THE STEPWISE REGRESSION ALGORITHM SEEN FROM THE STATISTICIAN'S POINT OF VIEW

by

Harry LÜTJOHANN

Research Memorandum No. 11
January 1968

CONTENTS

1. Introduction

1.1 Purpose. The purpose of this paper is to explain in statistical terms how the algorithm works which is used in computer programmes for Stepwise Regression. Stated more precisely, the purpose is to do this for one of several possible, closely related variants of the algorithm. The paper will contain nothing substantially new but may yet perhaps shed some new light on well-known facts.

1.2 Motivation. There are two reasons why the author has thought it worthwhile writing this paper. One is that Stepwise Regression programme descriptions are usually written by computer people who tend to use flow-chart terminology rather than statistical terminology. An example in case is the standard reference, Efroymson (1960).

The other reason for writing the paper is the author's belief that the computer algorithm may serve as a vehicle for teaching students some important aspects of Multiple Regression. Section 7 below is written with this latter point of view in mind.

1.3 Acknowledgements. The particular variant of the Stepwise Regression algorithm which will be presented and analysed in sections 5 and 6 below, was introduced to the author in London in 1961, by Mrs. P. Harris of the Operational Research Branch of British Petroleum.

An earlier version of the present paper was written in 1964 and circulated internally in the Institute of Statistics of Stockholm University.

The present paper was written in 1967 during a visit to the Institute for Advanced Studies in Vienna. Thanks are due especially to Dr. H. Winter of the latter Institute for encouragement and for a number of suggestions of improvements and clarifications.

2. A Matrix Algebraic Theorem

2.1 Introduction and Reference. The Stepwise Regression algorithm can be considered to be based fundamentally upon one particular theorem concerning the inversion of a certain class of partitioned matrices. This goes for all variants of the algorithm, one of which will be analysed in this paper.

The matrix algebraic theorem in question is found i.a. in Hohn (1958), section 3.9.

2.2 Assumptions and Notation for the Theorem. The theorem will be presented here in no greater generality than that needed for the present purpose. Some of the assumptions are therefore unnecessarily restrictive, judged by a mathematician's standard.

Let M be a symmetric positive-definite matrix of order $p + q$. Let the rows of M be partitioned into a subset of $p \geq 1$ rows plus a subset of $q \geq 1$ rows, and let the columns of M be partitioned in the same way. The four sub-matrices are double-indexed as follows.

$$M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}$$

The inverse matrix of M is partitioned in the same way as M, and its four sub-matrices are double-indexed by top indices as follows

$$M^{-1} = \begin{bmatrix} M^{11} & M^{12} \\ M^{21} & M^{22} \end{bmatrix}$$

For example, the order of M_{12} and of M^{12} is $p \times q$, and $M_{21} = M'_{12}$.

2.3 The Theorem. The four sub-matrices of M^{-1} can be expressed in terms of the four sub-matrices of M as follows.

$$M^{11} = M_{11}^{-1} + M_{11}^{-1} M_{12} (M_{22} - M_{21} M_{11}^{-1} M_{12})^{-1} M_{21} M_{11}^{-1}$$

$$M^{12} = - M_{11}^{-1} M_{12} (M_{22} - M_{21} M_{11}^{-1} M_{12})^{-1}$$

$$M^{21} = - (M_{22} - M_{21} M_{11}^{-1} M_{12})^{-1} M_{21} M_{11}^{-1}$$

$$M^{22} = (M_{22} - M_{21} M_{11}^{-1} M_{12})^{-1}$$

2.4 Proof. The positive-definiteness of M is sufficient to guarantee the existence of inverses of M_{11} and $(M_{22} - M_{21} M_{11}^{-1} M_{12})$. A hint of a proof of this statement will be given in the next subsection.

The inverse of a matrix is unique. It therefore suffices as a proof of the theorem to multiply out MM^{-1} in partitioned form and see that the result is the unit matrix of order $p + q$. This is easily done.

2.5 Hint of Auxiliary Proof. The positive-definiteness of M means, by definition, that for any $(p + q)$-vector $v \neq 0$, $v' Mv > 0$. Hence in particular for any $(p + q)$-vector $v \neq 0$ whose last q elements are all zero, $v' Mv = w' M_{11} w > 0$, where w is the p-element subvector of v coming before the q zeroes. Thus M_{11} is positive-definite.

Define an auxiliary square matrix H as follows.

$$H = \begin{bmatrix} I_p & O_{pq} \\ -M_{21} M_{11}^{-1} & I_q \end{bmatrix}$$

The determinant of this triangular matrix is unity. Further one easily finds that

$$HMH' = \begin{bmatrix} M_{11} & O_{pq} \\ O_{qp} & M_{22} - M_{21} M_{11}^{-1} M_{12} \end{bmatrix}$$

This latter matrix is block-diagonal. It follows from that and from $\det(H) = 1$ that

$$\det(M) = \det(HMH') = \det(M_{11}) \times \det(M_{22} - M_{21}M_{11}^{-1}M_{12})$$

But by assumption, $\det(M) > 0$, and the assumption has been shown to imply also $\det(M_{11}) > 0$. Thus it also implies that $\det(M_{22} - M_{21}M_{11}^{-1}M_{12}) > 0$.

The positive-definiteness of M implies the positive-definiteness both of M_{11} and of $M_{22} - M_{21}M_{11}^{-1}M_{12}$.

3. A Set of Multiple Regressions

3.1 Introduction and Assumptions. In this section, matrix algebraic formulas for a set of multiple regressions, each with a separate dependent variable but all with a common set of independent variables, will be presented. The reason for regarding such sets of regression will become apparent later, in section 4 below, especially subsection 4.2.

Let there be an $nx(p+q)$ data matrix $Z = [X \vdots Y]$. Each row corresponds to an observation and each column to a variable. The multiple regression will be computed of each of the $q \geq 1$ variables Y with respect to all the $p \geq 1$ variables X. The rank of the sub-matrix X is assumed to be p. All the variables are assumed to be measured from their respective averages over the n observations, i.e. the data matrix is in deviation form.

3.2 Standard Formulas. The matrix of the coefficients in the least squares regressions of each Y variable on all X variables is $(X'X)^{-1}X'Y$ of order $p \times q$.

The matrix of the sums of products and squares of the residuals after the least squares regressions of each Y variable on all X variables is $Y'Y - Y'X(X'X)^{-1}X'Y$ of order $q \times q$.

Assume the standard multiple regression model with uncorrelated, homoscedastic error terms and with X the set of independent variables, for one of the Y variables, say y. Then the corresponding column of the coefficient matrix, say $(X'X)^{-1}X'y$ contains best linear estimates of the parameter vector of the model. The covariance matrix of the estimators of which these estimates are an "outcome" is proportional to $(X'X)^{-1}$ of order $p \times p$.

3.3 A Remark and a Reference. If one column of the X matrix is reserved for a dummy variable which is identically equal to unity, the standard formulas are valid for a data matrix Z in "raw form" too.

The formulas given are straight-forward generalizations of the formulas for the case $q = 1$ found in many text-books. The generalized versions are given by Goldberger (1964), section 4.11.

3.4 Summarizing Regression Results. The computation of all the regression coefficients, residual sums of products and squares, and estimator covariances and variances (apart from a scalar factor) mentioned in subsection 3.2 starts from the cross-product sum matrix $Z'Z$.

$$Z'Z = \begin{bmatrix} X'X & X'Y \\ Y'X & Y'Y \end{bmatrix}$$

Each row and column of $Z'Z$ is associated with a particular variable.

All the computation results mentioned can be arranged into a matrix R which is of the same order as $Z'Z$.

$$R = \begin{bmatrix} (X'X)^{-1} & (X'X)^{-1}X'Y \\ Y'X(X'X)^{-1} & Y'Y - Y'X(X'X)^{-1}X'Y \end{bmatrix}$$

Each row and column of R is associated with the same variable as the corresponding row and column of $Z'Z$.

From the point of view of a computer programmer, the essential task involved in computing a set of multiple regressions is to transform the initial information $Z'Z$ into the required information R.

4. Description of Stepwise Regression

4.1 Basic Data and Task. Let there be a raw form data matrix Z^* of n observations on k variables. A necessary preliminary step in a Stepwise Regression computer programme is to read the data and compile the matrix of sums of products and squares of deviations $Z'Z$.

By a steering parameter or in some other way the programme user designates one of the k variables as the dependent one. All the other variables are candidates for the role of an independent variable in a multiple regression with the designated variable as the dependent variable. It is the task of the computer to find a "suitable" set of independent variables. The selection of that set from among the list of candidates is made according to programmed-in rules to be indicated briefly later.

4.2 Basic Computations. In the process of searching for "good" independent variables, the computer passes through a sequence of intermediate provisory selections of a set of independent variables. For any such provisory selection, the computer does not just compute the multiple regression of the designated dependent variable on the provisorily selected independent variables. Rather, it regards the total set of variables Z as partitioned into an X set and a Y set as in subsection 3.1 and computes all the result information for the corresponding set of multiple regressions needed to establish the matrix R of subsection 3.4 above. The X set is the set of provisorily selected independent variables, and the Y set is the rest of Z, i.e. the designated dependent variable and all the "candidate variables" not at the particular stage included in the provisory set of independent variables.

The computation of a number of additional regressions, meaningless in themselves, in addition to the immediately interesting one, is motivated by the way the Stepwise Regression functions, as will be seen in sections 5 and 6 below.

4.3 The Stepwise Process. On the basis of "R" information computed at one step, the computer decides according to its programmed-in rules what the next step will be, that is to say what partitioning of the total set of variables Z into an X set and a Y set to try next.

The change from the partitioning used at one step to that used at the next one always consists in the shifting over of just one variable from the Y set to the X set or in the opposite direction. For the multiple regression with the designated dependent variable this means either adding a new independent variable to those already used in the regression, or removing from the regression one of the independent variables.

4.4 Steering the Stepwise Process. Several different sets of rules may be used for steering the process of selecting successive partitionings of the total set of variables. One simple steering mechanism is to start with no independent variables at all, to add into the set of independent variables at any step that candidate variable whose inclusion yields the greatest reduction of the residual sum of squares, and never to remove any independent variable from the regression.

Another simple steering mechanism is to start with all the variables, except the designated dependent variable, in the X set, to throw out of the X set at any step that variable whose removal causes the smallest increase of the residual sum of squares, and never to let a variable re-enter the X set.

Either of the two simple mechanisms may be modified by a decision, for example, that a certain subset of variables must always be in the set of independent variables.

4.5 Refinements by Statistical Testing. The rules steering the stepwise selection of sets of independent variables may be supplemented by additional rules based on the computation of formal F-tests for the "marginal contribution" or "significance" of the independent variables in the multiple regression with the designated dependent variable. For example, the first set of rules of the preceding subsection may be supplemented by the rule never to enter a variable into the regression when it would not be significant. A further supplementary rule might be to remove at once from the regression any variable which is (no longer) significant in it. The levels of significance applied may be left for the programme user to decide upon, via steering parameters.

It should be noted that Stepwise Regression programmes making a systematic use of such F-tests do not thereby yield results firmly based on standard statistical theory. The several F-tests made during the running of such a programme are based on a number of different Multiple Regression models, which are not logically compatible. For example at some stage X_1, X_3 and X_6 are assumed to be the only variables that are allowed to occur in the true regression model, and the hypothesis that the parameter for X_6 is zero is tested. At a later stage, similarly, the only allowed independent variables are X_1, X_3, X_4, X_6 and X_7 and the hypothesis that the parameter for X_4 is zero is tested. The "maintained hypothesis" on the five acceptable variables at the later stage clearly contradicts that on only three acceptable variables at the earlier stage. Of course this statement does not imply that built-in F-tests are useless in practice, only that their theoretical meaningfulness is not self-evident and may even be doubted.

4.6 References. For a more detailed description of some specified Stepwise Regression steering mechanisms, with an attempt at an evaluation, see Draper and Smith (1966), chapter 6. A mathematical approach to the problem of evaluating different steering mechanisms is found in Wiezorke (1967).

5. A Stepwise Regression Algorithm

5.1 The Purpose of the Algorithm. A stepwise regression computer programme deals with a given set of variables (subsection 4.1). It passes through several stages, where at each stage the variables are split into one independent set and one remaining, dependent set, and where the set of multiple regressions of each of the latter with respect to all of the former is computed (subsection 4.2). The results of the computation at any stage may be summarized in a matrix "R", where each row and column is associated with one of the variables (subsection 3.4).

The purpose of the Stepwise Regression algorithm is to derive the "result matrix" for any stage from that of the preceding stage in a computationally convenient manner.

5.2 A Technical Modification. Simplicity is a highly desirable property of the algorithm leading from one "result-matrix" to the next one. In fact worthwhile simplification of the algorithm may be obtained by the simple device of exchanging the result registration matrix R of subsection 3.4 above for some related, somewhat different matrix. More than one such technical modification has been proposed, each leading to a slightly different algorithm. The particular modification, and the algorithm to go with it, which will be presented in this paper, were introduced to the author in 1961; see subsection 1.3 above. They are found i.a. also in Wiezorke (1967).

The modified result matrix will be denoted $S(X;Y)$ where X is the set of independent variables and Y the set of dependent variables considered. The formula is following; note the minus sign.

$$S(X;Y) = \begin{bmatrix} -(X'X)^{-1} & (X'X)^{-1}X'Y \\ Y'X(X'X)^{-1} & Y'Y - Y'X(X'X)^{-1}X'Y \end{bmatrix}$$

Matrices $S(X;Y)$ can of course also be defined for partitionings of the total set of variables where neither X nor Y is a set of variables with consecutive indices. For simplicity, however, S matrices will always be written here as if the variables had been re-ordered if necessary to let the S matrix take the simple form above.

5.3 A Generalization. The Stepwise Regression algorithm used in computer programmes brings variables into or out of the independent (and thus, conversely, also the dependent set) one by one. For the sake of theoretical interest, the algorithm that will be presented in this paper will be a generalized version, which brings variables into or out of the independent set by sets, where each set may contain one, two or more variables. Accordingly, the algorithm will be formulated in matrix algebra.

Let there be an $n \times k$ data matrix Z in deviation form. Let the variables be partitioned in some way into an X set and a Y set. (Only such partitionings are considered for which $X'X$ is non-singular, and in practice only partitionings where Y contains at least one variable, the designated dependent variable.) Let some subset Z_p of variables be moved over from the Y set to the X set or vice versa. The generalized algorithm serves to re-compute the S matrix to fit the new partitioning.

5.4 Assumptions, Notation and Terminology. The total set of variables Z is assumed to be partitioned into $r \geq 2$ sets Z_i, each with at least one member. The variables are always moved between X and Y in whole Z_i sets. For any partitioning with a non-singular X set, the corresponding S matrix exists. One example may be

$$S\,(Z_1, Z_r;\ Z_2, Z_3, \ldots, Z_{r-1})$$

where the independent set is the union of the sets Z_1 and Z_r. Denoting the empty set $\emptyset$ one may extend the S notation to the "extreme cases" $S(\emptyset; Z) = Z'Z$ and $S(Z;\emptyset) = -(Z'Z)^{-1}$, provided the latter exists.

Consider an arbitrary S matrix. For ease in writing, an abbreviated notation is introduced. The S matrix chosen is denoted M, with r^2 sub-matrices M_{ij} $(i, j = 1, 2, \ldots, r)$ corresponding to the sets of variables Z_i. Let one set of variables Z_p be moved over from the Y set to the X set or in the opposite direction. Denote the S matrix for the new partitioning M^* with sub-matrices M^*_{ij}. The algorithm defines each new sub-matrix M^*_{ij} as a function of some of the old sub-matrices M_{ij}.

The diagonal sub-matrix M_{pp} corresponding to the set Z_p of variables which changes sides, plays a prominent role in the algorithm, and is called the pivot sub-matrix. The algorithm is a little different according to the way Z_p changes sides. If Z_p moves from the Y set to the X set, application of the algorithm is called pivoting on M_{pp}, or for short pivoting on the set Z_p. If Z_p moves from the X set to the Y set, application of the somewhat different algorithm variant for such a case may be called antipivoting on M_{pp} or on Z_p.

Pivoting on a set of variables brings them into the regression with the designated dependent variable. Antipivoting on a set of variables throws them out of that regression.

5.5 The Generalized Algorithm. There are three phases.

Phase 1. For all M_{ij} with $i \neq p$ and $j \neq p$

$$M^*_{ij} = M_{ij} - M_{ip} M^{-1}_{pp} M_{pj}$$

Phase 2. For all M_{ip} with $i \neq p$

when pivoting $\quad M^*_{ip} = M_{ip} M^{-1}_{pp}$

when antipivoting $\quad M^*_{ip} = - M_{ip} M^{-1}_{pp}$.

For all M_{pj} with $j \neq p$

when pivoting $\quad M^*_{pj} = M^{-1}_{pp} M_{pj}$

when antipivoting $\quad M^*_{pj} = - M^{-1}_{pp} M_{pj}$.

Phase 3. $\quad M^*_{pp} = - M^{-1}_{pp}$.

5.6 The Computer Algorithm. When all the sets of variables Z_i have just one member each, all the operations of the algorithm are operations in scalar algebra, and M_{pp} is a pivot element.

If the operations are taken in the order the three phases are numbered above, each M^*_{ij} may be placed in the memory call of the computer previously occupied by the corresponding M_{ij} without any information needed for completion of the computation of M^* being thereby lost. It is therefore sufficient to reserve memory space for one S matrix, which is an attractive feature of the algorithm. Further, due to the symmetry of the S matrices, only the upper (or lower) half of the matrix need be stored.

5.7 Comment. So far, the generealized Stepwise Regression algorithm has simply been presented. It remains to demonstrate that it does in fact function as it ought to, i.e. that it always produces the relevant S matrix. This is the subject of the next section.

6. Four Propositions Concerning the Algorithm.

6.1 Assumptions and Notation for the Propositions. Let A be a symmetric positive-definite matrix of order k, partitioned symmetrically into $3 \times 3 = 9$ submatrices A_{ij} ; $i,j = 1, 2, 3$. The matrix A will be pivoted on A_{11} and the resulting matrix will be called B with sub-matrices B_{ij} of the same orders as A_{ij}. The matrix B will be pivoted on B_{22} and the resulting matrix will be called C. Finally, minus the inverse matrix of A will be denoted D.

The propositions will be formulated in language appropriate for the situation where A is a cross-product sum matrix $Z'Z$ with the set of variables partitioned into three sets Z_i ; $i = 1, 2, 3$. The concept of an "S" matrix introduced in the preceding section will be appealed to repeatedly. $A = S(\emptyset;\ Z_1,\ Z_2,\ Z_3)$.

6.2 First Proposition. Pivoting A on A_{11} leads to a matrix B which is an S matrix with Z_1 as the set of independent variables.

Direct application of the pivoting algorithm to A below obviously yields the result B below.

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix}$$

$$B = \begin{bmatrix} -A_{11}^{-1} & A_{11}^{-1}A_{12} & A_{11}^{-1}A_{13} \\ A_{21}A_{11}^{-1} & A_{22}-A_{21}A_{11}^{-1}A_{12} & A_{23}-A_{21}A_{11}^{-1}A_{13} \\ A_{31}A_{11}^{-1} & A_{32}-A_{31}A_{11}^{-1}A_{12} & A_{33}-A_{31}A_{11}^{-1}A_{13} \end{bmatrix}$$

Evidently B is the S matrix referred to. $B = S(Z_1; Z_2, Z_3)$.

6.3 Second Proposition. Pivoting B on B_{22} leads to a matrix C which is an S matrix with $\left[Z_1 \vdots Z_2 \right]$ as the set of independent variables. $C = S(Z_1, Z_2; Z_3)$.

(i) Application of the pivoting algorithm to the four top left sub-matrices of B yields the following sub-matrices of C.

$$C_{11} = - A_{11}^{-1} - A_{11}^{-1}A_{12}(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} A_{21}A_{11}^{-1}$$

$$C_{12} = A_{11}^{-1}A_{12}(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}$$

$$C_{21} = (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} A_{21}A_{11}^{-1}$$

$$C_{22} = - (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}$$

Comparison with the matrix algebraic theorem of section 2 immediately verifies the following equation.

$$\begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} = - \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}^{-1}$$

(ii) On further application of the pivoting algorithm, the two top right sub-matrices of C are found to turn out as follows.

$$C_{13} = A_{11}^{-1}A_{13} - A_{11}^{-1}A_{12}(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}(A_{23} - A_{21}A_{11}^{-1}A_{13}) =$$

$$= K_{11}A_{13} + K_{12}A_{23}$$

$$C_{23} = (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}(A_{23} - A_{21}A_{11}^{-1}A_{13}) =$$

$$= K_{21}A_{13} + K_{22}A_{23}$$

Collecting terms, one finds that $K_{ij} = -C_{ij}$ for all four (i, j). Thus, by reference to the result under (i),

$$\begin{bmatrix} C_{13} \\ C_{23} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}^{-1} \begin{bmatrix} A_{13} \\ A_{23} \end{bmatrix}$$

(iii) The two bottom left sub-matrices of C are the transposes of the symmetrically placed top right sub-matrices. This is seen from the symmetry of B and that of the algorithm.

(iv) The bottom right sub-matrix of C, finally, is

$$C_{33} = (A_{33} - A_{31}A_{11}^{-1}A_{13}) -$$

$$- (A_{32}-A_{31}A_{11}^{-1}A_{12})(A_{22}-A_{21}A_{11}^{-1}A_{12})^{-1}(A_{23}-A_{21}A_{11}^{-1}A_{13}) =$$

$$= A_{33} - A_{31}L_{11}A_{13} - A_{31}L_{12}A_{23} - A_{32}L_{21}A_{13} - A_{32}L_{22}A_{23}$$

Collecting terms, one finds that $L_{ij} = K_{ij} = -C_{ij}$ for all four (i, j). Thus,

$$C_{33} = A_{33} - \begin{bmatrix} A_{31} & A_{32} \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}^{-1} \begin{bmatrix} A_{13} \\ A_{12} \end{bmatrix}$$

The arguments (i) - (iv) together prove the second proposition.

6.4 Third Proposition. Antipivoting C on C_{22} brings back B. In terms of the sub-matrices of B, those of C can be written:

$$C = \begin{bmatrix} B_{11}-B_{12}B_{22}^{-1}B_{21} & B_{12}B_{22}^{-1} & B_{13}-B_{12}B_{22}^{-1}B_{23} \\ B_{22}^{-1}B_{21} & -B_{22}^{-1} & B_{22}^{-1}B_{23} \\ B_{31}-B_{32}B_{22}^{-1}B_{21} & B_{32}B_{22}^{-1} & B_{33}-B_{32}B_{22}^{-1}B_{23} \end{bmatrix}$$

This follows from the pivoting algorithm. It is easy to verify that antipivoting on C_{22} replaces each C_{ij} by the corresponding B_{ij}.

6.5 Fourth Proposition. Antipivoting $-A^{-1} = D$ on the sub-matrix D_{33} yields the S-type matrix C. — $D = S(Z_1, Z_2, Z_3; \phi)$.

The partitioning of Z is now collapsed into $Z = Z_4 \quad Z_3$. The new sub-matrices A_{43} and A_{44} are, consequently,

$$A_{43} = \begin{bmatrix} A_{13} \\ A_{23} \end{bmatrix} \qquad A_{44} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

By the matrix algebraic theorem, the four sub-matrices of D can be written as follows. (The sub-matrices of D are numbered like those of A.)

$$D_{44} = -A_{44}^{-1} - A_{44}^{-1}A_{43}(A_{33} - A_{34}A_{44}^{-1}A_{43})^{-1}A_{34}A_{44}^{-1}$$

$$D_{43} = A_{44}^{-1}A_{43}(A_{33} - A_{34}A_{44}^{-1}A_{43})^{-1} \qquad \text{and } D_{34} = D'_{43}$$

$$D = -(A_{33} - A_{34}A_{44}^{-1}A_{43})^{-1}$$

Direct application of the antipivoting algorithm is seen to produce the C matrix, partitioned into four parts analogously with A and D.

6.6 Conclusions from the Four Propositions. As the partitionings of the initial $A = Z'Z$ matrix are quite arbitrary, the four propositions imply some quite general conclusions.

The first and second proposition jointly constitute a proof by induction that any S matrix can be arrived at by application of the generalized pivoting algorithm in an arbitrary number of steps. This guarantees the validity of any sequence of Stepwise Regression steps where variables are added into the list of independent variables at every step and are never removed from it.

The first, second and third proposition jointly constitute a proof by induction that any S matrix can be arrived at by application of the generalized pivoting and antipivoting algorithms in an arbitrary number of steps. This guarantees the validity of any sequence of Stepwise Regression steps where variables are put into and thrown out of the regression in any conceivable way.

The fourth proposition is not needed in the context of Stepwise Regression. It is included in order to complete the proof that the generalized algorithms both work also as matrix inversion algorithms.

6.7 On Other Algorithms. For other variants of the Stepwise Regression algorithm, the above conclusions can be established by means of propositions similar to those four presented in this section.

7. Statistical Implications of the Algorithm

7.1 Introduction. The purpose of the present section is to show how certain important properties of Least Squares quantities of various kinds can be derived, once the algorithm is understood. The results are well-known; the idea is just to advocate a way of demonstrating them which may have some advantage to speak for it.

7.2 Arithmetical Means as Regression Coefficients. Let $Z = [X \vdots Y]$ be a data matrix in "raw form" where X is the column of units (cf. subsection 3.3) and Y is the "proper" data matrix, with the observations all in their original forms, i.e. not converted to deviations from their sample averages. Formally pivoting the matrix $Z'Z$ on the sub-matrix $X'X$ produces "regression coefficients" $(X'X)^{-1}X'Y$ which are the arithmetical means of each Y variable. The matrix of "residual sums of products" $Y'Y - Y'X(X'X)^{-1}X'Y$ can be seen to be the cross-product sum matrix of the Y variables in deviation form.

The calculation of the arithmetical mean of a sample of observations on a variable can in fact be regarded as an application of the Least Squares principle, for the arithmetical mean is the constant from which the sum of the squared deviations is the smallest. If the "regression" model $Y_{ij} = \mu_i + \varepsilon_{ij}$ with homoscedastic and uncorrelated random elements is assumed for the i'th Y variable, the variance of the "regression coefficient" $\bar{Y}_i$ is further equal to $\sigma_i^2 (X'X)^{-1}$ where σ_i^2 is the model variance and $X'X = n$.

So far the algorithm has not been invoked. It is however clear that it can be used to transform a "raw" product-sum matrix $Z'Z$ into the corresponding deviations cross-product sum matrix.

7.3 "Raw Form" Multiple Regression. As was said in subsection 3.3, the standard matrix formulas of Multiple Regression, the formulas incoporated in the S matrix in subsection 5.1, remain valid also if $Z = [X \mid Y]$ is a data matrix in "raw form" where one column of X has all its elements equal to unity. The corresponding regression coefficient is the intercept of the fitted regression equation. Hence the new "unit" X column may be called the column of the "intercept variable".

All the results of section 6 concerning the S matrix computed in a stepwise fashion remain valid. If the X set of the variables always includes the intercept variable, the statistical interpretation of all the elements of any S matrix are quite as before. (In fact if the intercept variable is not in the X set, S is still interpretable much as before, but then it refers to regressions with the intercept forced to be zero.)

By the properties proved for the Stepwise Regression algorithm in section 6, the S matrix containing the information concerning any desired multiple regression may be arrived at in two steps as follows. First pivot the "raw" Z'Z on the intercept variable. Then pivot the resulting S matrix on the "other" independent variables. The first step brings about an S matrix consisting of the deviations Z' Z matrix bordered by the mean vector and a diagonal element $-1/n$.

Suppose the matrix formulas of Multiple Regression had been established for "raw form" X and Y as textbooks often do. The Stepwise Regression algorithm then provides an easy proof of the validity of the similar formulas for matrices X and Y of deviations.

Further, the algorithm makes the "fitting" of arithmetical averages to variables appear naturally as simply a special case of regression, since it may always be made to occur as the first step in the stepwise computation of some multiple regression.

7.4 A Note on Computer Programmes. Computer programmes for Stepwise Regression do not normally store the row (and/or column) of the S matrices corresponding to the intercept variable, as they

do not contain useful information after the first step. Such programmes are therefore more naturally described as starting from Z'Z in deviation form, as has been done here. Unfortunately the omission of the intercept row/column also means losing the standard error of the intercept.

7.5 Regression of Residuals Upon Residuals. Consider the set of multiple regressions of each of the variables in a set Z_3 with respect to all of the variables of a set of independent variables X. Let the set of independent variables be partitioned in some arbitrary way into two sets Z_1 and Z_2, and let Z_2 be called the primary independent variables and Z_1 the secondary independent variables. (Please accept the somewhat unhappy relation between the set indices and the set names.)

The results from the set of multiple regressions are summarized in a matrix $S(Z_1, Z_2 ; Z_3)$ which can be identified with the matrix C of section 6, if the A matrix of that section is identified with the total initial Z' Z matrix. The matrix of sums of products and squares of the residuals of the Z_3 variables after regression on the Z_1 and Z_2 variables is thus the sub-matrix C_{33}.

Regard the two sets of multiple regressions whose dependent sets are Z_2 and Z_3 and whose common independent set is Z_1, that is the regressions of each of the primary and of the dependent variables on all of the secondary variables. Denote the matrices of residuals from these two sets of regressions U_2 and U_3 respectively. By the standard formula for a matrix of regression coefficients (subsection 3.2)

$$U_i = Z_i - Z_1 (Z_1' Z_1)^{-1} Z_1' Z_i \qquad i = 2, 3$$

It is easy to show algebraically that

$$U_i' U_j = Z_i' Z_j - Z_i' Z_1 (Z_1' Z_1)^{-1} Z_1' Z_j \qquad i = 2, 3 \qquad j = 2, 3$$

Identifying $Z_1' Z_j$ with A_{ij}, it follows immediately from the first expressions given for C_{23} and C_{33} in subsection 6.3 above that

$$C_{23} = (U_2' U_2)^{-1} U_2' U_3$$

$$C_{33} = U_3' U_3 - U_3' U_2 (U_2' U_2)^{-1} U_2' U_3$$

Thus C_{23} and C_{33} can be interpreted as the matrix of regression coefficients and of residual sums of products, respectively, in the set of regressions of one set of residuals on another set of residuals. The dependent set are the residuals of the dependent variables Z_3 after regression on the secondary set Z_1. The independent set are the residuals of the primary set Z_2 after regression on the secondary set Z_1.

It is important to note that the selection of a secondary set Z_1 from among the original independent variables is quite arbitrary. The results obtained are thus very general and as it were flexible. In any set of multiple regressions, any arbitrarily selected subset of the independent variables may be regarded as the secondary set, and the coefficients for the other (primary) independent variables as well as the residual second-order moments correspondingly re-interpreted. This is an extremely important property of Least Squares regression, and it is proved with almost no effort by scruting of the Stepwise Regression algorithm.

The special case where each of the three subsets Z_i has only one member is given by Johnston (1963), section 2.5. The less restricted special case where only Z_1 and Z_2 are restricted to have only one member each is given by Yule and Kendall (1950), sections 12.11 and 12.12.

7.6 The Chain Rule of Residual Variances. The partial correlation between two variables x and y, partial with respect to some set of variables z, is defined as the simple (total) correlation between the residuals of x and y after regression of each on z. (At least, this is one of the ways a partial correlation may be defined.)

From the preceding subsection it follows that the residual sum of squares of a variable Z, after regression on $Z_2, Z_3, \ldots, Z_k$ may be obtained as the residual sum of squares in the (one-variable) regression of the residual $e_{1.23\ldots(k-1)}$ upon the residual $e_{k.23\ldots(k-1)}$ where Yule's notation for residuals has been used in a slightly modified form. Thus,

$$\Sigma e^2_{1.23\ldots k} = (\Sigma e^2_{1.23\ldots(k-1)})(1 - r^2_{1k.23\ldots(k-1)})$$

using the well-known relation between residual sum of squares and correlation coefficient in simple regression, and the definition of a partial correlation as a simple correlation between residuals.

The same argument can be used to express the sum of squares in the right number as in terms of the sum of squares of a residual of the next lower order, and so on until the result is obtained :

$$\Sigma e^2_{1.23\ldots k} = (\Sigma z_1^2)(1 - r^2_{12})(1 - r^2_{13.2})\ldots(1 - r^2_{1k.23\ldots(k-1)})$$

The numbering of the variables, except Z_1, is of course arbitrary.

This result is given in several textbooks.

7.7 Partial Correlations of Different Orders. By means of the algorithm, it is also possible to derive the standard formula for the relation between a partial correlation of some order and those of the next lower order.

Using the notation of section 6, the procedure is to transform the sub-matrix of B referring to the variable sets Z_2 and Z_3, into a matrix of partial (with respect to Z_1) correlations. This is done by pre- and postmultiplication by a certain diagonal matrix. The partial correlation matrix is pivoted on Z_2 and the resulting C^*_{33} is "standardized" again into a matrix with unit main diagonal.

The result is compared with the "standardized" C_{33} which is the matrix of partial (with respect to Z_1 and Z_2) correlations for the variables Z_3, and can be seen to agree with it.

The following formula is thereby established,

$$r_{ij.12} = \frac{r_{ij.1} - r_{i2.1}\, r_{j2.1}}{\sqrt{1 - r^2_{i2.1}}\ \sqrt{1 - r^2_{j2.1}}}$$

where i and j denote any two members of Z_3, and "1" is a "collective index" for Z_1, and Z_2 has been assumed to have only one member. The formula is found in Yule and Kendall (1950), section 12.15.

8. Literature Cited

Draper, N. and Smith, H.: Applied Regression Analysis.
New York, Wiley, 1966.

Efroymson, M.A.: Multiple Regression Analysis.
Chapter 17 of Ralston and Wilf, see below.

Goldberger, A.S.: Econometric Theory.
New York, Wiley, 1964

Hohn, F.E.: Elementary Matrix Algebra.
New York, Macmillan, 1958.

Johnston, J.: Econometric Methods.
New York, McGraw-Hill, 1963.

Ralston, A. and Wilf, H.S., (eds.): Mathematical Methods for Digital Computers.
New York, Wiley, 1960.

Wiezorke, B.: Auswahlverfahren in der Regressionsanalyse.
Metrika, XII, 1967, 68-79.

Yule, G.U. and Kendall, M.G.: An Introduction to the Theory of Statistics, 14th edition.
London, Griffin, 1950.

9. Appendix: A Numerical Example

Introduction.

The paper is complemented by a numerical example. This was suggested to the author by Dr. A. Stanzel of the Institute for Advanced Studies.

The interested reader is recommended to work through the example step by step with pencil and paper. He may also check any of the results by computing them in some other, more direct way.

The algorithm illustrated by the example is of course the computer version where the cross-product sum matrix is partitioned maximally. The pivot element for the next pivoting has been underlined in each "S" matrix.

Data Matrix.

X_1	X_2	X_3	$X_4 = Y$
7	3	5	10
4	- 4	5	5
4	- 2	7	7
4	- 1	1	6
1	- 2	- 3	0
5	3	- 3	5
4	- 3	4	5
3	- 2	1	2
4	- 1	1	5

A column "X_0" with all elements equal to unity may be placed to the left of the X_1 column.

Cross Product Sum Matrix.

	X_0	X_1	X_2	X_3	X_4
X_0	9	36	- 9	18	45
X_1	36	164	- 16	92	213
X_2	- 9	- 16	57	- 38	- 19
X_3	18	92	- 38	136	142
X_4	45	213	- 19	142	289

This matrix can be denoted $S(\emptyset;\ X_0, X_1, X_2, X_3, X_4)$

$S(X_0;\ X_1, X_2, X_3, X_4)$

This matrix is computed from the preceding one by means of the pivoting algorithm, with $M_{00} = (9)$ as the pivoting element.

Examples of the computational steps.

$$M^*_{30} = M_{30}/M_{00} = 18/9 = 2$$

$$M^*_{32} = M_{32} - M_{30} \cdot M_{02}/M_{00} = -38 - 18 \cdot (-9)/9 = -20$$

	X_0	X_1	X_2	X_3	X_4
X_0	$-\frac{1}{9}$	4	- 1	2	5
X_1	4	20	20	20	33
X_2	- 1	20	48	- 20	26
X_3	2	20	- 20	100	52
X_4	5	33	26	52	64

The sub-matrix for X_1, X_2, X_3 and X_4 is the cross-product sum matrix for the four variables in deviation form. The X_0 row/column contains the means.

Choosing a pivot.

The "best" first explanatory variable for a regression with X_4 as the dependent variable is the one that reduces the unexplained sum of squares M_{44} the most. If X_p is chosen, the new M_{44} is

$$M^*_{44} = M_{44} - M_{4p} \cdot M_{p4} / M_{pp}$$

The "best" pivot element M_{pp} is thus that for which $M_{4p} \cdot M_{p4} / M_{pp}$ is the largest.

$$p = 1 : (33)^2/20 \cong 54$$
$$p = 2 : (26)^2/48 \cong 14$$
$$p = 3 : (52)^2/100 \cong 27$$

The matrix is therefore pivoted on $M_{11} = (20)$.

$S (X_0, X_1; X_2, X_3, X_4)$

	X_0	X_1	X_2	X_3	X_4
X_0	$-\frac{41}{45}$	$\frac{1}{5}$	- 5	- 2	$-\frac{8}{5}$
X_1	$\frac{1}{5}$	$-\frac{1}{20}$	1	1	$\frac{33}{20}$
X_2	- 5	1	28	- 40	- 7
X_3	- 2	1	- 40	80	19
X_4	$-\frac{8}{5}$	$\frac{33}{20}$	- 7	19	$\frac{191}{20}$

For the next pivoting, there are two candidates for the role of a pivoting element. As $(19)^2/80 > 7^2/28$ the next pivoting will be on $M_{33} = (80)$.

$S(X_0, X_1, X_3; X_2, X_4)$

	X_0	X_1	X_2	X_3	X_4
X_0	$-\frac{173}{180}$	$\frac{9}{40}$	-6	$-\frac{1}{40}$	$-\frac{9}{8}$
X_1	$\frac{9}{40}$	$-\frac{1}{16}$	$\frac{3}{2}$	$\frac{1}{80}$	$\frac{113}{80}$
X_2	-6	$\frac{3}{2}$	8	$-\frac{1}{2}$	$\frac{5}{2}$
X_3	$-\frac{1}{40}$	$\frac{1}{80}$	$-\frac{1}{2}$	$-\frac{1}{80}$	$\frac{19}{80}$
X_4	$-\frac{9}{8}$	$\frac{113}{80}$	$\frac{5}{2}$	$\frac{19}{80}$	$\frac{403}{80}$

Checking the results.

The matrix now arrived at should contain the results for the multiple regressions of X_2 and of X_4 upon X_1 and X_3. We check the former regression. In deviation form,

$$(X'X)^{-1} = \begin{bmatrix} 20 & 20 \\ 20 & 100 \end{bmatrix}^{-1} = \frac{1}{80}\begin{bmatrix} 5 & -1 \\ -1 & 1 \end{bmatrix}$$

These four numbers occur with the sign changed in M_{11}, M_{13}, M_{31} and M_{33}.

$$(X'X)^{-1}(X'Y) = \frac{1}{80}\begin{bmatrix} 5 & -1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 20 \\ -20 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 3 \\ -1 \end{bmatrix}$$

These are the elements M_{12} and M_{32}.

The intercept is $-1 - \frac{3}{2} \cdot 4 + \frac{1}{2} \cdot 2 = -6$, element M_{02}.

The residual sum of squares is element M_{22}:

$$Y'Y - \hat{\beta}' X'Y = 48 - \frac{3}{2} \cdot 20 + \frac{1}{2}(-20) = 8.$$

$S(X_0, X_1, X_2, X_3; X_4)$

	X_0	X_1	X_2	X_3	X_4
X_0	$-\frac{983}{180}$	$\frac{27}{20}$	$-\frac{3}{4}$	$-\frac{2}{5}$	$\frac{3}{4}$
X_1	$\frac{27}{20}$	$-\frac{11}{32}$	$\frac{3}{16}$	$\frac{17}{160}$	$\frac{151}{160}$
X_2	$-\frac{3}{4}$	$\frac{3}{16}$	$-\frac{1}{8}$	$-\frac{1}{16}$	$\frac{5}{16}$
X_3	$-\frac{2}{5}$	$\frac{17}{160}$	$-\frac{1}{16}$	$-\frac{7}{160}$	$\frac{63}{160}$
X_4	$\frac{3}{4}$	$\frac{151}{160}$	$\frac{5}{16}$	$\frac{63}{160}$	$\frac{681}{160}$

With a common denominator : $1440\ M =$

	X_0	X_1	X_2	X_3	X_4
X_0	- 7864	1944	- 1080	- 576	1080
X_1	1944	- 495	270	153	1359
X_2	- 1080	270	- 180	- 90	450
X_3	- 576	153	- 90	- 63	567
X_4	1080	1359	450	567	6129

Checking the new results.

The rows and columns for X_0, X_1, X_2 and X_3 should contain minus the inverse of the original cross-product sum matrix for those variables. That it does in fact do so is most easily verified by multiplying the matrices together. For example, in the product matrix MM^{-1}, the element "00" is

$$
\begin{aligned}
9 \cdot (-7864) + &= -70776 + \\
+36 \cdot 1944 + &\quad +69984 + \\
+(-9) \cdot (-1080) + &\quad +9720 - \\
+18 \cdot (-576) = &\quad -10368 = -1440
\end{aligned}
$$

The elements no. 04, 14, 24 and 34 in the new result matrix should be of the type $(X'X)^{-1}X'Y$, which is easily verified. For example, the element "24" can be obtained as minus one 1440th of

$$
\begin{aligned}
-1080 \cdot 45 + &= -48600 + \\
+270 \cdot 213 + &\quad +57510 + \\
+(-180) \cdot (-19) + &\quad +3420 - \\
+(-90) \cdot 142 = &\quad -12780 = -450
\end{aligned}
$$

Of course (partial) checking may also be performed by direct computation of the regression in deviation form.

Summary of the X_4 regressions.

Coefficients with approximate standard errors below them. After each regression the residual sum of sqares.

$$X_4 = \underset{}{-1.6} + \underset{(1.12)}{1.65}\ X_1 + e_{4.1} \qquad (9.55)$$
$$\qquad\quad (0.26)$$

$$X_4 = \underset{}{-1.13} + \underset{(0.90)}{1.41}\ X_1 + \underset{(0.23)}{0.24}\ X_3 + e_{4.13} \qquad (5.04)$$
$$\qquad\quad (0.10)$$

$$X_4 = \underset{}{0.75} + \underset{(2.16)}{0.94}\ X_1 + \underset{(0.54)}{0.31}\ X_2 + \underset{(0.33)}{0.39}\ X_3 + e_{4.123} \qquad (4.25)$$
$$\qquad\quad (0.19)$$

The increase of the squared standard errors from the second to the third regression is "explained" by the near-multicollinearity of the variables X_1, X_2 and X_3. In fact, as can be seen from the proper matrices,

$$R^2_{2.13} = 1 - \frac{8}{48} = \frac{5}{6}$$

An example of antipivoting.

Antipivoting $S(X_0, X_1, X_3 ; X_2, X_4)$ on the element M_{11} leads to the following matrix.

	X_0	X_1	X_2	X_3	X_4
X_0	$-\frac{136}{900}$	$\frac{18}{5}$	$-\frac{3}{5}$	$\frac{1}{50}$	$\frac{198}{50}$
X_1	$\frac{18}{5}$	16	24	$\frac{1}{5}$	$\frac{113}{5}$
X_2	$-\frac{3}{5}$	24	44	$-\frac{1}{5}$	$\frac{182}{5}$
X_3	$\frac{1}{50}$	$\frac{1}{5}$	$-\frac{1}{5}$	$-\frac{1}{100}$	$\frac{13}{25}$
X_4	$\frac{198}{50}$	$\frac{113}{5}$	$\frac{182}{5}$	$\frac{13}{25}$	$\frac{924}{25}$

It can be directly verified that this is the same result as would be obtained upon pivoting $S(X_0; X_1, X_2, X_3, X_4)$ upon M_{33}. It is the matrix $S(X_0, X_3; X_1, X_2, X_4)$.