DOCUMENT RESUME

ED 327 573 TM 016 020

AUTHOR Thompson, Bruce; And Others
TITLE Stepwise Methods Lead to Bad Interpretations: Better Alternatives.
PUB DATE 13 Jan 91
NOTE 18p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, January 25, 1991).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Admissions Officers; *Data Interpretation; *Effect Size; Higher Education; Medical Schools; *Predictor Variables; Research Methodology; *Research Problems; Statistical Significance
IDENTIFIERS *Analytical Methods

ABSTRACT

Problems with using stepwise analytic methods are discussed, and better alternatives are illustrated. To make the illustrations concrete, an actual data set, involving responses of 91 medical school admissions directors to 30 variables, was used. The 30 variables involved perceptions of barriers to medical school with respect to characteristics of medical students. The propensity of researchers to apply statistical significance tests to evaluate how many steps should be implemented is considered. The problems of using statistical significance testing in conjunction with stepwise methods are elaborated, but the emphasis is on better alternatives to stepwise methods. A two-stage approach to variable selection is recommended. If variables must be eliminated, a better procedure is to compute effect sizes for every possible predictor set using readily available computer software. In the two-stage process, variable selection does not depend on the results in previous steps. The first step is to determine the desired size, "k," of the predictor variable set. The second step is to select the best predictor set of the desired size, consulting the effect size; however, a better approach would be to select the predictor set based on theory, previous empirical results, or the accessibility of variables in a given set. A 19-item list of references is included. Three tables and one graph illustrate the analysis. (Author/SLD)

lesliem3.wp1 1/13/91

STEPWISE METHODS LEAD TO BAD INTERPRETATIONS: BETTER ALTERNATIVES

Bruce Thompson

Texas A&M University 77843-4225
and
Baylor College of Medicine

Quentin W. Smith
Leslie M. Miller
William A. Thomson

Baylor College of Medicine

Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, January 25, 1991.

ABSTRACT

The present paper discusses and illustrates the problems with using stepwise analytic methods and illustrates better alternatives to these methods. To make the illustrations concrete, an actual data set involving responses by 91 subjects to 30 variables is employed. Though data illustrating the problems with stepwise methods in a more dramatic fashion can be formulated, these data have the appeal of being real. In any case, the emphasis here is on better alternatives to stepwise methods, as against the problems, since the problems with these methods have been so fully elaborated elsewhere. A two-stage approach to variable selection is recommended. Problems with using statistical significance testing in conjunction with stepwise methods are also elaborated in some detail.

As Huberty (1989, p. 43) notes,

> The conduct of analytical procedures in "steps" is quite common... Although regression analysis and discriminant analysis problems are, without a doubt, the most popular contexts for the use of step-type computational algorithms, these approaches have also been suggested in multivariate analysis of variance (Stevens, 1973) and in canonical correlation analysis (Thompson, 1984, pp. 47-51; Thorndike & Weiss, 1983).

Various researchers have emphatically criticized the use of conventional stepwise methods (e.g., Huberty, 1989; Huberty & Wisenbaker, in press; Snyder, 1991; Thompson, 1988b, 1989).

Three major criticisms have been presented. First, conventional stepwise methods dramatically inflate Type I error rates. Snyder (1991) presents an impressive concrete example of how strongly stepwise methods can be influenced by sampling error. One reason why stepwise methods "are positively satanic in their temptation toward Type I errors" (Cliff, 1987, p. 185) involves the fact that computer programs use the wrong denominator degrees of freedom and sum-of-squares in their calculations. Indeed, so do most books, with the notable exception of Keppel and Zedeck's (1989, pp. 402-405) recent offering, as suggested by Thompson (in press).

Second, the variables identified after k steps of analysis may not include all or even any of the variables in the best predictor

set of size k. For example, in a problem involving 10 predictor variables, variables A and B may be entered in the first two steps, but the best predictor set of size k=2 for the same data may well be variables C and D. Third, order of entry provides very limited information regarding the relative importance of the variables, as Huberty (1989) explains in more detail.

The purpose of the present paper is to discuss and illustrate these problems and to illustrate better alternatives. Those who would like more detail regarding these issues are urged to consult Snyder (1991). To make the illustrations concrete, an actual data set involving responses by 91 subjects to 30 variables is employed. Though data that illustrate the problems with stepwise methods in a more dramatic fashion can be formulated, these data have the appeal of being real. In any case, the emphasis here is on alternatives to stepwise methods, as against the problems, since the problems with these methods have been so fully elaborated elsewhere.

Discriminant analysis is used as the analytic method in the present heuristic example. However, all analytic methods are correlational and are related (Knapp, 1978; Thompson, 1988a), and therefore the present discussion generalizes to other stepwise methods, e.g., stepwise regression as well.

The Heuristic Data Set

The 30 variables involved perceptions of barriers to education in medical schools with respect to medical students' characteristics. The variables were developed using a delphi study,

an approach that has proven useful in previous instrument development efforts (e.g., Lester & Thomson, 1989; Thomson & Ponder, 1979). The delphi process that ultimately produced the 30 items involved a national invitational conference; 34 professionals participated based on being nominated by one of the four sponsoring organizations (the Josiah Macy, Jr., Foundation, the Rockefeller Foundation, the Robert Wood Johnson Foundation, and the Baylor College of Medicine) as individuals who had had significant leadership roles in promoting minority citizens' access to careers in the health professions. The delphi study resulting in the isolation of the 30 items is described in Baylor College of Medicine (1986).

For the purposes of the present study admissions officials at 144 medical schools in the United States and Canada were asked to rate extent of agreement with the 30 statements using 1 to 5 (5 = strong agreement) Likert scales. Admissions officials from 58 schools (40.3%) returned completed questionnaires after the first mailing, and an additional 33 officials (22.9%) completed questionnaires sent in a follow-up mailing to the admissions officials at the 86 schools not responding to the first mailing. Thus, representatives from 91 out of 144 medical schools completed questionnaires, and the response rate was 63.2%. This response rate was considered acceptable, especially given that the average response rate in survey research is typically about 33% (Kerlinger, 1986).

The 91 admissions officials rated extent of agreement that

each of the 30 statements involved problems encountered by elementary and secondary educators working with each of four referent student populations: (a) black students, (b) hispanic students, (c) other minority students, and (d) non-minority students. The focus of the analysis in the present study was on explaining variance in perceptions of the four referent groups.

Initially items not explaining an appreciable portion of variance in perceptions of the four referent groups were deleted. The 10 variables with the smallest effect sizes (expressed as Wilks' lambda or one minus the ANOVA correlation ratio) with respect to discriminating the four referent groups were omitted: variables 29, 24, 16, 4, 21, 14, 15, 26, 20, and 30, respectively. The mean lambda for these 10 variables was 98.1% (SD=1.6%). Thus, on these average the four referents explained only about 2% of the variance in each of these 10 predictors.

Classical Stepwise Results

The first analysis involved conducting a conventional stepwise discriminant analysis with the four referent student groups as the dependent variable and the 20 remaining statements as predictor variables. Variables were only added in this stepwise analysis for these data (for some data the stepwise algorithm will also delete variables at certain steps), and 13 variables were entered: variables 1, 7, 2, 8, 10, 17, 6, 22, 12, 13, 18, 25, and 5, respectively. The F-to-enter for each of the remaining seven variables were all less than one, so the improvement in the model resulting from adding any of these variables would not have been

statistically sigrificant. In conventional fashion, the stepwise analysis was terminated at this point.

Table 1 presents the variables entered at each of the 13 steps, and the associated lambda effect sizes. Lambda is similar to r^2 in that it is an effect size function, and ranges between zero and one. However, the largest effect size for r^2 is one, while the largest effect size for lambda is zero, i.e., the two estimates are inversely related.

INSERT TABLE 1 ABOUT HERE.

A Better Alternative to Stepwise

If one must eliminate variables, a better procedure than stepwise is to compute the effect size for *every possible* predictor set of size k=1, size k=2, size k=3, and so forth. This sounds tedious, but can be done rapidly, accurately, and painlessly by readily available computer software. In the multiple regression case, the SAS program PROC RSQUARE is available. For discriminant analysis applications, the FORTRAN program written by McCabe (1975) is available.

Table 2 presents the lambda effect size for the best predictor variable combinations for sizes k = 1 through 12, as computed by McCabe's program. The program actually presents the lambdas for several variable combinations at various values of k, but only the best combination for each size is presented here.

INSERT TABLE 2 ABOUT HERE.

For example, the best predictor set of size k = 9 included variables 1, 7, 2, 8, 10, 17, 6, 13, and 12 (lambda = .4100). This example makes clear that stepwise methods do not isolate the best predictors for a given variable set size k, since variable 22 is entered in the eighth step of the stepwise analysis, but is not part of the best predictor set of size k = 9. That is, the stepwise analysis wrongly indicates that the best predictor set of size k = 9 includes variables 1, 7, 2, 8, 10, 17, 6, 22, and 12.

The better selection of predictors is made in a two-stage process in which *variable selection is not conditioned upon the results in previous steps*. Stepwise results are conditioned in this manner, e.g., if variable 6 had not been entered in step 7, then a variable other than 22 might well have been selected in step 8. Stepwise methods have the disadvantage of being tied to the limited context of the variables in the study and previously entered in the analysis. This limits the generalizability of conclusions, since the results are conditional upon the context of previous entries.

The first step of the procedure endorsed here is to initially determine the desired size, k, of the predictor variable set This can be done by computing the changes in lambda (or in R^2 in regression analysis) as new predictors are added, or by plotting lambda in a "scree" plot fashion, as illustrated in Figure 1. For the data in hand the optimal predictor set size appears to be size k = 7, since the addition of other variables results in relatively

negligible contributions to predictive power.

INSERT FIGURE 1 ABOUT HERE.

The second step in the two-stage process is to select the best predictor set of the selected size, k. The effect size should be consulted for this purpose. However, this tends to be just as "atheoretical" and "mechanical" as conventional stepwise methods (Keppel & Zedeck, 1989, pp. 398, 407). A better approach is to then select the predictor set based on theory or previous empirical results, or based on the accessibility of the variables in a given set.

Summary

Cliff (1987, pp. 120-121) notes that "a large proportion of the published results using this [stepwise] method probably present conclusions that are not supported by the data." As conventionally applied in regression and discriminant analysis, stepwise applications usually create serious problems.

One particular problem with stepwise analyses involves the propensity of researchers to apply statistical significance tests to evaluate how many steps to implement, as in the 13 step example presented here. Additional problems with statistical significance tests have been elaborated in detail elsewhere (Carver, 1978; Thompson, 1987; Welge-Crow, LeCluyse & Thompson, 1990), but this aspect of the problem may warrant some explanation.

Science is the business of creating and cumulating knowledge. This becomes possible only when results are reasonably

commensurable across studies. The problem can be illustrated with a series of hypothetical regression studies, each involving four predictor variables.

Say four researchers conduct identical studies, but with three separate samples of subjects, each varying in size (n_1 = 100, n_2 = 18, n_3 = 16, n_4 = 15). Say also that the researchers have <u>exactly identical</u> results with respect to the bivariate correlation matrices from which the regression results are extracted. Table 3 presents results that fit this description.

INSERT TABLE 3 ABOUT HERE.

For the Table 3 data, researcher one will conduct four steps of analysis, interpreting results involving an effect size of R^2 = 70% (F = 54.42, df = 4/95, $p < .05$) and predictors A, B, C, and D. Researcher two will conduct three steps of analysis, interpreting results involving an effect size of R^2 = 60% (F = 7.00, df = 3/14, $p < .05$) and predictors A, B, and C. Researcher three will conduct two steps of analysis, interpreting results involving an effect size of R^2 = 45% (F = 5.32, df = 2/13, $p < .05$) and predictors A and B. Researcher four will conduct one step of analysis, and conclude that an effect size of R^2 = 45% (F = 4.67, df = 1/13, $p > .05$) is not statistically significant and that <u>no</u> predictors are useful.

Yet all these divergent interpretations are based on exactly the same correlation matrix, and emerge solely as an artifact of the use of statistical significance testing in conjunction with

stepwise analysis! "Unfortunately," as Pedhazur (1982, p. 168) notes, "social science research is replete with misinterpretations of this kind."

References

Baylor College of Medicine. (1986). Enhancing opportunities in science, mathematics, and the health professions: Invitational conference. Houston: Baylor College of Medicine.

Carver, R. P. (1978). The case against statistical significance testing. Harvard Educational Review, 48(3), 378-399.

Cliff, N. (1987). Analyzing multivariate data. San Diego: Harcourt, Brace, and Jovanovich.

Huberty, C.J. (1989). Problems with stepwise methods--better alternatives. In B. Thompson (Ed.), Advances in social science methodology (Vol. 1, pp. 43-70). Greenwich, CT: JAI Press.

Huberty, C.J, & Wisenbaker, J.M. (in press). Discriminant analysis: Potential improvements in typical practice. In B. Thompson (Ed.), Advances in social science methodology (Vol. 2). Greenwich, CT: JAI Press.

Keppel, G., & Zedeck, S. (1989). Data analysis for research designs. New York: Freeman.

Kerlinger, F.N. (1986). Foundations of behavioral research (3rd ed.). New York: Holt, Rinehart and Winston.

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance testing system. Psychological Bulletin, 85, 410-416.

Lester, R.C., & Thomson, W.A. (1989). Perceptions of CRNAs: Current and future roles--part II. Journal of the American Association of Nurse Anesthetists, 57, 417-425

McCabe, G.P., Jr. (1975). Computations for variable selection in

discriminant analysis. Technometrics, 17, 103-109.

Pedhazur, E. (1982). Multiple regression in behavioral research: Explanation and prediction. New York: Holt, Rinehart and Winston.

Snyder, P. (1991). Three reasons why stepwise regression methods should not be used by researchers. In B. Thompson (Ed.), Advances in educational research: Substantive findings, methodological developments (Vol. 1, pp. 99-106). Greenwich, CT: JAI Press.

Thompson, B. (1987, April). The use (and misuse) of statistical significance testing: Some recommendations for improved editorial policy and practice. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C. (ERIC Document Reproduction Service No. ED 287 868)

Thompson, B. (1988a, April). Canonical correlation analysis: An explanation with comments on correct practice. Paper presented at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. ED 295 957)

Thompson, B. (1988b, November). Common methodology mistakes in dissertations: Improving dissertation quality. Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville, KY. (ERIC Document Reproduction Service No. ED 301 595)

Thompson, B. (1989). Why won't stepwise methods die?. Measurement

and Evaluation in Counseling and Development, 21(4), 146-148.

Thompson, B. (in press). Review of Data analysis for research designs by G. Keppel & S. Zedeck. Educational and Psychological Measurement.

Thomson, W.A., & Ponder, L.D. (1979). Use of delphi methodology to generate a survey instrument to identify priorities for state allied health associations. Allied Health and Behavioral Sciences, 2, 383-399.

Welge-Crow, P., LeCluyse, K., & Thompson, B. (1990, June). Looking beyond statistical significance: Result importance and result generalizability. Paper presented at the annual meeting of the American Psychological Society, Dallas.

Table 1
Stepwise Discriminant Analysis Results
(Y = 20)

Step	Variable Added	lambda
1	Q1	.728
2	Q7	.603
3	Q2	.542
4	Q8	.497
5	Q10	.467
6	Q17	.439
7	Q6	.426
8	-- Q22 --	.417
9	Q12	.411
10	Q13	.402
11	Q18	.397
12	Q25	.392
13	Q5	.388

Table 2
The Best Predictors for Variables Sets of Size k = 1 to 12

lambda	delta	% delta	n Var	Variables Selected
.7281	.7281	100.00%	1	1
.6033	.1248	17.14%	2	1 7
.5416	.0617	10.23%	3	1 7 2
.4971	.0445	8.22%	4	1 7 2 8
.4670	.0300	6.05%	5	1 7 10 2 8
.4393	.0277	5.92%	6	1 7 10 2 8 17
.4263	.0130	2.96%	7	1 7 6 10 2 8 17
.4168	.0095	2.24%	8	1 7 6 10 2 8 22 17
.4100	.0068	1.64%	9	1 7 6 10 12 2 8 17 13
.4023	.0077	1.87%	10	1 7 6 10 12 2 8 22 17 13
.3973	.0050	1.25%	11	1 7 6 10 12 2 8 22 17 13 18
.3923	.0050	1.25%	12	1 7 6 10 12 2 8 22 17 25 13 18

Table 3

Bivariate r Matrix and Stepwise F's for Four Sample Sizes

	Y		A		B		C		D
Y	1.000								
A	0.500	25.0%	1.000						
B	0.447	20.0%	0.000	0.0%	1.000				
C	0.387	15.0%	0.000	0.0%	0.000	0.0%	1.000		
D	0.316	10.0%	0.000	0.0%	0.000	0.0%	0.000	0.0%	1.000

$$F_1 = \frac{(R^2_2 - R^2_1) / (k2 - k1)}{(1 - R^2_2) / (n - k2 - 1)} = \frac{(.25 - .00) / (1 - 0)}{(1 - .25) / (100 - 1 - 1)} =$$

$$\frac{(.25) / (1)}{(.75) / (98)} = \frac{(.25)}{(.007653)}$$

$$= 32.6666 > c.\ 3.94 \quad \therefore p < .05$$

$$\frac{(.45 - .25) / (2 - 1)}{(1 - .45) / (100 - 2 - 1)} = 35.2727 > c.\ 3.94 \quad \therefore p < .05 \quad \underline{df} = 1/97$$

$$\frac{(.60 - .45) / (3 - 2)}{(1 - .60) / (100 - 3 - 1)} = 36.0000 > c.\ 3.94 \quad \therefore p < .05 \quad \underline{df} = 1/96$$

$$\frac{(.70 - .60) / (4 - 3)}{(1 - .70) / (100 - 4 - 1)} = 31.6666 > c.\ 3.94 \quad \therefore p < .05 \quad \underline{df} = 1/95$$

$$F_2 = \frac{(.25 - .00) / (1 - 0)}{(1 - .25) / (18 - 1 - 1)} = 5.3333 > 4.49 \quad \therefore p < .05 \quad \underline{df} = 1/16$$

$$\frac{(.45 - .25) / (2 - 1)}{(1 - .45) / (18 - 2 - 1)} = 5.4545 > 4.54 \quad \therefore p < .05 \quad \underline{df} = 1/15$$

$$\frac{(.60 - .45) / (3 - 2)}{(1 - .60) / (18 - 3 - 1)} = 5.2500 > 4.60 \quad \therefore p < .05 \quad \underline{df} = 1/14$$

$$\frac{(.70 - .60) / (4 - 3)}{(1 - .70) / (18 - 4 - 1)} = 4.3333 < 4.67 \quad \therefore p > .05 \quad \underline{df} = 1/13$$

$$F_3 = \frac{(.25 - .00) / (1 - 0)}{(1 - .25) / (16 - 1 - 1)} = 4.6667 > 4.60 \quad \therefore p < .05 \quad \underline{df} = 1/14$$

$$\frac{(.45 - .25) / (2 - 1)}{(1 - .45) / (16 - 2 - 1)} = 4.7273 > 4.67 \quad \therefore p < .05 \quad \underline{df} = 1/13$$

$$\frac{(.60 - .45) / (3 - 2)}{(1 - .60) / (16 - 3 - 1)} = 4.5000 < 4.75 \quad \therefore p > .05 \quad \underline{df} = 1/12$$

$$\frac{(.70 - .60) / (4 - 3)}{(1 - .70) / (16 - 4 - 1)} = 3.6667 < 4.84 \quad \therefore p > .05 \quad \underline{df} = 1/11$$

$$F_4 = \frac{(.25 - .00) / (1 - 0)}{(1 - .25) / (15 - 1 - 1)} = 4.3333 < 4.67 \quad \therefore p > .05 \quad \underline{df} = 1/13$$

$$\frac{(.45 - .25) / (2 - 1)}{(1 - .45) / (15 - 2 - 1)} = 4.3636 < 4.75 \quad \therefore p > .05 \quad \underline{df} = 1/12$$

$$\frac{(.60 - .45) / (3 - 2)}{(1 - .60) / (15 - 3 - 1)} = 4.1250 < 4.84 \quad \therefore p > .05 \quad \underline{df} = 1/11$$

$$\frac{(.70 - .60) / (4 - 3)}{(1 - .70) / (15 - 4 - 1)} = 3.3333 < 4.96 \quad \therefore p > .05 \quad \underline{df} = 1/10$$

Note. From Thompson (1991), with permission.

Figure 1
Plot of Δlambda in "scree" Form