Biometrika (1967), **54**, 3 *and* 4, *p.* 357
Printed in Great Britain

The discarding of variables in multivariate analysis

BY E. M. L. BEALE, M. G. KENDALL AND D. W. MANN
CEIR Ltd, London

SUMMARY

In many multivariate situations we are presented with more variables than we would like, and the question arises whether they are all necessary and if not which can be discarded. In this paper we consider two such situations.

(*a*) *Regression analysis.* The problem here is whether any variables can be discarded as adding little or nothing to the accuracy with which the regression equation correlates with the dependent variable.

(*b*) *Interdependence analysis.* The problem is whether a constellation in p dimensions collapses, exactly or approximately, into fewer dimensions, and if so whether any of the original variables can be discarded.

We may define the best solution to (*a*) using any given number of variables as the one that maximizes the multiple correlation between the selected variables and the dependent variable, and similarly for (*b*) as the one that maximizes the smallest multiple correlation with any of the rejected variables. In practice it is usual to accept an approximate solution to (*a*) based on 'step-wise' multiple regression: we know of no standard program for (*b*). We have developed cut-off rules that enable us to find the best solution to both problems by partial enumeration. The paper discusses the details of this approach, and computational experience.

1. INTRODUCTION

In a great many multivariate situations we are presented with more variables than we would like and there arises the question of whether they are all necessary and, if not, which of them can be discarded. This arises as follows:

(*a*) *Regression analysis.* The problem here is the familiar one whether any variable can be discarded as adding little or nothing to the accuracy with which the regression equation correlates with the dependent variable.

(*b*) *Interdependence analysis.* The problem is whether a constellation in p dimensions collapses, exactly or approximately, into fewer dimensions and, if so, whether any of the original variables can be discarded.

(*c*) *Discriminant analysis.* The problem is whether any variables can be discarded without seriously impairing the discriminating power of the set.

(*d*) *Classification and cluster analysis.* Again the problem is whether any of the variables are not contributing to the classification or clustering technique.

In this paper we are concerned with (*a*) and (*b*).

The question of how many variables should be used is often a delicate one, and is decided on grounds that are only partly statistical. But for any given number of variables we may define the best solution to (*a*) as the one that maximizes the multiple correlation between the selected variables and the dependent variable. Similarly for (*b*) we may regard all the rejected variables as dependent variables and maximize the smallest multiple correlation coefficient between the selected variables and any of the rejected variables.

The difficulty with these definitions is that there is no simple way to compute the solutions and a complete enumeration of all combinations, which has been implemented by Garside (1965) for up to about a dozen independent variables, become impractical for much larger numbers of variables. Therefore people have used 'step-wise regression' (Efroymson, 1960), to get an approximate solution to (*a*). Problem (*b*) can be tackled by finding principal components and associating variables to be rejected with the eigenvectors corresponding to small eigenvalues, but this process becomes increasingly arbitrary as the number of variables to be rejected increases, and it also introduces an irrelevant problem of deciding on the commensurability of the different variables.

The purpose of this paper is to develop cut-off rules for an enumeration scheme that enable one to ignore combinations of variables that could not possibly improve on the best solution found so far. These rules have been implemented to provide an alternative version of the CEIR Multiple Regression program. This program can handle up to 120 independent variables. The largest problem solved so far with the new program had 16 independent variables and took 8 min. on the CDC 3200 computer. The step-wise program will therefore continue to be useful for exploratory analyses of larger problems; but by changing one control card the data are converted to a suitable format for a more detailed analysis using the new program.

Although we have applied our methods to both problems, (*a*) and (*b*), our detailed description concentrates on Multiple Regression. In §2 we review the pivoting process used in most Multiple Regression work, and in §3 we explain our cut-off rules. The details of the algorithm are deferred to the Appendix. Computational experience to date is summarized in §4.

In §5 we turn to interdependence analysis. We point out that only minor changes in the optimum multiple regression program are needed to solve this problem. We then discuss an application of the procedure, and compare it with the results obtained on the same problem using component analysis.

2. REGRESSION

We shall discuss the regression of a continuous scalar $y = x_{p+1}$ on a set of p continuous, regressor, scalar variables $x_1, \ldots, x_p$. We shall use the multiple correlation coefficient R^2 (or its complement, the proportional residual sum of squares) as a criterion of goodness of fit of the regressed y and the regression equation. It is convenient to define a r-subset as a subset of r independent variables. One r-subset is then regarded as being better than another r-subset if the regression equation based on the former yields a larger R^2 than that based on the latter. It does not matter if some of the regressands are functions of others, so that our method is applicable to curvilinear regression; though it is not applicable to genuine non-linear regression, i.e. to models that are non-linear in the unknown parameters.

If there are n observations and we denote the value of the jth variable on the ith observation as x_{ij}, we form the $(p+1) \times (p+1)$ matrix (a_{jk}), where

$$a_{jk} = \sum_{i=1}^{n} x_{ij} x_{ik}, \quad \text{or} \quad a_{jk} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

if we are fitting a constant term.

We can now find the best-fitting equation involving any subset of the variables by pivoting on the corresponding diagonal elements. The regression coefficient of x_j is given by minus the transformed coefficient $\bar{a}_{j,p+1}$, and the residual sum of squares is given by the transformed coefficient $\bar{a}_{p+1,\,p+1}$ (Efroymson, 1960).

The symmetry of the matrix $(\overline{a}_{jk})$ can be preserved by changing the sign of the pivot (Stiefel, 1963, p. 65), so that the formulae for the new array $(\overline{a}'_{jk})$ in terms of the previous array $(\overline{a}_{jk})$ when introducing the variable x_q are

$$\left.\begin{aligned} \overline{a}'_{qq} &= -1/\overline{a}_{qq} \\ \overline{a}'_{jq} = \overline{a}'_{qj} &= \overline{a}_{jq}\overline{a}'_{qq} \\ \overline{a}'_{jk} = \overline{a}'_{kj} &= \overline{a}_{jk} + \overline{a}_{jq}\overline{a}'_{qq} \end{aligned}\right\} \quad (j, k \neq q).$$

When removing the variable x_q the formulae become

$$\left.\begin{aligned} \overline{a}'_{qq} &= -1/\overline{a}_{qq} \\ \overline{a}'_{jq} = \overline{a}'_{qj} &= -\overline{a}_{jq}\overline{a}'_{qq} \\ \overline{a}'_{jk} = \overline{a}'_{kj} &= \overline{a}_{jk} - \overline{a}_{jq}\overline{a}'_{qk} \end{aligned}\right\} \quad (j, k \neq q).$$

In practice it is usual to scale the variables before starting the pivoting process so that the non-zero diagonal elements are all equal to unity. The effect of this scaling is removed by suitable multiplications before printing out any regression coefficients. The element $\overline{a}_{p+1,p+1}$ then represents the proportion of the original sum of squares of the dependent variable that is not accounted for in the current regression equation, and similarly, if x_j is not in the current equation, $\overline{a}_{jj}$ represents the proportion of the original sum of squares of deviations of x_j that is not accounted for as a linear function of the variables currently in.

Using these formulae it is easy to produce a number of alternative regression formulae. In particular it is easy to operate in a 'step-wise' manner, bringing in the variables one at at time, always choosing the variable that maximizes $\overline{a}^2_{j,p+1}/\overline{a}_{jj}$, i.e. the one that maximizes the reduction in $\overline{a}_{p+1,p+1}$, but taking care to avoid pivoting on an element $\overline{a}_{jj}$ that is nearly zero, since this would mean that x_j is an almost linear function of the variables already considered, and numerical difficulties would arise from attempting to fit an independent regression coefficient for this variable. One can then find a possibly different sequence of alternative equations, each involving one fewer coefficient than the previous one, by pivoting back on the diagonal elements one at a time, always choosing the variable that minimizes the increase in $\overline{a}_{p+1,p+1}$.

These are the well-known techniques of forward selection and backward elimination. They can be used, in series or in parallel, to produce an equation that cannot be significantly improved by bringing in any one variable and that cannot be simplified without an excessive increase in the residual sum of squares by removing any one variable. Unfortunately we cannot guarantee that we shall achieve the best equation using r variables in this way, unless $r = 1, p-1$ or p. This has been known for some time. J. Oosterhoff in an unpublished report from the Mathematical Centre, Amsterdam emphasizes the point by constructing examples where this is so even where forward selection and backward elimination give the same sequence of equations.

We would therefore like to have a method of regression analysis which generates optimum r-subsets. Conceptually the problem is very simple. Each of the p regressors can appear in a regression equation. There are thus 2^p possible regression equations. All we need to do is to work them all out and choose the best r-subset for each value of r. Garside discusses this approach. He points out that if the 2^p equations are enumerated in a particular systematic order then each can be derived from the previous one by a single pivot step.

In this way he has solved a problem with $p = 13$ in 4 min. on the Atlas computer. The difficulty is that the problem inevitably doubles in size for each additional variable and it is this factor which has barred progress along these lines.

3. CUT-OFF RULES FOR A PARTIAL ENUMERATION

Our regression program considers all possible sets of independent variables, other than those which are demonstrably worse than the best equation found so far. The cut-off rules to exclude these useless combinations are based on 3 ideas:

(*a*) Having made a possibly arbitrary selection of $r-1$ variables, and manipulated the matrix $(\bar{a}_{jk})$ so that only these variables are pivoted in, it is easy to see which single additional variable will improve the situation most. We simply choose the variable x_q such that $\bar{a}^2_{q,p+1}/\bar{a}_{qq}$ is maximized for all q such that $\bar{a}_{qq} > \epsilon$, where ϵ is some small tolerance. It is therefore unnecessary to carry out this final pivot operation, until one has found the very best combination of variables and wants to print out the corresponding equation.

(*b*) For each independent variable x_j we define an unconditional threshold T_j. This is the residual sum of squares achieved by using all linearly independent variables except x_j. Any subset which does not contain x_j cannot have a residual sum of squares less than T_j. Therefore if at some stage we find that a certain variable has an unconditional threshold greater than or equal to the smallest residual sum of squares achieved with any r-subset considered so far, which we call the *current best sum of squares*, then this variable must be in any subset which gives a better residual sum of squares, and we can confine our attention to such subsets.

(*c*) We can extend the concept of thresholds and compute conditional thresholds while exploring the possibility of excluding specified variables from the equation. So the conditional threshold of x_j if, say, x_{j_1} and x_{j_2} are excluded is the residual sum of squares achieved by using all linearly independent variables other than x_j, x_{j_1} and x_{j_2}. If the current best sum of squares exceeds the conditional threshold, we can confine our search of r-subsets excluding x_{j_1} and x_{j_2} to those including x_j.

Details of how the search of all subsets is carried out in our algorithm and of how the above ideas are incorporated are to be found in the Appendix.

4. COMPUTATIONAL EXPERIENCE

One problem used to test the performance of our program was the forecasting of the number of bookings on all U.K. airlines in any given month, from the corresponding numbers in each of the previous 12 months. We are not necessarily advocating the use of Multiple Regression for this particular forecasting problem, but we think that optimum regression may be useful for deciding which lagged variables to use in many forecasting applications. The results were compared with the corresponding results using the conventional backward elimination procedure, and, while the differences in the residual sum of squares for any given number of non-zero regression coefficients were all small, different selections were made for 3, 4, 5 and 7 variables. The point is illustrated most clearly by comparing the best 2 with the best 3 variables. The best 2 variables are lagged by 1 and 7 months, but the best 3 variables are lagged by 1, 5 and 8 months: the optimum program takes advantage of the possibility of re-spacing the variables when one is omitted. The complete problem, including an analysis of residuals on 2 of the equations, took 4 min. on the CDC 3200.

The largest problem that we have tackled so far involved 16 independent variables. Only 5 off-diagonal elements of the correlation matrix exceeded 0·5 in magnitude, so we did not expect our answers to differ significantly from those given by the step-wise method. Nevertheless, the 5 and 7 subsets were different. The complete problem, again including an analysis of residuals on 2 of the equations, took 7 min. on the CDC 3200.

It is of some interest to subdivide the time into time taken to find the best r-subset for each value of r and into other time. The results are shown in Table 1.

Table 1. *Analysis of computer time (min) on CDC* 3200 *for optimal regression program*

	Problem 1 (12 independent variables)	Problem 2 (16 independent variables)
Time to find 1-subset	0·08	0·09
2	0·06	0·16
3	0·17	0·34
4	0·48	0·57
5	0·26	0·43
6	0·14	1·01
7	0·11	0·17
8	0·11	0·29
9	0·07	0·48
10	0·07	0·63
11	0·08	0·21
12	0·08	0·08
13	—	0·08
14	—	0·08
15	—	0·08
16	—	0·08
Other time	2·29	2·22
Total	4·00	7·00

If all possibilities had to be enumerated, we would expect these times to be binomially distributed with $\varpi = \frac{1}{2}$. In fact the distributions are highly skew, with peaks when $r \simeq \frac{1}{3}n$. This implies that our cut-off rules are of considerable value.

Having developed the optimum regression program, we found ourselves in a situation where we could tackle a multiple regression problem by one or both of two extreme methods. We could carry out the straightforward process of forward selection and backward elimination, or we could use the optimum regression program to find the very best subsets. The optimum program may take a considerable time on a problem with more than 20 independent variables. We have therefore modified the optimum regression program to allow the option of terminating the search for the best r-subset after at most N possibilities have been examined, and printing the best solution found so far. The program will also print the number of possibilities that had been examined when the quoted solution was found. This should help the user to choose a suitable value of N in any future runs on similar problems.

It is important that a multiple regression program for routine use should be cheap to run, if only because some data errors will not be revealed until one does an analysis of the residuals from a preliminary analysis. But the conventional step-wise procedure was originally developed for the IBM 650 computer, and on most modern computers the time for a step-wise

program is dominated by input and output. It therefore seems reasonable to use the arithmetic speed of the computer to investigate some more possibilities even in routine work.

5. INTERDEPENDENCE ANALYSIS

It is easy to modify our optimum multiple regression program to turn it into an optimum interdependence analysis program in the sense defined in the introduction. This means that we select the set of r variables that maximizes the minimum multiple correlation of the selected variable with any rejected variable. All that is required is a re-definition of the details of the cut-off rules expounded in §3, as follows:

(*a*) To find the best rth variable given $r-1$, we pivot the $(r-1)$ variables in and keep the others out, and then compute, for each possible last pivot, what each diagonal element associated with a rejected variable would become. We choose our last pivot to minimize the largest of these quantities. The formula is to choose j to minimize

$$\max_k (\bar{a}_{kk} - \bar{a}_{kj}^2/\bar{a}_{jj}),$$

where k ranges over all variables other than the selected $r-1$ variables, and j ranges over all such variables for which $\bar{a}_{jj} > \epsilon$.

(*b*) The unconditional threshold T_j for any variable x_j is the value of $\bar{a}_{jj}$ when x_j is pivoted out and as many of the other variables as possible are pivoted in.

(*c*) The conditional threshold on x_j when say x_{j_1} and x_{j_2} are excluded is the largest of $\bar{a}_{jj}$, $\bar{a}_{j_1 j_1}$ and $\bar{a}_{j_2 j_2}$, when x_j, x_{j_1} and x_{j_2} are pivoted out and as many of the other variables as possible are pivoted in.

We illustrate our interdependence analysis calculations by considering 17 measurements on 67 lubricating oil basestocks, using data kindly supplied by the British Petroleum Research Laboratories, Sunbury. This material, collected in 1966, is thought to be homogeneous and not to contain any odd or outlying observations. There are no missing values.

The 17 variables are as follows:

1. Kinematic viscosity at 100 °F.
2. Kinematic viscosity at 140 °F.
3. Kinematic viscosity at 210 °F.
4. Viscosity index.
5. Refractive index.
6. Molecular weight.
7. Sulphur (%).
8. Specific gravity 60 °F./60 °F.
9. Flash point.
10. Pour point.
11. Neutralization value mg. KOH/g.
12. Carbon residue.
13. % N + P (weight % naphthalenes plus paraffin).
14. % A + S (weight % aromatic plus sulphur compounds and other polar materials).
15. % C_A (percentage carbon in aromatic rings).
16. % C_N (percentage in naphthal rings).
17. % C_p (percentage not in ring structure, i.e. in paraffin + alkyl groups).

These data are not linearly independent. In fact variables 13 and 14 always add up to 100, and variables 15, 16 and 17 always add up to 100. Variables 1, 2 and 3 are also nearly linearly dependent, log viscosity being a linear function of temperature.

We did not make any special provisions in the analysis for these dependencies.

The results are summarized in Table 2.

Table 2. *Interdependence analysis of lubricating oil basestock measurements*

No. of variables selected	Variables selected															Min R^2 with rejected variables	Max R^2 between selected variables
1	16	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0·0177	—
2	6	12	—	—	—	—	—	—	—	—	—	—	—	—	—	0·2010	0·4367
3	4	5	11	—	—	—	—	—	—	—	—	—	—	—	—	0·4722	0·6999
4	1	2	4	5	—	—	—	—	—	—	—	—	—	—	—	0·5092	0·9800
5	3	7	10	11	17	—	—	—	—	—	—	—	—	—	—	0·7327	0·4256
6	3	5	7	10	11	17	—	—	—	—	—	—	—	—	—	0·8077	0·9094
7	7	8	9	10	11	12	16	—	—	—	—	—	—	—	—	0·8591	0·6674
8	5	7	9	10	11	12	13	16	—	—	—	—	—	—	—	0·8895	0·8548
9	3	5	7	9	10	11	13	13	16	—	—	—	—	—	—	0·9527	0·9014
10	3	7	8	9	10	11	12	13	15	16	—	—	—	—	—	0·9605	0·9322
11	1	6	7	8	9	10	11	12	13	16	17	—	—	—	—	0·9655	0·9640
12	2	4	6	7	8	9	10	11	12	13	15	16	—	—	—	0·9869	0·9644
13	2	4	5	6	7	8	9	10	11	12	13	15	16	—	—	0·9931	0·9869
14	1	3	4	5	6	7	8	9	10	11	12	13	15	16	—	0·9997	0·9874
15	1	2	3	4	5	6	7	8	9	10	11	12	13	15	16	1·0000	0·9997

The first columns of this table are self explanatory. The penultimate column gives the value of the objective function being maximized. We see that nothing very useful can be done with less than 5 variables, but that 6 variables can account for over 80 % of the variability in every variable, 9 can account for over 95 %, and 13 for over 99 %.

The last column may be of some interest: it gives the largest value of R^2 by regressing one of the selected variables on the others, and indicates that the selected variables are usually not highly correlated among themselves. With less than 9 variables selected only on 2 occasions did this number exceed 90 %.

It is of some interest to compare those results with those obtained earlier using component analysis. A component analysis carried out on the correlation matrix for all variables revealed 2 zero eigenvalues associated with the linear relationships noted earlier, and one eigenvalue of 0·00016 associated with the nearly linear relationship between variables 1, 2 and 3. At this point variables 2, 14 and 17 were eliminated from the problem, giving the same answer for 14 variables as the optimum analysis described in the table. There can be no objection to removing variable 14, since variable 13 is fully equivalent to it, but the optimum analysis shows the undesirability of rejecting any of 15, 16 and 17 at this stage since if we are subsequently reduced to taking just one of these quantities we cannot tell which will be most useful.

A further component analysis was carried out on the correlation matrix for the remaining 14 variables. This produced 4 small eigenvalues of 0·00045, 0·00073, 0·00160 and 0·00206. It was then decided to eliminate four more variables associated with the eigenvectors corresponding to these eigenvalues. The principle adopted was to eliminate the variable with the largest coefficient in the vector expressed as a linear function of the variables. The variables

removed in this way were 3, 4, 5 and 6, which left 1, 7, 8, 9, 10, 11, 12, 13, 15 and 16. This selection gives a minimum R^2 with the rejected variable of 0·9501, which is nearly as good as the optimum set. Judged by the criterion of the minimum R^2 with an unselected variable, 10 is not a particularly useful number of variables to select, and indeed the optimum program does marginally better with 9 variables than the component analysis program with 10.

If on the other hand one looks at the average R^2 with the unselected variables, which is more in the spirit of component analysis, the component analysis 10 gives a value of 0·9819 as compared with 0·9832 using the 10 from our program or 0·9730 using the 9 from our program. So our program is still marginally better but 10 seems a less illogical number of variables.

A third component analysis was carried out on the correlation matrix for the 10 remaining variables. This produced 3 small eigenvalues of 0·00372, 0·00653 and 0·00949, which were associated with the variables 1, 8 and 13, leaving 7, 9, 10, 11, 12 and 15. This selection gives a minimum R^2 with the rejected variables of 0·8531 and an average R^2 of 0·9078, as compared with 0·8591 and 0·9068 for our program. There is therefore not much to choose between these selections.

A fourth component analysis was carried out on the correlation matrix for the 7 remaining variables. This produced the following eigenvectors; 0·365, 0·237, 0·179, 0·094, 0·059, 0·035 and 0·030. None of these is particularly small, but if more variables had to be rejected it appeared best to choose 9, 15 and 11 leaving 7, 10, 12 and 16.

This selection gives a minimum R^2 with the rejected variables of 0·2403 and an average R^2 of 0·6685, as compared with 0·5092 and 0·7662 for our program.

For this problem the component analysis approach produced reasonable results, although they were not optimum in any precise sense. It is therefore relevant to ask whether they required significantly less computation. The answer is that the component analysis required 4 runs of between 1 and 3 min. on the CDC 3200 computer as compared with a single 15-min. run for the optimum program. But the saving in machine time is hardly significant when one considers the fact that the optimum program produces all the results with the associated statistics without the need for manual intervention and the formulation of *ad hoc* rules by the statistician.

References

Garside, M. J. (1965). The best subset in multiple regression analysis. *Appl. Stat.* **14**, 196–200.

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

Stiefel, E. L. (1963). *An Introduction to Numerical Mathematics*. New York and London: Academic Press.

[*Received March* 1967. *Revised May* 1967]

Appendix

An algorithm for optimum regression. In principle the algorithm incorporates a search of all possible subsets which is based on the following lemma. If r variables can be either in or out of the regression equation, and they are arranged in any order, say $v_1, v_2, \ldots, v_r$, then one of the following $(p+1)$ mutually exclusive possibilities must hold:

(1) v_1 is out;
(2) v_1 is in but v_2 is out;
(3) v_1 and v_2 are in but v_3 is out;
...
(i) $v_1, \ldots, v_{i-1}$ are in but v_i is out;
...
(r) $v_1, \ldots, v_{r-1}$ are in but v_r is out;
($r+1$) all are in.

Terminology. A convenient way of carrying out the search is to introduce the idea of variables being in or out of the equation at various levels. We do this as follows. Variables in the equation are at positive or zero levels; variables out of the equation are at negative levels. Variables are in at level 0 if their unconditional thresholds exceed the current best residual sum of squares.

A variable may be in at a higher even level $2n$ if its conditional threshold, computed with the excluded set being all those variables out of the equation at some level above $-2n$, exceeds the current best residual sum of squares.

Variables in at level 3 have been introduced to make up the numbers initially.

Variables in at a higher odd level $2n+1$ *have been* introduced to make up the numbers when the (unique) variable has been set out at level $-(2n-1)$.

Variables may be out at level -1, which means simply that they have not been considered, or else at a negative odd level, which indicates a branch of the tree of possibilities being investigated.

The Algorithm. The algorithm is illustrated by the flow chart. Here N is the number of variables required in the regression equation.

The algorithm develops optimum regression equations for all values of N from NMIN to NMAX inclusive.

At any point in the search: NØIN is the number of variables at present in the equation, MLEVEL is an array containing the level at which each variable is in or out of the equations, MAXØD is the maximum odd level at which any variable is in the equation encountered so far, and BRSSQ is the *rssq* corresponding to this best equation.

Table 3. *Flow chart*