TECHNOMETRICS VOL. 9, NO. 4 NOVEMBER 1967

Selection of the Best Subset in Regression Analysis

R. R. HOCKING AND R. N. LESLIE

Texas A & M University

The problem of selecting the best subset or subsets of independent variables in a multiple linear regression analysis is two-fold. The first, and most important problem is the development of criterion for choosing between two contending subsets. Applying these criteria to all possible subsets, if the number of independent variables is large, may not be economically feasible and so the second problem is concerned with decreasing the computational effort. This paper is concerned with the second question using the C_p-statistic of Mallows as the basic criterion for comparing two regressions. A procedure is developed which will indicate 'good' regressions with a minimum of computation.

1. INTRODUCTION

The classical multiple linear regression problem is concerned with estimating the coefficients in the linear model

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + e \tag{1}$$

from a set of n responses Y to the various values of the input variables X_i in the presence of errors e assumed to be independent $N(0, \sigma^2)$. It is well known that if the data satisfy the linear model assumptions, (see e.g., Graybill (1961)) then estimates with many optimality properties are obtained by minimizing the residual sum of squares. The estimates are given by solving the 'normal equations' given symbolically by

$$X'X\beta = X'y \tag{2}$$

Frequently, having obtained the above estimates, say b_i, the investigator may ask if the data might not have been adequately explained by using a subset of the original input variables $X_1, X_2, \cdots, X_k$.

If all 2^k possible regressions are computed then a natural procedure used by many investigators is to plot the residual mean squares for each regression against the number of variables in the regression. From such a plot it is usually possible to identify a few regressions which are apparently superior and, depending on how it is to be used, possibly select a single best regression.

Recently, an improvement on this method was suggested by Mallows (1964) and (1966) and investigated further by Gorman and Toman (1966). Mallows

Received March 1967.

suggests that the 'standardized total squared error' be used as a criterion and he developed an estimate C_p of this quantity given by

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} - (n - 2p). \tag{3}$$

In this equation, p is the number of variables in the regression, RSS_p is the residual sum of squares for the particular p-variate regression being considered and $\hat{\sigma}^2$ is an estimate of σ^2, frequently the residual mean square from the complete regression. Mallows has shown that regressions with small bias will have C_p's nearly equal to p and so this as well as the magnitude of C_p is used to judge a particular subset regression.

If the total number of input variables, k, is not too large then the residual sum of squares can be computed for each of the 2^k regressions and compared via the C_p-graph. With present computing equipment and efficient schemes for doing the computation such as that suggested by Garside (1965), this is not an expensive task if $k \leq 10$. However, the amount of computation does increase exponentially with k and since regressions with 20 or more variables are not rare some thought should be given to decreasing the computational effort. This is the purpose of the Gorman and Toman (1966) paper and of the present paper.

It is clear that of the $\binom{k}{p}$ possible regressions of size p only a few will be considered to be good regressions. Of particular interest is that subset of size p for which the residual sum of squares and hence C_p is minimal. In the next section, a procedure is described which frequently allows this subset to be identified after having computed the residual sum of squares for only a small fraction of the possible $\binom{k}{p}$ subsets. The method has the added feature that those regressions which must be computed to identify this particular subset generally include most of the regressions with small C_p.

2. Description of the Method

It will be convenient in this discussion to refer to the $r = k - p$ variables which are to be removed from the regression rather than the p variables which are to be retained. We shall also refer to the 'reduction in the regression sum of squares' due to removing a set of r variables. It is clear that the set of r variables for which this reduction is minimum determines that set of p variates to be retained for which the residual sum of squares is minimum. The C_p statistic can also be computed directly from this reduction. In fact, if σ^2 is estimated by the residual mean square for the complete regression, and Red_p is used to indicate the reduction, the statistic is given by

$$C_p = \frac{\text{Red}_p}{\hat{\sigma}^2} + (2p - k). \tag{4}$$

If a single variable, say the ith, is removed from the regression, it is well known that the reduction is given by $\hat{\sigma}^2 t_i^2$ where

$$t_i^2 = \frac{(b_i)^2}{\hat{\sigma}_{b_i}^2} \tag{5}$$

is the square of the usual t-statistic associated with the ith regression coefficient. Since these univariate reductions will be referred to frequently we introduce the symbols θ_i $i = 1 \cdots k$ defined as

$$\begin{aligned}\theta_i &= \hat{\sigma}^2 t_i^2 \\ &= \text{Reduction due to eliminating } i\text{th variable.}\end{aligned} \tag{6}$$

The first step of the procedure is to compute the full regression by solving the normal equations (2) and then evaluate the k univariate reductions, θ_i . Without loss of generality, assume that the variables are labeled according to the order on the θ_i . That is,

$$\theta_1 \leq \theta_2 \leq \cdots \leq \theta_k \tag{7}$$

Thus, with this labeling, the subset of size $k - 1$ with minimum residual sum of squares is obtained by deleting the first variable.

Before proceeding to the general development we shall state a fundamental property of quadratic forms which is the basis for the method:

> If the reduction in the regression sum of squares due to eliminating any set of variables for which the maximum subscript is j is not greater than θ_{j+1} , then no subset including any variables with subscripts greater than i can result in a smaller reduction.

The method which we shall now describe is a sequential one consisting of at most $p + 1$ stages for each value of $p = 1, \cdots, k - 2$. To fix the ideas the first three stages for given p will be described then the general sequential step will be described.

Stage 1 consists of computing the reduction due to eliminating variables $1, 2, \cdots, r$ for $r = k - p$ under the relabeling indicated in expression (7). If this reduction does not exceed θ_{r+1} then application of the above property allows the process to be terminated and the regression consisting of the p variables $r + 1, \cdots, k$ is seen to be the 'best' regression of size p in the sense of minimum residual sum of squares. (For the remainder of this discussion we shall, for brevity, use 'best' in this sense.)

If the reduction computed in stage 1 exceeds θ_{r+1} then no decision can be made at this point and we proceed to stage 2 and include variable $r + 1$ among the candidates for elimination. The $\binom{r}{1}$ reductions due to eliminating any set of r variables selected from the first $r + 1$ but containing the $(r + 1)$st variable are then computed. If the smallest of the $1 + \binom{r}{1}$ reductions computed to this point does not exceed θ_{r+2} the iteration is terminated and the corresponding regression is 'best'. If not, no decision can be made at stage 2 and we proceed to stage 3.

In stage 3 the reductions are computed for all subsets of size r selected from the first $r + 2$ variates which contain variable $r + 2$, a total of $\binom{r+1}{2}$ computations. The minimum of the $1 + \binom{r}{1} + \binom{r+1}{2}$ reductions from the first three stages is now compared with θ_{r+3} and the iteration either terminated or continued to the next stage.

In general, at any stage, say the qth, a total of $\binom{r+q-2}{q-1}$ reductions must be computed and checked to see if the 'best' subset can be identified. In the qth stage the largest subscript on any variable being considered is $r + q - 1$ and hence the search can be terminated if the minimum of the $\sum_{j=1}^{q} \binom{r+j-2}{j-1}$ reductions computed in the first q stages does not exceed θ_{r+q} and the corresponding subset is 'best'. If not, we proceed to stage $q + 1$ where subsets of size r containing variable $r + q$ are considered.

It is possible that all $p + 1$ stages must be completed and hence all of the $\sum_{j=1}^{p+1} \binom{r+j-2}{j-1} = \binom{k}{p}$ regressions evaluated. However, it has been observed that this rarely happens except for very small values of p which are not likely to be of interest and in any case do not require much computation.

3. Computational Considerations

The computation of the reduction due to eliminating a set of $r = k - p$ variables from the regression will require the inversion of a matrix whose dimension is the minimum of r and p. Thus, if $p \leq r$, it is most efficient to extract the appropriate $p \times p$ principal minor from the coefficient matrix for the complete regression and the corresponding right-hand sides and solve the resulting equations. The reduction is then computed directly. If $p > r$, it is more efficient to use the quadratic form which is the generalization of equation (6). Thus if C is the $r \times r$ principal minor of the inverse of the coefficient matrix for the complete regression and b is the r-vector of estimates from the complete regression corresponding to the variables being eliminated then the reduction is given by

$$b'C^{-1}b \tag{8}$$

Thus it is seen that the largest matrix to be inverted is not greater than the largest integer contained in $k/2$.

Although the method described in section 2 specifies the order in which variables are to be considered, the order in which computations are performed within a stage is not specified and an opportunity to make the computations more efficient is available. For a given $r = k - p$, a desireable sequence in which to compute the possible subsets of size r within the order prescribed in Section 2 is to proceed in such a way that only one variable changes from one subset to the next. That is, the new matrix of size $s = \min(r, p)$ to be inverted, say A for simplicity, differs from the previous one, say B, in only one row (and column by symmetry) say the ith. Given the inverse of B, the inverse of A is obtained quite efficiently by computing

$$A^{-1} = DB^{-1}E \tag{9}$$

Here D is the inverse of the elementary matrix which is an identity except for the ith column. This column is given by $B^{-1}a_i$ where a_i is the new ith column. The matrix D is then quite easily obtained for if $B^{-1}a_i = (u_1u_2 \cdots u_s)'$ then D is an identity except for the ith column which is given by

$$\left(-\frac{u_1}{u_i}, \cdots, -\frac{u_{i-1}}{u_i}, \frac{1}{u_i}, -\frac{u_{i+1}}{u_i}, \cdots -\frac{u_s}{u_i}\right)' \tag{10}$$

Similarly E is the inverse of the elementary matrix which is an identity except for the ith row. This row is given by $a_i'DB^{-1} = (v_1v_2 \cdots v_s)$. The matrix E is then given by the elementary matrix whose ith row is

$$\left(-\frac{v_1}{v_i}, \cdots, -\frac{v_{i-1}}{v_i}, \frac{1}{v_i}, -\frac{v_{i+1}}{v_i}, \cdots, -\frac{v_s}{v_i}\right) \tag{11}$$

The number of operations (multiplications and divisions) required to determine the new inverse by applying these formulae is $2(s^2 + 2s)$ as opposed to a number of order s^3 for determining the inverse directly.

There are many sequences for generating the subsets of size r so that the above computational procedure can be employed. Possibly the simplest to generate on the computer is given by borrowing from the sequence described by Garside (1965) for computing all possible subsets. If we consider the sets of variables which he generates as those to be removed from the regression, it is seen that imbedded in his sequence is, for any p, the sequence which is needed for our purpose.

4. Discussion

In evaluating the method described above there are several points to be considered. We have indicated in the introduction that it is necessary to consider only a small fraction of the 2^k regressions in order to determine the best subset of size p for $p = 1, \cdots k$. The size of this fraction will, of course, depend on the model from which the observations are taken, but to support the claim a limited Monte Carlo study was performed. In this study, data were simulated for 100 different regression models with $k = 10$, the models being selected at random. The 'best' subsets were determined for *all* p for each set of data with the following results:

(i) Of the 1024 possible subsets, the average number computed was 299. This includes the 10 subsets of size $k - 1$ which are always evaluated to determine the θ_i.

(ii) The efficiency of the procedure was very high for large p and decreased with p. With $p = 5$, for example, on the average only 55 of the 252 possible regressions were evaluated.

(iii) Since the regressions are of different sizes it is of more interest to observe that an average of 10, 700 operations were necessary for each problem as compared to over 43, 700 if all regressions were evaluated.

An added feature of the method is that the regressions which must be evaluated before the 'best' subset of any size is identified generally have small reductions.

This property which makes the method well suited for use with the C_p statistic was not unexpected when the order of computation is considered and has been observed in the application of the method.

As has been observed earlier, it may on occasion be feasible or desirable to compute C_p for all 2^k regressions. The method proposed here with the termination criterion removed has approximately the computational efficiency of the Garside procedure for computing all regressions.

5. An Example

In this section we present in some detail an example which will serve to illustrate the procedure as well as to demonstrate some of the properties mentioned above.

The data for this example are taken from the ten-variate example used by Gorman and Toman (1966) to illustrate their factorial approach hence it affords a comparison of the two methods. We refer to their paper for the data. Note that in our notation, $k = 11$, but since the constant is always included we have only 1024 possible regressions. To inspect only those regressions for which the constant is included, we need only apply the procedure described above to the correlation matrix.

The first step of the procedure is to evaluate the univariate reductions, θ_i, as described in Section 2 and reorder them as in expression (7). The θ_i and the corresponding C_p are shown in Table I along with the relation between our labeling of the variables, $1, 2, \cdots, 10$ and that of Gorman and Toman, $a, b, \cdots, k$. The reductions θ_i shown in Table I as well as those shown in later tables are computed from the correlation matrix and hence represent the reduction relative to a total sum of squares of unity. Only the relative magnitudes of the sums of squares are necessary for the selection procedure. The corresponding residual mean square is $\hat{\sigma}^2 = .004136$.

Referring to Table I we see that the best subset of size ten is obtained by deleting the first variable (or variable k) but that the subset obtained by deleting the second variable, d, is nearly as good.

Table I
Univariate Reductions and Relabeling

Variable	G.T.Label	θ_i	C_p
1	k	.00551	10.332
2	d	.00560	10.354
3	j	.00666	10.611
4	a	.01072	11.593
5	b	.02449	14.922
6	g	.02762	15.678
7	e	.03550	17.583
8	c	.05282	21.770
9	h	.06110	23.771
10	f	.06872	25.616

TABLE II
Results for $r = 2$, $p = 9$

Stage 1			Stage 2			Stage 3		
Subset	Reduction	C_p	Subset	Reduction	C_p	Subset	Reduction	C_p
12	.01345	10.251	23	.01132	9.738*	34	.02024	11.893
			13	.01266	10.062	24	.01897	11.587
						14	.01588	10.841

The determination of the best subset of size 9, that is, the best pair of variables to be deleted is illustrated in Table II.

In this table we show the three stages which had to be completed before the algorithm would terminate. The column labeled subset denotes those variables to be deleted from the regression. The other two columns for each stage show the reduction in the regression sum of squares (relative to unity) and the corresponding C_p. The subset with minimum reduction occurs in Stage 2, namely, the one in which variables (23) are deleted. The iteration could not be terminated at the end of Stage 2 because this reduction (.01132) was not smaller than $\theta_4 = .01072$. The iteration can be terminated at the end of Stage 3 since the minimum reduction to that point is smaller than $\theta_5 = .02449$. The algorithm required the completion of Stage 3 before it could be certain that (23) was the best pair to be removed, but note that Stage 3 contains regressions which should be considered. Thus by evaluating only six of the possible 45 regressions of size $p = 9$ we have identified the regression with minimum C_p as well as five additional good regressions.

In the application of this method, no other regressions would be evaluated and hence it is natural to ask if there are other good regressions which would be overlooked. To answer this question, the remaining regressions were evaluated and the results are summarized in the C_p-graph shown in Figure 1. Referring to the column for $p = 9$ the six white circles represent the C_p's shown in Table II and the two solid lines represent two groups of C_p's one of size 12 and another of size 27 which were not computed. In this case the results are excellent since only the six best regressions were evaluated. In general it has been observed that good regressions are rarely missed but it is not in general true that only those with smallest C_p are evaluated.

Note that the order in which the subsets are listed in Table II is just the desired order for application of the inversion method given in Section 3. That is, only one variable changes from one subset to the next. Garside's procedure for generating sequences of variables generates subsets of size 2 in precisely this order. Tables III and IV show the appropriate order for subsets of size 3 and 4.

The determination of the best subset of size $p = 8$ and $p = 7$, i.e. $r = 3$ and $r = 4$ is illustrated in Tables III and IV. From Table III it is observed that the best subset of size 3 to be deleted is the first one. That is, the best regression of size 8 is that one for which variables (123) are deleted. In this case the iteration was terminated at the end of Stage 2 since $.01951 < \theta_5 =$

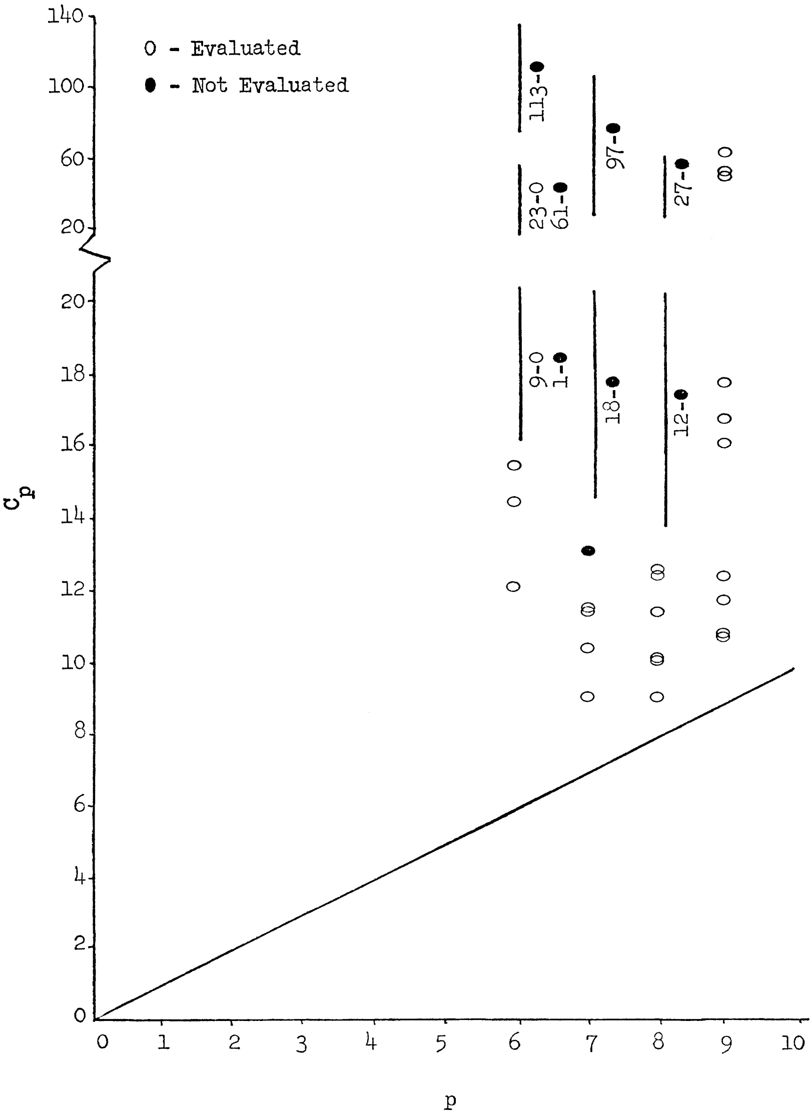

FIGURE 1
C_p-graph

TABLE III
Results for $r = 3$, $p = 8$

Stage 1			Stage 2		
Subset	Reduction	C_p	Subset	Reduction	C_p
123	.01951	9.717*	134	.02592	11.267
			234	.02742	11.630
			124	.02685	11.493

TABLE IV
Results for $r = 4, p = 7$

Stage 1		Stage 2		Stage 3		Stage 4	
Subset	C_p	Subset	C_p	Subset	C_p	Subset	C_p
(1234)	11.639*	1245	22.713	1256	22.991	1267	19.330
Reduction		2345	15.205	2356	13.896	2367	22.623
=.03573		1345	23.534	1356	24.748	1367	25.281
		1235	18.668	3456	18.618	3467	25.484
				2456	16.368	2467	18.218
				1456	27.413	1467	20.415
				1246	20.545	4567	20.359
				2346	17.452	3567	24.041
				1346	23.111	2567	17.759
				1236	19.490	1567	25.416
						1257	24.382
						2357	24.122
						1357	29.703
						3457	27.786
						2457	20.795
						1457	27.246
						1247	17.347
						2347	24.677
						1347	23.813
						1237	22.550

.02449. Thus a total of 4 out of a possible 120 regressions were evaluated. Reference to Figure 1 again shows that these are the four best subsets.

Inspection of Table IV reveals that, again, the first subset (1234) has minimum C_p but that it was necessary to complete four stages and hence a total of 35 of a possible 210 evaluations before the iteration was terminated. Reference to Figure 1 shows that those subsets which were not evaluated have large C_p but that in this case several regressions with large C_p had to be evaluated. It should be observed, however, that the iteration failed to terminate at Stage 3 because the minimum reduction .0357 was not less than $\theta_7 = .0355$. This suggests that a slight modification of the termination criterion might be considered. For example, the iteration might be terminated if the minimum reduction encountered through Stage q is less than $\theta_{r+q} + \epsilon$, for some small ϵ. For example, if $\epsilon = 0.1\hat{\sigma}^2$ we would be assured that even if we had not yet evaluated the regression with minimum C_p , the best C_p encountered would exceed the minimum C_p by less than 0.1. Our experience has shown that we would seldom overlook any good subsets and almost never overlook the best subset by such a modification.

We do not show the results for $p = 6$ in detail but point out that the minimum $C_p = 18.08$ occurred in Stage 2 as the result of the deletion of variables

(23456). As pointed out earlier, the best subset will invariably occur in one of the early stages, but the identification of the best subset becomes more difficult as p decreases. In this case a total of 126 of the possible 252 regressions had to be evaluated. This is reduced to 56 if a modified termination criterion is used.

It is clear that the locus of minimum C_p turns sharply upward as p is decreased. In general if the minimum C_p for, say $p = p^*$, is substantially greater than p^* there is no point in looking at smaller values of p. This suggests that the procedure be monitored and altered whenever (min $C_p - p$) exceeds a specified value. Alternately it is possible to decide in advance on the smallest value of p to be considered. For example, referring to Table I it is clear that for $r = 5$, i.e., $p = 6$ the minimum C_p is guaranteed to be greater than

$$C_p = \frac{\theta_5}{\hat{\sigma}^2} + 1 = 6.922$$

and in general will be much greater than this. Thus it is clear that there is little point in considering values of p less than 7.

In summary, using the suggested modifications it is seen that the determination of minimum C_p for $p = 7, 8, 9$ required the evaluation of only 25 regressions and most of these are contenders for the best regression.

6. Summary

The method described here for elimination of variables should not be judged alone on the example presented here or on the limited simulation study which was conducted. However, the method has been applied to several other regression problems where elimination of variables was suggested and in every case the performance has been comparable to the one discussed here. At this time we can only say that out limited experience has been encouraging.

7. Remarks

The Associate Editor has pointed out two, as yet unpublished, papers on this topic, "Best Models in Multiple Regression Analysis" by H. C. Kirton, and "The Discarding of Variables in Multivariate Analysis" by E. M. L. Beale, M. G. Kendall, and D. W. Mann. Both of these papers use the fundamental optimality principal given in Section 2, but are less specific about the order in which subsets are evaluated.

References

1. Garside, M. J., 1965. The Best Subset in Multiple Regression Analysis. *Applied Stat. Journ. of the Royal Statist. Society*, Series C., *14*.
2. Graybill, F. A., 1961. *An Introduction to Linear Statistical Models*. McGraw-Hill.
3. Gorman, J. W. and R. J. Toman, 1966. Selection of Variables for Fitting Equations to Data. *Technometrics*, *8*, pp. 27–51.
4. Mallows, C. L., 1964. Choosing Variables in a Linear Regression: A Graphical Aid. Presented at the Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, Kansas.
5. Mallows, C. L., 1966. Choosing a Subset Regression. Presented at the Joint Statistical Meeting, Los Angeles, Calif.