JLTIPLE LINEAR REGRESSION VIEWPOINTS
)LUME 15, NUMBER 1, SUMMER 1986

Testing Different Model Building Procedures Using Multiple Regression

Jerome D. Thayer
Andrews University

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.

<u>Data Sets Used</u>

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 (1966)
8. ENR1-ENR5 -- 1986 freshman enrollment data from Andrews University
9. LONG -- Data from Longley (1967)
10. HALO -- 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	-.068
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 date 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 thet selected by the forward method. If F-to-enter/remove limits of 4.0 had been used, the stepwise end backward methods would have removed one edditional variable giving a 4 predictor model 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 8 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.

SUP -- The stepwise and forward methods did not allow any variables to enter the model. The largest F-to-enter value was 1.99. The backward and best subsets models were the same with 3 predictors and an R^2 of .967. The lowest F-to-remove value of the 3 predictors was 85.16 which if removed would bring the R^2 down to .506. Each variable acting alone did not predict enough to be included but only showed its high predictive power in combination with the other variables.

Conclusions

If models chosen by different selection methods were relatively similar in the number of variables in the model, the variables included, and the amount of variance explained (R^2), and the model was to be used primarily for prediction, not explanatory purposes, it would seem that the suggestion of Draper and Smith (1981) that the stepwise method might be preferred because of its practical nature would seem reasonable. The results of this study suggest, however, that in some cases models that are severely inadequate are selected by the stepwise method and other consistent, but less important differences between the models selected by the different methods also appear.

Forward/stepwise comparison

It would be expected that the forward method would be more similar to the stepwise method than the backward or best subsets methods because the stepwise method is an extension of the forward method with the additional procedure of removing variables previously entered if they no longer contribute to the model. In both of the data sets in which a difference existed between these 2 methods, the forward method gave a larger data set by including a variable that became redundant when later variables were added by both methods.

Backward/stepwise comparison

The backward method differed in a consistent manner from the stepwise method in 2 ways. In each of the 5 data sets in which they differed the backward method selected a model with more predictors. If an F-to-enter limit of 4.0 had been used, the backward method also would have frequently given a larger number of predictors. Where the same number of predictors were selected but with different combinations, the stepwise method was more efficient, generally having the higher R^2. In 2 of the 5 cases in which they differed the R^2 values were fairly close but for the other 3 the R^2 values were markedly different (.347/.150, .509/.176, and .967/.000) with the backward method selecting the better model in each case. These 3 data sets all had a combination of variables that acted jointly to predict well but none of the variables entered the model individually in the stepwise or forward methods. These data sets illustrate that in certain circumstances the stepwise and forward methods can select very inadequate models.

Backward/best subsets comparison

On 12 of the 16 data sets the same model was selected by the backward and best subsets methods. The worst discrepancy between the models selected by the two methods was on the GMB1 data set in which the models had 5 and 8 predictors and R^2 of .716 and .739. It seems as if the backward and best subsets methods can be counted upon to give models that are reasonably similar in number of predictors and amount of variance explained, although if there is a difference the backward method generally will give a larger model. In the 4 data sets in which the 2 methods gave different models, the backward method selected a larger model 3 times and a smaller model once (although this was due to a tolerance problem).

Stepwise/best subsets comparison

Excluding the 3 cases in which the stepwise method was very inadequate and the case with the tolerance problem, the number of predictors selected by the stepwise method was the same as that selected by the best subsets method in all but 3 cases where the stepwise method gave 1 additional predictor in each case. The additional variable in each of the larger models barely entered over the F-to-enter of 2.00 level and the discrepancy should not be considered important but more of an indication that the F-to enter level of 2.00 was not exactly equivalent to the criterion of the lowest C_p value that was used in the BMDP9R program.

Best subsets summary

The algorithm used in BMDP9R, which admittedly does not compare all possible models, will not always list all "good" models. In the GMB1 data set, the 8 predictor model chosen by the backward method was not even listed as one of the alternatives in the BMDP9R output even though it had a higher R^2 than all but one of the alternatives that were mentioned. The best subset method, however, does seem to work the best of all of the prediction methods with the data sets used here. It is especially recommended because it encourages a non-mechanical selection process by giving many suggested models.

Backward summary

The backward method can be counted on to give a model which will explain about as much variance as models chosen by any other method but it may include more variables than are necessary to get a "good" model. A major danger occurs with this method, however, when there is high multicollinearity. In this case, computational inaccuracies may occur, so tolerance problems should be considered before running a backward solution.

Stepwise summary

The stepwise method will generally give a model that comes close to maximizing the amount of variance explained for a given number of predictors. If conditions of multicollinearity, suppression, and sets of variables working jointly do not occur, the models selected by the stepwise method can be expected to be as good as the models selected by the backward and best subsets methods. If these conditions do occur, however, the stepwise method may give a model that is completely inadequate. To guard against this occurrence, the stepwise method should never be used alone to select a model, but only in conjunction with the backward and/or best subsets methods.

Forward summary

The forward method, although discussed in almost all regression textbooks, is rarely, if ever recommended as a reasonable alternative to the stepwise method, and this paper supports the idea that the method has little merit if the stepwise method is available.

Selection process summary

When a model is to be selected, it is important to consider more than one procedure. If one method is to be used, it would appear that the best subsets method is the best of the methods examined here since the computer program generates many models from which a "best" one can be selected. The virtue of running a backward and/or stepwise solution in addition to the best subsets method would be to identify differences in the models that point out characteristics of the variables and/or data set that might be overlooked otherwise. Using the best subsets or backward procedures, it is unlikely that an extremely poor model would be chosen, but this is a real possibility with the stepwise and forward methods. For this reason it is recommended that the stepwise and forward methods NEVER be used alone in selecting a model for any purpose.

Table 1

Regression Models Selected by Different Selection Methods

DATA SET	N	IV's	Stepwise #	Stepwise R^2	Forward #	Forward +	Forward -	Forward R^2	Backward #	Backward +	Backward -	Backward R^2	Best Subsets #	Best Subsets +	Best Subsets -	Best Subsets R^2
GMA1	49	6	3	.497												
GMA3	13	6	4	.999	5	1	0	.999								
GMA6	50	14	2	.150					7	6	1	.347	7	6	1	.347
GMA8	33	9	4	.666									3	0	1	.677
GMR1	60	14	6	.727					8	4	2	.739	5	0	1	.716
GMB2A	40	6	4	.676												
GMB2B	40	6	1	.176					5	4	0	.509	5	4	0	.509
TAL	505	16	9	.404												
ENR1	579	7	2	.049												
ENR2	579	14	7	.316												
ENR3	579	26	9	.521					11	3	1	.525	8	0	1	.520
ENR4	579	14	5	.089												
ENR5	579	26	14	.361												
LONG	16	6	3	.965									4	1	0	.995
HALD	13	4	2	.979	3	1	0	.982								
SUP	10	3	0	.000					3	3	0	.967	3	3	0	.967

Column groups: Number of Predictors Selected/Differences from Stepwise/R^2

= number of predictors selected using F=2.0 for entry and F=1.99 for deletion for the stepwise, forward and backward models and C_p=2.0 for the best subsets model.

+ = number of predictors selected in this model that were not in the stepwise model

- = number of predictors in the stepwise model that were not selected in this model

References

Berenson, M. L. et al. (1983). Intermediate Statistical Methods and Applications. Englewood Cliffs, New Jersey: Prentice-Hall.

Chatterjee, S. & Price, B. (1977). Regression Analysis by Example. New York: John Wiley & Sons.

Cohen, J. & Cohen, P. (1975). Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Dixon, W. J. & Brown, M. B. (1979). BMDP-79 Biomedical Computer Programs P-Series. Berkeley, California: University of California Press.

Draper, N. R. & Smith, H. (1981). Applied Regression Analysis, Second Edition. New York: John Wiley & Sons.

Freund, R. J. & Minton, P. D. (1979). Regression Methods. New York: Marcel Dekker.

Gunst, R. F. & Mason, R. L. (1980). Regression Analysis and its Application. New York: Marcel Dekker.

Kleinbaum, D. G. & Kupper, L. L. (1978). Applied Regression Analysis and Other Multivariable Methods. North Scituate, Massachusetts: Duxbury Press.

Lohnes, P. R. & Cooley, W. W. (1968). Introduction to Statistical Procedures: with Computer Exercises. New York: John Wiley & Sons.

Longley, J. W. (1967). An appraisal of least squares programs for the electronic computer from the point of view of the user. Journal of the American Statistical Association, 62, 819-831.

Lutz, J. G. (1983). A method for constructing data which illustrate three types of suppressor variables. Educational and Psychological Measurement, 43, 373-377.

Morris, J. D. (1975). A computer program to create a population with any desired centroid and covariance matrix. Educational and Psychological Measurement, 35, 707-710.

Morrison, D. F. (1983). Applied Linear Statistical Methods. Englewood Cliffs, New Jersey: Prentice-Hall.

Neter, J. et al. (1983). Applied Linear Regression Models. Homewood, Illinois: Richard D. Irwin.

Pedhazur, E. J. (1982). Multiple Regression in Behavioral Research. New York: Holt, Rinehart, Winston.

Wilkinson, L. (1984). SYSTAT. Evanston, Illinois: SYSTAT, Inc.

Younger, M. S. (1979). A Handbook for Linear Regression. North Scituate, Massachusetts: Duxbury Press.