A Subset Selection Procedure for Regression Variables

by

George P. McCabe, Jr.* and James N. Arvesen**

Purdue University

Department of Statistics

Division of Mathematical Sciences

Mimeo Series # 316

February 1973

* Research supported in part by National Science Foundation grant GJ30269.
** Research supported in part by the Office of Naval Research Contract No. 0014-67-A-0226-00014 at Purdue University.

A Subset Selection Procedure for Regression Variables

George P. McCabe, Jr.* and James N. Arvesen**

Purdue University

West Lafayette, Indiana

Given a regression model with p independent variables, several methods are available for selecting a subset of size $t < p$ which gives an adequate description of the dependent variable. By using the capabilities of the computer, one can now determine the subset corresponding to the largest sample multiple correlation coefficient or equivalently the smallest residual mean square. Due to sampling variation, however, there is no guarantee that this corresponds to the smallest value of the expected residual mean square.

A procedure is presented to determine a collection of subsets, each of given size t, having the property that the probability of including the subset corresponding to the smallest value of the expected residual mean square is bounded below by some prespecified constant, $1 - \alpha$. An example using real data is examined to illustrate the technique.

INTRODUCTION

In the analysis of data which is assumed to be modeled by a regression equation, a problem of interest to many practicioners is the selection of an appropriate subset of the independent variables which adequately describe the variability of the dependent variable. Among the techniques most commonly employed are the forward selection, backward selection and stepwise procedures.

* Research supported in part by National Science Foundation grant GJ30269.
** Research supported in part by the Office of Naval Research Contract No. 0014-67-A-0226-00014 at Purdue University.

Computer programs implementing one or more of these methods are available in most statistical packages. A description of the algorithms is given in Draper and Smith (1963). These commonly used procedures, while requiring little computer time even for problems with a large number of independent variables, do not perform optimally in the sense of finding subsets of given size having the largest sample multiple correlation coefficient.

LaMotte and Hocking (1970) developed an algorithm for finding optimal subsets in the above sense while Furnival (1971) has proposed procedures for efficiently calculating all possible regressions. Although these methods obviously require more computer time than the sub-optimal procedures, the additional time required is not great for problems involving as many as twenty independent variables.

Suppose that for a given data set, one has found the subset of size t having the largest sample multiple correlation coefficient. It is possible that there are one or more other subsets of the same size with sample residual mean square nearly as small as the minimum. Certainly, if one views the sample values as estimates of population parameters, consideration must also be given to these subsets. In this paper, a method is investigated for selecting a collection of subsets which contains the subset corresponding to the minimum expected residual mean square with probability at least $1 - \alpha$, where α is a prespecified constant.

It should be noted that there are many alternative formulations for this problem. A thorough Bayesian approach incorporating the measurement cost of the variables is given by Lindley (1968). Other views are considered by Allen (1971), Hoerl and Kennard (1970a,b) and Mallows (1964).

In Arvesen and McCabe (1972), theoretical aspects of the procedure

described in this paper are discussed in detail. The application of the procedure requires, however, the computation of constants relating α to the selection rule. The present work presents an algorithm for calculating the constants.

DEFINITIONS AND NOTATION

Consider the model

$$Y = X\beta + \epsilon \tag{1}$$

where X is an Nxp known matrix of rank $p \leq N$, β is a pxl parameter vector and ϵ is an N-dimensional normal random vector with mean zero and covariance matrix $\sigma^2 I_N$. It is assumed that (1) represents the "true" model but that one is interested in fitting submodels of the following type:

$$Y = X_i\beta_i + \epsilon_i \tag{2}$$

where X_i is an Nxt submatrix of X obtained by deleting p-t columns, β_i is a txl parameter vector and ϵ_i is an N-dimensional normal random vector. All of the $k = \binom{p}{t}$ possible subsets are considered by letting $i = 1,\ldots,k$.

The residual sum of squares for each submodel in (2) is denoted by

$$SS_i = Y'Q_iY \tag{3}$$

where

$$Q_i = I_N - X_i(X_i'X_i)^{-1}X_i'. \tag{4}$$

It is easy to show (see, e.g. Searle (1971)) that SS_i/σ^2 has a non-central chi-quare distribution with N-t degrees of freedom and non-centrality parameter given by

$$(X\beta)'Q_i(X\beta)/2\sigma^2.$$

Thus, letting σ_i^2 denote the expected value of $SS_i/(N-t)$, it follows that

$$\sigma_i^2 = \sigma^2 + (X\beta)'Q_i(X\beta)/(N-t). \tag{5}$$

Let

$$\sigma^2_{[1]} \leq \sigma^2_{[2]} \leq \cdots \leq \sigma^2_{[k]}$$

denote the ordered values of the $\{\sigma_i\}$.

A procedure will consist of a rule for selecting a collection of subsets, i.e. a collection of models of the type in (2). Let P(CS) denote the probability that the collection will contain the subset corresponding to $\sigma^2_{[1]}$. Given a value of $\alpha > 0$, the goal of the procedure is to obtain a collection of subsets with the property that $P(CS) \geq 1 - \alpha$ for any configuration of the unknown parameters of the model.

Clearly, sample estimates of the σ_i^2 are easily calculated, e.g. using Furnival's (1971) method. However, due to the complex nature of the dependence relationships, the joint distribution of these statistics is not obtainable in useful form. An analysis of these difficulties is presented in Arvesen and McCabe (1972).

The procedure proposed for constructing the collections of subsets is of the same form as those developed by Gupta and Sobel (1962). They investigated the case in which the sample estimates of the variances are independent.

THE SELECTION PROCEDURE

Let c be a constant, $0 < c < 1$. The subset corresponding to X_i is to be included in the collection if

$$SS_i \leq SS_{[1]}/c \tag{6}$$

where

$$SS_{[1]} \leq SS_{[2]} \leq \cdots \leq SS_{[k]}. \tag{7}$$

Note that the collection is never empty.

The central problem in applying this procedure is the determination of the value of c which will be sufficient to insure the inequality $P(CS) \geq 1 - \alpha$ for all possible configurations of the parameters in the model equations (1) and (2).

The configuration of parameters which minimizes P(CS) in an asymptotic sense is discussed in Arvesen and McCabe (1972). This corresponds to β and $\{\beta_i\}$ in (1) and (2) being zero. Some computational experience indicates that the overall procedure is not particularly sensitive to this approximation. The computational steps involved in this determination are described in the next section. The results of this computation give an index K with the property that P(CS) assumes a minimum value (asymptotically) for

$$P^* = P(SS_K \leq SS_{[1]}/c) \tag{8}$$

where the Y vector in (3) is normal with mean zero and covariance matrix I_N. Note that without loss of generality, the parameter σ^2 may be taken to be unity. Setting $1 - \alpha = P^*$ in (8) gives the functional relationship between α and c. In Arvesen and McCabe (1972) analytic and numerical approaches to finding c given a value of α are investigated. Unfortunately, due to the complexity of the dependence structure of the random variables $\{SS_i\}$, these approaches

fail to give numerically useful results for practical problems. Therefore, an algorithm using simulation techniques is proposed as a solution. The details of this algorithm are presented in the next section.

COMPUTATIONAL PROCEDURES

A. DETERMINATION OF K

1. For $i = 1,\ldots,k$ and $i < j$ calculate

$$\rho_{ij} = \mathrm{tr}(Q_i Q_j) \tag{9}$$

where tr (M) denotes the trace of the matrix M. From (4),

$$Q_i Q_j = I - X_i (X_i'X_i)^{-1} X_i' - X_j (X_j'X_j)^{-1} X_j' + X_i (X_i'X_i)^{-1} X_i'X_j (X_j'X_j)^{-1} X_j'.$$

Hence,

$$\mathrm{tr}(Q_i Q_j) = N - 2t + \mathrm{tr}(X_i (X_i'X_i)^{-1} X_i'X_j (X_j'X_j)^{-1} X_j') \tag{10}$$

since X was assumed to be of full column rank. Direct evaluation of the last part of (10) would involve the calculation of the inverses and the evaluation of the trace of an NxN matrix. This would be inefficient both in terms of the amount of computation and numerical accuracy. With the aid of an efficient linear equation routine for symmetric matrices, the evaluation can be simplified as follows:

Use the linear equation routine to find matrices A(pxp) and B(pxp) where

$$(X_i'X_i)A = X_i'X_j$$

and

$$(X_j'X_j)B = X_j'X_i.$$

Then,

$$\mathrm{tr}(X_i (X_i'X_i)^{-1} X_i'X_j (X_j'X_j)^{-1} X_j') = \mathrm{tr}(AB).$$

Substitution into (10) gives ρ_{ij}.

2. Set $\rho_{ii} = N-t$ and $\rho_{ji} = \rho_{ij}$ for $i < j$. Note that these definitions are consistent with those above.

3. Let $\Gamma = (\rho_{ij})$ and let τ_i denote the diagonal elements of Γ^{-1}. Note that it is not necessary to calculate the full inverse.

4. K is the index associated with the largest of the $\{\tau_i\}$, i.e.

$$\tau_K = \max \{\tau_i: \quad i = 1,\ldots,k\}. \tag{11}$$

In the unlikely case of a tie, any K satisfying (11) may be chosen. In the example studied, the value of K had little effect on the estimate of c^{-1} obtained.

B. ESTIMATION OF c^{-1}

In this section, an algorithm for estimating the parameter c^{-1} which satisfies (8) is presented. Steps 1-5 describe the generation of the sample covariance matrix M under the assumption that the vector Y is normal with mean zero and covariance matrix I_N. The matrix M is then used to find the maximum squared multiple correlation coefficient ($R^2_{[k]}$) and R^2_K. The remaining steps indicate the procedure for processing the correlations to obtain the estimate of c^{-1}.

1. Calculate the Cholesky decomposition of the matrix X'X, i.e. find a lower triangular matrix P of dimension pxp such that

$$X'X = PP'.$$

2. Generate $Z = (Z_1,\ldots,Z_p)'$ where the Z_i are iid normal (pseudo) random variables with mean zero and variance one.

3. Let

$$W = PZ. \tag{12}$$

Note that W has the same distribution as X'Y when $\beta = 0$.

4. Generate G, a chi-square (pseudo) random variable with N-p degrees of freedom. For $N-p \geq 15$, the Wilson-Hilferty (1931) transformation may be used:

$$G = n(w(2/9n)^{\frac{1}{2}} + 2/9n + 1)^3$$

where $n = N-p$ and w is normal with mean zero and variance one. For smaller n, sums of squares of normals can be taken. Additional information concerning the generation problem is found in Newman and Odell (1971) and Mathur (1961).

Let

$$H = G + Z'Z, \tag{13}$$

where Z was generated at step 3. Note that (W,H) has the same distribution as (X'Y,Y'Y).

5. Form the array

$$M = \begin{pmatrix} Y'Y & Y'X \\ X'Y & X'X \end{pmatrix} \tag{14}$$

where X'Y = W and Y'Y = H from (12) and (13). Of course, this matrix should be stored in symmetric mode. The matrix M provides the input to a routine designed to calculate the maximum squared multiple correlation coefficient ($R^2_{[k]}$ for subsets of size t. The value of R^2_K is also determined. (Recall K was determined by (11)).

In the program described in McCabe, Arvesen and Pohl (1973), the Furnival (1971) routine was used. Alternatively, the LaMotte-Hocking algorithm could be used to find $R^2_{[k]}$ and a separate calculation incorporated for R^2_K.

Note that in terms of the R^2 values, (8) becomes

$$P^* = P((1-R^2_K)/(1-R^2_{[k]}) \leq c^{-1}). \tag{15}$$

6. Form the ratio

$$A = (1-R_K^2)/(1-R_{[k]}^2)$$

and store A.

7. Repeat steps 1-6, m times retaining the values $A_1, A_2, \ldots, A_m$.

8. Denote the ordered values of $\{A_i\}$ as

$$A_{[1]} \leq A_{[2]} \leq \cdots \leq A_{[m]}$$

Then the estimate of c^{-1} is

$$\hat{c}^{-1} = A_{[(1 - \alpha)m]} \tag{16}$$

where $1 - \alpha \simeq P^*$, the desired correct selection probability bound.

The choice of m in the above algorithm involves many considerations. Clearly, if $\alpha = 10^{-3}$ then $m = 10^3$ will be inadequate. For a given value of m, one can easily calculate a nonparametric confidence region for c^{-1} using standard methods. For example, see Hogg and Craig (1970, p. 352). From (6), it is clear that a conservative procedure would be to overestimate c^{-1}. Accordingly, an upper bound on c^{-1} can be determined by constructing a confidence region with coefficient $1-\delta$. Solving the equation

$$1-\delta \geq P(A_{[j]} \geq c^{-1}) = \Sigma_{i=0}^{j-1} \binom{m}{i} (1 - \alpha)^i \alpha^{m-i} \tag{17}$$

for j gives the desired upper bound as $A_{[j]}$. In practice, one could fix δ (e.g. $\delta = .10$) and then compare $A_{[(1 - \alpha)m]}$ with $A_{[j]}$. If the relative difference is small, then the value of m may be considered to be sufficiently large.

EXAMPLE

Longley (1967) examined the performance of several regression programs in terms of numerical accuracy for a particular set of economic data. Because of the colinearity of the independent variables and the interesting numerical properties of the data, this set was chosen to test and illustrate the procedures proposed in this paper. The complete date set can be found in Longley (1967).

There are N=16 observations on p=6 independent variables and 8 different dependent variables. The size of the subsets has arbitrarily been set at t=3. Thus, there are 20 possible subsets for each dependent variable. The program automatically includes the intercept for all models. Note that the estimate of c^{-1} is calculated from the X matrix alone and hence can be used for each dependent variable.

Table I gives the results of the calculations for c^{-1}. For this data set it would appear that m=1000 gives adequate estimates. Table II summarizes the subsets selected for the different dependent variables for α = .10 and .50. For n=10,000 the selection rules include a subset if

$$(1-R_i^2)/(1-R_{[1]}^2) \leq 1.432 \qquad (\alpha = .10)$$

and

$$(1-R_i^2)/(1-R_{[1]}^2) \leq 1.118 \qquad (\alpha = .50).$$

TABLE I

Estimates of c^{-1}

m	$1 - \alpha$	$\hat{c}^{-1}$	$1 - \delta$	j	$A_{[j]}$
1000	.90	1.443	.90	913	1.478
1000	.50	1.113	.90	521	1.220
10000	.90	1.432	.90	9071	1.449
10000	.50	1.118	.90	5065	1.120

TABLE II

Selected Subsets

Dependent Variables															
Y_1		Y_2		Y_3		Y_4		Y_5		Y_6		Y_7		Y_8	
SUBSET	R^2	SUBSET	R^2	SUBSET	R^2	SUBSET	R^2	SUBSET	R^2	SUBSET	R^2	SUBSET	R^2	SUBSET	R^2
346	.992**	145	.984**	134	.940**	456	.907**	245	.951**	236	.994**	246	.945**	145	.998**
234	.985	234	.982**	145	.934**	346	.904**	456	.945**	356	.991*	134	.930*	245	.997*
245	.983	134	.981*	124	.932**	146	.897**	124	.944*	136	.988	234	.927*	456	.997*
125	.982	146	.981*	234	.932**	246	.896**	346	.944*	346	.988	146	.926*	345	.996
136	.982	345	.980*	345	.931*	124	.895*	234	.943*	235	.979	245	.925*	234	.995
236	.982	456	.980*	146	.931*	134	.894*	235	.943*	234	.977	345	.925*	236	.992
356	.982	124	.979*	245	.925*	245	.894*	246	.943*	123	.976	124	.919	246	.992
235	.981	245	.979*	346	.925*	145	.891*	146	.941*	125	.975	346	.919	124	.990
123	.980	346	.979*	456	.925*	234	.890*	134	.935*	135	.973	145	.918	146	.990
256	.979	246	.978*	246	.920*	345	.869*	145	.935*	245	.973	456	.917	356	.990
135	.975	135	.967	236	.781	235	.859	236	.934*	256	.971	123	.783	346	.989
126	.973	136	.967	356	.720	136	.859	123	.933*	134	.965	135	.782	135	.988
246	.973	123	.966	235	.655	236	.859	126	.928	345	.962	136	.779	235	.988
134	.970	125	.965	135	.651	356	.859	136	.928	246	.958	235	.751	125	.987
124	.969	126	.965	123	.641	123	.858	156	.928	124	.952	356	.751	256	.987
345	.969	156	.965	125	.630	126	.858	256	.928	126	.936	236	.741	123	.985
146	.948	236	.959	136	.618	256	.858	356	.928	146	.932	256	.740	156	.985
156	.947	356	.959	256	.618	156	.857	125	.924	456	.932	125	.732	134	.978
456	.947	256	.958	156	.476	135	.854	135	.922	145	.929	126	.699	136	.971
145	.946	235	.956	126	.355	125	.849	345	.916	156	.906	156	.630	126	.962

* included in collection if $1 - \alpha = .90$

** included in collection if $1 - \alpha = .50$

From Table II, one could conclude that for predicting Y_1, variables 3, 4 and 6 appear to be the best set of three. However, for Y_3, there are several candidates which should be considered. In this case, other factors (e.g. cost) could be taken into account. Additional data if obtainable could be used to try to reduce the number of subsets in the selected collection.

PROGRAM

A FORTRAN program implementing the procedures described in this paper has been developed. A listing and instructions on the use of the program (McCabe, Arvesen and Pohl (1973)) is available from the authors. Running times for the CDC-6500 at the Purdue University Computer Center for the above example are about 10 secs. for the determination of K, 40 sec. for the estimates of c^{-1} when m=1000 and 380 sec. when m=10,000.

ACKNOWLEDGEMENT

The authors would like to thank Professor Herman Rubin for many helpful discussions.

References

Allen, D. M. (1971), "The prediction sum of squares as a criterion for selecting predictor variables", University of Kentucky Dept. of Statistics Report No. 23.

Arvesen, J. N. and G. P. McCabe (1972), "Subset selection problems for variances with applications to regression analysis". Purdue Univ. Dept. of Statistics Mimeo Series 309.

Draper, N. R. and H. Smith (1966), Applied Regression Analysis, Wiley, New York.

Furnival, G. M. (1971), "All possible regressions with less computation". Technometrics 13, 403-408.

Gupta, S. S. (1965), "On some multiple decision (selection and ranking) rules". Technometrics 7, 225-245.

Gupta, S. S. and M. Sobel (1962), "On selecting a subset containing the population with the smallest variance". Biometrika 49, 495-507.

Hoerl, A. E. and R. W. Kennard (1970a), "Ridge regression: applications to nonorthogonal problems". Technometrics 12, 69-82.

Hoerl, A. E. and R. W. Kennard (1970b), "ridge regression: biased estimation for nonorthogonal problems". Technometrics 12, 55-67.

Hogg, R. V. and A. T. Craig (1970), Introduction to Mathematical Statistics, Macmillan, New York.

LaMotte, L. R. and R. R. Hocking (1970), "Computational efficiency in the selection of regression variables". Technometrics 12, 83-93.

Lindley, D. V. (1968), "The choice of variables in multiple regression". Journal of the Royal Statistical Society, Series B, 30, 31-53.

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, 819-841.

Mallows, C. (1964), "Choosing variables in a linear regression; a graphical aid". Presented at the Central Regional Meeting of the Institute of Mathematical Statistics.

Mathur, R. K. (1961), "A note on the Wilson-Hilferty transformation of χ^2". Bulletin of the Calcutta Statistical Association 10, 103-105.

McCabe, G. P, J. N. Arvesen and R. Pohl (1973), "A computer program for subset selection in regression analysis". Purdue Univ. Dept. of Statistics Mimeo Series No. 317.

Newman, T. G. and P. L. Odell (1971), The Generation of Random Variates, Hafner, New York.

Searle, S. R. (1971), Linear Models, Wiley, New York.

Wilson, E. B. and M. M. Hilferty (1931), "The distribution of chi-square". Proceedings of the National Academy of Sciences 17, 684-688.