ED 335 422 TM 017 152

AUTHOR Loftin, Lynn B.
TITLE The Importance of Evaluating Whether Results Will Generalize: Application of Cross-Validation in Discriminant Analysis.
PUB DATE Jan 91
NOTE 17p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, January 24, 1991).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Computer Simulation; *Discriminant Analysis; Generalizability Theory; *Mathematical Models; Reliability
IDENTIFIERS *Cross Validation; *Invariance Principle

ABSTRACT

Cross-validation, an economical method for assessing whether sample results will generalize, is discussed in this paper. Cross-validation is an invariance technique that uses two subsets of the data sample to derive discriminant function coefficients. The two sets of coefficients are then used with each data subset to derive discriminant function scores. The scores are correlated to assess the stability of the discriminant function coefficients across samples. The closer the set of results are to each other, the greater the stability of the coefficients is across samples. One advantage of cross-validation is the ease of calculation and adaptability to other statistical procedures. A weakness is that the sample size must be large in order for division of data into data subsets to be meaningful. Six steps are outlined that demonstrate a double cross-validation invariance procedure in a discriminant analysis. The data set used consists of 64 subjects with two continuous predictor variables "x" and "y" and one criterion variable "Group" with four levels. It is noted that to evaluate the stability of discriminant function coefficients, the discriminant function scores themselves must be examined. Two tables are included, and an outline of the computer commands for simulating data in the first table is appended. (Author/RLC)

THE IMPORTANCE OF EVALUATING WHETHER RESULTS WILL GENERALIZE: APPLICATION OF CROSS-VALIDATION IN DISCRIMINANT ANALYSIS

Lynn B. Loftin
University of New Orleans 70148

Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX, January 24, 1991.

ABSTRACT

Cross-validation is an invariance technique that utilizes two subsets of the data sample to derive discriminant function coefficients. The two sets of coefficients are then used with each data subset to derive discriminant function scores. The scores are correlated to assess the stability of the discriminant function coefficients across samples. One advantage of cross-validation is ease of calculation and adaptability to other statistical procedures. A weakness is that the sample size must be large for division of data into data subsets to be meaningful.

Statistics are collected primarily from samples of populations in order to make predictions or to construct theories regarding the population as a whole. Repeating the study over many samples is the best way to increase the generalizability of the findings to the larger population. But, realistically, repeated studies are often impractical or impossible. However, there are statistical techniques designed to assess whether sample results will generalize.

Invariance techniques such as cross-validation are economical methods that address this question of generalizability. Unfortunately, these techniques are not utilized in many statistical analyses. When this source of valuable information goes untapped, the researcher runs a greater risk of unknowingly overinterpreting nonreplicable results. There are three statistical results that are commonly assessed in published research reports. However much these results contribute to interpretation, none of them produce information about external validity.

The statistical test most often included in research is the test of statistical significance. All too often, a test of statistical significance is the end of a statistical analysis. However, a binary test of the null hypothesis does not test the generalizability of the results. A statistical significance test does not yield an estimated probability that the results will be repeated with other samples, as erroneously believed by some (Carver, 1978). The primary feature that determines statistical significance is the size of the sample used in the study (Thompson, 1989). Stability of the underlying constructs that produce the findings is not addressed by tests of statistical significance. Too often the most meaningful implication of the research may not be addressed due to the overreliance on statistical significance testing (Fish, 1986b).

Thompson (1989) suggests that other results be examined in addition to statistical significance to gain some insight regarding result interpretation. Evaluation of <u>effect sizes</u> may produce information regarding result importance. Large effect sizes, in spite of a lack of statistical significance, suggest further study, whereas small effect sizes with statistical significance may not trigger such interest. However, reproducibility of the effect size in other samples can not be predicted by evaluating the effect size in one sample.

Another aspect of a statistical analysis to be evaluated are the derived statistical coefficients or weights (e.g., beta weights or discriminant function coefficients). The absolute size of these coefficients is indicative of the predictive power of the individual independent variables. For instance, in discriminant analysis, evaluation of the discriminant function coefficients suggests to the researcher which variables are the most important predictors of the dependent variable. However, large weights are not necessarily stable across multiple samples. Information about the stability is crucial if the results of a discriminant analysis are to be accurately generalized to other samples. Testing the stability of these coefficients is the objective of invariance procedures. Invariance procedures are an "attempt to determine how stable the statistical results are likely to be across different samples" (Fish, 1986a) and should be included in every analysis.

Cross-validation is one invariance procedure that can be used with several statistical techniques. It is one of several "hold-out" procedures which involves using one part of the data sample to generate function coefficients or weights to be used on the other part of the sample. In a cross-validation procedure, discussed in Cooil et al. (1987), the data sample is split, function coefficients are derived for each part of the split sample, the function coefficients for each subsample are applied to the opposite subsample, then the results are correlated across subsamples. The closer the sets of results are to each other, the greater is the stability of the coefficients across samples.

Cross-validation has the advantage of simplicity and versatility. The technique can be applied to any parametric procedure, such as regression analysis, discriminant analysis, canonical correlation analysis, and the various analyses of variance (Fish, 1986a; Sandler, 1987; Thompson, 1989). The procedure many be used repeatedly, but double cross-validation, described here, is considered adequate for most predictive studies (Kerlinger & Pedhazur, 1973). Cross-validation is limited in that the original sample must be large enough so that the size of the subsamples is adequate for the number of variables.

A concrete example will be used to demonstrate a double cross-validation invariance procedure in a discriminant analysis. The data set consists of 64 subjects with two continuous predictor variables "x" and "y" and one criterion variable, "Group," with four levels. The heuristic cross-validation illustrated here is performed on only the first discriminant function, however, it may be profitable to continue the cross-validation on the next functions if they are important to the research. The cross-validation procedure is the same for each function. The SPSS-X statistical package was used to conduct the empirical invariance analysis for these data on the first discriminant function (see Appendix for commands).

INSERT TABLE 1 ABOUT HERE

Steps for a Cross-Validation in Discriminant Analysis

1. The original sample of 64 subjects was randomly split into two invariance subsamples by using random numbers "cxsam". Preferably, the two subsamples are of somewhat unequal size to counter the objection that a satisfactory measure of invariance is dependent on a particular sample size (Fish, 1986b). The split for the sample data was n=30 for subsample 1 and n=34 for subsample 2. The split sample size was sufficient for a meaningful cross-validation, given the number of predictor variables.

2. A discriminant analysis was computed on the original sample and on each subsample in three separate operations. The three sets of generated discriminant function coefficients and

discriminant function scores were saved.

3. For the purpose of illustration, the raw data within each of the two subsamples, were separately transformed **not** into z-scores which utilize the subsample standard deviations, but into standardized scores that utilize the square root of the pooled within-group variances of each subsample. These values are used to define the separation between the groups in a particular direction as the distance between the means of the groups standardized for the within-group variance (Hand, 1981), and are also used in cross-validation. In this example these discriminant standardized scores are denoted "dSx_{1i}" and "dSy_{1i}" for x_i and y_i values from subsample 1 and "dSx_{2i}" and "dSy_{2i}" for x_i and y_i from subsample 2. The calculated pooled within-group variances can be obtained from the discriminant procedure for each subsample. For subsample 1, the pooled within-group variance for x_1 was 4.006410 and for y_1 was 3.00000; for subsample 2 x_2 was 4.606389 and y_2 was 4.213333. The procedure for calculation of discriminant standardized scores for each case x_i and y_i in each subsample is as follows:

$$dSx_{11} = (x_{11}-x_{1mean})/(\text{pooled within-group variance } x_1)^{1/2}$$
$$dSy_{11} = (y_{11}-y_{1mean})/(\text{pooled within-group variance } y_1)^{1/2}$$
$$dSx_{12} = (x_{12}-x_{2mean})/(\text{pooled within-group variance } x_2)^{1/2}$$
$$dSy_{12} = (y_{12}-y_{2mean})/(\text{pooled within-group variance } y_2)^{1/2}$$

4. Obtain the standardized discriminant function coefficients for x and y from the discriminant function analysis on subsample 1

performed in Step 2 above. The standardized discriminant function score for the first function for subsample 1 for x_1 "DFCx_1" was 1.38185 and for y_1 "DFCy_1" was -1.22482. These standardized function coefficients are used with the standardized scores from subsample 1, "dSx_{11}" and "dSy_{11}" to compute actual discriminant function scores for each case i <u>and</u> these same standardized discriminant function coefficients are used with the standardized scores from subsample 2, "dSx_{12}" and "dSy_{12}," to compute "crossed" discriminant function scores for each case i in subsample 2. Repeat the process using standardized discriminant function coefficients from subsample 2, x_2 "DFCx_2" was 1.71359 and y_2 "DFCy_2" was -1.29555, to compute another set of actual and crossed discriminant function scores. You will have two sets of actual discriminant function scores and two sets of crossed discriminant function scores. An actual discriminant function score for case i of subsample 1 is computed using standardized discriminant function coefficients (DFC) derived from subsample 1:

$$\text{discriminant function score}_{i11} = (dSx_{i1} \times DFC_{x1}) + (dSy_{i1} \times DFC_{y1})$$

An example of the procedures to produce a crossed discriminant function scores using subsample 1 with subsample 2 discriminant function coefficients:

$$\text{discriminant function score}_{i12} = (dSx_{i1} \times DFC_{x2}) + (dSy_{i1} \times DFC_{y2})$$

5. The four sets of scores are then correlated in a Pearson correlation matrix. The SPSS-X procedure CORRELATIONS will produce the matrix (See Appendix A). There would normally be four sets of scores to correlate, two sets of crossed and two sets of actual scores. However, for illustration, computer generated scores will also be included. The sets of scores are: (a) subsample 1 cases (i) using subsample 1 discriminant function coefficients (discriminant function $score_{111}$) labelled "dS1C1"; (b) subsample 1 cases (i) with subsample 2 discriminant function coefficients (discriminant function $score_{112}$) labelled "dS1C2"; (c) subsample 2 cases (i) using subsample 2 discriminant function coefficients (discriminant function score $_{122}$) labelled "dS2C2"; (d) subsample 2 cases using subsample 1 discriminant function coefficients (discriminant function score $_{121}$) labelled "dS2C1"; (e) original data with discriminant function scores generated from all the cases with discriminant function coefficients derived from all the cases (discriminant function $score_{1T}$) labelled "DSCORE1"; (f) the computer generated discriminant function scores generated from cases in subsample 1 with subsample 1 coefficients computed in step 2 labelled "DSCORE11"; (g) the computer generated discriminant function scores generated from cases in subsample 2 with subsample 2 coefficients computed in step 2 labelled "DSCORE21."

INSERT TABLE 2 ABOUT HERE

Of particular note are the scores partially hand computed for subsample 1 using subsample 1 coefficients and subsample 2 using subsample 2 coefficients versus the computer generated scores for subsamples 1 and 2. This comparison illustrates how the discriminant function scores are calculated. The actual computer generated scores and the hand computed scores should be very close and the correlation 1.00.

6. To assess the invariance, the correlations between the actual scores and the crossed scores are examined. Each correlation coefficient is squared for comparison. In our example the correlation coefficient between the actual scores and crossed scores from subsample 1 (discriminant $scores_{11}$ and discriminant $scores_{12}$), $r=.9882$ (r^2 = .9765392) reflects a very high degree of correlation. This would suggest the discriminant scores are very stable across samples. For a double cross-validation the actual and crossed scores from subsample 2 are correlated (discriminant $scores_{22}$ and discriminant $scores_{21}$). The correlation r = .9811, (r^2 =.9525572) also reflects a very high degree of correlation. This second cross-validation reinforces the indication of the first correlation that the coefficients are very stable. Thompson (1981) terms estimates such as these "invariance coefficients," since they evaluate the replicability or the invariance of results.

Summary

Invariance is an underutilized technique and its application is not standardized. The methods can be left to the creativity of the researcher keeping in mind the characteristics and goals of the study. The researcher must use judgement in determining the amount of change in the invariance coefficients that is acceptable.

The cross-validation invariance procedure in this example concerned the stability of the discriminant function coefficients which indicate the relative contribution of each variable to each function. Such an invariance procedure might be performed in an attempt to find out if particular variables are stable contributors to the analysis. In this example, the discriminant functions were very similar and the scores were also highly correlated. This suggests that for other samples, the coefficients would be similar, that external validity has been established. However, one important point to be made is that to evaluate the stability of discriminant function coefficients, the discriminant function scores themselves must be examined. The discriminant function coefficients may look different but render similar synthetic scores and hence be stable predictors. An invariance procedure, such as cross-validation, by empirically examining the discriminant function scores can reveal important insights regarding the stability of results.

References

Carver, R.P. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Cooil, B., Winer, R. S., & Rados, D. L. (1987). Cross-validation for prediction. Journal of Marketing Research, 24, 271-279.

Fish, L. (1986a). Relationships between gender, proposed teaching level, and several background variables to motives for selecting a career in teaching. Unpublished master's thesis, Dept. of Educational Leadership & Foundations, University of New Orleans.

Fish, L. (1986b, November). The importance of invariance procedures as against tests of statistical significance. Paper presented at the annual meeting of the Mid-South Educational Research Association, Memphis, TN. (ERIC Document Reproduction Service No. ED 278 707)

Hand, D. J. (1981). Discrimination and classification New York: John Wiley & Sons.

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

Sandler, A. B. (1987, January). The use of invariance and bootstrap procedures as a method to establish the reliability of research results. Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, TX. (ERIC Document Reproduction Service No. Ed 280 902)

Thompson, B. (1981) Utility of invariance coefficients. Perceptual and Motor Skills, 52, 708-710.

Thompson, B. (1989). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. Measurement and Evaluation in Counseling and Development, 22, 2-6.

TABLE 1

Hypothetical Data Set

Case	Group	X	Y	Cxsam*
1	1	4	2	2
2	1	5	3	2
3	1	4	4	2
4	1	4	5	2
5	1	3	4	1
6	1	6	5	2
7	1	5	6	2
8	1	7	5	2
9	1	6	6	1
10	1	8	6	2
11	1	7	6	2
12	1	9	7	2
13	1	8	7	1
14	1	8	8	2
15	1	9	9	1
16	1	9	9	2
17	2	1	2	2
18	2	3	3	2
19	2	3	5	1
20	2	3	5	1
21	2	2	5	2
22	2	4	6	1
23	2	4	5	2
24	2	5	6	2
25	2	6	6	2
26	2	6	6	2
27	2	6	7	2
28	2	7	7	2
29	2	7	7	1
30	2	8	9	1
31	2	8	9	1
32	2	9	9	2
33	3	4	1	2
34	3	4	2	1
35	3	3	2	1
36	3	2	4	2
37	3	5	3	2
38	3	7	4	1
39	3	4	5	1
40	3	5	4	2
41	3	7	5	1
42	3	9	5	1
43	3	6	5	1
44	3	5	6	2
45	3	7	6	2
46	3	9	7	2
47	3	8	6	1
48	3	8	5	2
49	4	1	7	1
50	4	1	2	1
51	4	1	1	2
52	4	2	2	2
53	4	2	3	1
54	4	2	3	1
55	4	2	3	1
56	4	3	3	1
57	4	3	4	2
58	4	4	5	1
59	4	4	4	1
60	4	4	5	1
61	4	4	6	1
62	4	5	6	1
63	4	5	7	1
64	4	5	7	2

* Cxsam is a random number used to split the data sample and was not used in the analysis.

TABLE 2

Invariance Statistics

	DS1C1	DS1C2	DS2C2	DS2C1	DSCORE1	DSCORE21	DSCORE11
DS1C1	1.000	.9882	(0)	(0)	.9788	(0)	1.0000
DS1C2	.9882	1.000	(0)	(0)	.9788	(0)	.9882
DS2C2	(0)	(0)	1.000	.9812	.9989	1.000	(0)
DS2C1	(0)	(0)	.9812	1.000	.9892	.9812	(0)
DSCORE1	.9788	.9986	.9989	.9892	1.000	.9989	.9788
DSCORE11	1.000	.9882	(0)	(0)	.9788	(0)	1.000
DSCORE21	(0)	(0)	1.000	.9812	.9989	1.000	(0)

APPENDIX

Example SPSS-X Commands* for Table 1 Data

```
TITLE "CROSS-VALIDATION IN DISCRIMINANT ANALYSIS"
FILE HANDLE DT/NAME=DT/CASE 1-2 GROUP 7 X 12 Y 17 CXSAM 22
LIST VARIABLES CASE GROUP X Y CXSAM
Subtitle "Computing Discriminant Standard Scores (dS1) for Sample1"
If (cxsam=1) dSx1=(x-4.83333)/(4.006410**.5)
If (cxsam=1) dSy1=(y-5.10000)/(3.000000**.5)
Subtitle "Computing Discriminant Standard Scores (dS2) for Sample2"
If (cxsam=2) dSx2=(x-5.32353)/(4.606389**.5)
If (cxsam=2) dSy2=(x--5.05882)/(4.213333**.5)
Subtitle "Computing 1st Discriminant Function Scores"
Subtitle "Sample 1 data with Discriminant Function Coefficient 1"
if (cxsam=1) f1dS1c1=(1.38185*dSx1)+(-1.22482*dSy1)
Subtitle "Sample 1 data with Discriminant Function Coefficient 2"
if (cxsam=1) f1dS1c2=(1.71359*dSx1)+(-1.29555*dSy1)
Subtitle "Sample 2 data with Discriminant Function Coefficient 2"
if (cxsam=2) f1dS2c2=(1.71359*dSx2)+(-1.29555*dSy2)
Subtitle "Sample 2 data with Discriminant Function Coefficient 1"
if (cxsam=2) f1dS2c1=(1.38185*dSx2)+(-1.22482*dSy2)
SUBTITLE "DISCRIMINANT ANALYSIS USING ALL DATA"
DISCRIMINANT GROUPS=GROUP (1,4)/VARIABLES=X,Y/METHOD=DIRECT
   /SAVE=SCORES=DSCORE
STATISTICS 1,2,3,
TEMPORARY
SUBTITLE "DISCRIMINANT ANALYSIS USING SPLIT DATA SAMPLE 1"
SELECT IF (CXSAM=1)
DISCRIMINANT GROUPS=GROUP(1,4)/VARIABLES=X,Y/METHOD=DIRECT
   /SAVE=SCORES=DSCORE1
STATISTICS 1,2,3
TEMPORARY
SUBTITLE "DISCRIMINANT ANALYSIS USING SPLIT DATA SAMPLE 2"
SELECT IF (CXSAM=2)
DISCRIMINANT GROUPS=GROUP(1,4)/VARIABLES=X,Y/METHOD=DIRECT
   /SAVE=SCORES=DSCORE2
STATISTICS 1,2,3
list variables
   case,group,x,y,dscore1,dscore11,f1ds1c1,f1ds1c2,f1ds2c2,f1ds2c1,
   dscore21
subtitle "invariance results"
correlations variables=dscore1,dscore11,f1ds1c1,f1ds1c2,f1ds2c2
   f1ds2c1
```

* The analysis requires three runs. The first utilizes the cards typed in all capitals and is conducted to derived the needed pooled-within covariances which were added in the second run typed in capitals with lower case. The third run in all lowercase is to list the variables to check calculations and to correlate the scores.