

DOCUMENT RESUME

ED 275 730

TM 860 644

AUTHOR Thayer, Jerome D.
TITLE Testing Different Model Building Procedures Using Multiple Regression.
PUB DATE Apr 86
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; Computer Simulation; *Mathematical Models; *Multiple Regression Analysis; *Predictor Variables; *Regression (Statistics); *Research Design; Statistical Studies

ABSTRACT

The stepwise regression method of selecting predictors for computer assisted multiple regression analysis was compared with forward, backward, and best subsets regression, using 16 data sets. The results indicated the stepwise method was preferred because of its practical nature, when the models chosen by different selection methods were similar in number of variables, variables included, and amount of variance explained. The best subset method worked very well for these data sets, and was recommended for encouraging a non-mechanical selection process by giving many suggested models. The backward method provided a model which explained about as much variance as models chosen by any other method, but this model may have included more variables than necessary. It was not recommended when there is high multicollinearity. The stepwise method was generally adequate except when conditions of multicollinearity, suppression, and sets of variables working jointly do not occur; then it should be used in conjunction with other methods. The forward method was not recommended if the stepwise method is available. It was concluded that the best subsets and backward procedures were the best, and that the stepwise and forward methods should never be used alone in selecting a model. (GDC)

 * Reproductions supplied by EDRS are the best that can be made *
 * from the original document. *

ED 275 730

**Testing Different Model Building Procedures
using Multiple Regression**

Jerome D. Thayer

Department of Educational and Counseling Psychology

Andrews University

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

This document has been reproduced as
received from the person or organization
originating it.

Minor changes have been made to improve
reproduction quality.

* Points of view or opinions stated in this docu-
ment do not necessarily represent official
OERI position or policy.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

J. D. Thayer

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

**Paper Presented at the Annual Meeting
of the American Educational Research Association**

San Francisco, California

April 16-20, 1986

TM 860 644

One of the most appealing aspects of multiple regression to beginning multiple regression students is the amazing feat performed by a stepwise regression computer program. The process of selecting the "best" combination of predictors so effortlessly and efficiently creates an overwhelming urge to use this procedure and the computer program that accomplishes it for a multitude of tasks for which it is ill suited. Many textbooks on multiple regression claim that abuse of this technique is common. Draper and Smith (1981) give a mild statement that "the stepwise procedure is easily abused by amateur statisticians (p. 310), while Wilkinson (1984) is much more dramatic:

Stepwise regression is probably the most abused computerized statistical technique ever devised. If you think you need stepwise regression to solve a particular problem you have, it is almost certain that you do not. Professional statisticians rarely use automated stepwise regression. (p. 196)

Cohen and Cohen (1975) suggest that model building should proceed according to dictates of theory rather than relying on the whims of a computer. But since in the social and behavioral sciences theoretical models are relatively rare (Neter et al., 1983), Cohen and Cohen suggest that the stepwise method is a "sore temptation" to replace theory in these situations (p. 103).

The authors of current multiple regression textbooks suggest the following considerations for selecting a subset of predictors for a regression model:

1. Selection of variables for a regression model should not be a mechanical process (Chatterjee and Price, 1977; Draper and Smith, 1981; Neter et al., 1983; Younger, 1979).
2. No one process will consistently select the "best" model (Berenson et al., 1983; Gunst and Mason, 1980; Kleinbaum and Kupper, 1978; Morrison, 1983; Pedhazur, 1982; Younger, 1979)

3. There is no one "best" model according to any common criterion such as the maximum R^2 (Chatterjee and Price, 1977; Freund and Minton, 1979; Neter et al., 1983).
4. The stepwise method should not be used to build models for explanatory research (Cohen and Cohen, 1975; Pedhazur, 1982).

In addition many authors point out that the stepwise method has limited usefulness when the predictors are highly correlated (Chatterjee and Price, 1977; Kleinbaum and Kupper, 1978; Neter et al., 1983), if a key set of variables work in combination (Younger, 1979), or when suppression exists (Cohen and Cohen, 1975). Chatterjee and Price (1977) suggest that with multicollinearity the backward method is preferred although other authors suggest that the backward method should not be used in this case because of computational inaccuracy that may occur if multicollinearity is severe and a near singular matrix is inverted.

In spite of these suggestions, there are still many research studies reported in the literature in which these guidelines are violated. Results are reported of a model "selected" by the computer, usually using the stepwise method with no indication that this model might not be the "correct" or "best" one. The discussion of the selected model is done in a mechanical fashion with no indication given of a careful critique of the adequacy of the computer-selected model. Explanatory interpretations are frequently made (Pedhazur, 1982) which often take the form of considering variables selected by the computer to be "good" predictors of the dependent variable because they have a "significant relationship" and variables not selected by the computer are considered to be "poor" predictors because they do not have a "significant relationship". A variable that may be one of the best predictors when studied individually and that fits nicely into an existing theory will be considered to be a "poor" predictor simply because it does not occur in the selected model even though its omission may be due to predicting the same variance as

other predictors already in the model that are no better predictors than it is.

There are many other competing procedures that can be used to select variables for a regression model other than the stepwise method. Three major ones mentioned in many regression textbooks are the forward, backward, and best subsets methods. This paper will endeavor to compare the stepwise method with these selection methods to determine the types of models that each would be likely to select and in so doing determine the strengths and weaknesses of each method.

Method

The procedure used was to apply each of the common selection methods to a number of data sets of various types and evaluate the differences between the models chosen. The source for each of the data sets used in the analysis is described below. In Table 1 the number of subjects and number of predictors for each data set is listed.

Data Sets Used

1. GMA1 -- Data Set A1 from Gunst and Mason (1980)
2. GMA3 -- Data Set A3 from Gunst and Mason (1980)
3. GMA6 -- Data Set A6 from Gunst and Mason (1980)
4. GMA8 -- Data Set A8 from Gunst and Mason (1980)
5. GMB1 -- Data Set B1 from Gunst and Mason (1980)
6. GMB2A-GMB2B -- Data Set B2 from Gunst and Mason (1980)
7. TAL -- Project Talent data from Lohnes and Cooley (1968)
8. ENR1-ENR5 -- 1986 freshman enrollment data from Andrews University
9. LONG -- Data from Longley (1967)
10. HALD -- Data from Draper and Smith (1981)
11. SUP -- Data generated from a contrived correlation matrix

Nine of the data sets were selected from textbooks that used the data sets to illustrate interesting and/or unusual applications of regression that would be brought out by the data. All of the variables were not included in some of the sets. Some of the variables in the GMA3 set were not used because there were more variables than subjects. One variable was removed from the GMB1 set due to tolerance problems (its tolerance was below .01, and thus was automatically excluded from the BMDP2R program although it would not have been included in any of the models if tolerance had been ignored). The categorical variables from the TAL set were not used.

The SUP data was generated using a program described in Morris (1975) from a contrived correlation matrix described below that included variables that illustrated suppression. To get a correlation matrix with suppression, three variables were constructed composed of random numbers with the first variable designated as the dependent variable and the other two designated as independent variables. A fourth variable was then constructed which did not have a high correlation with the dependent variable by itself but yielded a high multiple correlation with the dependent variable when combined with the two previously chosen independent variables. The correlation matrix from this data was then used as input to the Morris program which generated a new set of data which gave the same correlation matrix but was "marginally normal." The correlation matrix used was:

	1	2	3	4
1	1.000	.446	.292	.397
2		1.000	-.195	-.088
3			1.000	-.527
4				1.000

An alternate approach that would have given an equivalent matrix would have been to use the method suggested by Lutz (1983).

GMB2 was run twice using a different dependent variable each time. The ENR data was analyzed with 5 different sets of predictors. The variables used for the ENR data sets were selected from 86 variables which in turn were selected from a larger data base that included 499 variables. A principal components factor analysis was conducted using the 86 variables and the variables loading on the 14 factors with the highest eigen values (all above 1.3) were used in the 5 sets of predictors.

ENR1 had 1 predictor from each of the first 7 factors.

ENR2 had 2 predictors from each of the first 7 factors.

ENR3 had 4 predictors from each of the first 7 factors.

ENR4 had 1 predictor from each of the 14 factors.

ENR5 had 2 predictors from each of the 14 factors.

The computer programs used to select the best model from each data set were BMDP2R for the stepwise, forward and backward solutions, and BMDP9R for the best subsets solution. The stepwise and forward methods used an F-to-enter limit of 2.0 and the stepwise method used an F-to-remove limit of 1.99. These limits are in line with recommendations made for proper use of stepwise regression which suggest that the F-to-enter limit selected should be fairly low so as to allow more variables a chance to show their worth in the final model. The backward method used a comparable F-to-remove limit of 2.0. The BMDP9R program selected the model with the lowest C_p value, which is the default value of the program. An ideal C_p value is one that is equal to or lower than the number of parameters in the model (predictors + 1). Dixon and Brown (1979) suggest that this criterion will give models in which the variables in the model have F-to-remove values above 2.0, making this criterion similar to that used in the other three methods. Of course, the specific models selected would differ if other criteria were used, but the overall characteristics of the four selection methods should not change. To evaluate a different criterion, on some comparisons it will be noted what the

results would have been if an F-to-enter/remove level of 4.0 had been used rather than 2.0.

Table 1 reports the characteristics of the subsets selected by the 4 selection methods with the 16 data sets. For the stepwise method the number of predictors selected is reported along with the R^2 for the selected model. For the other methods information is only presented if the model selected was different from the model selected by the stepwise method. Additional information provided for these models includes the number of predictors in that model that were not in the stepwise model and the number of predictors in the stepwise model not included in that model.

Results

On 9 of the 16 data sets, the 4 methods chose different models using the initial criteria of a F-to-enter/remove of 2.0 and the lowest C_p . In comparison with the stepwise method, the forward method chose a different model on 2 data sets, the backward method chose a different model on 5 data sets, and the best subsets method chose a different model on 7 data sets. The backward method and best subsets method differed on 4 data sets. For each of the data sets on which differences were found, the differences will be described in detail.

GMA3 -- The stepwise, backward and best subsets methods selected the same model which had 1 less variable than that selected by the forward method. If F-to-enter/remove limits of 4.0 had been used, the stepwise and backward methods would have removed one additional variable giving a 4 predictor - while the model chosen by the forward method would not have changed, thus having 2 more predictors than the stepwise and backward methods.

GMA6 -- The backward and best subsets methods gave the same model which had an R^2 more than twice as much as that found by the stepwise and forward methods which gave the same model. The R^2 values found were .150 and .347.

The stepwise/forward model had 2 predictors and the backward/best subsets model had 7 predictors. The stepwise/forward methods did not enter a third variable because the highest F-to-enter was 1.96. The worst variable in the 7 variable backward and best subsets model had a F-to-remove of 3.25. If an F-to-enter limit of 4.00 had been used, there would have been no variables included in the stepwise/forward model since the first variable entered had an F-to-enter of 2.50 while the backward method would have removed the seventh variable leaving a 6 variable model with an R^2 of .300. The stepwise method gave much lower R^2 values at F-to-enter limits of both 2.0 and 4.0. The C_p value for the backward/best subsets model was 4.02 for 7 predictors while the stepwise/forward model had a C_p value of 5.54 for 2 predictors, indicating the 7 predictor model chosen by the backward and best subsets methods was a much better model.

GMA8 -- The stepwise, forward, and backward methods produced the same model which was different from that chosen by the best subsets method. The best subsets model had 1 less predictor, the last variable chosen by the stepwise/forward methods and the variable which would have been the next to be deleted by the backward method. The R^2 values for the 2 models were .886 and .877. The C_p values for the 2 models were about identical (1.51 for the stepwise/forward/backward model and 1.50 for the best subsets model). The F-to-remove for the fourth variable included in the larger model was 2.28.

GMB1 --The 4 methods produced 3 models, with the stepwise and forward methods selecting the same model. The R^2 values for the models were .716 for the 5 predictor best subsets model, .727 for the 6 predictor stepwise/forward model, and .739 for the 8 predictor backward model. All of the variables in the best subsets model were included in the stepwise/forward model with the additional variable in the stepwise/forward model having an F-to-enter of 2.02. The backward model used 4 of the 6 predictors in the stepwise/forward model and 4 additional predictors. The C_p values were 3.27 for the

stepwise/forward model and 3.14 for the best subsets model. The backward model was not listed as one of the 10 best 8 predictor models in the BMDP9R best subsets selection even though it had an R^2 of .737 which was higher than 9 of the 3 variable models listed. If the F-to-enter and F-to-remove limits had been 4.0, both the stepwise/forward and backward models would have included 5 variables but only 3 would have been common to both. The 5 variable model R^2 would have been .716 for the stepwise/forward model and .697 for the backward model.

GMB2B -- The model selected by the stepwise and forward methods had only 1 predictor with an R^2 value of .176. No variable was even close to being considered for entry as the F-to-enter value for the best additional second variable was 0.76. The backward and best subsets models were the same with 5 predictors and an R^2 of .509. The worst variable in the 5 predictor model had an F-to-remove value of 6.82. The reason for the discrepancy between the models was that 2 of the variables were only good predictors in combination. In the stepwise solution, one of this pair would have been the second variable added with an F-to-enter of 0.76 and increasing the R^2 from .176 to .193. The third variable added would have been the other member of the pair which would have increased the R^2 to .371. The better predictor of the pair in the second step added only .017 (.193-.176) while together as steps 2 and 3, the pair added .195 (.371-.176). The fourth and fifth predictors increased the R^2 from .371 to .509.

TAL -- All of the methods selected the same model but the order of entry of the variables in the stepwise/forward and backward methods were different. The last variable entered in the stepwise and forward methods was not the same as the variable that would have been removed next in the backward method. If the F-to-enter/remove limit had been 4.0, the models would have been different with the stepwise/forward method model having 4 variables with an R^2 of .388 and the backward model having 6 variables with an R^2 of .396. The additional

2 variables for the backward model were included because these 2 variables would not have been good enough to enter alone in the stepwise/forward methods, but together they were good predictors, making them remain in the backward method.

ENR3 -- The 4 methods produced 3 models, with the stepwise and forward methods selecting the same model. The R^2 values for the models were .520 for the 8 predictor best subsets model, .521 for the 9 predictor stepwise/forward model, and .525 for the 11 predictor backward model. All of the variables in the best subsets model were included in the stepwise model with the additional variable of the stepwise model having an F-to-enter of 2.02. All but one of the variables in the stepwise/forward model were included in the backward model with 3 additional variables added. The 3 models selected were the best, second best, and tied for third best in the best subsets method with C_p values of 5.88, 5.89, and 6.05. The other model with a C_p of 6.05 was the second best 8 predictor model selected by the best subsets method. This model had 1 predictor different from the best model selected. It appears as if the additional 2 or 3 variables of the backward model were not needed to select a good model but other combinations of variables would have given equally good smaller models. If an F-to-enter limit of 4.00 had been used, the stepwise/forward model would have contained 5 predictors with an R^2 of .510 and the backward model would have had 7 predictors with an R^2 of .517 with only 3 of the same predictors as the stepwise/forward model.

ENR5 -- All of the methods produced the same model but the stepwise/forward and backward models had a different order of entry. If the F-to-enter/remove limit had been 4.00, the stepwise/forward model would have had 8 predictors with a R^2 of .338 and the backward model would have had 9 predictors with a R^2 of .343 with 6 variables the same as those in the stepwise/forward model. If the ninth predictor of the backward model had been

removed, the remaining 8 variables would have had the same R^2 as the stepwise/forward model (.338) with 2 variables being different.

LONG -- The stepwise, forward and backward methods chosen by BMDP2R gave the same 3 predictor model with an R^2 of .985 and the best subsets model had 4 predictors with an R^2 of .995. The additional predictor in the best subsets model was not included in the other models due to its high intercorrelation (tolerance=.002) with the first 3 predictors in the model. BMDP9R (best subsets) allows a greater degree of multicollinearity than BMDP2R, so this problem was not encountered with the model chosen by that program. The F-to-remove value of the fourth variable in the best subsets model was 5.95 indicating it deserved to be in the model if the low tolerance could be ignored. The C_p value for the 4 predictor model was 3.24 compared to the 3 predictor value of 21.66. The first variable entered in the stepwise and forward methods was the variable that contributed the most to the high tolerance value for the fourth variable in the model (the correlation between them was .995). If a 3 predictor model had been chosen by all methods ignoring the tolerance problem, the backward and best subset methods would have chosen the same model with a higher R^2 than that chosen by the stepwise/forward method (.993 to .985). The C_p value for the 3 predictor backward/best subsets model would have been 6.24 compared to the stepwise/forward value of 21.66. The backward/best subsets model is better because the second and third variables entered in the stepwise/forward method in combination pair much better with the fourth variable than the first variable entered. The model chosen by the backward and best subsets methods was never evaluated in the stepwise and forward methods.

HALD --The stepwise, backward, and best subsets chose the same 2 predictor model while the forward method selected a 3 predictor model, including a variable that was the first one entered but that later became redundant with the addition of the second and third variables.