Association, Washington, D.C., 1967.

References

Cattin, P. A note on the estimation of the squared cross-validated multiple correlation of the regression model. Psychological Bulletin, 1980, 87, 63-65.

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

Cureton, E. Validity, reliability, and baloney. Educational and Psychological Measurement, 1950, 10, 94-96.

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

Herzberg, P. A. The parameters of cross-validation. Psychometrika Monograph Supplement, 1969, 34, No. 16.

Horst, P. An overview of the essentials of multivariate analysis methods. In R. B. Cattell (Ed.) Handbook of multivariate experimental psychology. Chicago: Rand-McNally, 1966.

McNemar, Q. Psychological statistics, 4th Ed. New York: Wiley, 1969.

Mosier, C. I. Problems and designs of cross-validation. Educational and Psychological Measurement, 1951, 11, 1-11.

Murphy, K. R. Fooling yourself with cross-validation; Single-sample designs. Personnel Psychology, In Press.

Norman, W. T. Double-split cross-validation: An extension of Mosier's design, two undesireable alternatives, and some enigmatic results. Journal of Applied Psychology, 1965, 49, 348-357.

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

Schmitt, N., Coyle, B., & Rauschenberger, J. A Monte Carlo evaluation of three formula estimates of cross-validated multiple correlation. Psychological Bulletin, 1977, 84, 751-758.

Footnotes

Requests for reprints should be sent to: Kevin R. Murphy, Dept of Psychology, New York University, 6 Washington Place, New York, NY 10003.

An earlier version of this paper was presented at the American Psychological Association convention, Washington, D.C., 1982.

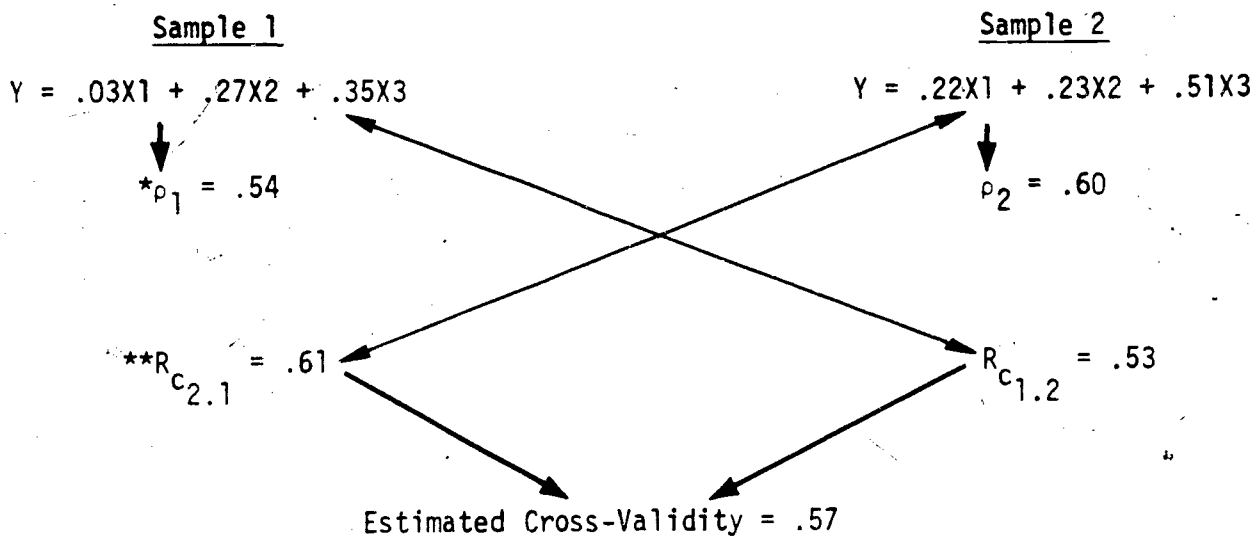
1. When sample regression weights are applied to a set of predictors, ρ refers to the population correlation between this linear combination of predictors ($\hat{Y}$) and the criterion (Y). ρ_c refers to the expected value of the correlation between Y and Y in a number of random samples from that population.

2. It is assumed throughout that there is no pre-selection of predictors. Statistical pre-selection of predictors greatly increases the necessity of empirical cross-validation (Cureton, 1950; McNemar, 1969).

3. To date, none of the most widely used regression programs include simple options for empirical cross-validation.

Figure Caption

Figure 1. A Double Cross-Validation Study



*ρ_1 Population Validity of Sample 1 Regression Equation

**$R_{c_{2.1}}$ Cross-Validity When Weights From Sample 2 Are Applied in Sample 1