DOCUMENT RESUME

ED 382 635 TM 023 064

AUTHOR Thompson, Bruce
TITLE Stepwise Regression and Stepwise Discriminant Analysis Need Not Apply.
PUB DATE 20 Apr 95
NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Educational Research; *Error of Measurement; Heuristics; *Psychological Testing; *Regression (Statistics); *Research Methodology; Sampling
IDENTIFIERS Research Replication; *Stepwise Regression

ABSTRACT

Stepwise methods are frequently employed in educational and psychological research, both to select useful subsets of variables and to evaluate the order of importance of variables. Three problems with stepwise applications are explored in some detail. First, computer packages use incorrect degrees of freedom in their stepwise computations, resulting in artifactually greater likelihood of obtaining spurious statistical significance. Second, stepwise methods do not correctly identify the best variable set of a given size, as illustrated by a concrete heuristic example. Third, stepwise methods tend to capitalize on sampling error, and thus tend to yield results that are not replicable. (Contains 22 references, 4 tables, and 1 figure.) (Author)

stepbad.wp1

STEPWISE REGRESSION AND STEPWISE DISCRIMINANT ANALYSIS NEED NOT APPLY

Bruce Thompson

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

Paper presented at the annual meeting of the American Educational Research Association (training session #25.16), San Francisco, April 20, 1995.

Abstract

Stepwise methods are frequently employed in educational and psychological research, both to select useful subsets of variables and to evaluate the order of importance of variables. Three problems with stepwise applications are explored in some detail. First, computer packages use incorrect degrees of freedom in their stepwise computations, resulting in artifactually greater likelihood of obtaining spurious statistical significance. Second, stepwise methods do not correctly identify the best variable set of a given size, as illustrated by a concrete heuristic example. Third, stepwise methods tend to capitalize on sampling error, and thus tend to yield results that are not replicable.

It is the practice within Educational and Psychological Measurement and other journals to present occasional supplementary guidelines for authors that complement general APA style requirements. For example, Thompson (1994b) discussed requirements involving both statistical significance testing and language regarding score reliability. The present paper focuses on major problems with stepwise analyses, and suggests that these methods ought to be avoided in favor of more suitable alternatives.

Huberty (1994) recently noted that, "It is quite common to find the use of 'stepwise analyses' reported in empirically based journal articles" (p. 261). However, various authors have presented scathing indictments of many of these applications (cf. Huberty, 1989; Snyder, 1991; Thompson, 1989). Three major problems can be noted.

The heuristic examples employed here to illustrate these three problems involve stepwise regression analysis. However, since all commonly applied analytic methods are correlational (Cohen, 1968), and are special cases of canonical correlation analysis (Knapp, 1978; Thompson, 1991), the present discussion generalizes across the full family of these various applications.

Some researchers employ stepwise methods to select a subset of better variables from among a larger constellation of predictors, for use in present or future research (i.e., so-called "variable selection"). The methods are also sometimes used to interpret data dynamics, under a premise that selected variables are more important than predictors that are not selected, or that entry order reflects variable importance (i.e., so-called "variable ordering"). Stepwise methods are not usually useful for either purpose.

Horrendously Wrong Degrees of Freedom

Problem

Degrees of freedom in statistical analyses reflect the number of unique pieces of information present for a given research situation. These degrees of freedom constrain the number of inquiries we may direct at our data, and are the currency we spend in analysis.

Regrettably, commonly used statistical packages incorrectly compute the degrees of freedom in stepwise analyses. The use of incorrect degrees of freedom in practice often has dire consequences as regards the accuracy of our inferences.

Table 1 presents an illustration. Presume that we have data from 101 subjects on a dependent variable ("Y") and 50 predictor variables. After five steps of stepwise regression analysis, the five entered predictor variables may "explain" 20% of the variability in the Y scores (i.e., 20/100 = 20% = R^2), as illustrated in Table 1.

INSERT TABLE 1 ABOUT HERE.

Computer packages compute the degrees of freedom correctly, as n-1. However, the degrees of freedom "explained" (also variously called "model", "regression", "between", etc.) is computed as the number of "entered" predictor variables (i.e., p_Y). The degrees of freedom "unexplained" (also variously called "error", "residual", "within", etc.) is then computed as n-1-pv. These calculations yield a statistically significant (α=.05) result in the Table 1 illustration.

However, various researchers (cf. Snyder, 1991) have correctly noted that these degrees of freedom calculations for the explained and unexplained variance partitions are simply wrong. If the five entered predictor variables had been randomly selected, an explained degrees of freedom of 5 might be arguably correct.

But our five predictors were selected by, at each step, looking at the results for all the predictor variables not yet entered! Viewed differently, at each step all 50 predictors variables were entered, though we may have constrained the b and β weights for most of the predictors to be 0 at each step (Cliff, 1987, p. 187). Thus, the computer packages are erroneously not charging us any degrees of freedom for consulting our data in this manner.

This statistical welfare system may cause us to radically overestimate the atypicality of our results, i.e., create an artifactually small $p_{CALCULATED}$. Table 1 dramatically illustrates how the use of the incorrect degrees of freedom can (a) radically inflate $MS_{EXPLAINED}$, (b) radically deflate $MS_{UNEXPLAINED}$, and consequently (c) very radically inflate $F_{CALCULATED}$ (e.g., 4.75 versus 0.25). No wonder Cliff (1987, p. 185) noted that "most computer programs for [stepwise] multiple regression are positively satanic in their temptations toward Type I errors."

Caveats

Of course, it is important in evaluating statistical practices not to make what in logic is termed an "is/ought" or a "should/would" error (Hudson, 1969; Hume, 1957). As Strike (1979) explains,

> To deduce a proposition with an "ought" in it from premises containing only "is" assertions is to get something in the conclusion not contained in the premises, something impossible in a valid deductive argument. (p. 13)

The fact that most researchers "are" using the wrong degrees of freedom in their stepwise analyses does not mean that we therefore "should" abandon these methods. Instead, logically we ought simply to use the correct degrees of freedom.

We need not even somehow persuade the software companies to fix their computer programs; we need only use the printed sums-of-squares instead with the correct degrees of freedom we derive ourselves to then recalculate the remaining statistical tests. Doing so merely requires a willingness to believe that computer programs are not infallible, because computer programs were written by fallible people and not by higher beings.

It is important to note that all stepwise applications are not equally evil as regards the inflation of Type I error. For example, the stepwise results after one step for a problem involving only two predictors might not be so seriously distorted. Some readers may protest that no one would ever invoke stepwise methods with a small number of predictor variables. However, a colleague only a few days ago described a manuscript for which he was serving as a referee, and in that study submitted to a prominent national journal the authors conducted several dozen stepwise methods for problems each involving only three predictor variables!

The seriousness of problems with wrong degrees of freedom being used, as with most statistical (and life) issues, is situationally conditional. Stepwise methods will be somewhat less evil, for example, when (a) the sample size is very large, (b) the number of predictor variables is small, and/or (c) the sum of squares explained remains near zero across steps.

<u>Does Not Identify the Best Predictor Set of Size "q"</u>

<u>Problem</u>

Unfortunately, many researchers erroneously believe that conducting two or five steps of analysis will identify the best predictor set of size two or five. This simply is not what stepwise methods typically do.

Ignoring for present purposes the variable deletion aspect of a true stepwise analysis, at step number five forward stepwise methods address the question, "Given the four predictors already entered, which one additional predictor will most improve the analysis?". Thus, the question is conditioned on the presence of the first four predictors, and yields a *situation-specific conditional* answer in the context (a) only of the specific variables already entered and (b) only those variables used in the particular study but not yet entered.

If the first variable entered was different, so the variable entered in the remaining steps might differ. Furthermore, even if the first four entered variables remained constant, deleting or adding predictors from the study certainly might also yield a different answer to the context-specific stepwise question.

But if we wish to determine the best set of predictor variables of size *q*, the question, "what is the best set of $q=5$ predictors?", does <u>not</u> ask a conditional question invoking a linear sequence of variable entry. Of course, if we desire this second question to be answered, it is not reasonable to invoke the answer to a question one is not posing!

Thus, the five predictors entered in five steps of forward entry will not typically answer the question as to what are the best $q=5$ predictors, and it is even conceivable that *none* of the five variables selected by stepwise will be included in the best subset of five predictors.

Figure 1 presents the Venn diagram of a heuristic example to make this dynamic concrete. Since Venn diagrams are two-dimensional representations of multi-dimensional phenomena, they must be interpreted as only figurative portrayals of simultaneous relationships among three or more variables (Craeger, 1969). However, bivariate relationships can be literally presented in this manner.

INSERT FIGURE 1 ABOUT HERE.

The example involves a dependent variable, <u>Y</u>, and four predictor variables. Table 2 presents sums-of-squares variance partitions associated with Figure 1, e.g., X_1 explains 100 of the 400 sums-of-squares units associated with the individual differences (i.e., variability) in the $\underline{Y}$ scores. Table 3 translates the sums of squares into correlation coefficients.

INSERT TABLES 2 AND 3 ABOUT HERE.

Table 4 presents the regression analyses for the data. If a stepwise analysis was conducted, predictor X_1 would be entered first, because this variable has the largest squared bivariate correlation (r^2 = 25%) with $\underline{Y}$. In the second step, predictor X_2 would be entered, and the resulting R^2 would be 45.00%.

INSERT TABLE 4 ABOUT HERE.

However, if an all-possible-subsets analysis is conducted with the same data, the best predictor set of size q=2 is determined to be predictors X_2 and X_4, with an R^2 of 47.5%. The best predictor set of size q=2 does not include either of the two predictors entered in the two steps of the stepwise analysis!

Caveats

Again, few behaviors either in life or in statistics are always wrong. Some behaviors are only usually wrong, and we have to think about whether special exceptions have arisen. This is what makes teaching methodology so difficult--we must teach our students to think rather than only to memorize universal principles of lock-step rote behaviors.

First, our two questions ("which one additional predictor...?"

and "what is the best set...?") are logically equivalent when we are investigating the subset, q=1. Stepwise analysis does correctly identify the best single predictor.

Second, the two types of analyses do yield the same answers whenever the predictors are perfectly uncorrelated. This occurs when we use orthogonally-rotated principal components scores in an analysis, for example. Of course, 30 steps of stepwise with such predictors tells us nothing we don't already know, if we already know the 30 correlation coefficient involving Y and each of the 30 uncorrelated component scores.

Tendency to Yield Non-replicable Results

Problem

Stepwise methods tend to yield conclusions that will not replicate in future research. This is because stepwise methods tend to capitalize outrageously on sampling error. Sampling error is variability in sample data that is unique to the given sample, and therefore cannot be reproduced in subsequent samples. Snyder (1991) presents an excellent heuristic example of these dynamics.

At a given step, the determination of which single variable to enter will enter variable X_1 over variables X_2, X_3, and X_4, even if X_1 is only infinitesimally superior to the other three variables. It is entirely possible that this infinitesimal advantage of variable X_1 over another variable is sampling error, given that the competitive advantage of X_1 is so small.

Stepwise analysis is a linear series of conditional decisions, not unlike the choices one makes in working through a maze. An early mistake in the sequence will corrupt the remaining choices. If X_1 is incorrectly entered first in the analysis due to an infinitesimal advantage representing only a small amount of sampling error, all remaining conditional entry decisions may also therefore be incorrect.

Since small differences may reflect sampling error, but these small differences can greatly effect the sample results, stepwise sample results often do not generalize. Thus, Cliff (1987, pp. 120-121) suggested that, "a large proportion of the published results using this method probably present conclusions that are not supported by the data."

Caveats

Obviously, less sampling error tends to be present in data sets involving (a) larger samples, (b) fewer predictor variables, and (c) larger effect sizes, as reflected in the factors involved in most statistical corrections for positive bias in uncorrected variance-accounted-for effect sizes (Snyder & Lawson, 1933; Thompson, 1990). Thus, use of stepwise methods in these circumstances might be somewhat less sinful. And again, if the predictor variables are uncorrelated, the analysis is not distorted by the sampling error in the relationships among the predictors.

Summary

Stepwise methods do not do what most researchers believe the methods do. Stepwise methods are especially problematic when statistical significance tests are invoked to determine stopping positions, because the methods have all the problems associated with conventional statistical significance applications (Carver, 1978; Cohen, 1994; Thompson, 1993, 1994a, 1994b, 1994c), in spades.

As a general proposition, there are readily available software programs to assist with appropriate *variable selection* efforts by conducting almost instantly-available and painless all-possible-subsets analyses. Thus, stepwise analyses should be eschewed in favor of programs such as those offered by McCabe (1975),the Morris program distributed within Huberty's (1994) book, or SAS procedure RSQR. As regards interpretations involving the origins of explained variance, i.e., *variable ordering*, a useful alternative is simply to consult standardized weights (called different names across analyses to confuse graduate students, e.g., beta weights, factor pattern coefficients, standardized discriminant function coefficients) and structure coefficients (Thompson & Borrello, 1985). Huberty (1994) summarizes a variety of other helpful variable ordering strategies for the discriminant analysis case.

References

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

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

Cohen, J. (1968). Multiple regression as a general data-analytic system. Psychological Bulletin, 70, 426-443.

Cohen, J. (1994). The earth is round ($p < .05$). American Psychologist, 49, 997-1003.

Craeger, J. (1969). The interpretation of multiple regression via overlapping rings. American Educational Research Journal, 6, 706-709.

Huberty, C. (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. (1994). Applied discriminant analysis. New York: Wiley and Sons.

Hudson, W.D. (1969). The is/ought question. London: MacMillan.

Hume, D. (1957). An inquiry concerning human understanding. New York: The Liberal Arts Press.

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

McCabe, G. P. (1975). Computations for variable selection in discriminant analysis. Technometrics, 17, 103-109.

Snyder, P. (1991). Three reasons why stepwise regression methods should not be used by researchers. In B. Thompson (Ed.), (1991). Advances in educational research: Substantive findings, methodological developments (Vol. 1, pp. 99-105). Greenwich, CT: JAI Press.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. Journal of Experimental Education, 61, 334-349.

Strike, K.A. (1979). An epistemology of practical research. Educational Researcher, 8(1), 10-16.

Thompson, B. (1989). Why won't stepwise methods die?. Measurement and Evaluation in Counseling and Development, 21(4), 146-148.

Thompson, B. (1990). Finding a correction for the sampling error in multivariate measures of relationship: A Monte Carlo study. Educational and Psychological Measurement, 50, 15-31.

Thompson, B. (1991). A primer on the logic and use of canonical correlation analysis. Measurement and Evaluation in Counseling and Development, 24(2), 80-95.

Thompson, B. (Ed.). (1993). Special issue on statistical significance testing, with comments from various journal editors, Journal of Experimental Education, 61(4).

Thompson, B. (1994a). The concept of statistical significance testing (An ERIC/AE Clearinghouse Digest #EDO-TM-94-1). Measurement Update, 4(1), 5-6. (ERIC Document Reproduction Service No. ED 366 654)

Thompson, B. (1994b). Guidelines for authors. Educational and Psychological Measurement, 54, 837-847.

Thompson, B. (1994c). The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results. Journal of Personality, 62(2), 157-176.

Thompson, B., & Borrello, G.M. (1985). The importance of structure coefficients in regression research. Educational and Psychological Measurement, 45, 203-209.

Table 1
Hypothetical Five-Step Regression Model
With 101 Subjects and 50 Predictor Variables

Analysis	Source	SOS	df	MS	*F*calc	*F*crit	R^2
1	Explained	20	5	4.0000	4.75	4.41	20.00%
	Unexplained	80	95	0.8421			
	Total	100	100				
2	Explained	20	50	0.4000	0.25	***[a]	20.00%
	Unexplained	80	50	1.6000			
	Total	100	100				

[a]Since *F*critical at infinite and infinite degrees of freedom equals 1, an *F*calculated less than 1 can not be statistically significant.

step.wk1 3/22/95

Table 2
Variance Partitions of the Predictive
Abilities of the Four Predictor Variables

Single Partitions		Partitions in Combinations								
Partition	SOS	Predictor		Partitions						Total
A	20	X_1	=	E	+	F	+	G		
B	50		=	21	+	49	+	30	=	100
C	27	X_2	=	B	+	C	+	D		
D	3		=	50	+	27	+	3	=	80
E	21	X_3	=	A	+	B	+	E		
F	49		=	20	+	50	+	21	=	91
G	30	X_4	=	D	+	G	+	H		
H	66		=	3	+	30	+	66	=	99

Table 3
Pairwise r Values

Variable Pair	Common SOS	r^2	r
X_1, X_2	0	.0000	.0000
X_1, X_3	30	.0750	.2739
X_1, X_4	60	.1500	.3873
X_1, Y	100	.2500	.5000
X_2, X_3	185	.4625	.6801
X_2, X_4	3	.0075	.0866
X_2, Y	80	.2000	.4472
X_3, X_4	0	.0000	.0000
X_3, Y	91	.2275	.4770
X_4, Y	99	.2475	.4975

Note. r^2 = Common SOS / 400. For example, $r^2_{X1,Y}$ = 100/400 = +.2500, while $r_{X1,Y}$ = the square root of $r^2_{X1,Y}$ = the square root of +.2500 = +.5000.

Table 4
Calculation of β's and R^2's for the
Six Pairwise Combinations of the Four Predictors

Predictors		r1	r2	rxx	β	β(r1) + β(r2) = R^2
1,2	1	.5000	.4472	.0000	.5000	
	2	.4472	.5000	.0000	.4472	.2500 + .2000 = .4500
1,3	1	.5000	.4770	.2739	.3993	
	3	.4770	.5000	.2739	.3676	.1997 + .1753 = .3750
1,4	1	.5000	.4975	.3873	.3616	
	4	.4975	.5000	.3873	.3575	.1808 + .1778 = .3586
2,3	2	.4472	.4770	.6801	.2285	
	3	.4770	.4472	.6801	.3215	.1022 + .1534 = .2556
2,4	2	.4472	.4975	.0866	.4072	
	4	.4975	.4472	.0866	.4622	.1821 + .2300 = .4121
3,4	3	.4770	.4975	.0000	.4770	
	4	.4975	.4770	.0000	.4975	.2275 + .2475 = .4750

Note. β = (r1 - (r2 * rxx)) / (1 - rxx^2). For example, for predictor pair X_1 and X_3, β_1 =

(.5000 - (.4770 * .2739)) / (1 - $.2739^2$)
(.5000 - .1306) / (1 - .0750)
.3694 / .9250 = .3993

R^2 = β(r1) + β(r2). For example, for predictor pair X_1 and X_3, R^2 =

(.3993 * .5000) + (.3676 * .4770)
.1997 + .1753 = .3750

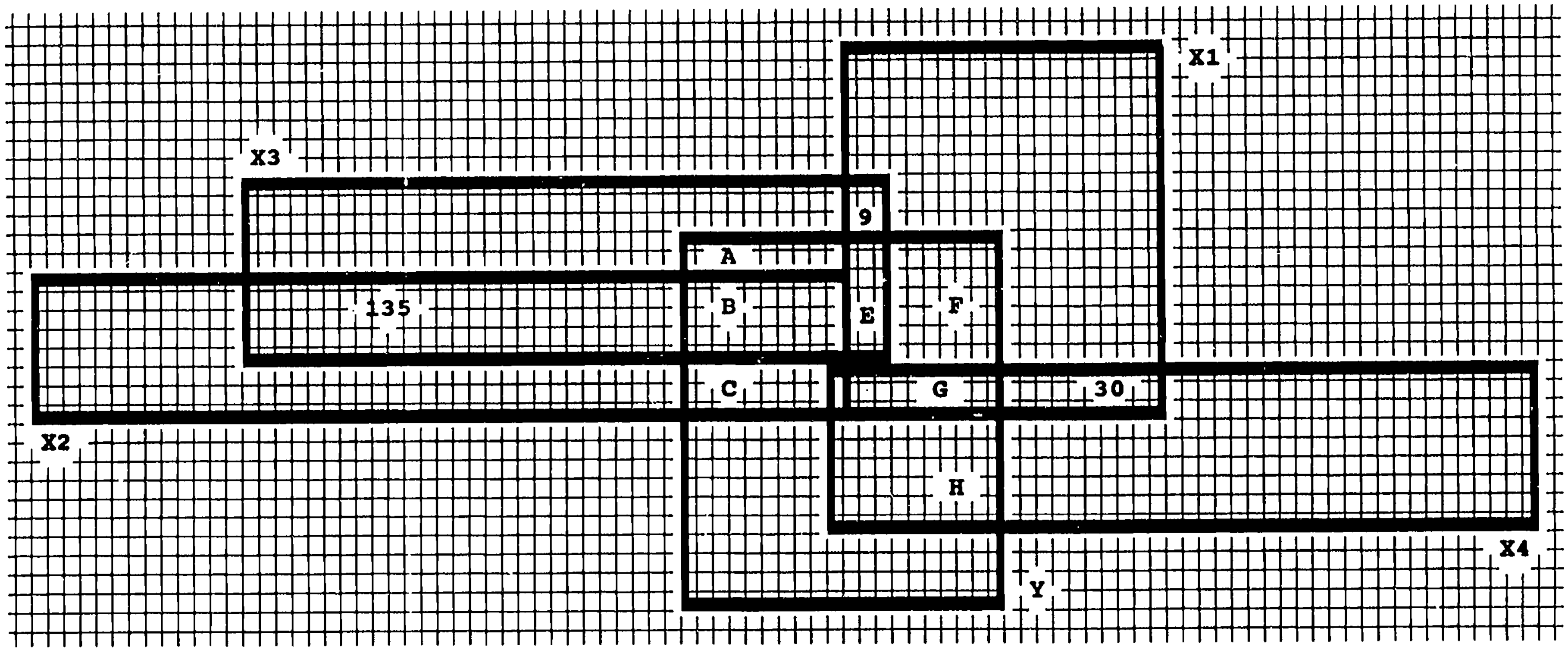
step6666.wk1 3/24/95

graph666.wp1

Figure Caption.

Figure 1

Venn Diagram of Relationships Among Five Variables



Note. The common areas of the four predictor variables with Y are labelled with letters. The areas out of the 400 sums-of-squares for Y that are common are:

A = 20 B = 50 C = 27 E = 21 F = 49 G = 30 H = 66.

The area common to Y, X_2 and X_4 is designated D, and D = 3.