MULTIPLE LINEAR REGRESSION VIEWPOINTS
VOLUME 18, NUMBER 1, FALL 1991

Considerations, Issues and Comparisons In Variable Selection and Interpretation In Multiple Regression

Susan Tracz, Ric Brown, and Rebecca Koprive
California State University, Fresno

The selection of independent variables when utilizing multiple linear regression in a study is an involved and complex process. The availability of a variety of computer programs usually referred to as "stepwise" procedures affords users numerous options about which they often have little understanding. The purpose of this paper, then, is twofold: first, to present the major uses of regression analyses, the advantages and disadvantages of selection procedures and some caveats for researchers and those who teach statistics, and secondly, to present, compare and contrast several variable selection techniques using two data set.

Huberty (1989) suggests that the concept of variable selection may have some worth in terms of parsimony, explaining relationships, lowering the cost of data collection, and, sometimes, parameter estimation. Variable selection procedures called stepwise procedures are available on all the major statistical computing packages including SAS, SPSS, and BMDP. Even novice researchers can easily run numerous stepwise procedures. Huberty (1989), however, continues by saying that stepwise analyses have been basically used for three purposes: 1) selection and deletion of variables, 2) assessing relative variable importance, and 3) a combination of selection and variable ordering.

Given this information, it is not surprising to find numerous articles in the literature and theses and dissertations in university libraries that have used and misused stepwise procedures despite the many published caveats concerning its appropriateness. Perhaps one reason for the frequent misuse of stepwise procedures is the mistaken perception that the results of a stepwise procedure will yield the "best" equation. According to Hocking (1983), "there is not likely to be a best equation in multiple regression" (p. 226). This is because the use of differing criteria may result in the selection of different sets of variables (Draper & Smith, 1981). Pedhazur (1982) more specifically stated that such methods as all possible regressions, forward selection, backward elimination, stepwise selection and blockwise selection can be utilized with differing criteria which will result in differing solutions depending on the criteria. Morris (1989) sums up these ideas by saying that "there is little theoretical justification for expecting any stepwise procedure to be best" (p. 2).

The goal of stepwise regression is to choose a subset of variables from a larger set for the purpose of parsimony, prediction, explanation, and/or theory-building. However, since the criteria used in selecting variables are statistical, measurement error or randomness may lead to the selection of one variable instead of an equally viable alternative variable. Cohen and Cohen (1975) expounded on this issue saying that "problems include capitalization on chance because of simultaneous tests, sample specificity and trivial differences in partial relationships leading to choosing one variable over another" (p. 103).

When predictor variables are intercorrelated, "there is no satisfactory way to determine relative contributions of the variables on R squared" (Edwards, 1984, p. 107) and "the idea of independent contribution to variance has no meaning" (Darlington, 1968, p. 169). Huberty (1989) reiterates these points by noting that various subsets of a given size can yield nearly the same R^2 value. Pedhazur (1982) states that the R^2 in variance partitioning is sample specific and that nearly identical regression equations can have radically different R^2 values. Furthermore, an incremental R^2 may be statistically significant but substantially meaningless. Pedhazur (1982) argues that the incremental partitioning of variance may be used to control one variable while studying another variable only in causal modeling, and even then the results are of limited value in determining policy.

Another problem to be dealt with is the interpretation of the regression coefficients. Huberty (1989) cautions that the order in which a variable is entered into a model should not be used to assess its relative importance. "The interpretation of regression coefficients as indices of effects of independent variables on the dependent variable appeals to researchers because it holds the promise for unraveling complex phenomena. Examination, however, is important because the apparent simplicity is deceptive" (Pedhazur, 1982, p. 221). Pedhazur (1982) warns that the absence of a theoretical model makes the meaningful interpretation of the estimated regression coefficients impossible. The types of specification errors that can occur are numerous including omission of relevant variables, inclusion of irrelevant variables, interactions among variables, and the hierarchy of polynomial terms (Cohen & Cohen, 1975; Pedhazur, 1982; Peixoto, 1990).

When so many caveats against it have been published, the continued wide usage of stepwise procedures is difficult to understand. Variable selection techniques in regression analysis can be discussed in terms of parsimony, prediction, explanation and theory-building, and selection techniques are problematic in all of these areas.

Parsimony involves finding "a smaller set of predictor variables that do an accurate job of predicting, nearly as well as the total set of variables" (Morris, 1984, p. 1). Obviously, parsimony is helpful to researchers who reap benefits in terms of economy of data collection costs and time. However, the criteria for the selection of the best variables must be weighed on a continuum between internal (parsimony) and external (cross validation) accuracy (Morris, 1984). A prior decision made in the name of parsimony can have a tremendous impact on the results of regression analyses used for prediction, explanation and theory-building.

Pedhazur (1982) states that "for prediction, the goal of regression is to optimize prediction of criteria" (p. 136). The selection of variables for this purpose should account for as much of the variance as possible while balancing practical

considerations such as cost and ease of administration. While Morris (1989) finds "particularly 'pernicious' ... a situation with a naive researcher ascribing the best prediction equation from the results of a stepwise program" (p. 1), Pedhazur (1982) argues that "prediction may be accomplished in the absence of theory, but explanation is inconceivable without theory" (p. 174).

The goals of many researchers in terms of explanation have been to identify major variables and determine their relative importance (Pedhazur, 1982). This suggests that stepwise techniques may be plausible initially. The stepwise programs basically perform a hypothesis formulation function (McNeil, Kelly, & McNeil, 1975). However, "problems arise with the stepwise approach, since a great many hypotheses are being tested the resulting best model will most likely be drastically overfit with replication relatively unlikely" (p. 364).

Cohen and Cohen (1975) state that "a research strategy of treating all independent variables simultaneously is most appropriate when no logical or theoretical basis for considering any variable to be prior to any other either causal or relevant in terms of research goals" (pp. 97-98). However, despite this seeming endorsement, they continue by saying "a dim view is taken of stepwise in exploratory research because orderly advance is more likely in the social sciences when researchers use theory to provide hierarchical ordering formed by causal hypotheses rather than computers ordering independent variables" (p. 103).

Given all the problems of sample specificity, interpretation of regression weights, and varying R values, the question arises when is it actually appropriate to use stepwise procedures. Huberty (1989) says that in cases where a large ratio of sample size to variables exists, generalizability of stepwise regression is enhanced, but an external analysis or a cross validation should also be conducted. Thorndike (1978) agrees arguing that "when a fairly large number of predictor variables are available it is advisable to use a stepwise approach, but cross validate" (p. 167). Finally, Cohen and Cohen (1975) state that the distrust of stepwise procedures decreases if: "1) the research goal is predictive not explanatory; 2) N is very large for a given number of independent variables (40 to 1); and, 3) cross validate" (p. 104). Perhaps, Huberty (1989) offers the best advise when he says that "thorough study and sound judgement are suggested for choosing variables at the outset" (p. 62), and that "the data analyst should allow the findings at each stage to influence the direction through subsequent stages" (Allen & Cody, cited in Huberty, 1989, p. 65).

The numerous stepwise procedures available in the major statistical computing packages are so easy to execute, however, that users quickly learn to rely on them, and there is a great temptation for researchers, especially novice researchers, to assume that a stepwise procedure will yield the best model which will stand up to the test of cross validation. Again, this is simply not true. Stepwise procedures actually yield many best

models depending on the procedure used and the criteria employed, and it is up to the researcher to decide which one to use and why. In short, stepwise procedures are no substitution for thinking and theorizing. This paper, will now present, compare and contrast several variable selection techniques using two data sets. In the first example, the results of various stepwise techniques from the SAS package will be compared. In the second example, the results of several stepwise regressions used to answer various research questions will be compared.

The first example consists of a dummy data set of 30 subjects used for classroom teaching purposes. The dependent variable is graduate grade point average [GPA], and the four independent variables are the Graduate Record Exam Quantitative subscale [GREQ], the Graduate Record Exam Verbal subscale [GREV], the Miller's Analogy Test [MAT], and a faculty rating of graduate student performance [RAT]. (This data set is available from the authors upon request).

The intercorrelations among these variables and the associated probabilities are presented in Table 1.

Table 1 Correlations and probabilities (N = 30)

Variables		GREQ	GREV	MAT	RAT
GPA	(r)	.61	.58	.60	.62
	(p)	.0003	.0008	.0004	.0003
GREQ	(r)		.47	.27	.51
	(p)		.009	.15	.004
GREV	(r)			.43	.41
	(p)			.02	.03
MAT	(r)				.52
	(p)				.003

As can be seen the dependent variable GPA is highly correlated with all of the independent variables. All the independent variables are also highly correlated with each other except for the combination of GREQ and MAT (r = .27) and possibly GREV and RAT (r = .41). Therefore, pairs of unique information have been set up between GREQ and MAT and between GREV and RAT.

Five different analyses were run using this data set. The first was a full model with all four dependent variables using the forced solution, PROG REG. This model was significant ($F_{4, 25}$ = 11.13, $p < .0001$, R^2 = .64, adjusted R^2 = .58). The parameter estimates, t values and probabilities appear in Table 2. In this model the t values for GREQ and MAT are significant, while those for GREV and RAT are not.

Table 2 Results of full model using the forced solution in PROC REG to predict GPA from all independent variables

Variable	Parameter Estimate	t	p
Intercept	-1.738	-1.83	.08
GREQ	.004	2.18	.04
GREV	.002	1.45	.16
MAT	.021	2.19	.04
RAT	.144	1.28	.21

The next analysis which was performed was a forward selection. This program identifies a subset of variables which will be as efficient as the entire set of variables for predicting GPA. In this case, the significance level for entering a variable into the model has been set on the lenient side to .15. The variables were entered into the model in the following order: RAT, GREV, MAT, and GREQ. The R^2 values for each new model and the change in R^2 are presented in Table 3. The R^2 for the full stepwise model is .64, as in the full model, since all the variables were entered into the model.

Table 3 Resulting R^2s and changes in R^2s from the forward selection method to predict GPA from allindependent variables

Variable Entered into the Model	R^2	Change in R^2
RAT	.39	-
GREV	.52	.13
MAT	.57	.05
GREQ	.64	.07

The third analysis was a backward elimination. The procedure starts with all the variables entered into the model and then eliminates variables. The significance level for retaining a variable in the model has been set to .05. Again the full model had an R^2 of .64. The variable, RAT, was removed first (R^2 = .62) and then GREV (R^2 = .58), so the best model with GREQ and MAT only included has an R^2 of .58. The results appear in Table 4.

Table 4 Resulting R^2s and changes in R^2s from the backward elimination method to predict GPA from all independent variables

Variables Included	Variables Removed	R^2	Change in R^2
GREQ, GREV, MAT, RAT	-	.64	-
GREQ, GREV, MAT	RAT	.62	.02
GREQ, MAT	RAT, GREV	.58	.04

The fourth analysis used the stepwise method. This procedure differs from the forward selection method in that variables entered on earlier steps do not necessarily remain in the model on subsequent steps. After a variable is added, other variables in the model are inspected to determine if they still produce a significant F statistic. If the F is not significant, the variable is deleted from the model on that step. For this case, the significant level for entry into the model was set to 15, and the significance level for remaining in the model was set to .05. The results for this analysis appear in Table 5. The variable, RAT, was entered into the model first (R^2 = .39), then GREV (R^2 = .52) and then MAT (R^2 = .57). Finally, MAT (R^2 = 52) was removed from the model because the F value for that variable was not significant, so the resulting best model included RAT and GREV (R^2 = .52).

Table 5 Resulting R^2s and changes in R^2s from the stepwise procedure to predict GPA from all independent variables

Step	Variable Entered	Variable Removed	R^2	Change in R^2
1	RAT	-	.39	-
2	GREV	-	.52	.13
3	MAT	-	.57	.05
4	-	MAT	.52	.05

Finally, the last stepwise procedure used was the maximum R^2 method. This procedure adds variables that maximizes R^2. The results of this procedure are presented in Table 6. This procedure went through five steps and arrived at a model which included all four independent variables (R^2 = .64). However, it could be argued that the best model is determined on the basis of the C(P) statistic. The optimal model is the one for which the C(P) statistic approaches the number of predictors. In this case, the researcher should stop at step 4 since the C(P) statistic is then equal to 4.63 which is closest to the number of predictor variables or four.

Table 6 Resulting R^2 and C(P) from the maximum R^2 method to predict GPA from all independent variables

Step	Variables in the model	R^2	C(P)
1	RAT	.39	16.74
2	GREV, RAT	.52	9.69
3	GREV, Mat, Rat	.57	7.77
4	GREQ, GREV, MAT	.62	4.63
5	GREQ, GREV, MAT, RAT	.64	5.00

Table 7 presents a summary of the results of all the procedures. The full model, forward selection, and maximum R^2 method all include all four predictor variables and give an R^2 of .64. What is curious is that for the procedures which select only two variables the solutions are quite different. The stepwise procedure ends up with RAT and GREV (R^2 = .52), while the backward elimination ends up with GREQ and MAT (R^2 = .58). The forward, stepwise and maximum R^2 methods all enter RAT into the model first because this variable has the highest correlation with GPA (4 = .62). The next variable entered is GREV. The correlation between RAT and GREV is .41. In the other "best" two variable solutions the correlation between the two predictors, GREQ and MAT is .27. It is important to note that these are the lowest two correlations among all the predictor variables. When variables are highly intercorrelated and one variable is entered into a model first, the next variable entered will add the most unique information, i.e., has the lowest correlation with the first variable. In other words, variables are really entered as pairs (GREQ & MAT, R^2 = .58; GREV & RAT, R^2 = .52). Also, in some situations the procedures, namely forward selection, stepwise, maximum R^2, did not produce the maximum R^2 for the two variable models even though most users think they do. This is because the algorithms in these procedures don't really check all the possibilities.

Table 7 Comparison among the best models of the full model and stepwise results

Procedure	Variables in the model	R^2
Full model	GREQ, GREV, MAT, RAT	.64
Forward selection	RAT, GREV, MAT, GREQ	.64
Backward elimination	GREQ, MAT	.58
Stepwise procedure	RAT, GREV	.52
Maximum R^2	GREQ, GREV, MAT, RAT	.64

In light of this information, what advise can be given to researchers using stepwise procedures? First of all, users of computer packages should know the limitations of the procedures

they use. Secondly, researchers should always study the correlation matrix before looking at other results. A thorough knowledge of the intercorrelations may lead researchers to force certain variables into their models first.

In the next example, the results of stepwise regressions are used to answer different research questions. In this example, data from 65 first time, post-myocardial infarction and first time, post-coronary bypass patients were used to study attributions, self-efficacy, and outcome expectations as predictors of depression. The dependent variable was a 20 item scale called the Center for Epidemiological Studies - Depression [CES-D]. Attribution was measured by two instruments: a 9 item behavioral attribution scale [BEHATT] measuring the causes of heart disease that an individual can change, such as smoking, drinking, etc., and an 8 item nonbehavioral attribution scale [NONBATT] measuring the causes of heart disease that are less controllable, such as heredity, luck, etc. The self-efficacy scale [SELFEFF] has 19 items and measures behaviors that individuals have some degree of confidence that they can change. Outcome expectancy 1 [OUTEXP1] was a 19 item scale rating how important patients believe changing particular behaviors are in preventing future heart attacks. Outcome expectancy 2 [OUTEXP2] was a 19 item scale rating the extent of a patient's belief that if behaviors are changed future heart disease will be prevented. A series of four research questions was asked by individual members of a group of researchers and medical practitioners who each advocated a different modelling approach. The data was then analyzed using combinations of forced and stepwise procedures.

In the first analysis, the question was asked whether the set of attribution or the set of self-efficacy and outcome expectation yielded the largest R^2. The results of this analysis consisting of two regression models which entered all variables simultaneously appears in Table 8. These two regression models produce very similar R^2 values (.28 for the attribution variables and .32 for the self-efficacy and outcome expectation variables), and the weights for four of the five variables were significant. In general, it was found that individuals were less depressed about their heart condition if they believed they had some control in the matter.

Table 8 Analysis 1 - A comparison of outcome expectancy/self-efficacy and attribution regression analyses to predict depression

Variable	Beta	F
Regression Model 1		
OUTEXP2	-.48	15.6*
SELFEFF	-.26	4.6*
OUTEXP1	-.12	1.0
R^2 = .32		
Regression Model 2		
BEHATT	-.37	9.6*
NONBATT	.31	6.5*
R^2 = .28		

*p < .05

In the second analysis, the question was asked which set of variables explains the most variance after one set was already forced into the model. When the self-efficacy and outcome expectation variables were entered into the model first, the R^2 was .32. After the attribution measures were added the R^2 increased by .08 to .40. When the attribution measures were forced into the model first, the R^2 was .28. After the self-efficacy and outcome expectation variables were added, the R^2 increased by .12 to .40. The results of both analyses were fairly similar.

The third analysis was a forward stepwise regression using all five independent variables. These results appear in Table 9. In this case, the two behavioral attributions added significantly to outcome expectancy 2 in predicting depression.

Table 9 Analysis 3 - Resulting R^2s and changes in R^2s using stepwise regression to predict depression with all independent variables

Variables	R^2	Change in R^2
OUTEXP2	.19	.19*
BEHATT	.31	.12*
NONBATT	.37	.06*
SELFEFF	.40	.03
OUTEXP1	.40	.00

* p< .05

The fourth analysis took a more theoretical approach. Some theory suggests that attributions precede behaviors. Following this reasoning two analyses were performed. For the first model, the behavioral attribution variable was forced into the model followed by the stepwise addition of the self-efficacy and

outcome expectation variables. For the second model, the nonbehavioral attribution scale was forced into the model followed by the stepwise addition of the self-efficacy and outcome expectation variables. The results appear in Table 10. Only the significant additions of the stepwise procedures are reported. In both cases, outcome expectancy 2 was the only significant contribution to the attribution variable in predicting depression. Again the resulting R^2 values (.31 and .27) from these two models are quite similar.

Table 10 Analysis 4 Two combinations of forced attribution and stepwise outcome expectancy/self-efficacy regression analyses to predict depression

Variables	R^2	Change in R^2
Regression Model 1		
BEHATT	.18	.18
OUTEXP2	.31	.13
Regression Model 2		
NONBATT	.14	.14
OUTEXP2	.27	.13

In summary, although one could argue in favor of each of these four analyses, the last analysis seems most reasonable since it was based on theory. This example does show, once again, that the research question must dictate the research methodology.

It is hoped that researchers will realize that although multiple linear regression is a powerful and flexible statistical technique and although stepwise computer procedures are potentially useful and facilitative, using these techniques and procedures to meaningfully explain data is a complex process.

For non-experimental research, it is difficult if not impossible to untangle the effects of various variables. Sound thinking, theoretical framework and understanding of the analytical methods are necessary to avoid illogical or unwarranted conclusions" Pedhazur, 1982, p. 175). "Any meaningful analysis applied to complex problems is never routine. The clarifying of controversies in social science research will not be enhanced by applying all sorts of techniques" (Pedhazur, 1982, p. 171).

References

Cohen, J., & Cohen, P. (1975). Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, New Jersey: John Wiley & Sons.

Darlington, R. B. (1968). Multiple regression in psychological research and practice. Psychological Bulletin, 69(3), 161-182.

Draper, N., & Smith, H. (1981). Applied regression analysis (2nd Ed.). New York: Wiley.

Edwards, A. L. (1984). An introduction to linear regression and correlation (2nd Ed.). New York: W. H. Freeman and Company.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. Biometrica, 32, 1-49.

Huberty, C. J. (1989). Problems with stepwise methods -- better alternatives. Advances in Social Science Methodology, 1, 43-70.

McNeil, K. A., Kelly, F.J., & McNeil, J.T. (1975). Testing research hypotheses using multiple linear regression. Carbondale, IL: Southern Illinois University Press.

Morris, J. D. (1989). Alternative variable selection strategies in classification problems. Paper presented at the meeting of the American Educational Research Association, San Francisco.

Nie, N. H., Huss, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. (1975). Statistical package for social sciences (2nd Ed.). New York: McGraw Hill.

Pedhazur, E. J. (1982). Multiple regression in behavioral research. (2nd Ed.) New York: Holt, Rinehart and Winston.

Peixoto, J. (1990). A property of well-formulated polynomial regression models. American Statistician, 44(1), 26-30.

Thorndike, R. M. (1978). Correlational procedures for research. New York: Gardner Press, Inc.