Cambridge Scientific Center

IBM
Data Processing Division

36, Y12
November 1966

Efficient Calculation of All Possible Regressions

Efficient Calculation of All Possible Regressions

S. Fienberg, M. Schatzoff, R. Tsao

IBM Cambridge Scientific Center Report

International Business Machines Corporation
Cambridge Scientific Center
Cambridge, Massachusetts

November, 1966

36. Y12
November, 1966
Scientific Center Report
Limited Distribution

EFFICIENT CALCULATION OF ALL POSSIBLE REGRESSIONS

S. Fienberg, M. Schatzoff and
R. Tsao

International Business Machines
Corporation

Abstract

This paper describes efficient computational procedures for calculating all possible 2^k regressions of a dependent variable upon subsets of k independent variables. The principal result of the paper is contained in theorem 1, which provides a constructive proof that all 2^k possible regressions can be accomplished by "sweeping" the cross-products matrix exactly 2^k times.

It is shown also that further economies in computation can be achieved by taking advantage of a general symmetry property of the cross-products matrix at each stage, and applying the sweep operation itself to a minimal sized submatrix at each step. The paper provides an example involving four independent variables, as well as a Fortran routine for generating an optimal sequence of sweeps.

Index Terms for the IBM Subject Index

Statistics
Regression Analysis
05-Computer Application
16-Mathematics

LIMITED DISTRIBUTION NOTICE

This report has been submitted for publication elsewhere and has been issued as a Technical Report for early dissemination of its contents. As a courtesy to the intended publisher, it should not be widely distributed until after the date of outside publication.

TABLE OF CONTENTS

I. INTRODUCTION

Numerous procedures have been proposed for the selection of a subset of k independent variables in fitting a multiple regression equation to a set of data. Included among these are various forward and backward stepwise selection techniques (Efroymson 1960, Hamaker 1962, Oosterhoff 1963) and the C_p statistic proposed by C. Mallows (1964.) Theoretically speaking, none of these methods can claim to achieve optimality, so that the sure way of finding that regression which is best according to some criterion is to carry out all possible 2^k regressions and use that criterion to select the "best" of the 2^k possible regressions. Because the number of calculations increases exponentially with k, it is particularly important to have an efficient algorithm for carrying out the necessary computations. However, if k is sufficiently large, it may not be practical to carry out all of the possible regressions. In such cases, Gorman and Toman (1966) have suggested the use of fractional factorial designs for selecting a subset of the 2^k regressions. They then use the C_p statistic as a criterion for selecting the "best" regression. In this situation as well, it is important to carry out the computations efficiently and to extract as much information as possible from the data. In this paper, we describe an efficient way of fitting all possible regressions, and offer some comments on the fractional factorial case.

The calculations required to fit a regression equation to a given set of data are essentially those of solving a set of simultaneous linear equations. A commonly used, and usually efficient direct method of carrying out such calculations is the Gaussian elimination (or pivotal inversion) method. (Wilkinson, 1965).

In section 2 of this paper we describe a variant of the usual Gaussian elimination method, hereinafter referred to as sweep (Beaton 1964), and show how it can be used to add and delete independent variables from a fitted regression. We then go on to show, in section 3, how to carry out all possible regressions efficiently. An example of the procedure is given in section 4.

In section 5, we offer some comments about the problem of selecting balanced fractions of all the possible regressions, as proposed by Gorman and Toman.

II. APPLICATION OF SWEEPING TO MULTIPLE REGRESSION

Following Beaton (1964), a square matrix $A = (a_{ij})$ is said to have been swept on the r^{th} row and column (or r^{th} pivotal element) when it has been transformed into a matrix $B = (b_{ij})$ such that

$$
\begin{aligned}
b_{rr} &= 1/a_{rr} \\
b_{ir} &= -a_{ir}/a_{rr} \qquad i \neq r \\
b_{rj} &= a_{rj}/a_{rr} \qquad j \neq r \\
b_{ij} &= a_{ij} - a_{ir}\, a_{rj}/a_{rr} \qquad i, j \neq r
\end{aligned}
\tag{2.1}
$$

It is easy to check from (2.1) that the sweep operator possesses the following useful properties:

1. Sweep is reversible.

   That is, sweeping a matrix twice on the same row and column is equivalent to not having swept the matrix at all.

2. Sweep is commutative.

   That is, sweeping a matrix first on the r^{th} and then on the s^{th} pivotal element is equivalent to sweeping the matrix in the opposite order.

In terms of regression analysis, consider the normal equations of regression theory in their matrix representation

$$(2.2) \qquad \underset{\sim}{X}' \underset{\sim}{X} \hat{\underset{\sim}{B}} = \underset{\sim}{X}' \underset{\sim}{Y}$$

where $\underset{\sim}{X} = \begin{pmatrix} X_{10} & X_{11} & \cdots & X_{1k} \\ \vdots & \vdots & & \\ X_{n0} & X_{n1} & \cdots & X_{nk} \end{pmatrix}$ $\underset{\sim}{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$ $\hat{\underset{\sim}{B}} = \begin{pmatrix} \hat{B}_o \\ \hat{B}_1 \\ \vdots \\ \hat{B}_n \end{pmatrix}$

and $X_{10} = \ldots = X_{n0} = 1$.

If the cross product matrix

$$(2.3) \qquad \underset{\sim}{C} = \left(\begin{array}{c|c|c} \underset{\sim}{X}' \underset{\sim}{X} & \underset{\sim}{X}' \underset{\sim}{Y} & \begin{matrix} 0 \\ \vdots \\ k \end{matrix} \\ \hline \underset{\sim}{Y}' \underset{\sim}{X} & \underset{\sim}{Y}' \underset{\sim}{Y} & k+1 \end{array}\right)$$

is swept on the first $k+1$ pivotal elements, then provided that $(\underset{\sim}{X}'\underset{\sim}{X})^{-1}$ exists, the result will be

$$(2.4)\qquad \underset{\sim}{C}^{*} = \left(\begin{array}{c|c} (\underset{\sim}{X}'\underset{\sim}{X})^{-1} & (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\,\underset{\sim}{X}'\underset{\sim}{Y} \\ \hline -\underset{\sim}{Y}'\underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-1} & \underset{\sim}{Y}'\underset{\sim}{Y} - \underset{\sim}{Y}'\underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{Y} \end{array}\right)$$

$$= \left(\begin{array}{c|c} (\underset{\sim}{X}'\underset{\sim}{X})^{-1} & \hat{\underset{\sim}{B}} \\ \hline -\hat{\underset{\sim}{B}}' & (n-k-1)\,\hat{\sigma}^{2} \end{array}\right)$$

where $\hat{\underset{\sim}{B}}$ and $\hat{\sigma}^{2}$ are the least squares estimates for the regression coefficients and error variance respectively. In general, sweeping $\underset{\sim}{C}$ on any subset of the first $k+1$ pivotal elements will yield the estimates of the regression coefficients and error variance corresponding to the regression of Y on that subset of the X's. Thus, by the reversibility and commutativity properties of sweep, each application of sweep to a particular row and column of the cross-products matrix either introduces the variable corresponding to that row and column into the fitted regression equation, or removes it if it was already in the equation. This property of sweep suggests that it would provide an efficient method for calculating all regressions.

III. DOING ALL POSSIBLE REGRESSIONS

In using the sweeping method to do all possible regressions, computational efficiency can be achieved in two ways. First, it is important to obtain the 2^k regressions with a minimum number of sweeps. Having achieved this, we would like to carry out each sweep as efficiently as possible. In theorem 1 below, we give a constructive proof that the 2^k possible regressions can be carried out in a sequence of just 2^k sweeps, which is of course the minimum number of required sweeps. We then go on to show how to effect a significant

reduction in the number of calculations needed for each sweep in the sequence. These economies are brought about by sweeping the smallest possible submatrix at each stage and using the inherent symmetry properties of the matrix in such a way as to operate only on the upper triangular part of the submatrix.

THEOREM 1. All possible 2^k regressions of a dependent variable on a set of k independent variables can be obtained through a sequence of exactly 2^k sweeps of the (k+2) x (k+2) cross product matrix $\underset{\sim}{C}$.

Proof: (By mathematical induction)

First, sweep $\underset{\sim}{C}$ on the zeroth pivotal element to produce the regression $\hat{Y} = \hat{B}_o$. Denote the resulting matrix by $\underset{\sim}{W}$. We proceed to show that the additional $2^k - 1$ regressions can be obtained with exactly $2^k - 1$ sweeps of W.

When k=1, $2^k - 1 = 1$ and the single regression $\hat{Y} = \hat{B}_o + \hat{B}_1 X_1$ is performed by sweeping on the first pivotal element of $\underset{\sim}{W}$.

Assume that the $2^k - 1$ possible regressions on k-1 variables can be obtained in a sequence S_{k-1} of exactly $2^{k-1} - 1$ sweeps on the first k pivotal elements. Now, sweeping on the (k, k) pivotal element will produce a new regression which adds the k^{th} variable to the regression produced by the sequence S_{k-1}. Since the sequence S_{k-1} produced 2^{k-1} distinct regressions not including the k^{th} variable, repetition of the sequence S_{k-1} will now produce another 2^{k-1} distinct regressions including the k^{th} variable. The total number of sweeps in $S_k = (S_{k-1}, k, S_{k-1})$ is thus seen to be $(2^{k-1} - 1) + 1 + (2^{k-1} - 1) = 2^k - 1$. Together with the sweep which produced W and the regression $Y = B_o$, we see that all possible 2^k distinct

regressions will have been produced in exactly 2^k sweeps.

The proof suggests a recursive algorithm for constructing the sequence S_k, that is $S_k = (S_{k-1}, k, S_{k-1})$, as illustrated in Table 1.

TABLE I.

k	S_k
1	1
2	1 2 1
3	1 2 1 3 1 2 1
4	$\underbrace{1\ 2\ 1\ 3\ 1\ 2\ 1}_{S_3}$ 4 $\underbrace{1\ 2\ 1\ 3\ 1\ 2\ 1}_{S_3}$
5	
k	(S_{k-1}, k, S_{k-1})

However, it is not actually necessary to construct the sequence recursively, since the recursion formula for S_k reveals that the i^{th} pivotal element $(i = 1, 2, \ldots, k)$ is swept for the first time on the 2^{i-1}st. sweep and is swept on every 2^i th sweep thereafter. Table 2 provides a simple Fortran program for generating S_k directly by the above considerations.

Fortran Routine for Generating Sequence of All Regressions

TABLE 2

k = no. of variables

2^k-1 = no. of possible regressions (in addition to the regression $\hat{Y} = \hat{B}_o$)

input k

output I = vector of length $2^k - 1$ containing sequence of sweeps

```
      SUBROUTINE REGR (K, I)
      DIMENSION I (2048)
      DO 1 J = 1, K
      JJ = 2 ** (J-1)
      INCR = 2 * JJ
      I (JJ) = J
      JJJ = 2 **(K-J) -1
      IF (JJJ) 1, 1, 3
3     DO 2 M = 1, JJJ
      JJ=JJ + INCR
2     I(JJ) = J
1     CONTINUE
      RETURN
      END
```

Having obtained an optimum sequence of sweeps, we now turn to the problem of carrying out each sweep as efficiently as possible. We note from (2.1) that the symmetric property of the cross product matrix is partially destroyed by the sweep operation. That is, when a symmetric matrix A is swept on the r^{th} pivotal element, the resulting matrix B has the property $b_{ir} = - b_{ri}$, and $b_{ij} = b_{ji}$ for $i, j \neq r$. This particular property of sweep gives rise to the definition of an absolute symmetric matrix $|b_{ij}| = |b_{ji}|$ for all i and j.

In general, the relation between b_{ij} and b_{ji} may be found readily if we know how many times A has been swept on the i^{th} and j^{th} pivotal elements. Define a parity vector $T = (t_o, t_1, \ldots, t_k)$ where $t_i = 1$ if the matrix A has been swept an even number of times on the i^{th} pivotal element and $t_i = - 1$ otherwise. The property of absolute symmetry, combined with the parity vector T permits us to sweep only the upper triangular part of the matrix at each step, where the sweep operation[2, 1] is redefined by (3.1) below.

$$
\begin{aligned}
b_{rr} &= 1/a_{rr} \\
b_{ir} &= -a_{ir}/a_{rr} \qquad i < r \\
b_{rj} &= a_{rj}/a_{rr} \qquad j > r \\
b_{ij} &= a_{ij} - a_{ir}\, a_{rj}/a_{rr} \qquad j > i
\end{aligned}
$$

where $a_{uv} = t_u t_v a_{vu}$ if $u > v$.

It should be noted that $t_r = 1$ for all r initially (i.e. - for symmetric A). Each time the matrix is swept on the r^{th} pivotal element, the sign of t_r must be reversed. At any step in a sequence of sweeps, the

variable X_r is included in the regression equation if $t_r = -1$.

It should be noted that theorem 1 gives a sequence of sweeps to be applied to the entire $(k+2) \times (k+2)$ cross product matrix $\underset{\sim}{C}$. However, a further substantial savings in computing time can be achieved by applying the sweep operator only to that submatrix containing the relevant elements at the particular stage of the regression.

In the case of sweeping the matrix on a pivotal element which results in adding a variable, say X_i, to the regression equation, the submatrix to be swept must include all variables which are already in the equation, any variable which will be entered into the equation before X_i is removed, and the dependent variable. When the variable X_i is next deleted, the same submatrix must be swept. The procedure is summarized by the following simple rule, which gives the sequence of minimum sized submatrices associated with the successive sweeps.

Rule 1:

> When sweeping the i^{th} variable, the submatrix to be swept consists of the rows and columns with indices $0, 1, 2, \ldots,$ $i+1$, all $j > i+1$ such that $t_j = -1$, and $k+1$ (the index of the dependent variable).

To summarize then, the recommended procedure for calculating all 2^k regressions in a minimum number of computations is as follows:

1. Carry out a sequence of sweeps as given by theorem 1.
2. For each such sweep:
   a. Apply Rule 1 to obtain the minimum sized submatrix to be swept.

b. Use (3.1) to apply the sweep operator only to the upper triangular part of the submatrix defined in 2a above.

In the next section, we will apply the above procedure to an example in which $k = 4$.

IV. A FOUR VARIABLE EXAMPLE

The basis for the fundamental result of the previous section (i.e., Theorem 1) can be illustrated by figure 1, which gives a geometric representation of a four variable example in terms of travelling along edges of cubes, where the i^{th} coordinate of a vertex indicates the presence (1) or absence (0) of the i^{th} variable in the regression equation. Starting at (0, 0, 0, 0) the arrows indicate the successive regressions, and the number alongside each arrow indicates the pivotal element to be swept in order to move from the regression represented by the starting vertex to that of the ending vertex. Thus, for example, the vertex (1, 1, 0, 0) represents the regression $\hat{Y} = \hat{B}_0 + \hat{B}_1 X_1 + \hat{B}_2 X_2$, and sweeping variable X_1 will remove it from the regression, taking us to the vertex (0, 1, 0, 0), which represents the regression $\hat{Y} = \hat{B}^*_0 + \hat{B}^*_2 X_2$. Table 3 summarizes the sequence of sweeps, the resulting regressions, and the applicable submatrices corresponding to the sweep sequence depicted in figure 1.

Geometric Representation of 2^4 Regressions in 2^4 Sweeps

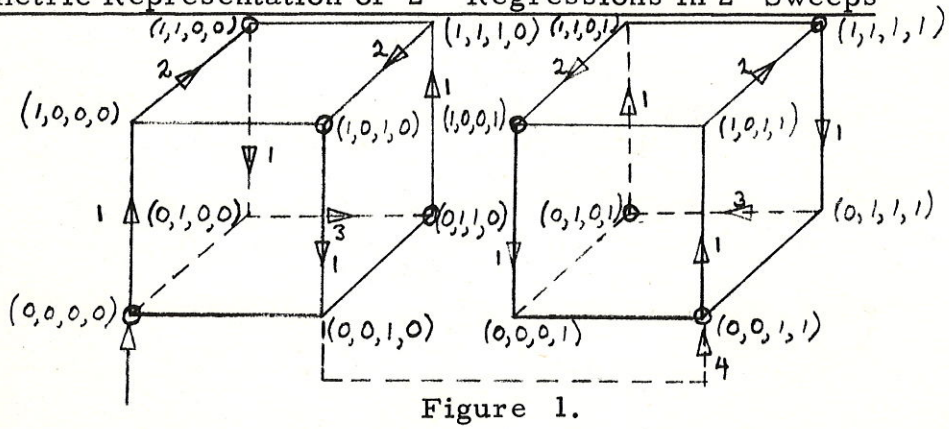

Figure 1.

KEY: 1. The i^{th} coordinate of each vertex indicates the presence (1) or absence (0) of the i^{th} variable in the regression.

2. Starting at (0, 0, 0, 0), arrows indicate successive regressions, and the number alongside each arrow indicates the pivotal element to be swept.

SEQUENCE OF SWEEPS AND CORRESPONDING REGRESSIONS

Order of Sweep	Sequence of Sweeps (S_k)	Vertex	Independent variables included in Regression	Submatrix to be swept
1	0	(0, 0, 0, 0)	none	0, 1, 2, 3, 4, 5
2	1	(1, 0, 0, 0)	1	0, 1, 2, 5
3	2	(1, 1, 0, 0)	1, 2	0, 1, 2, 3, 5
4	1	(0, 1, 0, 0)	2	0, 1, 2, 5
5	3	(0, 1, 1, 0)	2, 3	0, 1, 2, 3, 4, 5
6	1	(1, 1, 1, 0)	1, 2, 3	0, 1, 2, 3, 5
7	2	(1, 0, 1, 0)	1, 3	0, 1, 2, 3, 5
8	1	(0, 0, 1, 0)	3	0, 1, 2, 3, 5
9	4	(0, 0, 1, 1)	3, 4	0, 1, 2, 3, 4, 5
10	1	(1, 0, 1, 1)	1, 3, 4	0, 1, 2, 3, 4, 5
11	2	(1, 1, 1, 1)	1, 2, 3, 4	0, 1, 2, 3, 4, 5
12	1	(0, 1, 1, 1)	2, 3, 4	0, 1, 2, 3, 4, 5
13	3	(0, 1, 0, 1)	2, 4	0, 1, 2, 3, 4, 5
14	1	(1, 1, 0, 1)	1, 2, 4	0, 1, 2, 4, 5
15	2	(1, 0, 0, 1)	1, 4	0, 1, 2, 3, 4, 5
16	1	(0, 0, 0, 1)	4	0, 1, 2, 4, 5

Table 3.

V. SOME COMMENTS ON FRACTIONAL REPLICATION

If the number of independent variables under consideration is sufficiently large, it may be impractical to calculate all possible regressions. However, in such instances, it may be possible to identify the important variables by restricting the search to a subset of the possible regressions.

Gorman and Toman (1966) have proposed a procedure involving calculation of only a balanced fraction of the possible regressions, and use of the C_p statistic to select one of these. We shall, in this section, investigate the use of this procedure in terms of the sweep operation.

With reference to figure 1, if $X_1X_2X_3X_4$ is chosen as the defining contrast for a 1/2 replicate, we obtain the regressions corresponding to the circled vertices as listed in Table 4.

Order of Sweeps	Sequence of Sweeps	Vertex	Treatment Combination (independent Variables Included in Regression)
1	0	(0, 0, 0, 0)	none
2, 3	1, 2	(1, 1, 0, 0)	1, 2
4, 5	1, 3	(0, 1, 1, 0)	2, 3
6, 7	1, 2	(1, 0, 1, 0)	1, 3
8, 9	1, 4	(0, 0, 1, 1)	3, 4
10, 11	1, 2	(1, 1, 1, 1)	1, 2, 3, 4
12, 13	1, 3	(0, 1, 0, 1)	2, 4
14, 15	1, 2	(1, 0, 0, 1)	1, 4

Table 4.

It is immediately apparent from the above listing of treatment combinations that at least two sweeps are required to move from any vertex in the design to any other vertex in the design. Thus, the computation required to find the regressions corresponding to a half replicate of all the

possible regressions will actually produce all possible regressions. (In this example, one additional sweep is required in order to reach (0, 0, 0, 1) but if the other half-replicate had been selected, this sweep would have been required anyhow.) It is readily seen that regardless of the number of variables, any balanced half replicate requires us, at least, to move along two edges of a cube to get from one point to another in the design. Similarly, in the one quarter replicate case, one must again travel along at least three edges of a cube before reaching another admissible vertex. Thus, in general, at least 2^{k-p+1} sweeps will be required to carry out a balanced 2^{k-p} fraction of the possible 2^k regressions.

The situation may be summarized as follows:

1. In using a sequence of sweep operations to calculate a subset of the possible regressions, it is meaningless to consider a half-replicate since the required computations will automatically yield the entire 2^k regressions.
2. The computations required to carry out a quarter-replicate, automatically yield an additional (unbalanced) quarter-replicate.
3. In carrying out a 2^{k-p} fraction of all the possible regressions, it does not cost anything to look at the additional regressions which are automatically produced in the course of moving from one regression to another.

The authors do not know of any method for generating balanced fractions of all regressions in an optimal fashion. Since as noted earlier, the number of possible regressions increases exponentially with k, the number of independent variables, a solution to the above problem, if it can be obtained, would constitute a valuable contribution.

REFERENCES:

1. Beaton, Albert E., 1964. The Use of Special Matrix Operators in Statistical Calculus. Research Bulletin RB-64-51, Educational Testing Service, Princeton, New Jersey.

2. Gorman, J.W. and Toman, R.J., 1966. Selection of Variables For Fitting Equations to Data. Technometrics, 8, 27-51.

3. Hamaker, H.C., 1962. On Multiple Regression Analysis. Neerlandica, 16, 31-56. (in English).

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, May 7-9.

5. Oosterhoff, J., 1963. On the Selection of Independent Variables in a Regression Equation. Report S 319 (VP 23), The Mathematical Center at Amsterdam.

6. Ralston, A. and Wilf, H.S. (Editors), 1960. Mathematical Methods for Digital Computers. John Wiley and Sons, Inc., New York, 191-203.

7. Wilkinson, J.H., 1965. The Algebraic Eigenvalue Problem. Oxford University Press, London.

IBM