DOCUMENT RESUME

ED 048 353 TM 000 429

AUTHOR Gustafson, Richard A.
TITLE Multiple Regression Prediction Models in the Behavioral Sciences: Prediction of Federal Aid Allocations to Local School Districts.
PUB DATE Feb 71
NOTE 9p.; Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York, February 4-7, 1971

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS *Community Characteristics, *Federal Aid, *Models, *Multiple Regression Analysis, *Prediction, Predictive Measurement, Predictor Variables, Research Methodology, School Districts, Statistical Analysis

ABSTRACT

Twenty-nine community characteristics were studied to determine which were statistically most useful as predictors of per-pupil Federal aid to the 169 school districts of Connecticut. Three regression models were developed using community traits as predictors of Federal aid allocations. Cross-validation of regression models to predict future per-pupil Federal aid allocations introduced a number of problems which were generalizable to other research situations in psychology and measurement. Improving cross-validation of regression models by using restricted models, equating the means of vectors by using constants as multipliers, and examination of standard errors are discussed. Data was analyzed using these three techniques and results compared to those obtained from the traditional cross validation procedures. Implications of these findings are discussed in terms of improving predictive models in measurement and psychological research. See TM 000 419 for a report of the output of the regression models. (Author/LR)

Multiple Regression Prediction Models in the Behavioral Sciences:

Prediction of Federal Aid Allocations to Local School Districts

by

Richard A. Gustafson
University of Connecticut
and
The Center for Planning and Evaluation
1110 North Tenth Street
San Jose, California

A Symposium Presentation given at the
American Educational Research Association
New York, New York
February 4, 1971

Multiple Regression Prediction Models in the Behavioral Sciences: Prediction of Federal Aid Allocations to Local School Districts[1]

Overview of the Study

The main thrust of this presentation is to explore alternative procedures in the cross validation of regression models. However, a brief overview of the Federal aid prediction study would be helpful in setting the stage.

The purpose of the study was to determine which community characteristics, among the twenty nine studied, were statistically most useful as predictors of per-pupil Federal aid to the 169 school districts of Connecticut for the 1968 and 1969 fiscal years. Three regression models were developed using these community traits as predictors. The predictors and criterion variables used in these models were nearly perfect in their reliability as they were based upon dollar amounts, welfare records, or other equally measurable factors. This reliability of predictors and criterion was quite different than in the other studies presented here today.

Multiple correlation coefficients for all models were significant at the .01 level with the community characteristics reflecting need as defined by law proving to be the best predictors of aid allocations. Cross validation of these models provided the methodological concerns of this paper.

Cross Validation

The cross validation technique used was to predict future or past funding levels and compare these with actual Federal aid grants. This type of cross validation introduced certain problems not encountered when using the classical procedures.

Traditional procedures, and their limitations, for cross validating

---

1. This paper is based upon a portion of the author's doctoral dissertation; "The Development of Regression Models using Community Characteristics as Predictors of Federal Aid Allocations to Connecticut School Districts."

predictive equations derived from multiple linear regression and multiple correlation analysis have been discussed by Darlington (1968) and Kelley (1969). It is generally agreed that the statistical estimates of the Mult-R shrinkage using either the Wherry (1931) or the Lord (1950) - Nicholson (1948) formulas yield optimistic results. Empirical cross validation on an independent sample or developing the predictive equation on two-thirds of the original sample and cross validating on the remaining one-third are the alternatives usually proposed.

A single model is used here to illustrate the alternative procedures for cross validation.

Model III for the 1969 fiscal year has as its criterion the per-pupil aid granted under the sum of all major components of Federal aid administered by the State Department of Education. A multiple correlation coefficient of .810 was achieved when a total of 27 predictors were permitted to enter the model. The standard error of the estimate was 11.07.

Preliminary cross validation of this full model using the weights derived from the 1969 data to predict 1968 allocations indicated a cross validated correlation of .67. This was considerably lower than the shrinkage estimates obtained when using the Wherry or Lord - Nicholson formulas.

Empirical cross validation traditionally requires the selection of an independent sample on which to apply the derived predictive equation. Sample independence permits one to assume equality of means and homogeneity of variance.

In the Federal aid study, however, cross validation was performed not on an independent sample of communities, but on the same communities used in the developmental sample. Moreover, because of fluctuating Federal aid levels from year to year, one could not assume equality of means and homogeneity of variance between the developmental and cross validation samples.

For the above reasons, the usual cross validation correlation or its estimate from the Wherry and Lord - Nicholson formulas seemed inappropriate. Three

alternative techniques were examined in an effort to achieve a more meaningful and reliable cross validation procedure.

The Stepwise Multiple Regression routine (IBM, 1968) returns a new Mult-R and standard error of the estimate value for each successive step in the procedure. This standard error diminishes with each step to a point where the inclusion of additional predictors will cause it to increase. Examination of the following formula, where 'D' is the total sum of squares, 'Scum' is the cumulative sum of squares reduced through the i-th step, 'n' is the sample size, and 'k' is the number of predictors, explains this phenomenon.

$$SE_{y.1,\ 2,\ 3....i} = \sqrt{\frac{D - Scum}{n - k - 1}}$$

The first technique employed was to use only those predictors, which when included in the model, caused a reduction in the standard error of the estimate. This restricted model included thirteen predictors, giving a Mult-R of .806. As indicated above, empirical cross validation using the full model yielded a correlation coefficient of .67. The cross validated correlation for the restricted model was .77. This technique of using a restricted model, which had a slightly lower Mult-R with respect to the developmental sample, gave a much improved cross validated result. This finding has been replicated with respect to other models developed in the Federal aid study.

Insert Table 1 about here

While the cross validated correlation improved ten points when the restricted model was employed, the predicted means for both the full and restricted models were significantly different from the criterion mean. (See Table 1)

A second technique was then employed to handle this difference. Federal aid levels were about eleven percent higher in 1968 than in 1969. Accordingly, the predicted vectors for both the full and restricted models were multiplied by the constant 1.11. The means of these 'corrected' vectors were much closer to the criterion mean. 'T' tests (See Table 1) indicated these differences were not significant. The multiplication of the predicted vectors by a constant to reflect different levels of Federal funding yielded a better predictive model; however, this improvement would not be reflected in the correlation between the criterion and the 'corrected' vectors.

The third factor examined was the standard error of the estimate. A comparison between the standard error of the estimate and the standard deviation of the criterion provides a measure of how well the predictive model is performing. As the standard error approaches the standard deviation of the criterion, the Mult-R approaches zero. Conversely, as the standard error approaches zero, the Mult-R approaches one.

Insert Table 2 about here

Examination of Table 2 reveals that the best predictive model, in terms of cross validated results, was the restricted model using thirteen predictors corrected for differential levels of Federal funding. The cross validated correlation was .77, the mean of the predicted values was not significantly different from the criterion mean, and the standard error of the estimate was the lowest.

Summary and Implications:

It was found that restricted models using fewer predictors achieved higher cross validated correlations. These models also yielded lower standard errors. When using cross validation samples which are not independent of the developmental sample, one cannot assume equality of means and homogeneity of variance. 'T' tests were employed to identity differences with respect to the means. Where

significant differences existed, a factor reflecting the different levels of Federal funding from year to year was used to equate the means and yield better predictive models. Models using fewer predictors and including the correction factor yielded smaller standard errors.

When developing predictive models to be used to estimate future traits on the same or other non-independent sample, cross validation via standard empirical or shrinkage procedures may not yield optimal feedback on the effectiveness of the predictive models. Consideration of restricted models using only a few good predictors may yield more valid results in terms of higher cross validated correlations, and reduced standard errors. Equating means by the introduction of a constant may also improve model generalizability.

RAG:ss
CPE
12-16-70

Table 1

't' Tests Between Actual 1968 Funding and Predicted Funding Levels

	full model		restricted model	
	1968 predicted	1968 predicted and corrected	1968 predicted	1968 predict and correcte
Actual 1968 Funding	3.58*	.997	4.57*	1.61

* significant at the .01 level

Table 2

Cross Validation Results for Full and Restricted Models

Comparing Actual 1968 Federal Funding with Predicted Funding Levels

				Actual 1968		Predicted 1968		Predicted 1968 when corrected	
	Derived Mult-R	SE of Estimate	Cross Validated Correlation	Mean	S.D. Var.	Mean	Standard Error of Estimate	Mean	Standard Error of Estimate
Full Model	.810	11.07	.67	27.46	17.54 308	23.78	13.35	26.40	13.78
Restricted Model	.806	10.69	.77	27.46	17.54 308	23.49	11.23	26.07	11.14

REFERENCES

Darlington, R. B. Multiple Regression in Psychological Research and Practice. Psychological Bulletin. 1968, 69, (3), 161-182.

Gustafson, R. A. The Development of Regression Models Using Community Characteristics as Predictors of Federal Aid Allocations to Connecticut School Districts. Unpublished doctoral dissertation, University of Connecticut, 1970.

Kelley, F. J., D. L. Beggs, and K. A. McNeil. Research Design in the Behavioral Sciences: Multiple Regression Approach. Cartondale, Illinois: Southern Illinois University Press, 1969.

Lord, F. M. Efficiency of Prediction When a Progression Equation From One Sample is Used on a New Sample. Research Bulletin (No. 50-40), Princeton, N. J., Educational Testing Service, 1950.

Nicholson, G. E.,Jr. The Application of a Regression Equation to a New Sample. Unpublished Doctoral dissertation, University of North Carolina, 1948.

System/360. Scientific Subroutine Package, Version III. White Plains, N. Y.: International Business Machines, 1968.

Wherry, R. J. A New Formula for Predicting the Shrinkage of the Coefficient of Multiple Correlation. Annals of Mathematical Statistics, 1931, 2, 440-457.

RAG:ss
CPE
12-14-70