British Journal of Mathematical and Statistical Psychology (1992) 45, 265–282 Printed in Great Britain

Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables

Shelley Derksen and H. J. Keselman†

Department of Psychology, University of Manitoba, Winnipeg, Manitoba R3T 2N2, Canada

The use of automated subset search algorithms is reviewed and issues concerning model selection and selection criteria are discussed. In addition, a Monte Carlo study is reported which presents data regarding the frequency with which authentic and noise variables are selected by automated subset algorithms. In particular, the effects of the correlation between predictor variables, the number of candidate predictor variables, the size of the sample, and the level of significance for entry and deletion of variables were studied for three automated subset algorithms: BACKWARD ELIMINATION, FORWARD SELECTION, and STEPWISE. Results indicated that: (1) the degree of correlation between the predictor variables affected the frequency with which authentic predictor variables found their way into the final model; (2) the number of candidate predictor variables affected the number of noise variables that gained entry to the model; (3) the size of the sample was of little practical importance in determining the number of authentic variables contained in the final model; and (4) the population multiple coefficient of determination could be faithfully estimated by adopting a statistic that is adjusted by the total number of candidate predictor variables rather than the number of variables in the final model.

1. Introduction

Educational and psychological researchers frequently subject their data to least squares multiple linear regression (MLR) analyses in order to derive models of the phenomenon under investigation. Often, the decision about which variables are to be included in the model is based upon automated statistical 'best' subset search algorithms, where 'best' refers to that model, among all the models of a given size, which has the smallest residual mean square (MSRES). A perusal of the popular textbooks on MLR (e.g. Cohen & Cohen, 1983, pp. 123–125; Neter, Wasserman & Kutner, 1985, pp. 417–443; Pedhazur, 1982, pp. 150–171; Weisberg, 1980, pp. 190–202) indicates that the search algorithms that receive most attention are the backward elimination, forward selection, and stepwise algorithms [these algorithms often are referred to collectively as stepwise methods, see SAS' (SAS Institute, 1985) STEPWISE PROCEDURE, for example].

†Requests for reprints.

There are a number of factors that may contribute to this reliance on automated search algorithms. First, there are the 'data miners' (Lovell, 1983) who go about examining a research question by collecting data on virtually every variable that could possibly be related to the phenomenon under investigation. Because data miners do not have an *a priori* model of the phenomenon they are investigating which directs the nature of their inquiry, they rely on 'sifting' through their data in order to filter out a model. Second, researchers employ automated search algorithms for reasons of economy. These researchers have some model(s) in mind prior to data collection, that is, a theory has guided the initial selection of variables to be investigated, but the number of variables is felt to be too large and/or costly. Therefore, for reasons of economy smaller subsets are desired. Third, is the issue of parsimony, that is, the desire to identify the simplest representation of the phenomenon under investigation. Finally, there are statistical considerations relevant to building models that may motivate the use of these algorithms. That is, as Hocking (1976) has indicated, estimated regression coefficients and predicted values from subset models will have smaller variances than the estimates from the complete model. Strictly speaking, however, these statistical properties apply only when the subsets are chosen on the basis of information other than the data itself, which certainly is not the case when subset variable selection algorithms are used. Hence, users who rely on these statistical considerations to justify the use of subset selection techniques should do so cautiously.

For these reasons, automated search algorithms have become part of the educational and psychological researchers' collection of statistical tools. Unfortunately, educational and psychological researchers often expect more from the procedures than they can deliver and the results from the automated search algorithms frequently are interpreted inappropriately. Accordingly, one purpose of this paper is to describe the automated search algorithms and to review briefly the extant literature on these procedures. The second purpose of this paper is to present the results of a Monte Carlo investigation that addresses issues concerning the use of these automated subset search algorithms that have not, to date, been adequately addressed in the literature.

2. Subset selection algorithms

It should be noted there are many more automated subset search algorithms than are discussed in this paper. We choose to limit our presentation to those that are typically recommended and discussed in most MLR testbooks. However, we refer the interested reader to the discussions found in BMDP (Dixon, Brown, Engelman, Hill & Jennrich, 1988, pp. 357–388), SAS (1985, pp. 763–774), and SPSS (Norusis, 1985, pp. 42–48). [See also Miller (1984) and Younger (1985, pp. 479–573).] Additionally, the programs found in these packages offer many variations on the themes discussed in this paper (e.g. forcing variables into the model).

All possible regressions

This technique fits all $\binom{p}{k}$ $\{k=1,\ldots,p\}$ regression models, where p is the total number of predictor variables in the set and k is the subset size. Given that all models for each size are computed, one can locate the 'best' model of each size, that is, those where MSRES is minimal. If p is large, however, the $2^p - 1$ different models produced by this algorithm rapidly become too expensive to compute and too difficult to evaluate. For example, for $p = 10$ predictor variables there would be 1023 regression models. Indeed, some authors even feel it is 'unwarranted' to look at 'all the possibilities' since some models would not be meaningful (Draper & Smith, 1981, p. 302).

Optimal subsets

Several algorithms have been developed which build a 'best' subset of predictor variables without computing the all possible regressions (Furnival & Wilson, 1974; also see Hocking, 1976). Further, these algorithms are contained in the popular statistical packages [BMDP's (Dixon *et al.*, 1988, pp. 993–1012) 9R program and SAS' (SAS Institute, 1985, pp. 711–724) RSQUARE PROCEDURE]. Not only can these algorithms find a model of each size that has minimum MSRES, they can also provide information on models which are nearly as good as the best. That is, they provide a number of competing models. SAS' RSQUARE algorithm is a case in point; it finds many competing 'best' subsets of each size.† Like the all possible regressions algorithm, however, there can be a considerable amount of information that has to be sifted through when the number of predictor variables is large (Hoerl, Schuenemeyer & Hoerl, 1986).

Stepwise methods

Because the amount of computation can be considerable for the all possible and optimal subsets methods, other automated algorithms are available for searching for good models. These methods (stepwise) consider models where variables are added and/or deleted one at a time. Hence, the number of models evaluated with these algorithms is considerably less than those computed and evaluated by the all possible and optimal subsets procedures.

Forward selection. In this method, variables are added to a model one at a time. The first predictor variable to enter the model is that variable which has the highest correlation with the response scores. At each successive step, the variables in the remaining set are considered for inclusion in the current model. The variable that is included at each step is that which produces the largest reduction in the residual sum of squares. Forward selection continues until all variables are in the model or until a stopping rule is satisfied.

Backward elimination. This method starts with all p variables included in the model, with variables subsequently being eliminated one at a time. At each step, the variable

†Indeed, the user can specify the smallest and largest number of predictor variables to appear in a subset in addition to the number of subsets of each size to be selected.

that is deleted is that which results in the least inflation in the residual sum of squares. This method of deletion continues until only one variable is left in the model or until a stopping rule is satisfied.

Stepwise selection. In forward selection, it is possible that a variable selected at an early stage may become unimportant at a later stage, that is, as other variables enter the model. Similarly, in backward elimination, a variable deleted at an early stage could become important at a later stage, that is, as other variables are eliminated from the model. In response to these facts, Efroymson (1960) developed the procedure that is commonly referred to as the stepwise procedure. This procedure is basically a forward selection algorithm, however, at each step the procedure also checks to see whether variables can be dropped from the model using a backward elimination process.†

It is important to note that the models identified by these stepwise methods need not be the same. Further, these procedures do not attempt to identify 'best' subset models, in that they do not necessarily locate the model with the minimum residual sum of squares. Nevertheless, in a comparison between models selected by the optimal and stepwise algorithms (BACKWARD and FORWARD), Berk (1978) reported that the average difference between the residual sums of squares of the stepwise and optimal methods rarely exceeded 7 per cent. In addition, subset models selected by stepwise methods are less likely than optimal methods (e.g. RSQUARE) to include noise variables unrelated to the response variable rather than authentic variables, variables which have a non-zero regression coefficient in the population regression model.

Despite explicit discussions of the characteristics and limitations of these stepwise procedures (e.g. Hocking, 1976; Rawlings, 1988, p. 180), many users of these algorithms still attribute more to the methods than they should. First, many users still believe that these methods identify models with minimum MSRES. Second, 'importance' too often is attached to variables as a result of whether or not they are included and/or remain in the model. Similarly, 'relative importance' is often associated with the order of entry or deletion. For example, the first variable entered in forward selection often is regarded as the most important variable while the first variable deleted in backward elimination is regarded as the least important. As Hocking (1976) noted, however, the order of entry and/or deletion of variables does not relate to variable importance. In fact, it is possible that the first variable to enter the model in forward selection is the first variable deleted in backward elimination.‡

In short, stepwise methods were not designed to find 'best' models or to indicate the relative importance of variables. On the contrary, they were designed to select subsets from data sets 'padded with extraneous variables—for example, those that contain everything we could measure' (Hoerl *et al.*, 1986, p. 378). It remains to be seen, however, whether these methods can indeed filter out the authentic variables

†A backward stepwise algorithm can also be implemented [See BMDP (Dixon *et al.*, 1988, p. 373); Younger (1985, pp. 489, 501–502)].

‡Correlation (collinearity) between predictor variables certainly affects the order of entry and/or deletion of variables (see Younger, 1985). Thus, Darlington (1968) and Huberty (1989) feel that attempting to assess importance when predictor variables are correlated is futile.

from a large pool of predictor variables padded with noise variables. Indeed, this was the purpose of the Monte Carlo study to be reported subsequently.

3. Issues relevant to stepwise algorithms

Stopping rules

As previously indicated, the stepwise algorithms will systematically enter and/or delete variables from a model until they run out of variables to consider or until a criterion, by which entry and/or deletion occurs, is satisfied. These criteria, which clearly affect the size of the subset selected, are referred to as stopping rules (Hocking, 1976).

Stopping rules typically are operationalized by setting the significance level of an F to enter (FTE) statistic in forward selection, an F to delete (FTD) statistic in backward elimination, and both an FTE and an FTD statistic in the stepwise procedure (see Hocking, 1976). With these stopping rules, the number of variables in the final model can be controlled. That is, by making the FTE sufficiently large not all possible variables are entered in the model; by making the FTD sufficiently small, not all possible variables remain in the model.

Bendel & Afifi (1977) compared many stopping rules in forward selection. Their findings suggest that a significance level of between 0.15 and 0.25 yields an FTE that is large enough to keep noise variables from being included in the model yet small enough to allow authentic variables to enter the model, with the best overall results occurring when $\alpha=.15$. These findings are consistent with those reported by Kennedy & Bancroft (1971) who recommended setting $\alpha=.15$ for the FTE in forward selection and $\alpha=.10$ for the FTD in the backward algorithm. Bendel & Afifi (1977) also suggested that their findings may be applicable to the stepwise algorithm, however, they believed that the FTD value should be set at a value equal to one half of the FTE value. This recommendation was supported by Hoerl *et al.* (1986) who compared a number of subset selection algorithms (e.g. stepwise, ridge-selection methods). Draper & Smith (1981, p. 309), on the other hand, stated that the values should be equal. This recommendation was adopted by Flack & Chang (1987) who used FTE/FTD values of .15 in their investigation of the stepwise algorithm.†

Inflation of Type I errors

As is well known, when many tests of significance are computed in a given experiment, the probability of making at least one Type I error in the set of tests, that is, the maximum familywise Type I error rate (MFWER), is far in excess of the probability associated with any one of the tests. This multiplicity of testing problem

†Interestingly, the statistical packages employ very different FTE and FTD default values. SAS' (SAS Institute, 1985) STEPWISE PROCEDURE uses .50 and .15 for entry values for their forward and stepwise algorithms, respectively, and .10 and .15 for deletion values in their backward and stepwise procedures, respectively. SPSSX's (Norusis, 1985) FTE and FTD default values are .05 and .10, respectively, for all algorithms. BMDP (Dixon *et al.*, 1988) uses default critical values of 4.000 and 3.996 which correspond to FTE and FTD values of .05.

(Tukey, 1977) clearly applies when using automated subset search algorithms to build models (see Lovell 1983; SAS, 1985, p. 765; Wilkinson, 1979). Consequently, many authors recommend FTE and/or FTD values which are substantially less than the .15 value recommended by Bendel & Afifi (1977), in order to maintain the MFWER$\leqq$.05 [Aitkin, 1974; Lovell (1983); Wilkinson (1979)]. Further research is needed to resolve these conflicting recommendations.

The coefficient of multiple determination

Many criteria have been proposed for assessing subset size (see e.g. Hocking, 1976). Many of these criteria, however, are monotonic functions of MSRES. In this paper we will discuss the criterion of proportion of variance accounted for since: (1) a frequent goal of model selection is to find a model that minimizes MSRES or, correspondingly, maximizes the proportion of variance in the response measure accounted for by the model; and (2) this statistic played an important role in our Monte Carlo investigation.

One of the most frequently used statistics for evaluating this proportion of explained variance is the coefficient of multiple determination, R^2, where

$$R^2 = \frac{\text{SSR}}{\text{SST}},$$

and SSR and SST are the regression and total sum of squares, respectively.

There are problems, however, associated with the use of R^2 as a measure of the variance explained by a model. Specifically, because the model selection procedure capitalizes on chance variation in identifying good models, the obtained value of R^2 is an overestimate of the population value. That is, even if predictor variables are uncorrelated with the response measure in the population, the sample coefficients of correlation (and their corresponding squared values) will assuredly assume non-zero values simply as a result of random sampling variation (Cohen & Cohen, 1983, p. 105). In fact, the larger the number of predictor variables, the more inflated are the R^2 values found through automated subset search algorithms.† Under cross-validation, therefore, we can expect substantial shrinkage in R^2 (Cohen & Cohen, 1983, pp. 105–107; Wilkinson, 1979). (See also Copas, 1983, for a general discussion of shrinkage.)

An alternate and popular approach to estimating the proportion of variance accounted for by the model is given by the adjusted ('shrunken') $\tilde{R}^2$, where

$$\tilde{R}^2 = 1 - (1 - R^2) \cdot \frac{n-1}{n-k-1},$$

where n stands for the sample size. Unfortunately, as Cohen & Cohen (1983, pp. 106–107) noted, the size of $\tilde{R}^2$ is also inflated when utilizing stepwise algorithms,

†To assist in the problem of inflated R^2 values, Pedhazur (1982, p. 149) noted the use of sample sizes as large as 500!

again as a result of capitalizing on chance variation. Consequently, they suggested that a better estimate of the population value and one which results in a more realistic estimate of shrinkage, is obtained by substituting p for k in the calculation of $\tilde{R}^2$.

It is also important to note that the test of R^2 is biased. When no subset selection has taken place, a significance test of R^2 is

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)},$$

with k and $n-k-1$ degrees of freedom. On the other hand, when a subset selection algorithm is used to choose the k predictor variables, that is, when k is not fixed but is *determined from the data*, then this statistic is not distributed as a central F variable (Pope & Webster, 1972).†

Since evaluating a best subsample R^2 statistic with critical values from the sampling distribution of $F(k, n-k-1)$ results in positively biased tests, some authors (Rencher & Fu Ceayong, 1980; Wilkinson, 1979) have addressed this problem by tabling critical values of R^2 which apply when a subset of k predictors is selected from the complete set of p variables. Others have provided a function to determine statistical significance of R^2 (Diehr & Hoflin, 1974). Wilkinson (1979) tabled the 95th and 99th percentage points of R^2 which are applicable to the forward selection algorithm, while Diehr & Hoflin (1974) and Rencher & Fu Ceayong (1980) provided results for the all possible and stepwise algorithms, respectively. To date, however, the work of these authors has not been adopted enthusiastically.

Collinearity and subset selection algorithms

Collinearity (correlation) among the predictor variables affects the size and stability of estimated regression coefficients (see Belsley, Kuh & Welsch, 1980, pp. 85–191; Farrar & Glauber, 1967; Fox, 1984, pp. 138–141; Gordon, 1968; Rockwell, 1975). Until recently, however, little work has focused on the effect of collinearity on subset selection. Citing the serious distortions that are introduced in ordinary least squares MLR by collinear data, Chatterjee & Price (1977, p. 200) simply state that they do not recommend the use of stepwise procedures in a collinear situation.

Lovell (1983) was the first to examine the use of automated subset selection algorithms (specifically forward selection) to determine, among other things, how frequently authentic predictor variables are selected from a pool of correlated predictor variables containing both authentic and noise variables. Simulating economic time series models, where the response measure was personal consumption expenditures and the predictor variables were a series of fiscal, monetary, etc., variables, Lovell (1983) found that, for collinear data, the authentic variables were chosen 70 per cent of the time.

Most recently, Flack & Chang (1987) compared the all-subsets (RSQUARE) and

†The FTE and FTD statistics are biased similarly (see Miller, 1984).

stepwise algorithms with respect to the frequency with which they specified subset models containing authentic versus noise predictor variables. For the parametric conditions they investigated, Flack & Chang (1987) found that *both* algorithms typically selected a large percent of noise variables (e.g. 67–89 per cent). In fact, the optimal (RSQUARE) procedure resulted in subset models containing more noise than authentic variables than did the stepwise procedure.† In addition, they found that the median values of their adjusted estimate of the multiple coefficient of determination were always larger than the true value [see Darlington, 1990, p. 121 for the definition of the adjusted R^2 used by SAS (SAS Institute, 1985)].

4. Literature summary

From the comments and studies just cited, it is clear that further work needs to be done to clarify some important issues related to the use of automated subset search algorithms to build models. Specifically, research on the choice of the level of significance for inclusion and/or deletion of predictor variables has resulted in two diverse recommendations (i.e. set $\alpha = .15$ vs. set $\alpha < .05$) yet there has been no study assessing these recommendations when using the stepwise procedures to build models from collinear data. In addition, research on the effects of collinearity on the subset algorithms is limited and, more importantly, difficult to generalize to educational and psychological phenomena. That is, in the studies just cited, the conditions that were investigated were not representative of conditions likely to be encountered in educational and psychological research. For example, in the Flack & Chang (1987) study, which is most relevant to the investigation we will reported subsequently, the correlations between authentic variables and the response variable was set at a value of .50, a value which is typically much larger than those characteristic of behavioural science relationships (Cohen, 1969, pp. 74–78; Cohen & Cohen, 1983, pp. 160–161; Flack & Chang, 1987). In addition, Flack & Chang (1987) created correlations between the predictor variables through serial correlation, which has limited generalizability to behavioural science investigations. Similarly, Flack & Chang's (1987) autocorrelation values of .3 and .5 do not reflect the range of collinearity typically found in behavioural science investigations. Finally, the information regarding the biases of the sample estimates of the population coefficient of multiple determination is very limited, particularly for collinear data.

Accordingly, the goal of our simulation study was to extend the research on the selection of predictor variables with automated subset algorithms under parametric conditions more characteristic of behavioural science investigations. The results of our study should be relevant to those who use these algorithms to derive descriptive models. Indeed, the utility of these reduced descriptive models would depend on whether they contain authentic and/or noise variables.

†This was determined by comparing the RSQUARE and stepwise results when the correlation between the predictor variables equalled .30, as comparisons between these two methods could only be made at this value.

Table 1. Parameters of the study

$\rho_{X_j X_{j'}}$	$\beta_1 = \beta_2 = \cdots = \beta_6$	$\beta_j\ (j>6)$	ρ^2_{YX}
0.0	.147442	0.0	.130435
0.4	.049147	0.0	.043478
0.8	.029488	0.0	.026087

The regression coefficients are calculated from the relationship $\beta = (X^T X)^{-1} X^T Y$, where in correlation transformation form, $\beta = \rho^{-1}_{X_j X_{j'}} \rho_{YX_j}$. ρ^2_{YX} is obtained from the relationship: $\rho^2_{YX} = \beta_1 \rho_{Y1} + \beta_2 \rho_{Y2} + \cdots + \beta_k \rho_{Yk}$.

5. Methods of the Monte Carlo study

The linear model

The study investigated the linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i \qquad i = 1, \ldots, N, \tag{1}$$

where Y_i is the response observation, $X_{i1}, X_{i2}, \ldots, X_{ip}$ are the ith observations on the p predictor variables measured without error, β_0 is the Y intercept, $\beta_1, \beta_2, \ldots, \beta_p$ are the p regression parameter constants, where β_j $(j = 1, \ldots, p)$ measures the change in Y per unit change in X_j when all other independent variables are held constant. It is assumed the Y and $X_1, \ldots, X_p$ are randomly distributed and that their joint distribution follows a $p+1$ multivariate normal distribution with mean 0 and covariance (correlation) structure defined below.

Authentic variables for MLR

The simulations modelled the MLR model given in equation (1) where k predictor variables were chosen from p candidate variables by a subset selection algorithm. Among the p candidate variables, 6 were authentic and $p-6$ were noise. An authentic variable was defined as a predictor variable whose corresponding population regression coefficient from the full p model regression equation is non-zero. Predictor variables with corresponding zero population regression coefficients were defined as noise variables.

The values of the regression coefficients associated with the authentic variables were determined by assuming a medium effect size (ES) for the population multiple coefficient of determination, ρ^2_{YX}. That is, Cohen & Cohen's (1983) effect size index, f^2, for the coefficient of multiple determination is $f^2 = \rho^2_{YX}/(1 - \rho^2_{YX})$. In this study $f^2 = .15$, which Cohen & Cohen (1983, p. 161) operationally define as an effect of medium size. For $f^2 = .15$, $\rho^2_{YX} = .130435$. In order to obtain the simple correlations between authentic predictor variables and the response variable, ρ^2_{YX} was divided evenly among the authentic predictor variables; thus $\rho_{YX_1} = \rho_{YX_2} = \cdots = \rho_{YX_6} = .147442$. Though this choice is one among many and is therefore arbitrary, it coincides with one strategy used by Cohen & Cohen (1983, pp. 118–119) for determining sample size for *a priori* power determination. [In Flack & Chang's (1987) study $\rho^2_{YX} = .3125$, .3725, and .5000 while $\rho_{YX_1} = \rho_{YX_2} = .5$.] Table 1 contains the values of ρ^2_{YX} and β_j used in the simulations.

Data generation and parameters of the study

The study investigated five factors: (1) the intercorrelations between predictor variables ($\rho_{X_j X_{j'}}$), that is the degree of collinearity; (2) the number of candidate predictor variables (p); (3) the size of the sample (N); (4) the level of significance for the inclusion and/or deletion of candidate variables; and (5) the type of subset selection algorithm.

In order to obtain the empirical results, N observations were generated for each of p candidate predictor variables by the algorithm employed by Galarneau-Gibbons (1981), McDonald & Galarneau (1975), and Wichern & Churchill (1978), that is,

$$X_{ij}=(1-\pi^2)^{1/2}Z_{ij}+\pi Z_{i(p+1)}, \quad i=1,\dots,N; j=1,\dots,p, \tag{3}$$

where Z_{ij} and $Z_{i(p+1)}$ are independent identically distributed pseudorandom $N(0,1)$ variates and π is prespecified. The resulting candidate variables have a pairwise correlation of π^2. Of the p candidate predictor variables, the six authentic variables were generated using a non-zero value of π which reflects a collinearity pattern. The remaining $p-6$ noise candidate variables were uncorrelated and were generated by using a simplified version of (3) where $\pi=0$. The pseudorandom unit normal deviates were generated by the procedure due to Marsaglia, McLaren & Bray (1964).

Three degrees of correlation between predictor variables were studied. In particular, we let $\rho_{X_j X_{j'}}=0.00$, 0.40, and 0.80. The range of 0.40 to 0.80 typifies the magnitude of intercorrelations found in behavioural science test batteries (see Cronbach, 1970; Sax, 1989; Thorndike & Hagen, 1977). A value of $\pi=\rho^{1/2}_{X_j X_{j'}}$ corresponding to each value of $\rho_{X_j X_{j'}}$ was used to generate the predictor variables having the specified correlation structure. Having generated the X matrix, observations on the response variable were generated by

$$Y_i=\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\beta_3 X_{i3}+\beta_4 X_{i4}+\beta_5 X_{i5}+\beta_6 X_{i6}+\varepsilon_i \qquad i=1,\dots,N. \tag{4}$$

The predictor and response variables were then standardized so that $X^{\mathrm{T}}X$ and $X^{\mathrm{T}}Y$ were in correlation form (T is the transpose operator); hence β_0 was equal to zero.

The numbers of candidate predictor variables were $p=12$, 18 and 24. This represents the case where 50 per cent (50%), 33.3 per cent (66.7%) and 25 per cent (75%) of the available candidate predictor variables were authentic (noise), respectively.

In order to investigate the effects of sample size, three values of N were studied. A sample size of 60 was determined to be that which would yield a power value of .80 to detect the squared semipartial correlation coefficients (see Cohen & Cohen, 1983, pp. 116–119), assuming that $\rho_{X_j X_{j'}}=0.0$. In addition, sample sizes that were 50 per cent ($N=30$) and 150 per cent ($N=90$) of the $N=60$ sample size were also generated.

The three subset selection algorithms that were compared were those provided for by the SAS (SAS Institute, 1985, pp. 763–774) STEPWISE procedure, that is, BACKWARD ELIMINATION, FORWARD SELECTION, and STEPWISE. For each subset selection algorithm, the level of significance for entry and/or deletion of variables was set at 0.15, $\alpha_i = 1-(1-\alpha_{\mathrm{MFWER}})^{1/p}$, or .05. The 0.15 value was chosen as it reflected the recommendations of Bendel & Afifi (1977) and corresponded to the value used by Flack & Chang (1987). The α_i value was chosen to reflect a concern for the issue of multiplicity of testing. Lovell (1983) and Wilkinson (1979) documented how the MFWER was inflated when k predictor variables were chosen from p candidate predictor variables. For $p=12$, 18, and 24 candidate predictor variables, the MFWER equals .858, .946, and .980, respectively, when $\alpha=.15$. The value of α_i therefore was chosen to limit the MFWER to 0.15. For $p=12$, 18, and 24, the protected entry and deletion values were .0134519, .0089882, and .0067481, respectively. Finally, since many statistical software packages use .05 as a default level of significance [see for example BMDP (Dixon *et al.*, 1988, p. 381; SPSSX (Norusis, 1985, p. 57), and MINITAB (Ryan, Joiner & Ryan, 1981)], this value was also investigated. One should note however, that, for $\alpha=.05$, the MFWERs are .460, .603, and .708 for $p=12$, 18, and 24, respectively.

Computer simulation

For each combination of $\rho_{X_jX_{j'}}$, p and N, 250 samples were generated. This number is ambitious as compared to other studies where 25 to 50 replications were used (Berk, 1978; Flack & Chang, 1987; Lovell, 1983). Further, we felt that 250 replications were sufficient to reliably demonstrate the frequency of obtaining authentic and noise variables. Each set of data was generated according to equations (3) and (4) using FORTRAN and was stored on disk. The data were then read from disk by SAS and subjected to each of the three subset selection algorithms. Results from each of these algorithms were again stored on to disk so that SAS could read this information and strip off the values of R^2 and identify the predictor variables in the final model.

6. Results of the simulation study

For each simulation, the following measures were taken: (1) frequency counts of both the number of authentic and noise variables that entered the subset model selected by each algorithm; and (2) three estimates of the population p model multiple coefficient of determination. These estimates were:

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}}, \tag{5}$$

$$R^2_{A(k)} = 1-(1-R^2)\frac{N-1}{N-k-1}, \quad \text{and} \tag{6}$$

$$R^2_{A(p)} = 1 - (1 - R^2)\frac{N-1}{N-p-1}, \tag{7}$$

where SSR and SST are the respective regression and total sums of squares respectively for the final MLR model. $R^2_{A(k)}$ and $R^2_{A(p)}$ are the adjusted sample estimates of ρ^2_{YX}, where $R^2_{A(k)}$ is adjusted by the number of selected variables in the final model and $R^2_{A(p)}$ is adjusted by taking the total number of candidate predictor variables into account (see Cohen & Cohen, 1983, chapter 3). In addition, we computed means and standard deviations on each of the aforementioned measures.

The results that follow are based upon the following analyses of the dependent measures. First, the mean number of authentic and noise variables that were contained in the final subset models for each combination of algorithm × α-inclusion/deletion level and for the various values of $\rho_{X_jX_{j'}}$, p, and N were examined. These values are tabled in the Appendix of this paper. Secondly, for all combinations of the study factors, we examined the mean value of R^2, $R^2_{A(k)}$ and $R^2_{A(p)}$ associated with the final subset models. On the basis of these examinations, it was evident that trends due to the degree of collinearity, the number of candidate variables, and the sample size existed in the data.

In order to quantify the effects of these factors, for each combination of algorithm and α-inclusion/deletion level we computed trend effects (linear and quadratic) for each main and interaction effect in the three-way $\rho_{X_jX_{j'}} \times p \times N$ design. To determine the r^2_{effect} for each trend component, we then divided the linear and quadratic sums of squares for each effect by the model sum of squares. For each of the major linear effects the direction of the relationship was also noted.

The results of these analyses indicated that the degree of collinearity, the number of predictor variables, and the size of sample influenced the results of the subset selection algorithms; however, the effects of these factors with respect to their magnitude and direction were quite different for the dependent variables investigated. Specifically, the effects of these factors were as follows:

(1) The degree of correlation between the predictor variables affected the frequency with which authentic predictor variables found their way into the final model. This effect was negative in that fewer authentic variables gained entry into the final model as the correlation between the predictor variables increased in magnitude. Indeed, the degree of collinearity between predictor variables was the most important factor influencing the selection of authentic variables. The degree of collinearity also had a negative effect on the size of $R^2_{A(p)}$. This factor, however, was of secondary importance in determining the size of $R^2_{A(p)}$.

(2) The number of candicate predictor variables affected the number of noise variables that gained entry to the model. As expected, for a fixed number of authentic variables, as the number of candidate predictor variables increased the frequency with which noise variables entered the final models increased. The number of candidate predictor variables also affected the sizes of $R^2_{A(k)}$ and $R^2_{A(p)}$. With respect to the former, the effect was positive, while with respect to the latter, it was predictably negative. The number of candidate predictor variables was the most important factor affecting both the number of noise variables entering the model and the size of $R^2_{A(p)}$. Its effect was of secondary importance in determining the size of $R^2_{A(k)}$.

Table 2. Mean number of authentic and noise variables and percentage noise contained in the subset models

						$\rho_{X_j X_{j'}}$					
				0.0			0.4			0.8	
							VAR				
p	*N*	*p*/*N*	*A*	*N*	%*N*	*A*	*N*	%*N*	*A*	*N*	%*N*
	30	*.40*	.89	.58	39.	.63	.52	45.	.54	.53	49.
12	*60*	*.20*	1.28	.47	27.	.71	.46	39.	.60	.49	45.
	90	*.13*	1.70	.43	20.	.86	.47	35.	.72	.41	36.
	30	*.60*	.97	1.23	56.	.68	1.24	65.	.60	1.23	67.
18	*60*	*.30*	1.20	1.03	46.	.74	1.08	59.	.60	.95	61.
	90	*.20*	1.64	.96	40.	.87	.93	52.	.68	.89	57.
	30	*.80*	1.03	2.39	70.	.88	2.39	73.	.84	2.43	74.
24	*60*	*.40*	1.25	1.55	55.	.72	1.45	67.	.62	1.47	70.
	90	*.27*	1.66	1.44	46.	.83	1.36	62.	.68	1.50	69.

Key. *VAR* = Variable; *A* = Authentic; *N* = Noise; %*N* = Percentage noise.

(3) The size of the sample affected the number of authentic variables found in the final models as well as the size of $R^2_{A(k)}$. As expected, the size of the sample positively effected the number of authentic variables in the final model. However, this effect was surprisingly small. For example, in the STEPWISE procedure, the mean number of authentic variables was only increased from 1.164 to 1.683 when N increased from 30 to 90, even when $\alpha=0.15$. Sample size was found to be the major factor affecting the size of $R^2_{A(k)}$. Here the effect was negative, thus having an important effect on the inflation of this estimate.

In short, the major variables examined in this study, i.e. $\rho_{X_j X_{j'}}$, p, and N, affected the outcome of all subset selection algorithms. Further, the magnitude and direction of these effects were generally the same for both the FORWARD and STEPWISE algorithms, across the α-inclusion/deletion levels of significance. For the BACKWARD algorithm, however, the magnitude and direction of the effects as well as the pattern of these effects across the α-inclusion/deletion levels differed somewhat from those associated with the other two procedures. One consequence of these differences was that the final subset models generated by the BACKWARD procedure contained more predictor variables, with a greater percentage of these variables being noise variables. In addition, the BACKWARD procedure resulted in more inflated values of R^2 and $R^2_{A(k)}$ and less conservative values of $R^2_{A(p)}$.

To amplify on the major study findings, we refer the reader to Table 2 which presents the mean values, collapsed over subset algorithm, for the number of authentic and noise variables (and percentage of noise variables) contained in the final model, as a function of the major study factors. From the table it can be seen that the mean number of authentic and noise variables contained in the subset regression models was very much influenced by all three investigated factors. Increases in the degree of collinearity between predictor variables resulted in

Table 3. Mean R^2, $R^2_{A(k)}$, and $R^2_{A(p)}$ values

				p					
	12			18			24		
				N					
	30	*60*	*90*	*30*	*60*	*90*	*30*	*60*	*90*
				p/N					
	.40	*.20*	*.13*	*.60*	*.30*	*.20*	*.80*	*.40*	*.27*
				$\rho^2_{XX}=0.0(\rho^2_{YX}=.13)$					
R^2	.33	.19	.16	.40	.24	.18	.50	.27	.21
$R^2_{A(k)}$	.27	.16	.13	.34	.20	.15	.42	.22	.17
$R^2_{A(p)}$	.05	.04	.05	.03	.03	.03	.03	.02	.02
				$\rho^2_{XX}=0.4(\rho^2_{YX}=.04)$					
R^2	.29	.15	.11	.38	.20	.14	.49	.22	.16
$R^2_{A(k)}$	.24	.13	.09	.31	.17	.12	.41	.18	.13
$R^2_{A(p)}$	.03	.02	.02	.02	.02	.01	.03	.01	.01
				$\rho^2_{XX}=0.8(\rho^2_{YX}=.03)$					
R^2	.28	.15	.11	.37	.18	.12	.49	.22	.16
$R^2_{A(k)}$	.23	.12	.09	.30	.15	.10	.41	.18	.13
$R^2_{A(p)}$	.03	.02	.01	.02	.01	.00	.04	.01	.00

decreases in the number of candidate variables contained in the final model and increases in the proportion of these variables that were noise. Even in the most favourable case investigated ($\rho^2_{X_jX_{j'}}=0.0$, $p=12$, and $N=90$) 20 per cent of the variables finding their way into the model were noise. In the worst case ($\rho^2_{X_jX_{j'}}=0.8$, $p=24$, and $N=30$), 74 per cent of the selected variables were noise.

It is important to note that although our results generally are consistent with those reported by Flack & Chang (1987), our results do not convey a picture as grim as those reported by Flack & Chang (1987). The median percentage of noise variables contained in the final subset models ranged from 33 to 89 per cent in Flack & Chang's (1987) study (see their Table 3). In our study, our mean values for percentage of noise variables in the final model ranged from 20 to 74 per cent. This difference can be attributed to the fact that we (1) tabled values for $\rho_{X_jX_{j'}}=0.0$; and (2) investigated more favourable combinations of the number of candidate predictor variables and the sample size.

In order to further understand the effects of $\rho_{X_jX_{j'}}$, p, and N on the estimates of the population multiple coefficient of determination, we refer the reader to Table 3 where we have tabled the mean sample values of these estimates (R^2, $R^2_{A(k)}$, and $R^2_{A(p)}$) for all combinations of these factors. From Table 3, it can be seen that R^2 always overestimated the population values. Notice that as the value of p/N increased, the value of R^2 increased, progressively overestimating the population value. When the number of predictor variables was large relative to sample size, one should expect shrinkage in the R^2 estimate. The $R^2_{A(k)}$ adjusted value is indicative of this shrinkage. Although $R^2_{A(k)}$ was certainly a better estimate than R^2, it nonetheless also overestimated the population value for unfavourable combinations of p and N; $R^2_{A(k)}$ was reasonably close in value to ρ^2_{YX} for uncorrelated predictor variables and favourable

p/N ratios. On the other hand, the $R^2_{A(p)}$ statistic never overestimated the population values. For uncorrelated predictor variables, its conservative nature was quite pronounced; when correlation existed between the predictor variables, it provided a good approximation to ρ^2_{YX}.

7. Conclusions and recommendations

Based upon the results of our investigation and the findings reported by Flack & Chang (1987), it is evident that subset models selected through stepwise algorithms or optimal methods frequently will contain a *sizable* percentage of noise variables when the pool of predictor variables contain both authentic and noise variables. Indeed, over all conditions investigated, the average number of authentic variables found in the final subset models was always less than half the number of available authentic predictor variables. Furthermore, the incidence of noise variables in the final subset models is affected by the degree of collinearity among the predictor variables, the level of entry and/or deletion employed as a stopping rule, the sample size, the number of predictor variables in the pool, and various combinations of these factors.

The main conclusion to be drawn from these findings is that the initial set of predictors should be selected carefully, including for study only those variables that, according to theory/previous research, are known/expected to be related to the response variable. In short, and with respect to the number of variables to be included in a study, we subscribe to Cohen's (1990) position that 'less is more'. Our results indicate clearly that the 'data mining' approach to model building is likely to result in final models containing a large percentage of noise variables which will be interpreted incorrectly as authentic. Bluntly put '"If you torture the data for long enough, in the end they will confess." Errors of grammar apart, what more brutal torture can there be than subset selection? The data will always confess, and the confession will usually be wrong' (Copas in Miller, 1984, p. 412).

Secondly, educational and psychological researchers who use automated subset selection procedures should be aware of the operating characteristics of these procedures. Specifically, the issues ('best' models, criteria for good models, biased tests, entry and/or deletion levels, inflation of the Type I error rate, collinearity of predictor variables) that we have addressed should be kept in mind when applying these procedures. Finally, as was previously noted, while assessing variable importance is typically of interest to researchers, this assessment should not be based on the inclusion/non-inclusion of variables in the final model nor on the order of the entry/deletion of variables into/from the final model. For an excellent summary of the issue of assessing variable importance, see Huberty (1989). (Also see Darlington, 1990 and Rawlings, 1988.)

As a postscript, it is important to remember that once a model is selected, through whatever means, the data should be checked for deficiencies (i.e. outliers, non-normality, heteroscedasticity, and non-independence of errors).

Acknowledgements

We wish to thank Carl Huberty (University of Georgia), Joanne C. Keselman (University of

Manitoba), and two anonymous referees for their valuable comments. The research reported in this article was supported by the Social Sciences and Humanities Research Council of Canada (Award No. 410-88-0665) by a grant to the second author.

References

Aitken, M. A. (1974). Simultaneous inference and the choice of variable subsets in multiple regression. *Technometrics*, **16**, 221-227.

Belsley, D. A., Kuh, E. & Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.

Bendel, R. B. & Afifi, A. A. (1977). Comparison of stopping rules in forward 'stepwise' regression. *Journal of the American Statistical Association*, **72**, 46–53.

Berk, K. N. (1978). Comparing subset regression procedures. *Technometrics*, **20**, 1-6.

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

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, **45**, 1304–1312.

Cohen, J. & Cohen, P. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ: Erlbaum.

Copas, J. B. (1983). Regression, prediction and shrinkage. *Journal of the Royal Statistical Society*, Series B, **45**, 311-354.

Cronbach, L. J. (1970). *Essentials of Psychological Testing*, 3rd ed. New York: Harper & Row.

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

Darlington, R. B. (1990). *Regression and Linear Models*. New York: McGraw-Hill.

Diehr, G. & Hoflin, D. R. (1974). Approximating the distribution of the sample R^2 in best subset regressions. *Technometrics*, **16**, 317- 320.

Dixon, W. J., Brown, M. B., Engelman, L., Hill, M. A. & Jennrich, R. I. (1988). *BMDP Statistical Software Manual*, vol. 1. Berkeley: University of California Press.

Draper, N. & Smith, H. (1981). *Applied Regression Analysis*, 2nd ed. New York: Wiley.

Efroymson, M. A. (1960). Multiple regression analysis. In A. Ralston & H. S. Wilf (Eds), *Mathematical Methods for Digital Computers*, pp. 191–203. New York: Wiley.

Farrar, D. E. & Glauber, R. R. (1967). Multicollinearity in regression analysis: The problem revisited. *Review of Economics and Statistics*, **49**, 92-107.

Flack, V. F. & Chang, P. C. (1987). Frequency of selecting noise variables in subset regression. *The American Statistician*, **41**, 84–86.

Fox, J. (1984). *Linear Statistical Models and Related Methods*. New York: Wiley.

Furnival, G. M. & Wilson, R. B. (1974). Regressions by leaps and bounds. *Technometrics*, **16**, 499-511.

Galarneau-Gibbons, D. (1981). A simulation study of some ridge estimators. *Journal of the American Statistical Association*, **76**, 131-139.

Gordon, R. A. (1968). Issues in multiple regression. *The American Journal of Sociology*, **73**, 592-616.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics*, **32**, 1–49.

Hoerl, R. W., Schuenemeyer, J. H. & Hoerl, A. E. (1986). A simulation of biased estimation and subset regression techniques. *Technometrics*, **28**, 369-380.

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

Kennedy, W. J. & Bancroft, T. A. (1971). Model-building for prediction in regression based on repeated significance tests. *Annals of Mathematical Statistics*, **42**, 1273-1284.

Lovell, M. C. (1983). Data mining. *The Review of Economics and Statistics*, **65**, 1-12.

Marsaglia, G., MacLaren, M. D. & Bray, T. A. (1964). A fast procedure for generating normal random variables. *Communication of the ACM*, **7**, 4–10.

McDonald, G. C. & Galarneau, D. I. (1975). A Monte Carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association*, **70**, 407–416.

Miller, A. J. (1984). Selection of subsets of regression variables. *Journal of the Royal Statistical Society*, Series A, **147**, 389–425.

Neter, J., Wasserman, W. & Kutner, M. H. (1985). *Applied Linear Regression Models*, 2nd ed. Homewood, IL: Irwin.

Norusis, M. J. (1985). *SPSS^x Advanced Statistics Guide*. New York: McGraw-Hill.

Pedhazur, E. J. (1982). *Multiple Regression in Behavioral Research: Explanation and Prediction*, 2nd ed. New York: Holt, Rinehart & Winston.

Pope, P. T. & Webster, J. T. (1972). The use of an *F*-statistic in stepwise regression procedures. *Technometrics*, 14, 327–340.

Rawlings, J. O. (1988). *Applied Regression Analysis: A Research Tool*. Pacific Grove, CA: Wadsworth.

Rencher, A. C. & Fu Ceayong, P. (1980). Inflation of R^2 in best subset regression. *Technometrics*, **22**, 49–53.

Rockwell, R. C. (1975). Assessment of multicollinearity. *Sociological Methods & Research*, **3**, 308–320.

Ryan, T. A. Jr, Joiner, B. L. & Ryan, F. (1981). *Minitab Reference Manual*, University Park, PA: Pennsylvania State University.

SAS Institute (1985). *SAS User's Guide: Statistics*, 5th ed. Cary, NC: Author.

Sax, G. (1989). *Principles of Educational and Psychological Measurement and Evaluation*. Belmont: Wadsworth.

Thorndike, R. L. & Hagen, E. P. (1977). *Measurement and Evaluation in Psychology and Education*, 4th ed. New York: Wiley.

Tukey, J. W. (1977). Some thoughts on clinical trials, especially problems of multiplicity. *Science*, **198**, 679–684.

Weisberg, S. (1980). *Applied Linear Regression*. New York: Wiley.

Wichern, D. W. & Churchill, G. A. (1978). A comparison of ridge estimators. *Technometrics*, **20**, 301–311.

Wilkinson, L. (1979). Tests of significance in stepwise regression. *Psychological Bulletin*, **86**, 168–174.

Younger, M. S. (1985). *A First Course in Linear Regression*. Boston: Duxbury.

Received 10 September 1991; revised version received 14 May 1992

Appendix

Table A. Effect of collinearity, number of candidate variables and sample size (dependent variable: Mean number of authentic variables)

Method		S	S	S	B	B	B	F	F	F
E/D Level		α/p	.05	.15	α/p	.05	.15	α/p	.05	.15
$\rho_{x_j x_{j'}}$										
0.0		0.373	1.120	2.127	0.441	1.394	2.525	0.373	1.122	2.152
0.8		0.181	0.451	0.922	0.258	0.824	1.679	0.181	0.451	0.939
p										
12		0.332	0.720	1.402	0.388	0.884	1.744	0.332	0.720	1.412
24		0.227	0.732	1.452	0.276	1.154	2.236	0.227	0.735	1.488
N										
30		0.146	0.495	1.164	0.223	1.027	2.164	0.146	0.498	1.195
90		0.406	0.970	1.683	0.455	1.107	1.956	0.406	0.970	1.699
$\rho_{x_j x_{j'}}$	N									
0.0	30	0.184	0.679	1.556	0.253	1.207	2.332	0.184	0.682	1.587
	90	0.560	1.573	2.727	0.643	1.728	2.897	0.560	1.573	2.744
0.8	30	0.111	0.342	0.840	0.231	0.983	2.097	0.111	0.344	0.869
	90	0.267	0.577	1.017	0.328	0.759	1.439	0.267	0.577	1.029
p	N									
12	30	0.172	0.488	1.112	0.244	0.733	1.611	0.172	0.491	1.132
	90	0.493	0.972	1.684	0.545	1.083	1.919	0.493	0.972	1.689
24	30	0.124	0.511	1.240	0.220	1.404	2.833	0.124	0.512	1.289
	90	0.349	0.949	1.700	0.391	1.115	2.005	0.349	0.949	1.719

Note. S = STEPWISE; B = BACKWARD; F = FORWARD; E/D = Entry/Deletion. Interactions between factors that were found to account for at least a small ES (see Cohen, 1969, pp. 74–78) are also tabled. Main effect values are based on 2250 simulations while two-way interaction values are based on 750 simulations. To conserve space, we have excluded the empirical values associated with the intermediate level of each treatment variable examined. The complete table can be obtained from the second author.

Table B. Effect of collinearity, number of candidate variables and sample size (dependent variable: Mean number of noise variables)

Method		S	S	S	B	B	B	F	F	F
E/D Level		α/p	.05	.15	α/p	.05	.15	α/p	.05	.15
$\rho_{x_j x_{j'}}$										
0.0		0.101	0.674	2.062	0.168	1.195	3.000	0.101	0.681	2.097
0.8		0.102	0.640	2.012	0.171	1.174	2.985	0.102	0.645	2.035
p										
12		0.084	0.308	0.953	0.108	0.403	1.147	0.084	0.308	0.959
24		0.120	0.989	3.154	0.220	2.072	5.095	0.120	0.995	3.196
N										
30		0.095	0.708	2.253	0.223	1.854	4.310	0.095	0.713	2.293
90		0.104	0.618	1.906	0.128	0.739	2.209	0.104	0.622	1.923
p	N									
12	30	0.087	0.327	1.016	0.123	0.539	1.361	0.087	0.327	1.027
	90	0.077	0.284	0.889	0.097	0.319	1.004	0.077	0.284	0.889
24	30	0.097	1.096	3.609	0.329	3.636	7.995	0.097	1.111	3.680
	90	0.137	0.943	2.923	0.163	1.171	3.416	0.137	0.948	2.956

Note. See Table A note.