NOVEMBER 1961 BULLETIN 336

A Systematic Procedure for Determining Potent Independent Variables in Multiple Regression *and* Discriminant Analysis

AGRICULTURAL EXPERIMENT STATION
AUBURN UNIVERSITY
E. V. Smith, *Director*
Auburn, Alabama

CONTENTS

SUMMARY 3
INTRODUCTION 5
REVIEW OF LITERATURE 7
SIMILARITIES AND DIFFERENCES IN REGRESSIONS AND DISCRIMINANTS 8
PROCEDURES FOR SOLVING REGRESSIONS 10
Obtaining the Sums of Squares and Products 10
Coding the Sums of Squares and Products 12
The Solution Proper 13
Deletion of a Variable 17
IDENTIFYING THE MOST POTENT VARIABLES IN REGRESSION 19
Some Considerations in Choosing a Method 19
General Procedure 22
To Find the Most Potent Single Variable 22
To Find the Second Most Potent Variable 23
To Find the Third Most Potent Variable 24
Construction of a Mask to Aid in Computations 24
To Find the Fourth and Other Most Potent Variables. . 25
Testing Significance of the Variables 25
USING REGRESSION PROCEDURES IN DISCRIMINANT ANALYSIS 26
NUMERICAL EXAMPLES 30
Numerical Example of Regression 30
Finding the Most Potent Single Variable 31
Coding 32
Finding Other Potent Variables after the Most Potent . 33
Numerical Example of a Discriminant 39
OTHER TESTS OF SIGNIFICANCE 49
LITERATURE CITED 56
TABLES 1–18 58

First Printing, November 1961

SUMMARY

A method is presented for finding which few of a large number of independent variables are the most potent predictors of some dependent variable Y in the case of a multiple regression or are the most potent discriminators in the case of a discriminant function.

The most potent variable is defined as that independent variable most closely related to the dependent variable; the second most potent variable as that variable which *together with the most potent variable* makes the pair of independent variables most closely related to the dependent variable; the third most potent variable as that independent variable which *together with the most potent pair of variables* makes the trio of variables most closely related to the dependent variable, etc.

The success of the method is based on the intuitively reasonable idea that variables chosen according to the above definition cannot form a much poorer set than the absolutely most potent set; and, on the basis of some practical experience, no sets have as yet been uncovered that have been much better than those thus chosen. The practicability of the method is based primarily on the fact that, under the proposed scheme, it is necessary to examine only n regressions with one independent variable, $n - 1$ regressions with two independent variables, $n - 2$ regressions with three independent variables, and so on up to a maximum of usually not over $n - 5$ or $n - 6$ regressions with five or six independent variables. On the other hand, to obtain the absolutely most potent set would require examination of $n!/n!(n - 1)!$ regressions with one independent variable, $n!/2!(n - 2)!$ regressions with two independent variables, $n!/3!(n - 3)!$ regressions with three independent variables, and so on to $n!/5!(n - 5)!$ or $n!/6!(n - 6)!$. A further advantage is that only those sums of products involving Y and/or the independent variables actually chosen will have to be computed, whereas the absolutely most potent set will require that *all* sums of products existing among the variables be computed. The only other process that could be proposed as leading to the desired set of variables would be to work the regression with all n independent variables, identify and discard the weakest, work the regression with $n - 1$ independent variables, identify and discard the weakest, work the regression with $n - 2$ variables, and so on. This process does not yield the set of absolutely most potent variables either, yet it requires that all sums of squares and products be computed. Furthermore, the computation load will be exceedingly heavy when compared with the proposed method if there are more than about 10 independent variables. Experience indicates that rarely are there as many as five or six potent variables finally selected.

A systematic procedure with computational checks and some devices for reducing duplication of computations is described in detail and illustrated with worked examples. Computations and tests of reliability for many of the summarizing statistics of multiple regression are also described.

A Systematic
PROCEDURE for DETERMINING POTENT INDEPENDENT VARIABLES in MULTIPLE REGRESSION and DISCRIMINANT ANALYSIS*

E. FRED SCHULTZ, JR. Biometrician**
JAMES F. GOGGANS, Associate Forester

Researchers and their statistical advisors are often confronted with the problem of determining the relative potency of a large number of variables in accounting for the behavior of some dependent variable or in discriminating between two discrete groups. Their problem usually is not to assess the potency of all the variables singly, though this may be a beginning, but to find some satisfactorily small number of variables that will explain some satisfactorily large portion of the variability in the dependent variable, or discriminate satisfactorily between the groups. One would like for the chosen set to explain more variability or discriminate more certainly than any other set with this many or fewer variables. Such a set can be described as the absolutely most potent set.

To ensure finding the absolutely most potent set of r variables out of n, would require that all possible multiple regressions or discriminants with r predicting or independent variables be evaluated, with the set accounting for the most variability being chosen. The number of such sets is

$$C_r^n = \frac{n!}{r!\,(n-r)!}\ ,$$

the number of combinations of n things taken r at a time, where n is the total number of variables to be examined and r is the number of variables to be allowed as predictors in the multiple regression or discriminant at any one time. Even the procedure of examining all the C_r^n sets with r variables does not ensure, however, that some set with $r - 1$ or $r - 2$ variables would not do substantially as well. To ensure

* The investigations leading to this report were supported by Hatch and State Funds and carried out cooperatively by the Departments of Forestry and Botany and Plant Pathology. The data were gathered as a part of Alabama Project 509.

** Resigned.

that this situation does not arise would require evaluating all the possible regressions or discriminants with r or fewer variables. Another possibility, that inclusion of one more predictor variable would yield a considerably better prediction equation, can be examined only by evaluating all possible regressions or discriminants with $r + 1$ variables for all possible values of r. This is really the total of all possible regressions or,

$$\sum_{r=1}^{n} C_r^n.$$

If the number of variables of possible predicting value is at all large, say above 10, and especially if the final multiple regression or discriminant is to be allowed to have as many as 4 or more independent variables (if shown to be necessary or desirable by examination of all regressions or discriminants with fewer independent variables), it is apparent that the number of multiple regressions or discriminants to be evaluated would be so large as to economically prohibit such a search in many studies.

However, it is possible and with much less labor to find a set of variables that, though not guaranteed to be the absolutely most potent set, does have some probability of being the absolutely most potent set. In any event, the set may be regarded as most potent in the following sense: The absolutely most potent single predicting or discriminating variable is identified and selected. Following this selection, the most potent pair of variables of which one is the previously chosen most powerful single variable is identified and selected. In the next selection, the most potent trio of variables is identified and chosen, two of which are the previously chosen pair. This procedure continues in the case of regression until a satisfactorily large portion of the variability of the dependent variable is accounted for or until additional variables do not account for a significant amount of the remaining variability in the dependent variable. In the case of a discriminant, the procedure is followed until a satisfactory discriminating equation is obtained or until further variables do not significantly improve the discriminant function.

When it is realized from the start that such a search is to be made, it is possible to further reduce the work by systematization and short cuts. The purpose of this bulletin is: (1) to describe such a systematic search for potent predicting or discriminating variables, (2) to emphasize that a single computing procedure serves for both regression and discriminant, and (3) to bring together in one place directions for all the computations, operations, and tests necessary for such a search. It is intended that these directions shall be in sufficient detail to serve as computing instructions to a group whose members are not highly trained in statistics, but who in the aggregate account for much of the practical use made of statistical procedures. Examples of such persons are researchers with some but limited experience in statistics and mathematics, graduate students in fields other than statistics and mathematics, and clerks who must sometimes function with only very sketchy directions from the researcher whose data they process. The directions are specifically for use with desk calculators, although they could perhaps be adapted to other types of calculators.[1]

REVIEW OF LITERATURE

The literature concerning multiple regressions is voluminous and scattered throughout journals and textbooks in many fields of science. For this reason the authors make no pretense of having made a thorough search of all the literature to determine whether the procedure outlined in this bulletin or a similar procedure has been previously proposed. Since this report is directed primarily toward those users desiring computing instructions rather than development of theory, most references are to textbooks rather than journal articles.

In most textbooks the discussion of multiple regression and discriminant function analysis is limited to finding the regression equation or discriminant function, testing significance, and interpreting results, assuming that there is no uncertainty about the choice of independent variables to be used. Usually there is very little or no discussion of the problem of finding the best possible predicting variables. The reader is left to assume that determining the variables to be put in the regression is not a statistical matter. Some awareness of the larger problem of choosing the best predictor variables is acknowledged, however, as in the discussions on net and standard partial regression coefficients, Ezekiel (*7*), Croxton and Cowden (*4*), Mills (*15*), and Snedecor (*19*); on partitioning the total "determination" or sums of squares due to regression for several variables, Hendricks (*12*), Anderson and Bancroft (*1*), Goulden (*11*), Wert, Neidt, and Ahmann (*23*), Croxton and Cowden (*4*), Mills (*15*), and Snedecor (*19*); and on deleting or omitting variables, Villars (*22*), Anderson and Bancroft (*1*), Rao (*18*), Goulden (*11*), Wert, Neidt, and Ahmann (*23*), Friedman and Foote (*9*) and, Snedecor (*19*).

[1] As this bulletin was being prepared for publication, a computer program (Multiple Regression by Stepwise Procedure) that would make the procedures of this report applicable to use with electronic computers was listed by Leone (*13a*).

The finding of a regression equation or discriminant function requires the solution of simultaneous equations. There are several methods of solution, with many minor variations extant in the literature and textbooks. The abbreviated Doolittle solution is of particular interest. It has been described by Dwyer (*5*, *6*), Peach (*16*), Anderson and Bancroft (*1*), Goulden (*11*), and Friedman and Foote (*9*).[2]

Procedures in discriminant analysis are discussed by Cox and Martin (*3*), Mather (*14*), Fisher (*8*), Rao (*18*), Goulden (*11*), Quenouille (*17*), Tippett (*21*), Wert, Neidt, and Ahmann (*23*), and Bennett and Franklin (*2*).

SIMILARITIES AND DIFFERENCES IN REGRESSIONS AND DISCRIMINANTS

Regression analysis is widely known and used in research. Many research people have operating knowledge of some portion of the technique and associated computational procedures. This does not hold at least to the same degree for discriminant analysis, even though practically identical computational procedures may be used for the two techniques. For this reason it seems desirable to give a brief general description of a discriminant function and its use. This is done by comparing two situations, one suitable for a regression and the other suitable to a discriminant.

An educator might wish to know what items of information about students entering college would be useful in predicting degree of success or achievement during the freshman year, the degree of success in this situation being commonly measured by overall average grade for the year. A typical regression study might call for information on such potential independent or predictor variables as average high school grade, IQ, age, college entrance examination grades, and education of parents in order to investigate their effectiveness as predictors of average grade during the freshman year. Many state universities are required by law to admit all applicants with a diploma from an accredited high school within the state. In such cases 10 to 20 per cent of enrolling freshmen may not finish the year and, thus, would not have an average grade. In such universities this early attrition is a serious problem; thus, other educators might wish to know what items of information about entering college students would be useful in discriminating between those students who will drop out and those who will remain. The researcher might investigate the same set of independent variables

[2] Since the first draft of this Manuscript, the authors have become aware of an exposition by Kramer (*13*) embodying some of the same computational features of this report.

as for the regression study, high school grade, IQ, age, college entrance examination grades, and education of parents in order to determine their effectiveness in discriminating between the two types of students.

The distinction between the two cases lies wholly in the nature of the dependent variable, Y. In the former, or *regression case*, the dependent variable, success, is a *continuous variable taking infinitely many values*. In the latter, or *discriminant case*, the dependent variable is a *discrete variable taking two forms only*; the student either *finishes* or *does not finish* the year.

It is quite possible that the two investigators might each decide on the same set of r predictor variables. It might turn out that each investigator has records on p students and it could even be that some of the students are common to both studies.

Suppose that the number of independent variables is r, or 4, as here listed: high school grade $= X_1$, IQ $= X_2$, college entrance exam grade $= X_3$, and parents' education $= X_4$. The problem of regression is to obtain the constants and coefficients of the regression or prediction equation,

$$\hat{Y} = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$$

such that the *sum of squares of deviations of actual average grade, Y, from the predicted grade, $\hat{Y}$, is minimized,* i.e., $\sum (Y - \hat{Y})^2$ is less than with any other prediction equation that might be suggested or used. The problem of the discriminant function is to find the constants and coefficients of the discriminant or discriminating equation,

$$Z = \lambda_1X_1 + \lambda_2X_2 + \lambda_3X_3 + \lambda_4X_4$$

such that, if a value of Z is computed for every student, the *average difference between the Z values of the two groups, $D = \bar{Z}_{\text{I}} - \bar{Z}_{\text{II}}$,* is maximized. If this is done,

$$G = \frac{(\bar{Z}_{\text{I}} - \bar{Z}_{\text{II}})^2}{\sum (Z - \bar{Z}_{\text{I}})^2 + \sum (Z - \bar{Z}_{\text{II}})^2}$$

is a maximum. Equally as well, the t or F of a test of significance of a difference between the two groups of Z-values is greater than with any other equation that might be suggested or used.

Changing the forms of the equations, squaring, summing over all individuals, partially differentiating, and equating to zero, yields in each case a set of simultaneous equations which must be solved to obtain the needed coefficients, see Goulden (*11*) and Bennett and

Franklin (*2*). The simultaneous equations in regression analysis are given as:

$$\begin{aligned}
(\textstyle\sum x_1^2)b_1 + (\sum x_1x_2)b_2 + (\sum x_1x_3)b_3 + (\sum x_1x_4)b_4 &= \textstyle\sum x_1y \\
(\textstyle\sum x_2x_1)b_1 + (\sum x_2^2)b_2 + (\sum x_2x_3)b_3 + (\sum x_2x_4)b_4 &= \textstyle\sum x_2y \\
(\textstyle\sum x_3x_1)b_1 + (\sum x_3x_2)b_2 + (\sum x_3^2)b_3 + (\sum x_3x_4)b_4 &= \textstyle\sum x_3y \\
(\textstyle\sum x_4x_1)b_1 + (\sum x_4x_2)b_2 + (\sum x_4x_3)b_3 + (\sum x_4^2)b_4 &= \textstyle\sum x_4y.
\end{aligned}$$

Simultaneous equations in discriminant analysis are:

$$\begin{aligned}
(\textstyle\sum x_1^2)\lambda_1 + (\sum x_1x_2)\lambda_2 + (\sum x_1x_3)\lambda_3 + (\sum x_1x_4)\lambda_4 &= d_1 \\
(\textstyle\sum x_2x_1)\lambda_1 + (\sum x_2^2)\lambda_2 + (\sum x_2x_3)\lambda_3 + (\sum x_2x_4)\lambda_4 &= d_2 \\
(\textstyle\sum x_3x_1)\lambda_1 + (\sum x_3x_2)\lambda_2 + (\sum x_3^2)\lambda_3 + (\sum x_3x_4)\lambda_4 &= d_3 \\
(\textstyle\sum x_4x_1)\lambda_1 + (\sum x_4x_2)\lambda_2 + (\sum x_4x_3)\lambda_3 + (\sum x_4^2)\lambda_4 &= d_4.
\end{aligned}$$

One may observe that the two sets of equations are *identical in form.* The reader can be assured that the *left hand sides are identical* except for differences in symbols. The variables, b_1, b_2, b_3, and b_4 of the simultaneous equations for regression are identical with the variables, λ_1, λ_2, λ_3, and λ_4 of the simultaneous equations for discriminants. The only difference is at the right hand side. However, even this difference will not affect the solution procedure once the correct quantities are entered. *This means that the same procedures may be used in discriminant analysis and a search for potent discriminators as in regression analysis and a search for potent predictors—with certain minor modifications.*

PROCEDURES FOR SOLVING REGRESSIONS

Obtaining the Sums of Squares and Products

Since the development of a multiple regression or discriminant function demands the solution of simultaneous equations and since the practicability of the method to be presented for finding potent variables depends in part upon the method used for solving the simultaneous equations, the particular modification of the abbreviated Doolittle solution used at this station is given in Tables 1 and 2 (pages 58–61) for a multiple regression with four independent variables, X_1, X_2, X_3, and X_4, and one dependent variable, Y, called X_5 here for convenience and for relating these instructions to those for multiple correlation as given in many statistical textbooks. The procedure may readily be extended or reduced for the cases of more or fewer variables.

Table 1 indicates the steps in obtaining and coding the sums of squares of deviations and sums of products of deviations that are the coefficients of the b-values in the four simultaneous equations. The variable X_6 (used for checking purposes) is the sum of the values of the four possible predictor variables plus the dependent variable.

When the sums of squares and products are arranged as in Table 1, it turns out that the sums of squares lie along the diagonal and that the sums of products are symmetrically distributed about the diagonal, that is $\sum x_1x_3 = \sum x_3x_1$, $\sum x_2x_5 = \sum x_5x_2$, etc. Thus, it is possible to effect some sizable savings in work by listing only one side of the diagonal. This may lead to some confusion in operations that call for the sum over columns of all the values in a particular row. However, the confusion may be abated somewhat by adding the values *from right to left,* remembering when one reaches the diagonal the reason the remaining values were omitted from the row is that they have already appeared in the column above the diagonal value.

The following definitions hold throughout this report: X is an observation; $\bar{X}$ is a mean value; $x = X - \bar{X}$ is a deviation; and C is a correction factor to subtract from a sum of products or squares of observations, $\sum X_iX_j$, in order to yield the desired sum of products or squares of deviations, $\sum x_ix_j$. From the foregoing definitions Table 1 should be self explanatory to persons with a little experience in statistical analysis except for the column of X_i code, the row of X_j code, and the check values in the last column. To obtain the values in the check column, X_6, one must create for each and every set of values X_1, X_2, X_3, X_4, and X_5 a sixth value, $X_6 = X_1 + X_2 + X_3 + X_4 + X_5$. Thereafter this value is treated as an additional variable. The primary check is afforded in that each $\sum X_iX_6$ as calculated from the data and entered in Table 1 is equal to $\sum_j \sum X_iX_j$, the sum over all the different values of j(columns) of those $\sum X_iX_j$ in the ith row of products of observations. Remember that missing values in the ith row can be found in the jth column, $i = j$ for rows and columns that meet at the diagonal. Thus:

$$\sum X_iX_6 = \sum_j \sum X_iX_j \qquad [1]$$

and as an example, if $i = 3$,

$$\sum X_iX_6 = \sum X_3X_6 = \sum_j \sum X_3X_j = \sum X_3X_5 + \sum X_3X_4 + \sum X_3^2 + \sum X_2X_3 + \sum X_1X_3. \qquad [1a]$$

The next check operation checks all three lines together as follows:

$$C_{i6} + \sum x_ix_6 = \sum_j C_{ij} + \sum_j \sum x_ix_j = \sum X_iX_6. \qquad [2]$$

As an example, if $i = 3$,

$$\begin{aligned} C_{36} &+ \sum x_3x_6 \\ &= \sum_j C_{3j} + \sum_j \sum x_3x_j \\ &= C_{35} + C_{34} + C_{33} + C_{23} + C_{13} + \sum x_3x_5 + \sum x_3x_4 \qquad [2a] \\ &\qquad + \sum x_3^2 + \sum x_2x_3 + \sum x_1x_3 \\ &= \sum X_3X_6. \end{aligned}$$

Coding the Sums of Squares and Products

The sums of squares and products should be coded by powers of 10, which is merely a matter of shifting decimal points with the object of bringing the diagonal terms to values between 0.1 and 10.0 and other terms to values close to this range. This procedure gives the advantage of uniform number of decimal places in the use of the calculating machines without losing significant numbers. Coding by dividing each $\sum x_ix_j$ by $\sqrt{\sum x_i^2}\ \sqrt{\sum x_j^2}$ will also accomplish these same objectives, bringing the diagonal terms to unity and other terms to values lying between one and minus one (simple correlation coefficients or r values) and will also facilitate calculation of partial correlation coefficients and partial regression coefficients. However, coding by powers of 10 is very much quicker and easier; partial coefficients are needed for only a very few of all the multiple regressions solved; and when needed can still be found fairly easily, even after coding by powers of 10. After subtraction of the C_{ij} term from $\sum X_iX_j$ to yield $\sum x_ix_j$, the coding factors m_i and m_j are applied to yield the coded values of $\sum x_ix_j$ lying near the range 0.1 to 10.0. These are designated as a_{ij}. They are the elements of the information matrix that will be used in the abbreviated Doolittle solution proper.

The code values m_i and m_j are determined by the size of the $\sum x_ix_j$ along the diagonal where $\sum x_ix_j = \sum x_i^2$ because $i = j$. The determination is made in the following manner: For each $\sum x_i^2$ choose that *even* power of 10, which when multiplied by $\sum x_i^2$ will yield a value between 0.1 and 10.0. Take the square root of this even power of 10 and designate it as m_i. The coding factor for m_j is the same as for m_i when $i = j$. Enter this value in the appropriate row (i) and in the appropriate column (j) at the margins. The coding values are used as multipliers for the $\sum x_ix_j$ as explained in the stub for each row of a_{ij} values, Table 1. For example, if $\sum x_3^2 = 51273.6$, the *even* power of 10, which when multiplied by 51273.6 yields a value between 0.1 and

10.0 is 10^{-4} or 0.0001. The square root of this number $= 10^{-2}$ or 0.01; thus, $m_3 = 0.01$ in the X_i code and $m_3 = 0.01$ in the X_j code. *This type of coding is equivalent to coding the original X_j-values by multiplying by the appropriate m_j.* In this example, equivalent results could be obtained by multiplying each X_3 by m_3 or 0.01.

The Solution Proper

Table 2 outlines the procedural steps of an abbreviated Doolittle solution. The quantities a_{ij} entered in the first 5 rows and columns of Table 2 are the coded sums of squares and products of deviations as calculated in Table 1. The symbols in all other cells are directions for computing the values belonging in those cells. The h values in the check column are, as indicated, the sums of the a_{ij} values in that row, remembering that values not in a row may be found in the column turning up at the diagonal value.

The "forward" solution results in as many pairs of rows of values, A_{ij} and B_{ij}, as there are variables; thus, in Table 2 there are five pairs of rows. It is one of the labor-saving features of this solution that values A and B can be calculated in pairs. This feature exists because any B_{ij} is equal to the same A_{ij} divided by the leading A or A_{ii} of that row; hence the division may be made before the A_{ij} is cleared from the machine without having to later re-enter the numbers in the machine. There is also somewhat less rounding error introduced by this procedure than in copying the number and later re-entering it in rounded form. Since many of the A_{ij} values are negative, it should be pointed out that this result can be identified by the string of nines appearing in the add result dial of the calculator. The negative value desired is the complement of this number (the number which when added to the result causes all numbers, including the nines, to change to zeros). If this complement is entered on the keyboard and the keyboard locked so that the keys do not clear on depressing the add bar (or multiplying by unity), then a single depression of the add bar (or accumulative multiplication by one) will show all zeros in the result dial, verifying the complement. A second depression (or accumulative multiplication) will show the complement itself in the result dials at which time it may be copied as the desired A_{ij} and given its negative sign. Since it is in the proper dial of the machine for division, it may now be divided by the leading A_{ij} of that row and the result recorded as the desired B_{ij}. As checks on accuracy it may be noted that *every value of A_{ij} on the diagonal must be positive*; consequently every pair of values, A_{ij} and B_{ij}, must have the same sign.

The cells of Table 2 show that all the A_{ij} values, except those for the first row, A_{1j}, must be computed by subtracting one or more products, $A_{ij}B_{ij}$, from some a_{ij}. Both factors of any such product have the same i subscript, because they belong to the same pair of lines. The j subscript of B is the number of the column for which the particular A_{ij} is being computed and the j subscript of A is the number of the row for which the A_{ij} is being computed. The j subscript of A is also the number of the column at the diagonal value of the row from which the original a_{ij} was obtained.
Thus any A_{ij}, say $A_{i'j'}$, is given by

$$A_{ij} = A_{i'j'} = a_{i'j'} - \sum_i A_{ii'}B_{ij'} \qquad [3]$$

where $i < i'$. As an example, if $A_{ij} = A_{i'j'} = A_{34}$

$$\begin{aligned} A_{34} &= a_{34} - \sum_i A_{i3}B_{i4}, \qquad [3a] \\ &= a_{34} - A_{13}B_{14} - A_{23}B_{24} \end{aligned}$$

where $i = 1, 2$. In computing values of A_{ij} on the diagonal, the products, $A_{ij}B_{ij}$, must come from values in the same column since the column for which A_{ij} is being computed and the column at the diagonal of this row are the same column.

Except for rounding discrepancies, each of the products $A_{ij}B_{ij}$ may be calculated as the product of the B_{ij} corresponding to the A_{ij} actually used and the A_{ij} corresponding to the B_{ij} actually used. As examples:

$$A_{23}B_{24} = B_{23}A_{24} \quad \text{and} \quad A_{34}B_{3g} = B_{34}A_{3g}.$$

The pair nearer each other in size, disregarding sign, will yield the result with less rounding error. This leads to a rule: *Examine both ways of calculating each product, $A_{ij}B_{ij}$, and subtract the one that is obtained by the A and B values nearer each other in magnitude disregarding sign.* The operations are carried forward in the machine with nothing being entered on paper except the final results, A_{ij} and B_{ij}. To find A_{3g}, for example, if working to 8 decimal places, $a_{3g} = g_3$ is entered with decimal at 16th place followed by subtracting the products of $A_{13}B_{1g}$ (or $B_{13}A_{1g}$) and $A_{23}B_{2g}$ (or $B_{23}A_{2g}$) each entered with 8 decimal places. This manipulation can be done by use of the cumulative negative multiplication procedure. Whether the values will actually be cumulatively subtracted or cumulatively added will be decided by *considering the signs of all processes and quantities and following the algebraic rules for handling signs.*

As each pair of lines is completed, there is a check on arithmetic accuracy—each

$$A_{ih} = A_{ig} + \sum_j A_{ij} \quad \text{and each} \quad B_{ih} = B_{ig} + \sum_j B_{ij}. \qquad [4]$$

The vacant cells of these rows *are omitted* in summing, not because of symmetry, but because each A_{ij} or B_{ij} cell below the diagonal has the value zero. For example:

$$A_{3h} = A_{3g} + A_{34} + A_{33}. \qquad [4a]$$

As soon as all the pairs of lines A_{ij} and B_{ij} (there is a pair for each variable) have been computed, it is possible to evaluate the success of the multiple regression in explaining variability (or of the discriminant in discriminating). The residual sum of squares in Y not explained by the independent variables X_i is the quantity A_{gg} in the $X_5 = Y$ column. A_{gg} is the remainder from the original coded sum of squares of Y or g_g, after subtracting the sum of squares of deviations in Y accounted for by the regression. This latter quantity, called the regression or reduction sum of squares, is represented by the symbol $\sum \hat{y}^2$ and is given by

$$\sum \hat{y}^2 = \sum_i A_{ig}B_{ig} = A_{1g}B_{1g} + A_{2g}B_{2g} + A_{3g}B_{3g} + A_{4g}B_{4g}, \qquad [5]$$

where $A_{1g}B_{1g}$ is the sum of squares due to X_1; $A_{2g}B_{2g}$ is the sum of squares due to X_2 independent of X_1; $A_{3g}B_{3g}$ is the sum of squares due to X_3 independent of X_1 and X_2; and $A_{4g}B_{4g}$ is the sum of squares due to X_4 independent of X_1, X_2, and X_3.

The coefficient of *multiple determination* R^2, the proportion of the sum of squares of deviations in Y accounted for by the regression, is given by

$$R^2 = \frac{\sum \hat{y}^2}{g_g} = \frac{\text{Decoded} \sum \hat{y}^2}{\sum y^2}. \qquad [6]$$

The coefficient of *multiple regression* R is the square root of this value. If one's interest is in these measures only, then just this much of the solution (the so called "forward" solution) is needed.

It is possible now to perform a "back" solution for the b_i values (regression coefficients) and also a "back" solution for the table of c_{ij} values, which are actually an inverse matrix of the matrix of a_{ij} values. Ordinarily one would not do both.

After the b values have been found (either directly as in column

$X_5 = Y$ Table 2, or by means of the c_{ij}, bottom of Table 2), the regression equation can be written,

$$\hat{Y} = \bar{Y} + \sum_i b_i x_i, \qquad [7]$$

where $\hat{Y}$ is the predicted value of Y, $\bar{Y}$ is the average observed value of Y and x_i is the deviation in X_i, $x_i = X_i - \bar{X}_i$. This equation can be rewritten as

$$\hat{Y} = \bar{Y} - \sum_i b_i \bar{X}_i + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4. \qquad [8]$$

It is also possible now to find the sum of squares due to regression in another way,

$$\sum \hat{y}^2 = \sum_i b_i g_i = b_1 g_1 + b_2 g_2 + b_3 g_3 + b_4 g_4. \qquad [9]$$

Ordinarily the c matrix is calculated if one is interested in tests of significance other than the R test or F test of total reduction due to the several independent variables. In such a case b values would be calculated from c values. If one desires the regression coefficients but does not care to make any statistical tests other than R or F tests of reduction due to regression, the simplest method of calculation is that given in column $X_5 = Y$. A check on these computations is afforded at corresponding positions in the column $X_6 =$ Check $\sum$. The check: $b_i + 1 = h_{ib}$, thus $b_3 + 1 = h_{3b}$. *The back solutions start at the bottom and work up, thus b_4 is calculated before b_3 and c_{44} is calculated before either c_{33} or c_{34}.*

Customary tests of significance may be made without decoding; treat all values as if they were derived from uncoded sums of squares and products, in which case t and F will be the same as if decoded values were used. In calculating $\hat{Y}$, an estimated value of Y for a particular set of X values, and its confidence limits, it is probably best to code the X values to be used by multiplying by the appropriate power of 10, m_j, then use the coded values of b_i and c_{ij}. The final results can be very easily decoded by dividing them by m_y, the Y code. It is possible, of course, to first decode the b_i and c_{ij} values and then use actual X and Y values. To decode a b value:

$$\text{decoded } b_i = b_i^* = b_i \frac{1}{m_y} m_i = b_i \frac{m_i}{m_y}, \text{ thus} \qquad [10]$$

$$\text{decoded } b_3 = b_3^* = b_3 \frac{m_3}{m_y}. \qquad [10a]$$

The superscript asterisk is used to denote a decoded value. To decode a c value:

$$\text{decoded } c_{ij} = c_{ij}^* = c_{ij}m_im_j, \text{ thus} \qquad [11]$$

$$\text{decoded } c_{34} = c_{34}^* = c_{34}m_3m_4. \qquad [11a]$$

To decode a sum of squares in Y as $\sum \hat{y}^2$,

$$\text{decoded } \sum \hat{y}^2 = \sum \hat{y}^{2*} = \sum \hat{y}^2\left(\frac{1}{m_y}\right)^2. \qquad [12]$$

If it is also desired to calculate the various two-factor standard partial correlation and regression coefficients, it is necessary to have the c or inverse matrix of the matrix of correlation coefficients that would have resulted from coding the sums of squares and products of deviations by division of $\sum x_ix_j$ by $\sqrt{\sum x_i^2}\sqrt{\sum x_j^2}$. This matrix may be easily had from the matrix at hand, since any element, c_{ij}^{**}, of the matrix of correlation coefficients is given by

$$c_{ij}^{**} = c_{ij}m_im_j\sqrt{\sum x_i^2}\sqrt{\sum x_j^2}. \qquad [13]$$

A specific example is

$$c_{13}^{**} = c_{13}m_1m_3\sqrt{\sum x_1^2}\sqrt{\sum x_3^2}. \qquad [13a]$$

Some of the quantities that may be calculated and tested, using the c values are summarized in a later section.

After decoding the b values, there is one more check,

$$\sum_j b_j(\sum x_ix_j) = \sum x_iy, \qquad [14]$$

which merely says that the original simultaneous equations should be satisfied if we substitute the solution results. Thus, remembering that values missing from the ith row are in the column of that same number, we find the check of X_2 in Table 1 to be:

$$\sum x_2y = b_4 \sum x_2x_4 + b_3 \sum x_2x_3 + b_2 \sum x_2^2 + b_1 \sum x_1x_2. \qquad [14a]$$

Deletion of a Variable

After a regression has been completed including calculation of the regression coefficients b and the c or inverse matrix, it is possible to determine which independent variable is contributing least to the total regression. This is done by determining for each variable the additional sum of squares that it adds to the regression sum of squares over and above the amount already attributable to the other independent variables. This quantity can be called the sum of squares due to deleting or adding a variable depending on viewpoint. It is a measure of the variability in the dependent variable explained by the variable in question after all the variability that can be explained by the other variables is discounted. If expressed as a proportion of that variation not explained by the other independent variables, it is the coefficient of partial determination. It may be calculated as

$$\sum \hat{y}^2_{i.\text{all others}} = \frac{b_i^2}{c_{ii}}, \qquad [15]$$

which may be read as "the estimated reduction in sum of squares of dependent variable Y due to independent variable X_i when all other X's are held constant." This value may be decoded:

$$\text{decoded} \sum \hat{y}^2_{i.\text{all others}} = \sum \hat{y}^{2*}_{i.\text{all others}} = \sum \hat{y}^2_{i.\text{all others}} \left(\frac{1}{m_y}\right)^2. \qquad [16]$$

The variable contributing least to the regression of course, is that variable with the smallest sum of squares, $\sum \hat{y}^2_{i.\text{all others}}$.

If a variable is deleted, the regression coefficients b change, as do the elements of the c matrix. It is possible to recompute these elements, but it is also possible to estimate them somewhat more rapidly by the following formulas:

$$b_i \text{ after deleting } X_k = b_i - \frac{c_{ik}}{c_{kk}} b_k, \qquad [17]$$

and

$$c_{ij} \text{ after deleting } X_k = c_{ij} - \frac{c_{ik}c_{jk}}{c_{kk}}, \qquad [18]$$

where the subscripts i, j, k refer to the tabled values existing before deletion (not after). Remember that the table of c values is symmetrical so that $c_{ij} = c_{ji}$.

If it should be desirable to drop, say X_2, then $k = 2$, and as examples:

$$b_1 \text{ after deleting } X_2 = b_1 - \frac{c_{12}}{c_{22}} b_2, \qquad [17a]$$

and

$$c_{34} \text{ after deleting } X_2 = c_{34} - \frac{c_{32}c_{42}}{c_{22}} = c_{34} - \frac{c_{23}c_{24}}{c_{22}}. \qquad [18a]$$

IDENTIFYING THE MOST POTENT VARIABLES IN REGRESSION

Some Considerations in Choosing a Method

With the necessary calculating procedures explained, the primary objective of describing a method for finding a set of potent independent variables can be undertaken.

If there should be no more than 6 to 8 independent variables to be examined, it would not be too illogical to work the regression with all independent variables, find and discard the weakest, then work the regression with the weakest omitted, identify and discard the second weakest independent variable, and so on until omission of another variable would significantly reduce the information obtained. This system of dropping nonsignificant variables would result eventually in a set of potent variables. These variables would not necessarily be either the absolutely most potent set or the set that would be found by the method described in this bulletin.

As mentioned before, the only sure way of finding the r absolutely most potent variables out of n is to work all the C_r^n regressions having r independent variables and then choose the best. Any systematic method for finding the variable that is strongest, next strongest, etc., or weakest, next weakest, etc., makes the assumption that all the r most potent variables will be present in the list of $r + 1$ most potent variables. *This does not necessarily happen.* Practical experience indicates that sets decidedly better than those discovered by the procedure outlined in this bulletin are rare.

The decision of whether to choose the strongest, next strongest, etc., or eliminate the weakest, next weakest, etc., depends primarily upon the amount of work each will require, the usefulness of the intermediate results, and the psychological attitude engendered by the process.

The process of choosing the strongest single variable and then the strongest pair (including the strongest single), etc., has the desirable characteristic that the variable selected as the strongest single variable *really is* the strongest single variable. The work could be stopped at this stage with this useful piece of information. However, if the study can afford to introduce a second variable, then the best variable to use with the first one chosen is the one that the proposed method will yield. The process is continued for subsequent variables. The attitude toward a process as straightforward as this should be good. On the other hand, the process of eliminating the weakest variables must proceed *through the full discarding process* before much information of value is obtained. Further, the most potent single variable might not even be in the list of potent variables retained.

The amount of work involved in the solutions depends upon the number of independent variables to be examined. If there are no more than six or eight variables, with the likelihood that three, four, or five may be accepted, there may be little difference in the two methods. It is the experience of the authors that the number of independent variables to be considered will not be 6 to 8, but 12 to 20, or even more, and that the number of independent variables in the final regression will often be no more than 3 or 4, hardly ever more than 5 or 6. The reason for the large number of independent variables to be examined is that from those variables actually measured *additional variables are created* to account for possible curvilinearity and interaction. As an example, consider a study of volume of tree product produced per acre in which only three independent variables were measured, X_1 = height, X_2 = age, and X_3 = number of trees per unit area. Since all three variables might show the phenomenon of diminishing returns and since age .or X_2 might have cubic effects as well as interactions with both height and number of trees, it is apparent that the independent variables to be investigated are not three but nine in number X_1, X_1^2, X_2, X_2^2, X_2^3, X_3, X_3^2, X_2X_1, and X_2X_3. Such a study was actually made and the number of independent variables after all such considerations was 12, Goggans and Schultz (*10*).

It is apparent, especially in a case in which 20 to 40 independent variables are to be investigated with considerable likelihood that not more than 6 will be retained, that the appropriate method is *not* that of working the regression with all variables, deleting the weakest, and reworking and deleting until only the few best variables remain. In this situation the solution is much shorter to work all simple regressions, thereby identifying the most potent single variable; then work all regressions with two independent variables involving the variable previously identified as most potent. This procedure identifies the most potent pair of variables—subject to the condition that one of them is the previously identified most potent single variable. Extending the procedure involves working all the regressions with three independent variables in which two of the variables are those two previously found to be the most potent pair. This sequence results in finding the most potent triplet of variables—subject, of course, to the condition that two of these are the pair previously chosen of which one is the most potent single variable.

The process described can be extended until all the independent variables have been used and ordered. However, it is usually stopped when a satisfactorily large portion of the variability in Y has been accounted for, or when additional variables do not account for significant amounts of variability. For this purpose significance might be set at a level of chance, say 0.10, rather than the more conventional levels of 0.05 and 0.01. If there are n independent variables with r finally selected, there are n simple regressions to be evaluated, $n - 1$ regressions with two independent variables, $n - 2$ regressions with three independent variables and so on to $n - (r - 1)$ regressions with r independent variables.

In contrast with the method of eliminating weakest variables, which requires that the inverse or c matrix be computed so that the weakest variable can be identified, the only information needed about a set of regressions in order to decide which is the most potent is $\sum \hat{y}^2$, the reduction in $\sum y^2$ due to regression.

When solving for the variables in order of most potent, next most potent, etc., it is only necessary to carry the abbreviated Doolittle solutions through that part of the solution designated as the "forward" solution, or down to the second horizontal ruling of Table 2. At this stage the reduction due to regression may be calculated by [5] as

$$\sum \hat{y}^2 = \sum_i A_{ig}B_{ig}.$$

After the potent set is chosen, it may be desirable to make various tests of significance and perhaps find confidence limits, but on the *chosen set* only.

The factors that make feasible or practicable such a search as here described are:

(1) The end point can be recognized; either a satisfactorily large portion of the variability in Y is explained or further variables do not explain a significant amount of the variability in Y.

(2) The number of regressions with r independent variables to be solved is $n - (r - 1)$ rather than $n!/r!\,(n - r)!$.

(3) The matrices of simultaneous equations may be solved by the abbreviated Doolittle method and need not be carried farther than the "forward" part of the solution—followed, of course, by calculating the sum of squares in Y attributable to the regression, [5].

(4) It seems from some experience that the number of potent variables will usually not exceed five or six. Thus, the heavy computational load of evaluating regressions with more than five or six independent variables does not seem likely to exist.

(5) Only sums of squares, sums of products involving Y, and sums of products between those independent variables finally selected as potent will have to be computed.

(6) Since it is known from the start that all simple regressions will be examined followed by examination of all two-variable regressions involving some most potent single variable, etc., it is possible to organize and systematize the work to save duplication and, further, to use some mechanical tricks to reduce computations and copying.

GENERAL PROCEDURE

To Find the Most Potent Single Variable

For purposes of illustration assume that there are some definite number of independent variables, say 10, rather than the more general case of n variables. Prepare the outline of a table, such as Table 1, but extended to the case of 10 independent variables, Table 3 (page 62). In this case $X_{11} = Y$ and $X_{12} = X_1 + X_2 + \cdots + X_{10} + Y$. *Calculate only* $\sum X_j$, $\bar{X}_j$, $\sum x_i^2$ *in the cells of the diagonal,* $\sum x_i y$ *in column* $X_{11} = Y$, *and the single cell* $X_{11}X_{12}$ *in the* X_{12} *or check column*—see cells indicated by (1), Table 3. The $\sum X_j$ may be checked by $\sum X_{12} = \sum_i \sum X_j$. The work of column X_{11} may be checked by [1], [1a], [2], and [2a] as follows:

$$\sum X_{11}X_{12} = \sum_j \sum X_{11}X_j$$

$$C_{11,12} + \sum x_{11}x_{12} = \sum_j C_{11,j} + \sum_j \sum x_{11}x_j$$

$$= \sum X_{11}X_{12},$$

remembering to add from right to left and to turn up the column at the diagonal, which in this case is the first value added. The quantities in the diagonal cells (sums of squares) will have to be checked by calculating them a second time, but all future entries made in the table, as well as those now entered, will be susceptible to being checked by the check column before being used (or used again in the case of values now entered). The coding values, m_i and m_j, may now be determined from the diagonal entries, but may also be postponed until a later step.

The necessary sums of squares and products are now available to compute the reduction in $\sum y^2$ due to each of the simple regressions. This computation is better if done from uncoded values, since it is just as easy and saves the coding and decoding process. The reduction sum of squares for the ith variable may be calculated as

$$\sum \hat{y}_i^2 = \frac{(\sum x_i y)^2}{\sum x_i^2} \qquad [19]$$

and if $i = 3$

$$\sum \hat{y}_i^2 = \sum \hat{y}_3^2 = \frac{(\sum x_3 y)^2}{\sum x_3^2}. \qquad [19a]$$

The variable with the greatest $\sum \hat{y}^2$ is the most potent variable and is so designated. The computation time is about one minute for each $\sum \hat{y}^2$ computed.

To Find the Second Most Potent Variable

If, for example, it turns out that the variable X_6 is the most potent variable, the next step is to complete in Table 3 all cells involving X_6, both in the row for X_6 and in the column for X_6. See values indicated by (2). This row (column) of calculations may be checked by use of [1] and [2].

$$\sum X_6 X_{12} = \sum_j \sum X_6 X_j$$

and

$$C_{6,12} + \sum x_6 x_{12} = \sum_j C_{6j} + \sum_j \sum x_6 x_j = \sum X_6 X_{12}.$$

Coding values, m_i and m_j, should now be established from the diagonal values. Other a_{ij} values are then calculated for cells indicated as either (1) or (2) by the relationship,

$$a_{ij} = \sum x_i x_j m_i m_j. \qquad [20]$$

If the code fairly consistently yields values in a particular row (column) that are either greater than 10.0 or less than 0.1, the m_i and m_j of this row and column may be changed. Any change should be to the square root of some other even power of 10.

The a_{ij} values in cells indicated by (1) and (2) are sufficient to solve every multiple regression of two independent variables when one of the variables is the singly most potent variable, X_6. The regressions are solved by transferring the a_{ij} to information matrices and carrying out the abbreviated Doolittle solution through the "forward" solution, as described in Table 2, and then calculating the reduction due to regression, [5]:

$$\sum \hat{y}^2 = A_{1g} B_{1g} + A_{2g} B_{2g}.$$

This is done for each pair of independent variables that includes X_6. The largest reduction signifies the most potent pair, and the variable other than the most potent is designated as next most potent—subject to the definition of this report which allows that the chosen pair may not be the absolutely most potent pair. These solutions can be completed in 6 to 8 minutes by clerks who are sufficiently familiar with the process that they have no uncertainty about the next step.

To Find the Third Most Potent Variable

To continue the example, if it turns out that the variables X_6 and X_3 are the most potent pair, X_3 is designated as the second most potent variable. The next step is to complete in Table 3 all cells involving X_3. See values indicated by (3). Accuracy may be checked by use of the check column. After coding, the necessary values of a_{ij} are available to solve every multiple regression of three independent variables when two of the variables are the most potent pair, X_6 and X_3.

Construction of a Mask to Aid in Computations

It is still necessary to transfer the appropriate a_{ij} to information matrices and perform abbreviated Doolittle solutions through the "forward" solution. In being systematic about the work, however, it seems logical to let the most potent variable become X_1' and the second chosen variable be X_2', where the "prime" indicates that the subscript of X may not be the original subscript. Thus, it turns out that each of the eight three-variable regressions, which must be solved, has certain parts that are identical. Table 4 (page 63) indicates by identifying symbols those portions of the solutions that are the same in every solution and indicates by leaders (. . . .) those values that vary as the third variable varies.

The proportion of the solution that is constant is not large with just two out of three independent variables constant, but increases as the process is extended to solving several regressions of four, five, or six independent variables with all but one constant.

Since the only information that will ever be wanted from most of these regressions is $\sum \hat{y}^2$ (the reduction in sum of squares of deviations in Y that may be attributed to the regression), any device to save reworking or even recopying the constant part of these solutions would be worthwhile. The device used at this laboratory is a mask consisting of the constant values with the columns and cells in which new values are to be entered or computed cut out as shown by dotted lines in Table 4. To work a regression, the mask is laid over a fresh sheet of paper, values of a_{ij} entered according to the variable being evaluated as the third predictor, the necessary remaining calculations made, and the reduction in sum of squares calculated, [5]:

$$\sum \hat{y}^2 = A_{1g}B_{1g} + A_{2g}B_{2g} + A_{3g}B_{3g}.$$

The most potent trio of variables, of course, is the one that has the largest sum of squares attributable to regression, or the largest $A_{3g}B_{3g}$. Since two of these three variables are already designated as most potent and second most potent, the third is designated as third most potent. It must be remembered that this list *might not* include all (or indeed even any) of the three absolutely most potent variables, though experience indicates this is unlikely, or if it did happen, the difference in $\sum \hat{y}^2$ by which the absolutely most potent variables would displace these variables would not be large.

The values appearing on the mask for aiding in the search for the most powerful *trio* of variables are copied directly from the regression of the most potent *pair* of variables. (The sturdiness of the mask is increased if the single cell cut out of the Y column, the cell for $a'_{34} = g'_3$ of Table 4, is covered front and back with Scotch tape to provide a "window.") Each regression of three variables requires 8 to 10 minutes to compute, assuming that all necessary values in Table 3 have been computed and are ready for use.

To Find the Fourth and Other Most Potent Variables

If the most potent trio of variables should be X_6, X_3, and X_7, then to proceed further, it will be necessary to complete in Table 3 all cells involving X_7, labeled (4) in Table 3, and solve all regressions of Y on four independent variables in which X_6, X_3, and X_7 are present. The mask may be prepared from the most potent trio of variables. Solution of each regression will require from 12 to 15 minutes computational time. The process can be extended indefinitely. Regressions with five independent variables require from 20 to 25 minutes computational time after the mask has been prepared and the necessary values entered in Table 3.

Testing Significance of the Variables

The question of when to stop is related to how well the independent variables account for the variation in the dependent variable. In many fields of work, it is quite probable that if the investigator could find one or two variables that would account for 95 per cent of the variability in Y he would be quite willing to stop. This would be a satisfactorily large amount of variation explained by satisfactorily few variables. Usually the investigator is not this fortunate but must continue until all the potent variables are identified, that is, until further variables added to the potent list do not account for significant portions of the variability. The significance of the variability accounted for by additional variables can be tested by a series of F tests, as in Table 5 (page 64). Significant variables may be regarded as potent variables.[3]

USING REGRESSION PROCEDURES IN DISCRIMINANT ANALYSIS

If the problem is one of finding a discriminant function rather than a multiple regression, there will be certain changes, but the procedures are essentially the same. First the data will be divided into the two groups on the basis of the discrete variable Y. Again assuming a case in which four independent variables were measured, one will have to calculate the triangular array of $\sum x_1x_j$ as in Table 1 for *each* of the groups, I and II, which are to be discriminated. If a check column is to be used, there is a check column for *each* group based on adding together the 4 X's (but no Y). After computing the $\sum x_ix_j$ for both groups, *corresponding $\sum x_ix_j$ are added together* and summarized in a table similar to Table 1. Instead of a column of $\sum x_iy$ in this table, there is a column listing the mean differences, d_i, between groups I and II for the several X-variables:

$$d_i = \bar{X}_{i,\text{I}} - \bar{X}_{i,\text{II}}. \qquad [21]$$

Call this column d in discriminant analysis. The first difference is listed in line 1, $d_1 = \bar{X}_{1,\text{I}} - \bar{X}_{1,\text{II}}$, the second in line 2, $d_2 = \bar{X}_{2,\text{I}} - \bar{X}_{2,\text{II}}$, and so on, remembering to always take the differences in the same direction. There is no value of d in discriminant analysis corresponding to $\sum y^2$ or g_g of regression analysis. After the $\sum x_ix_j$ have been added together and summarized with the d-values, they may be coded by powers of 10 to a_{ij} values, after which the operations of Table 2, the abbreviated Doolittle solution, are in order. The coded values of d_i serve exactly as the coded values of $\sum x_iy$ in Table 2 and are given the same symbol, g_i, as in regression. The computations are exactly the same as in regression. The discussion about "forward" and "back" solutions, and decoding still holds. Since there is no value corresponding to $\sum y^2$ or g_g of regression analysis, there is no A_{gg} to be calculated in discriminant analysis.

The coefficients of X_i in the discriminant function are calculated in identically the same manner as the b's of multiple regression, though they are usually designated λ or L rather than b, and the quantity

[3] While the authors know of no formal investigations on the matter, there is some intuitive feeling that the significance level here should not be too stringent, say 0.10.

maximized is the average difference between the Z-values of the two groups, I and II, rather than a sum of squares due to regression, $\sum \hat{y}^2$. The difference *can* be computed as

$$D = \bar{Z}_{\mathrm{I}} - \bar{Z}_{\mathrm{II}} \qquad [22]$$

but is *more usually* calculated in the same manner as a sum of squares due to regression, after either [5] or [9]. If after [5],

$$D = \sum_i A_{ig} B_{ig} \qquad [23]$$

or if after [9],

$$D = \sum_i \lambda_i g_i. \qquad [24]$$

After the λ-values have been found (by either of the methods described for b-values), the discriminant function may be written as

$$Z = \lambda_1 X_1 + \lambda_2 X_2 + \lambda_3 X_3 + \lambda_4 X_4. \qquad [25]$$

where either the λ_i must be decoded or the X_i must be coded.

In the same manner as for a regression, the significance of the discriminant may be tested by means of an F-test. The sum of squares attributable to the independent variables is given by

$$SS \text{ due to variables} = \frac{n_{\mathrm{I}} n_{\mathrm{II}}}{n_{\mathrm{I}} + n_{\mathrm{II}}} D^2, \qquad [26]$$

where n_{I} and n_{II} are the numbers of sets of observations in groups I and II, respectively. D is itself the sum of squares for residual and has degrees of freedom, $n_{\mathrm{I}} + n_{\mathrm{II}} - (1 +$ the number of independent variables), as indicated in the analysis of variance tabulated following.

Source of variation	*Degrees of freedom*	*Sum of squares*	*Mean squares*
Variables	No. of variables	$\frac{n_{\mathrm{I}} n_{\mathrm{II}}}{n_{\mathrm{I}} + n_{\mathrm{II}}} D^2$	$\frac{n_{\mathrm{I}} n_{\mathrm{II}} D^2}{(n_{\mathrm{I}} + n_{\mathrm{II}})(\text{no. of var.})}$
Residual	$n_{\mathrm{I}} + n_{\mathrm{II}} - 1 -$ no. of variables	D	$\frac{D}{n_{\mathrm{I}} + n_{\mathrm{II}} - 1 - \text{no. of var.}}$

The residual is used for testing variables, and a significant F-value for variables indicates a significant discriminant.

The discussion of considerations in choosing a method for determining potent variables holds as certainly for discriminants as for regression. If the problem with 10 independent variables should be one of discriminant analysis rather than multiple regression, there must be a table like Table 3 (omitting the a_{ij}) for *each* of the two groups to be discriminated. The corresponding $\sum x_i x_j$ from groups I and II are added together and listed in a third table. The $d_i = \bar{X}_{i\text{I}} - \bar{X}_{i\text{II}}$ are also listed in this third table in the column for $X_{11} = d$, same column as $X_{11} = Y$ in regression. As with $\sum \hat{y}_i^2$ in simple regression, the quantities D_i for the case of single variable discriminants are best calculated without coding. Thus, similarly to [19] and [19a]

$$D_i = \frac{d_i^2}{\sum x_i^2},$$

and if $i = 3$

$$D_i = D_3 = \frac{d_3^2}{\sum x_3^2}.$$

The variable yielding the largest D is the most potent single variable to use as a discriminator. If the most potent variable in the discriminant analysis of the problem with 10 independent variables should turn out to be X_6, it will be necessary as in regression to complete through $\sum x_i x_j$ all the cells involving X_6. This must be done for *both* groups, I and II, adding corresponding $\sum x_i x_j$ together in the third table and then coding to a_{ij}. The necessary values of a_{ij} (coded $\sum x_i x_j$) are now available to solve every discriminant function of two independent variables when one of the variables is the singly most potent variable, X_6. The pertinent a_{ij} are removed to matrices as in Table 2 and the discriminants are solved through the "forward" solution and the evaluation of D by means of [23]

$$D = A_{1g}B_{1g} + A_{2g}B_{2g}.$$

The largest D signifies the most potent pair of variables to be used in a discriminant function. The resulting function [25] is

$$Z = \lambda_1 X_1 + \lambda_2 X_2.$$

If, as in the regression discussion, it should turn out that the variables X_6 and X_3 are the most potent pair, then X_3 is designated as the second most potent variable. To look for the third most potent variable, all cells in Table 3 involving X_3 (see values in Table 3 indicated by "3") must be completed for both groups. After combining and coding, the

a_{ij} can be removed to individual matrices for solution just as in regression. A mask is just as valuable here as in regression analysis.

The search for potent variables may continue to be extended in discriminants just as in regressions, except for the necessary modifications in first finding the $\sum x_i x_j$ of two separate groups and using the average group differences, d_i, of the X_i rather than sums of products, $\sum x_i y$.

The problem of when to stop a search for potent variables in discrimination is subject to the same considerations as in regression. The significance of additional variables can be tested by a series of F-tests as illustrated in Table 5 for regression, although the complexity of the testing process is increased because the sums of squares do not exist as such but must be computed from the values of D. The D calculated with any number of variables may be regarded as the residual sum of squares remaining after the sum of squares due to those variables has been subtracted from the total sum of squares. The sum of squares due to the variables is not known, but may be computed from [26] as

$$SS \text{ due to variables} = \frac{n_{\text{I}} n_{\text{II}}}{n_{\text{I}} + n_{\text{II}}} D^2,$$

where n_{I} and n_{II} are the numbers of sets of observations in the two groups, I and II, and $D = \sum_i A_{ig} B_{ig}$. The total sum of squares is given by

$$\text{Total } SS = \frac{n_{\text{I}} n_{\text{II}}}{n_{\text{I}} + n_{\text{II}}} D^2 + D. \qquad [27]$$

There is still a difficulty in that the basic variable, $D = \bar{Z}_{\text{I}} - \bar{Z}_{\text{II}}$, changes as variables are added to or deleted from the discriminant so that the total sum of squares with a two-variable discriminant is, for example, different from the total sum of squares with a three-variable discriminant. However, since the F-ratio of the mean square for variables to the mean square for residual is valid for each discriminant, the mean squares may all be made comparable to one another by applying the necessary factors to bring every total sum of squares (and its component parts) to some constant total sum of squares. Since the proposed analysis tests a single variable first, then a second variable, a third, and so on, it is proposed that the total sum of squares for the *single most potent variable* be accepted as the constant total sum of squares to which the total sums of squares of other discriminants are to be made equal. In most cases the only sum of squares of interest from the discriminant with r variables is the sum of squares due to the r variables.

This may be adjusted to equal the sum of squares of the single most potent variable as follows:

$$\text{Adj. } SS,\ r \text{ variables} = \frac{n_{\mathrm{I}}n_{\mathrm{II}}}{n_{\mathrm{I}} + n_{\mathrm{II}}} D_r^2 \frac{\text{Total } SS,\ 1 \text{ variable}}{\text{Total } SS,\ r \text{ variables}} \qquad [28]$$

$$= \frac{n_{\mathrm{I}}n_{\mathrm{II}}}{n_{\mathrm{I}} + n_{\mathrm{II}}} D_r^2 \frac{\dfrac{n_{\mathrm{I}}n_{\mathrm{II}}}{n_{\mathrm{I}} + n_{\mathrm{II}}} D_1^2 + D_1}{\dfrac{n_{\mathrm{I}}n_{\mathrm{II}}}{n_{\mathrm{I}} + n_{\mathrm{II}}} D_r^2 + D_r},$$

where D_1 and D_r are the largest D's for 1 and r variables, respectively. Significance of additional variables in discriminant analysis is tested in the same manner as for regression analysis in Table 5. The entries in the first three lines of the sum of squares column of the table resembling Table 5 are: line (2) = the sum of squares due to the most potent variable = $D_1^2\, n_{\mathrm{I}}n_{\mathrm{II}}/(n_{\mathrm{I}} + n_{\mathrm{II}})$, line (3) = the sum of squares among Z-values of the same group = D_1, and line (1) = the Total sum of squares = (2) + (3). Lines (4) and (7) are the sums of squares due to two and three most potent variables, respectively, calculated and adjusted to the total sum of squares of the most potent single variable as in [28]. Sums of squares for all other lines are calculated as indicated in Table 5.

NUMERICAL EXAMPLES

Because members of the group to whom this account is directed often feel a little uncertain as to whether their applications of symbolic representations are correctly made, worked numerical examples of both regression and discriminant analysis are added against which they may check themselves.

Numerical Example of Regression

The data for the regression example, set forth in Table 6 (page 65), consist of 40 sets of observations on one-tenth acre plots of planted longleaf pines. The dependent variable, Y, is the average height in feet of dominant and codominant trees. The independent variables or predictors, X_i, are defined as follows:

X_1 = silt plus clay content of topsoil in per cent,
X_2 = imbibitional water value of the most impervious soil horizon,
X_3 = silt plus clay content of B horizon in per cent, and
X_4 = age of planting in years.

The variables created to allow for curvilinearity and interaction of effects are respectively,

$$X_5 = X_4^2 = (\text{age})^2$$

and

$$X_6 = X_1X_4 = (\text{silt} + \text{clay of topsoil})\ (\text{age}).$$

As a matter of record, the original analysis studied 19 predictor variables, but for purposes of illustration only 6 are included here. The results of the study have been reported by Goggans and Schultz (*10*).

Finding the Most Potent Single Variable

Table 7 (page 66) is prepared in the manner of Table 3, filling in first the sums and means of the X's including $\sum X$, for the check column, the quantities in the diagonal cells, the values in the column $X_7 = Y$, and the values in the single cell X_7X_8 of the check column. Note the checks on computation,

$$18{,}231.8 = 1{,}139.2 + 9{,}190.6 + \cdots + 820.8.$$

Also by [1] and [1a]

$$550{,}444.53 = 34{,}171.52 + 279{,}200.63 + \cdots + 24{,}070.97$$

and by [2] and [2a]

$$\begin{aligned} 519{,}241.66 &+ 31{,}202.87 \\ &= 32{,}444.42 + 261{,}748.29 + \cdots + 23{,}376.38 \\ &\qquad + 1{,}727.10 + 17{,}452.34 + \cdots + 694.59 \\ &= 550{,}444.53. \end{aligned}$$

Values in diagonal cells must be checked by recomputing.

The sums of squares due to regression when regarding each variable as a simple predictor are calculated by [19] and [19a]

$$\sum \hat{y}_i^2 = \frac{(\sum x_i y)^2}{\sum x_i^2},$$

which for X_3 is

$$\sum \hat{y}_3^2 = \frac{(\sum x_3 y)^2}{\sum x_3^2} = \frac{(489.75)^2}{5{,}522.85} = 43.43.$$

The reductions for the six variables are listed in order of decreasing magnitude:

$\sum \hat{y}_4^2$ due to age, X_4, $=1{,}200.53$,

$\sum \hat{y}_5^2$ due to (age)^2, X_5, $=1{,}147.85$,

$\sum \hat{y}_6^2$ due to (age) (silt + clay of topsoil), X_6, $=641.79$,

$\sum \hat{y}_1^2$ due to silt + clay of topsoil, X_1, $=204.24$,

$\sum \hat{y}_3^2$ due to silt + clay of B horizon, X_3, $=43.43$,

and

$\sum \hat{y}_2^2$ due to imbibitional water value of the most impervious soil horizon, X_2, $=0.39$.

Note that no quantities have been coded in these computations. The greatest reduction is 1,200.53 due to X_4; thus, X_4 or age is the most potent single predictor of height. The significance of this reduction is tested in the first three lines of Table 8 (page 67), which is prepared in the manner of Table 5. $F = 86.62$ with 1 and 38 degrees of freedom occurs in not more than 0.001 of cases due to chance; hence, the effect of age on height is to be regarded as very highly significant.

Since it has turned out that X_4, or age, is the most potent single variable, it is necessary to complete Table 7 with regard to column X_4 and row X_4. As computational checks, [1] and [1a]

$$211{,}216.30 = 12{,}949.1 + 107{,}160.4 + \cdots + 4{,}985 + 15{,}579.0 + \cdots + 9{,}190.6$$

and, [2] and [2a]

$$199{,}182.42 + 12{,}033.88 = 12{,}445.8 + 100{,}407.3 + \cdots + 8{,}967.2 + 503.3 + 6{,}753.1 + \cdots + 223.4 = 211{,}216.30$$

Coding

Before evaluating the several regressions with two independent variables, the $\sum x_i x_j$ should be coded to a_{ij}. As an example of coding

$$(\sum x_3^2)(10^{-4}) = (5{,}552.85)(0.000{,}1) = 0.552{,}285,$$

a value between 0.1 and 10.0, so the coding factor for X_3 is

$$\sqrt{0.000{,}1} = 0.01 = 10^{-2}.$$

Enter this value as m_3 for both X_j code and X_i code factors for X_3. As another example

$$(\sum x_6^2)(10^{-6}) = (474{,}585.41)(0.000{,}001) = 0.474{,}585{,}41,$$

a value between 0.1 and 10.0, so code for X_6 is

$$\sqrt{0.000{,}001} = 0.001 = 10^{-3}.$$

Enter this value as m_6 for both the X_i and X_j code of X_6. Other code values are computed similarly with the exception now noted. If the X_7 code is chosen by the above directions, the X_7 code equals the Y code which equals 0.01 and

$$a_{17} = g_1 = 0.069{,}459,$$

$$a_{27} = g_2 = -0.014{,}96,$$

and

$$a_{37} = g_3 = 0.048{,}975,$$

which are all values lying on the low side of the recommended range of 0.1 to 10.0, ignoring signs. It will probably increase agreement between check column and the terms, which should check with the check column, to use 0.1 rather than 0.01 for the $X_7 = Y$ code. This change in coding is introduced and will be used. The diagonal value and one other value exceed 10.0. This result is preferable to values beginning 0.0.

Each a_{ij} is computed from its corresponding $\sum x_i x_j$ by [20].

$$a_{ij} = (\sum x_i x_j)(i \text{ code})(j \text{ code}).$$

As examples:

$$a_{14} = (\sum x_1 x_4)(X_1 \text{ code})(X_4 \text{ code}) = (223.4)(0.01)(0.1) = 0.223{,}4,$$

$$a_{44} = (\sum x_4^2)(X_4 \text{ code})(X_4 \text{ code}) = (211)(0.1)^2 = 2.11.$$

Other values of a_{ij} are computed similarly and entered in Table 7.

Finding other Potent Variables after the Most Potent

The quantities, a_{ij} or coded $\sum x_i x_j$, necessary for evaluating the five regressions of height on two independent factors when one of the independent factors is age are now available. As an example, consider X_4, age, and X_1, silt plus clay content of topsoil; transfer the a_{ij} involving X_1, X_4, and Y to a table of the same form as Table 2; and perform the "forward" part of the solution. This manipulation is shown in Table 9 (page 68).

The values in the first three rows of the first three columns of Table 9 are a_{ij} values entered from Table 7. The "primes" are reminders that the row and column subscripts of X's in Table 9 may not be the original subscripts of Tables 6 and 7. The first three values in the check column are the sums of the quantities in the row involved, and are so obtained. Thus, the value at row X'_2 of check column is given by

$$1.154{,}230{,}00 = 0.694{,}590{,}00 + 0.236{,}240{,}00 + 0.223{,}400{,}00.$$

All values in A_{1j} are a_{1j} values copied from row X'_1. All values in B_{1j} are corresponding A_{1j} values divided by the leading $A = A_{11}$; thus,

$$B_{13} = \frac{A_{13}}{A_{11}} = \frac{5.033{,}000{,}00}{2.110{,}000{,}00} = 2.385{,}308{,}06$$

and

$$B_{14} = \frac{A_{14}}{A_{11}} = \frac{7.366{,}400{,}00}{2.110{,}000{,}00} = 3.491{,}184{,}83.$$

The computational check in A_1 is

$$7.366{,}400{,}00 = 5.033{,}000{,}00 + 0.223{,}400{,}00 + 2.110{,}000{,}00$$

and in B_1 is

$$3.491{,}184{,}83 = 2.385{,}308{,}06 + 0.105{,}876{,}78 + 1.0.$$

These results should agree within rounding errors.

Following the guide provided in [3], [3a], and Table 2,

$$A_{22} = 0.236{,}240{,}00 - (0.223{,}400{,}00)(0.105{,}876{,}78) = 0.212{,}587{,}13.$$

This value divided by the leading A, itself, is equal to 1.0. Now,

$$A_{23} = 0.694{,}590{,}00 - (0.223{,}400{,}00)(2.385{,}308{,}06) = 0.161{,}712{,}18$$

is entered in the table and is divided by the leading A or A_{22}, 0.212,587,13, before removing from the machine; that is,

$$B_{23} = \frac{0.161{,}712{,}18}{0.212{,}587{,}13} = 0.760{,}686{,}59.$$

Remember that the product, (0.223,400,00)(2.385,308,06), which is subtracted from 0.694,590,00 could also have been obtained from the product (0.105,876,78)(5.033,000,00), but the rounding error is smaller on the average when the pair more nearly equal in absolute value is chosen. Further,

$$A_{24} = 1.154{,}230{,}00 - (0.223{,}400{,}00)(3.491{,}184{,}83) = 0.374{,}299{,}31,$$

and

$$B_{24} = \frac{A_{24,\text{ before clearing machine}}}{0.212{,}587{,}13} = 1.760{,}686{,}59.$$

Checks:

$$0.374{,}299{,}31 = 0.161{,}712{,}18 + 0.212{,}587{,}13$$

and

$$1.760{,}686{,}59 = 0.760{,}686{,}59 + 1.0.$$

These results should agree within rounding errors.

The sum of squares due to regression is given by [5]:

$$\sum \hat{y}^2 = \sum_i A_{ig}B_{ig}.$$

This calculation is indicated under Table 9 and the result is decoded there by dividing by the square of the Y code. Note that decoded $A_{1g}B_{1g} = 12.005{,}255{,}47(1/0.1)^2 = 1200.525{,}547$, the sum of squares due to X_1 as calculated by [19].

Similar solutions must be made for the four other two-factor regressions involving X_4 or age. The results of all five solutions are given in Table 10 (page 69) where it may be observed that the greatest $\sum \hat{y}^2$, reduction due to regression, is that due to X_1' and X_2', which are equal to X_4 and X_2, respectively, or age and imbibitional water value of the most impervious soil horizon. Thus, it is concluded that the second most potent predictor—subject to the condition that one of the predictors is X_4 or age—is X_2 or imbibitional water value of the most impervious soil horizon. The significance of this result may be tested by adding three more lines to Table 8. The sum of squares due to imbibitional water value independent of the most potent variable, age, is the difference in sums of squares for age plus imbibitional water value and age alone; that is,

$$1241.49 - 1{,}200.53 = 40.96$$

with 1 degree of freedom. The probability of $F = 3.12$ with 1 and 37 degrees of freedom due to chance alone is less than 0.1. At the 10 per cent significance level, it may be concluded that the height that young longleaf pine trees attain is related to the imbibitional water value of the most impervious soil horizon.

At this stage one could compute on the worksheet for $X_1' = X_4$ and $X_2' = X_2$ the partial regression coefficients, $b_2 = b_{Y2.1}$ and $b_1 = b_{Y1.2}$, for imbibitional water and age, respectively, to verify that their signs and size are reasonable. This computation is not necessary and is not done in this example.

Since X_2 is the second most potent variable, it is necessary to complete Table 7 with respect to column X_2 and row X_2 and then to solve the four regressions of Y on three independent variables when two of them are X_4, or age, and X_2, or imbibitional water. The usual checks [1], [1a], [2], and [2a] are made on the accuracy of entries in Table 7. The new entries are coded to a_{ij} by the existing coding factors.

The mask to reduce the duplication and copy work of solving several multiple regressions of three independent variables when two of the independent variables are always the same may be copied from the solution of the multiple regression on the two independent variables chosen to be constant. The two constant variables are the two most potent ones or, in this study, $X_1' = X_4$, age, and $X_2' = X_2$, imbibitional water value of most impervious horizon. The solution when $X_1' = X_4$ and $X_2' = X_2$ has not been shown, but is of the same form as Table 9.

The columns for $X_1' = X_4$ and $X_2' = X_2$ are copied from their solution table into columns $X_1'' = X_4$ and $X_2'' = X_2$ of Table 11 (page 69), where the double "primes" indicate possible further changes in subscripts. The next column is left open for $X_3'' = ?$; and $X_3' = Y$ of Table 9 is copied into $X_4'' = Y$ of Table 11. The check column, X_5'' or the check sum which includes the values entered in X_3'', is different for each variable and is also left open. Table 11 represents the mask to be used for these variables. The dotted lines show how the mask would be cut out to allow several different X_j to be evaluated as X_3''.

To solve the regression of height on $X_1'' = X_4$, $X_2'' = X_2$, and $X_3'' = X_1$, place the mask on a fresh sheet of paper, copy in the needed values of a_{ij}, and solve. The final results will appear as in Table 12 (page 70). The dotted line divides the portion on the mask from that copied directly to the new sheet from Table 7 and from that which must be worked out for a particular solution.

The first four values in the check column of Table 12 are obtained by adding quantities in the row in which the value is to be entered; that is, $5.438{,}860{,}00 = -0.149{,}600{,}00 + 0.535{,}760{,}00 + 5.745{,}700{,}00 - 0.693{,}000{,}00$, etc. Line A_{1j} is line X_1'' copied. Line B_{1j} is line A_{1j} divided by the leading A of that line, A_{11}; thus,

$$0.105{,}876{,}78 = \frac{0.223{,}400{,}00}{2.110{,}000{,}00}.$$

A_{2h} of check column is given by [3] and [3a]

$$5.438{,}860{,}00 - (-0.693{,}000{,}00)(3.162{,}748{,}82) = 7.630{,}644{,}93$$

which is divided by 5.518,093,84 before clearing machine to yield B_{2h} whose value is 1.382,840,73. Remember that (−0.693,000,00)(3.162,748,82) was chosen rather than (−0.328,436,02)(6.673,400,00) because the absolute magnitudes of factors in the first product are nearer each other in size. Further,

$$A_{33} = 0.236{,}240{,}00 - (0.223{,}400{,}00)(0.105{,}876{,}78) - (0.609{,}132{,}61)(0.110{,}388{,}23) = 0.145{,}346{,}06$$

Computational checks are available in that each value in the check column should be the same (within rounding) as the sum of the other quantities in the same line, [4] and [4a]. In the lines for A_{ij} and B_{ij}, missing entries are zero so that there is no need to turn up the column at the diagonal; that is,

$$6.673{,}400{,}00 = 5.033{,}000{,}00 + 0.223{,}400{,}00 - 0.692{,}000{,}00 + 2.110{,}000{,}00$$

and

$$1.382{,}840{,}73 = 0.272{,}452{,}50 + 0.110{,}388{,}23 + 1.0.$$

When working to eight decimal places, as in this example, rounding errors preventing checking in the seventh place may occur if a diagonal value of A_{ij} becomes less in absolute value than 0.010,000,00.

The results of the four regressions with three independent variables are summarized in Table 13 (page 71). Even the trio of independent variables with the largest regression sum of squares does not add a significant amount to the reduction in sum of squares of Y attributable to regression. (See Table 8.) It is probably safe to conclude at this stage that only age and imbibitional water value of the most impervious soil horizon are potent variables.

Attention is called to the fact that in Table 8 the sum of squares, "reduction due to age" which equals 1,200.53, is the same as

$$A_{1g}B_{1g}\left(\frac{1}{m_y}\right)^2 = (5.033{,}000{,}00)(2.385{,}308{,}06)\left(\frac{1}{0.1}\right)^2 = 1{,}200.526$$

of Table 12; and the sum of squares, "reduction due to imbibitional water value independent of age" which equals 40.96 of Table 8, is the same as

$$A_{2g}B_{2g}\left(\frac{1}{m_y}\right)^2 = (1.503{,}418{,}48)(0.272{,}452{,}50)\left(\frac{1}{0.1}\right)^2 = 40.961$$

of Table 12, both of which are found on the mask. The sum of squares, "reduction due to $(\text{age})^2$ independent of the others," is equal to 6.53 and the "residual" sum of squares after all the variables are accounted for is 479.08. These are calculated by $A_{3g}B_{3g}(1/m_y)^2$ and by $A_{4g}(1/m_y)^2$, respectively, of a table like Table 12 but including $(\text{age})^2$ as the third independent variable rather than "silt plus clay content of topsoil."

The results of this analysis fall into a pattern that seems to be quite common; that is, there are only one, two, or at most a few variables that are sufficiently strongly related to the dependent variable to be useful as predictors. Although not necessary, it is usual that as new variables are added to the list in order of potency, as assessed by the method described above, later-added variables contribute smaller sums of squares to regression and are of less significance. This is because of occasional interrelationships among the variables such that two or three variables jointly contribute sizably to the sums of squares of regression with very small increments from any one or two of the variables used without completing the set. For this reason it may be desirable to ascertain that two successive most potent variables are both nonsignificant before abandoning the search.

If one should wish to know whether the fourth most potent variable in this study is significant, it will be necessary to identify and test this variable by extending the process already described. Fill in the missing values in row and column X_4, solve the three regressions of Y on four independent variables when three of them are X_4, X_2, and X_5, and test the one with the greatest $\sum \hat{y}^2$ by means of three more lines in Table 8. For those who are curious, silt plus clay content of B horizon is the next most potent variable. However, the total sum of squares due to regression of these four variables is 1251.29, which is only 3.27 more than the reduction due to three variables, and actually less than the mean square of residual. It is concluded that, of this list of six independent variables, only two have genuine worth as predictors of height of dominant trees in young longleaf pine plantations. These variables are age and imbibitional water value of the most impervious soil horizon.

The solution for these particular two independent variables has not been shown as such. However, it has been reproduced on the mask portions of Tables 11 and 12. From these values, using Table 2 as a guide, it can be determined that

$$b_2'' = b_2 = 0.272{,}452{,}50$$

and

$$\begin{aligned} b_1'' = b_4 &= 2.385{,}308{,}06 - (0.272{,}452{,}50)(-0.328{,}436{,}02) \\ &= 2.474{,}791{,}27, \end{aligned}$$

hence, using [10],

$$b_2^* = (0.272{,}452{,}50)\left(\frac{0.1}{0.1}\right) = 0.272{,}5$$

and

$$b_4^* = (2.474{,}791{,}27)\left(\frac{0.1}{0.1}\right) = 2.475.$$

From the foregoing results, the means of Tables 6 or 7 and equations [7] and [8], the regression equation,

$$\hat{Y} = \bar{Y} + \sum_i b_i x_i = \bar{Y} - \sum_i b_i \bar{X}_i + b_1 X_1 + b_2 X_2$$

is found to be

$$\hat{Y} = 28.48 - (2.475)(10.92) - (0.272{,}5)(5.88) + 2.475X_4 + 0.272{,}5X_2.$$

Hence,

$$\hat{Y} = -0.15 + (2.475)(\text{age}) + (0.272{,}5)(\text{imbibitional water value}).$$

Note as checks on computation that every B_{ij} in Table 12 agreed with its A_{ij} in sign and that *all values on the diagonal of the solution were positive.* These conditions must always be true.

Numerical Example of a Discriminant

The data for illustrating the discriminant function, presented in Table 14 (page 71), are taken from Goulden (*11*), page 352, who abstracted them from the study by Cox and Martin (*3*) to determine a discriminant function for differentiating soils with and without *Azotobacter.* The sums of squares and products of deviations for each group are set forth in the upper part of Table 15 (page 72) and then added together in the lower part of the table. This is just one of several ways these sums can be obtained since any procedure yielding the sums of squares and products "Within Groups" would give the same results.

The differences, d_i, between the average values of the two groups for the several variables, X_i, are calculated by [21]

$$d_i = \bar{X}_{i,\text{I}} - \bar{X}_{i,\text{II}}$$

and listed in a column, $X_4 = d$, of the lower part of Table 15. This column takes the place of the column $X = Y$ of Table 1, in which the $\sum x_i y$ values of regression are recorded. The values in the lower part of Table 15 are coded by powers of 10 to the quantities a_{ij}, which are then removed to Table 16 where the abbreviated Doolittle solution is performed, including the back solution for the c_{ij} values and calculation of the discriminant coefficients, λ_1.

The upper part of Table 15 should not require much explanation. The sums of squares and products of the possible discriminating variables are obtained by the usual processes. The column of the check sum is used in the same way as illustrated for regression in Table 1. Thus, for example: in Group II, using [1], [1a], [2], and [2a] to check X_2:

$$64{,}677.8 = 20.928 + 37{,}979 + 5{,}770.8$$

and also

$$\begin{aligned} 61{,}189.7333 + 3{,}488.0667 \\ &= -186.6667 + 3{,}632.0000 + 42.7333 \\ &\quad + 5{,}728.0667 + 34{,}347.0000 + 21{,}114.6667 \\ &= 64{,}677.8000 \end{aligned}$$

In the lower part of Table 15, by [21]

$$d_i = \bar{X}_{i,\text{I}} - \bar{X}_{i,\text{II}},$$

and, for example,

$$d_2 = \bar{X}_{2,\text{I}} - \bar{X}_{2,\text{II}} = 87.8400 - 35.6667 = 52.1733.$$

The powers of 10 for coding factors are chosen as the square roots of the even powers of 10 that reduce the diagonal values of $\sum x_i x_j$ to values between 0.1 and 10.0 and are written in as X_i code and X_j code. For example,

$$(0.0001)(\sum x_2 x_2) = (0.0001)(88{,}897.3600) = 8.889{,}736{,}00,$$

a value between 0.1 and 10.0; so,

$$X_i \text{ code for } X_2 = \sqrt{0.0001} = 0.01.$$

and since the X_i code is equal to the X_j code when $i = j$,

$$X_j \text{ code for } X_2 = 0.01.$$

The a_{ij} values in the lower part of Table 15 are obtained by [20] from the $\sum x_i x_j$ values by multiplying each $\sum x_i x_j$ by the appropriate code factors. For example,

$$\begin{aligned} a_{13} &= \sum x_1 x_3 (X_1 \text{ code})(X_3 \text{ code}) \\ &= (148.2403)(0.1)(0.01) = 0.148{,}240{,}30. \end{aligned}$$

After transferring the a_{ij} to Table 16 (page 73), the "forward" portion of the Doolittle solution is performed as already illustrated several times. (See Table 2 for directions.) There is no value in the column,

$$X_4 = d,$$

corresponding to $\sum y^2$ or g_g so that there is no A_{gg} value to be computed. Using the directions of Table 2, the λ_i may now be calculated in the column, $X_4 = d$, exactly as the b_i are calculated in regression; that is,

$$\lambda_3 = B_{34} = B_{3g} = 0.054{,}935{,}95,$$

and

$$\begin{aligned}\lambda_2 = B_{2g} - \lambda_3 B_{23} &= 0.028{,}634{,}99 - (0.054{,}935{,}95)(0.076{,}497{,}57)\\ &= 0.024{,}432{,}52,\end{aligned}$$

etc. Checks are carried in column X_5, which is the Check $\sum$. The c_{ij} are calculated exactly as indicated in the appropriate cells of Table 2, as examples,

$$\begin{aligned}c_{12} &= -B_{13}c_{23} - B_{12}c_{22}\\ &= -(0.707{,}354{,}58)(-0.167{,}357{,}12) - (2.013{,}252{,}37)(0.137{,}175{,}72)\\ &= -0.157{,}788{,}52\end{aligned}$$

and

$$\begin{aligned}c_{11} &= \frac{1}{A_{11}} - B_{13}c_{13} - B_{12}c_{12}\\ &= \frac{1}{0.209{,}570{,}00} - (0.707{,}354{,}58)(-1.210{,}578{,}80)\\ &\qquad - (2.013{,}252{,}37)(-0.157{,}788{,}52)\\ &= 5.945{,}651{,}91.\end{aligned}$$

Remember in determining the c_{ij} to start at the bottom with

$$c_{33} = \frac{1}{A_{33}} = \frac{1}{0.457{,}091{,}82} = 2.187{,}744{,}25.$$

A check may be established on the calculation of the c values: the sum of all the a_{ij} values in some particular row, i, each multiplied by the c_{ij}at the same position of the c_{ij} table is unity; that is,

$$\sum_j a_{ij}c_{ij} = 1.0.$$

For example,

$$\begin{aligned}\sum_j a_{2j}c_{2j} &= a_{23}c_{23} + a_{22}c_{22} + a_{12}c_{12}\\ &= (0.913{,}509{,}33)(-0.167{,}357{,}12) + (8.889{,}736{,}00)(0.137{,}175{,}72)\\ &\qquad + (0.421{,}917{,}30)(-0.157{,}788{,}52)\\ &= 0.999{,}999{,}94,\end{aligned}$$

which is sufficiently close. These checks are outlined in Table 2.

The λ_i may also be computed from the c_{ij} in exactly the same manner as the b_i were, letting the column, $X = d$, serve in the place of $X = Y$. As an example,

$$b_2 = c_{23}g_3 + c_{22}g_2 + c_{12}g_1$$

and likewise

$$\lambda_2 = c_{23}g_3 + c_{22}g_2 + c_{12}g_1$$

so

$$\lambda_2 = (-0.167{,}357{,}12)(0.145{,}141{,}00) + (0.137{,}175{,}72)(0.521{,}733{,}00) + (-0.157{,}788{,}52)(0.144{,}790{,}00) = 0.024{,}432{,}52.$$

The λ_i are decoded in the same way as the b_i; thus, similarly to [10]

$$\text{decoded } \lambda_i = \lambda_i^* = \lambda_i \frac{m_i}{m_d}. \qquad [29]$$

As an example,

$$\text{decoded } \lambda_3 = \lambda_3^* = \lambda_3 \frac{m_3}{m_d} = 0.054{,}935{,}95 \frac{0.01}{1.0} = 0.000{,}549{,}36.$$

With the λ_i computed, it is possible to write the discriminant function [25]:

$$Z = \lambda_1 X_1 + \lambda_2 X_2 + \lambda_3 X_3,$$

and using the decoded λ_i,

$$Z = 0.060{,}284X_1 + 0.000{,}244X_2 + 0.000{,}549X_3.$$

But the units of λ are arbitrarily chosen, so divide each λ by the smallest λ, λ_2, which gives

$$Z = 246.7X_1 + X_2 + 2.248X_3,$$

(This is identically the answer Goulden would have obtained, except for what seems to have been an error in rounding off the divisor.) It might better bring out the interrelationships and relative contributions of the variables to make the coefficient, λ, of the most potent variable unity. To do this divide each coefficient by the λ of the most potent variable. This gives

$$Z = X_1 + 0.004{,}053X_2 + 0.009{,}113X_3.$$

The λ_i are really coefficients such that the difference, D, between the average Z-values of the two groups as given by [22], [23], or [24],

$$D = \bar{Z}_{\text{I}} - \bar{Z}_{\text{II}} = \sum_i A_{ig}B_{ig} = \sum_i \lambda_i g_i,$$

is maximized relative to the variability within the groups.

$$\begin{aligned} D &= A_{1g}B_{1g} + A_{2g}B_{2g} + A_{3g}B_{3g} \\ &= (0.144{,}790{,}00)(0.690{,}890{,}87) + (0.230{,}234{,}19)(0.028{,}634{,}99) \\ &\qquad + (0.025{,}110{,}77)(0.054{,}935{,}95) \\ &= 0.100{,}034{,}09 + 0.006{,}592{,}75 + 0.001{,}379{,}48 \\ &= 0.108{,}006{,}32, \end{aligned}$$

or, using the coded (not yet decoded) results:

$$\begin{aligned} D &= \lambda_1 g_1 + \lambda_2 g_2 + \lambda_3 g_3 \\ &= (0.602{,}842{,}84)(0.144{,}790{,}00) + (0.024{,}432{,}52)(0.521{,}733{,}00) \\ &\quad + (0.054{,}935{,}95)(0.145{,}141{,}00) = 0.108{,}006{,}33. \end{aligned}$$

From the calculation of $D = A_{1g}B_{1g} + A_{2g}B_{2g} + A_{3g}B_{3g}$, it can be seen that D is made up of three parts: 0.100,034,09 + 0.006,592,75 + 0.001,379,48. These are due to X_1, to X_2 independent of X_1, and to X_3 independent of X_1 and X_2, respectively. Using the same method as in multiple regression [15], it is possible to estimate the contribution of *each* variable to D independent of the other two despite the fact that X_3 is the only one occurring in the last position; that is,

$$D_{i.\text{all others}} = \frac{(\lambda_i)^2}{c_{ii}}. \qquad [30]$$

For the three variables these values are:

$$D_{1.23} = \frac{(\lambda_1)^2}{c_{11}} = \frac{(0.602{,}842{,}84)^2}{5.945{,}651{,}91} = 0.061{,}123{,}57$$

$$D_{2.13} = \frac{(\lambda_2)^2}{c_{22}} = \frac{(0.024{,}432{,}52)^2}{0.137{,}175{,}72} = 0.004{,}351{,}70$$

$$D_{3.12} = \frac{(\lambda_3)^2}{c_{33}} = \frac{(0.054{,}935{,}95)^2}{2.187{,}744{,}25} = 0.001{,}379{,}48.$$

Note particularly that $D_{3.12} = 0.001,379,48$ is the same as was estimated for X_3 independent of X_1 and X_2 by the term $A_{3g}B_{3g}$. It is apparent that X_3 contributes least to the discriminant so that if a two-variable discriminant should be desired, the appropriate two variables (if practical matters of measurement do not intervene) would be X_1 and X_2, or pH and available phosphate, respectively.

The potency of *individual, single-variable* discriminants may be quickly estimated from quantities available in Table 16. The procedure is analogous to [19] and [19a] for estimating the several simple regressions in regression analysis, but instead of computing

$$\sum \hat{y}_i^2 = \frac{(\sum x_i y)^2}{\sum x_i^2}$$

calculate

$$D_i = \frac{(d_i)^2}{\sum x_i^2} \qquad [31]$$

For the three variables these values are:

$$D_1 = \frac{(d_1)^2}{\sum x_1^2} = \frac{(0.144,790,00)^2}{0.209,570,00} = 0.100,034,09$$

$$D_2 = \frac{(d_2)^2}{\sum x_2^2} = \frac{(0.521,733,00)^2}{8.889,736,00} = 0.030,620,18$$

$$D_3 = \frac{(d_3)^2}{\sum x_3^2} = \frac{(0.145,141,00)^2}{0.609,001,19} = 0.034,590,92.$$

The largest D is due to X_1; hence the most potent single variable to use as a discriminator would be X_1, or pH.

When there are only three independent variables (in either regression or discriminant), a combination of operations consisting of finding the most potent single variable and then finding the weakest of the three variables in the three-variable relationship results in complete knowledge of order of potency, since the most potent pair comes from dropping the weakest. In this study the most potent variable is X_1, the next most potent is X_2, and the least potent is X_3. This is the same order in which the variables were tabulated but order had nothing to do with this result, a fact one may verify by arranging the variables in some other order and solving again.

Having ordered the potency of these three variables as discriminators, it is possible to make F tests of the significance of the amounts by which D is increased by the addition of each successive variable. For this purpose, use the arrangement of Table 5 and equations [26], [27], and [28]. The results are tabulated in Table 17 (page 74).

The value of D_1 is 0.100,034,09 and, being the largest such D, indicates that X_1 is the most potent single discriminator. The sum of squares due to this variable as a discriminator may be calculated by [26]

$$SS \text{ due to 1 variable } = \frac{n_{\text{I}}n_{\text{II}}}{n_{\text{I}} + n_{\text{II}}} D^2 = \frac{(25)(27)}{25 + 27} (0.100{,}034{,}09)^2$$
$$= 0.129{,}896{,}23.$$

This is entered in Table 17 at line (2). The residual or error sum of squares $= D$ is entered at line (3). The total sum of squares may be calculated by (2) + (3) in Table 17 or by [27]:

$$\text{Total } SS = \frac{n_{\text{I}}n_{\text{II}}}{n_{\text{I}} + n_{\text{II}}} D^2 + D = 0.129{,}896{,}23 + 0.100{,}034{,}09$$
$$= 0.229{,}930{,}32.$$

This is entered in Table 17 at line (1).

To find the sum of squares for the two most potent discriminators, it is first necessary to find D for the two most potent variables. This value can be found by solving the discriminant for these two variables, but it may also be had in this special case in which there are only three independent variables as D for three variables minus D for the weakest variable independent of the other two variables, in this example $D_{3.12}$; thus

$$D_{\text{omitting } X_3} = D_{2 \text{ variables}} = D_{3 \text{ variables}} - D_{3.12}$$
$$= 0.108{,}006{,}33 - 0.001{,}379{,}48 = 0.106{,}626{,}85.$$

From this value, the sum of squares due to two variables may be calculated from [26] as follows:

$$SS \text{ due to 2 variables } = \frac{n_{\text{I}}n_{\text{II}}}{n_{\text{I}} + n_{\text{II}}} D^2 = \frac{(25)(27)}{25 + 27} (0.106{,}626{,}85)^2$$
$$= 0.147{,}581{,}75.$$

The total sum of squares may be calculated by [27]

$$\text{Total } SS = \frac{n_{\text{I}}n_{\text{II}}}{n_{\text{I}} + n_{\text{II}}} D^2 + D = 0.147{,}581{,}75 + 0.106{,}626{,}85$$
$$= 0.254{,}208{,}60.$$

To be comparable with sums of squares already calculated for the single most potent variable X_1, this sum of squares due to the two variables, X_1 and X_2, must be adjusted for the fact that its total sum of squares is not the same total sum of squares as for X_1 only. From [28] this adjustment is

$$\text{Adj. } SS\text{, 2 variables} = (SS \text{ due to 2 variables}) \frac{\text{Total } SS\text{, 1 variable}}{\text{Total } SS\text{, 2 Variables}}$$

$$= 0.147{,}581{,}75 \frac{0.229{,}930{,}32}{0.254{,}208{,}60} = 0.133{,}486{,}90.$$

This value is entered in Table 17 at line (4).

The sum of squares due to 3 variables is

$$SS \text{ due to 3 variables} = \frac{n_{\mathrm{I}} n_{\mathrm{II}}}{n_{\mathrm{I}} + n_{\mathrm{II}}} D^2 = \frac{(25)(27)}{25 + 27} (0.108{,}006{,}33)^2$$

$$= 0.151{,}425{,}48.$$

The total sum of squares is

$$\text{Total } SS = 0.151{,}425{,}48 + 0.108{,}006{,}33 = 0.259{,}431{,}81.$$

Adjusted SS due to three variables is

$$\text{Adj. } SS\text{, 3 variables} = (\text{SS due to 3 variables}) \frac{\text{Total } SS\text{, 1 variable}}{\text{Total } SS\text{, 3 variables}}$$

$$= (0.151{,}425{,}48) \frac{0.229{,}930{,}32}{0.259{,}431{,}81} = 0.134{,}206{,}01.$$

This value is entered in Table 17 at line (7). After entering the foregoing values in Table 17, the remaining quantities may be calculated as indicated therein.

From Table 17 it is evident that the discriminant based on the three variables, $X_1 = pH$, $X_2 =$ phosphate, and $X_3 =$ nitrogen content is no better than the one based on two, pH and phosphate. The logical interpretation is to ignore nitrogen and compute the discriminant function based on the two most potent variables only. The λ_i could be recomputed leaving out X_3, but they can be much more quickly computed from data in Table 16 using [17], the formula for computing some b_i after deleting some variable, X_k.

$$\lambda_i \text{ after deleting } X_k = \lambda_i - \frac{c_{ik}}{c_{kk}} \lambda_k$$

In this case $X_k = X_3$ so that

$$\begin{aligned}\lambda_1 \text{ after deleting } X_3 &= \lambda_1 - \frac{c_{13}}{c_{33}}\lambda_3\\ &= 0.602{,}842{,}84 - \frac{-1.210{,}578{,}80}{2.187{,}744{,}25}\,0.054{,}935{,}95\\ &= 0.633{,}241{,}41\end{aligned}$$

and

$$\begin{aligned}\lambda_2 \text{ after deleting } X_3 &= \lambda_2 - \frac{c_{23}}{c_{33}}\lambda_3\\ &= 0.024{,}432{,}52 - \frac{-0.167{,}357{,}12}{2.187{,}744{,}25}\,0.054{,}935{,}95\\ &= 0.028{,}634{,}99.\end{aligned}$$

Since these are the values of λ that would be obtained considering just X_1 and X_2 as predictor variables, they may be used to obtain the discriminant function based on these two variables alone. Therefore, by [25] and [29] the decoded discriminant function is:

$$\begin{aligned}Z &= \lambda_1^* X_1 + \lambda_2^* X_2 = \lambda_1 \frac{m_1}{m_d} X_1 + \lambda_2 \frac{m_2}{m_d} X_2\\ &= 0.633{,}241{,}41 \frac{0.1}{1.0} X_1 + 0.028{,}634{,}99 \frac{0.01}{1.0}\\ &= 0.063{,}324{,}14X_1 + 0.000{,}286{,}35X_2.\end{aligned}$$

However, since these units are arbitrary, divide through by the smallest λ, or 0.000,286,35, to obtain

$$Z = 221.1X_1 + X_2$$

or divide through by the λ of the most potent variable to obtain

$$Z = X_1 + 0.004{,}522X_2$$

which is the discriminant function for the two most potent variables, pH and soil phosphate content. Division by λ of the most potent variable gives other λ values as proportions of the most potent.

Working the discriminant with three variables, as just concluded, has served to give numerical examples of many of the computational procedures used in regression and discriminant analysis. It has served also to show that these procedures are for the most part identical. The demonstrated procedures were used to find the order of potency of *three* independent variables.

The ordering of three independent variables according to potency, however, is a special case that does not call for use of procedures proposed earlier in this bulletin for finding potent variables. However, the proposed procedures *may* be used. To demonstrate their operation in discriminant analysis, this same problem will be re-examined using the proposed procedure.

As before, the first step would be the preparation of Table 15, but with this difference, **only** $\sum X_i$, $\bar{X}_i$, $\sum x_i^2$ (those quantities on the diagonal), and the d_i would be computed. From these, one would determine the potency of all individual, single-variable discriminants by [31]

$$D_i = \frac{(d_i)^2}{\sum x_i^2}$$

These have already been computed and will not be duplicated here. The result is noted that X_1 is the most potent variable with $D_1 =$ 0.100,034,09. The significance of this discriminant may be tested by means of the first three lines of Table 17 using [26] and [27]. Since D is the same as in the previous analysis the results in Table 17 will be the same.

To determine the next most potent variable, it would be necessary to complete in Table 15 all rows and columns associated with the most potent variable, X_1. This includes the value in the Check $\sum$ column to be used with [1], [1a], [2], and [2a] in checking work. The results would be sufficient to evaluate every two variable discriminant in which one of the variables is the most potent variable X_1. The entries of Table 15 would now be coded by powers of 10 and removed to individual matrices for the abbreviated Doolittle solution. The matrix for the discriminant using X_1 and X_2 is given in Table 18 (page 75). The result, $D =$ 0.106,626,84 compares with $D =$ 0.103,654,62 for the discriminant with X_1 and X_3. Thus, the two-variable discriminant with the most potency is that one with X_1 and X_2; hence X_2 is designated as the second most potent variable. Using [26], [27], and [28] three more lines may be added to Table 17, testing the significance of X_2 as the second most potent variable. Since the value, 0.106,626,84, obtained for the discriminant with X_1 and X_2 is the same as that obtained by deleting the weakest variable in three, the results of testing significance are the same as before.

To evaluate further independent variables, it is necessary to complete all rows and columns involving X_3 in Table 15, thus making possible the evaluation of all discriminants with three independent variables in which two of the variables are X_1 and X_2. In this particular example, there is only one such discriminant (which has already been evaluated), but the procedure as laid out is general; therefore, it applies no matter how many variables are in the study. Also, if there were more variables, the procedure could be *extended* until either sufficient discriminatory power is obtained or until additional variables added to the discriminant do not significantly increase the discriminatory power. The mask for reducing computational work in solving matrices of three or more independent variables when there are large numbers of variables is just as useful in discriminant analysis as in regression analysis.

In the present numerical example, the regression with three independent variables would turn out as previously worked so that Table 17 would be completed without change. This indicates that the proposed procedure for determining potent variables and the special case procedure used for three independent variables give the same results.

OTHER TESTS OF SIGNIFICANCE

Having identified the most potent predicting (or discriminating) variables in a regression (or discriminant), one would probably wish to complete the regression (or discriminant). This would probably involve calculating the regression coefficients, b_i (or discriminant coefficients, λ_i), the inverse matrix of c_{ij} values, and the regression formula, $\hat{Y}$ (or discriminant function, Z). A summary of calculation procedures for finding such quantities and testing significance of various propositions follows.

The sum of squares of deviations in Y, called $\sum y^2$, is calculated in the usual manner for a sum of squares in cell $X_5 = Y$ of column $X_5 = Y$ in Table 1. The coded value of $\sum y^2$ is a_{55} or g_g.

The reduction in sum of squares of deviations in Y attributable to regression may be calculated as either

$$\sum \hat{y}^2 = \sum_i A_{ig}B_{ig} \qquad [5]$$

or

$$\sum \hat{y}^2 = \sum_i b_i g_i \qquad [9]$$

and, similarly, the value of the discriminant may be calculated by either

$$D = \sum_i A_{ig}B_{ig} \qquad [23]$$

or

$$D = \sum_i \lambda_i g_i \qquad [24]$$

where i specifies a particular independent variable, X_i, and the other symbols are used in the same way as in Tables 1 and 2.

The regression sum of squares may be decoded as

$$\sum \hat{y}^{2*} = \sum \hat{y}^2 \left(\frac{1}{m_y}\right)^2, \qquad [12]$$

where m_y is the power of 10 used for Y code in Table 1, and the superscript asterisk denotes a decoded value. Since the units of D are perfectly arbitrary, there is no need to decode D.

The total sum of squares due to regression on any number of variables may be divided into (1) that due to X_1 alone or

$$\sum \hat{y}_1^2 = A_{1g}B_{1g},$$

(2) that due to X_2 independent of X_1 or

$$\sum \hat{y}_{2.1}^2 = A_{2g}B_{2g},$$

(3) that due to X_3 independent of X_1, and X_2 or

$$\sum \hat{y}_{3.12}^2 = A_{3g}B_{3g},$$

(4) etc. Each of these values may be decoded by multiplying by $(1/m_y)^2$.

The sum of squares due to adding or deleting any variable to or from a set may be calculated as

$$\sum \hat{y}_{i.\text{all others}}^2 = \frac{b_i^2}{c_{ii}}. \qquad [15]$$

This may be decoded by multiplying by $(1/m_y)^2$.

The proportion of $\sum y^2$ due to regression is

$$R^2 = \frac{\sum \hat{y}^2}{g_g} = \frac{\text{decoded} \sum \hat{y}^2}{\sum y^2}; \qquad [6]$$

none of these need to be decoded.

The multiple correlation coefficient,

$$R = \sqrt{\frac{\sum \hat{y}^2}{g_g}} = \sqrt{\frac{\text{decoded} \sum \hat{y}^2}{\sum y^2}},$$

does not need to be decoded.

In general, the sum of squares of deviations from regression (residual sum of squares, error sum of squares, or sum of squares of deviations in Y independent of X_1, X_2, X_3, $\cdots$, X_r) is

$$\sum d_{Y.123\ldots r}^2 = g_g - \sum \hat{y}^2 = A_{gg}.$$

This may be decoded as

$$\sum d_{Y.123\ldots r}^{2*} = \sum d_{Y.123\ldots r}^2 \left(\frac{1}{m_y}\right)^2,$$

where d is a deviation, $Y - \hat{Y}$, and r is the number of independent variables.

The mean square deviation from regression, or sample variance, or error in Y independent of X_1, X_2, X_3, $\cdots$, X_r is

$$s_{Y.123\ldots r}^2 = \frac{\sum d_{Y.123\ldots r}^2}{\text{degrees of freedom}} = \frac{\sum d_{Y.123\ldots r}^2}{n - (r + 1)} = \frac{A_{gg}}{n - (r + 1)}$$

$$= \frac{\sum y^2 - \sum \hat{y}^2}{n - (r + 1)} = \frac{\sum y^2 - \sum\limits_i A_{ig}B_{ig}}{n - (r + 1)}.$$

This may be decoded as

$$s_{Y.123\ldots r}^{2*} = s_{Y.123\ldots r}^2 \left(\frac{1}{m_y}\right)^2.$$

The regression coefficients, b_i, more properly called the partial regression coefficients, $b_{Yi\cdot other\ X's}$, and also the discriminant coefficients, λ_i, of discriminant analysis are calculated in two different ways in Table 2 (column $X_5 = Y$, and bottom section). The values of b_i and λ_i may be decoded as

$$b_i^* = b_i \frac{m_i}{m_y}, \qquad [10]$$

and

$$\lambda_i^* = \lambda_i \frac{m_i}{m_d} \qquad [29]$$

where the m-values are the powers of 10 used for coding. Values of λ_i may be further coded by dividing each λ_i by the smallest λ_i, or more meaningfully, perhaps, by the λ of the most potent variable.

The *standard* partial regression coefficient is

$$b_i\sqrt{\frac{a_{ij}}{g_g}},$$

where $j = i$. This does not need to be decoded.

The regression function or estimated value of Y for any particular set of X_i values is

$$\hat{Y} = \bar{Y} + \sum_i b_i x_i, \qquad [7]$$

where x_i is the deviation of some specified X, say i, from its mean, $\bar{X}$. This equation may be rewritten as

$$\hat{Y} = \bar{Y} - \sum_i b_i\bar{X}_i + b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_rX_r, \qquad [8]$$

where $\hat{Y}$ is the estimated value, $\bar{Y}$ is the mean of Y, $\bar{X}_i$ is the mean of the ith X, and X_1, X_2, X_3, $\cdots$, X_r are the particular set of X's of interest. This result may be decoded as

$$\hat{Y}* = \hat{Y}\frac{1}{m_y}.$$

The discriminant function is

$$Z = \lambda_1X_1 + \lambda_2X_2 + \lambda_3X_3 + \cdots = \lambda_rX_r, \qquad [25]$$

where either the λ_i must be decoded or the X_i must be coded.

The Gauss multipliers, c-values, or elements of the inverse matrix of sums of squares and products, called c_{ij}, are calculated in the lower portion of Table 2. These may be decoded as

$$c_{ij}^{**} = c_{ij}m_im_j. \qquad [11]$$

The estimated variance of $\bar{Y}$ is given by

$$s^2_{Y.123\ldots r} = \frac{s^2_{Y.123\ldots r}}{n}.$$

This may be decoded as

$$s^{2*}_{Y.123\ldots r} = s^2_{Y.123\ldots r}\left(\frac{1}{m_y}\right)^2.$$

The estimated variance of any b_i is given by

$$s^2_{b_i} = s^2_{Y.123\ldots r}c_{ij},$$

where $i = j$. This may be decoded as

$$s_{b_i}^{2*} = s_{b_i}^2\left(\frac{m_i}{m_y}\right)^2.$$

The estimated variance of a quantity, such as

$$\hat{Y} = \bar{Y} + b_i x_i,$$

if the x_i is a population characteristic, is the sum of the variances of the two terms $\bar{Y}$ and $b_i x_i$. These values are

$$s_{\bar{Y}.123\ldots r}^2 + s_{b_i x_i}^2 = \frac{s_{Y.123\ldots r}^2}{n} + s_{Y.123\ldots r}^2 c_{ii} x_i^2$$

$$= s_{Y.123\ldots r}^2\left(\frac{1}{n} + c_{ii} x_i^2\right).$$

This may be decoded as

$$s_{\hat{Y}}^{2*} = s_{Y.123\ldots r}^2\left(\frac{1}{n} + c_{ii} x_i^2 m_i^2\right)\left(\frac{1}{m_y}\right)^2.$$

The inverse of a matrix with just one independent variable is $c_{11} = 1/\sum x^2$ so that the estimated variance of $\bar{Y} + bx$ is given by

$$s_{Y.X}^2\left(\frac{1}{n} + \frac{x^2}{\sum x^2}\right),$$

which is the formula often given for the variance of an estimated Y in simple regression.

The estimated variance of some predicted value, such as $\hat{Y} = \bar{Y} + b_i x_i$, where x_i is not a population characteristic but is the result of a *single* observation, will have all the variability of the estimating procedure as just outlined for the case of a population x_i plus the variability of an individual Y value, $s_{Y.123\ldots r}^2$. Thus the variance of the prediction for a single individual is given by

$$s_{Y.123\ldots r}^2\left(1 + \frac{1}{n} + c_{ii} x_i^2\right),$$

which may be decoded as

$$s_{\hat{Y}}^{2*} = s_{Y.123\ldots r}^2\left(1 + \frac{1}{n} + c_{ii} x_i^2 m_i^2\right)\left(\frac{1}{m_y}\right)^2.$$

The estimated variance of a quantity, such as $\hat{Y} = \bar{Y} + \sum_i b_i x_i$ if the x_i are population characteristics, is the sum of the variances of the two terms, $\bar{Y}$ and $\sum_i b_i x_i$. The term $\sum_i b_i x_i$ is itself a linear combination of b's with their coefficients. The variance of any linear combination of b's with their coefficients as $b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_rx_r$ is given as

$$\begin{array}{llllll}
s^2_{Y.123}\{x_1^2c_{11} & + \; 2x_1x_2c_{12} & + \; 2x_1x_3c_{13} & + \; \cdots & + \; 2x_1x_rc_{1r} & \\
+ & \;\; x_2^2c_{22} & + \; 2x_2x_3c_{23} & + \; \cdots & + \; 2x_2x_rc_{2r} & \\
 & + & \;\; x_3^2c_{33} & + \; \cdots & + \; 2x_3x_rc_{3r} & \\
 & & & & \;\cdot \quad \cdot & \\
 & & & & \;\cdot \quad \cdot & \\
 & & & & \;\cdot \quad \cdot & \\
 & & & & + \quad x_r^2c_{rr}\}. &
\end{array}$$

One must be careful to observe all signs in the above. Decoding would be a matter of decoding each term.

From the foregoing discussion, the variance of a multiple regression estimate,

$$\hat{Y} = \bar{Y} + \sum_i b_ix_i,$$

is

$$(s^2_{Y.123\ldots r})\left(\frac{1}{n}\right) + \text{the variance of the linear combination of } b\text{'s.}$$

The variance of a prediction based on a single individual is

$$(s^2_{Y.123r})\left(1 + \frac{1}{n}\right) + \text{ the variance of the linear combination of } b\text{'s.}$$

All of the foregoing variances may be used in the customary manner of using variances to establish confidence limits within which it is believed that the true or population value lies. Thus, population parameter = sample estimate ± $(t_\alpha)(\sqrt{\text{variance of the estimate}})$, where t_α is Student's t at the α probability level of chance occurrence with the degrees of freedom of the error or residual mean square from which the variance was estimated.

If it is desired to test a hypothesis rather than establish a confidence interval, then calculate the test statistic

$$t = \frac{\text{estimate} - \text{hypothetical value.}}{\sqrt{\text{variance of the estimate}}}$$

Compare this calculated t-value with the distribution of Student's t in standard t tables, using the degrees of freedom of the mean square from which the variance estimate was obtained. The interpretation of such t-values is made in the same manner as for a t calculated for any other statistic.

Some texts give directions for calculating matrices of partial regression coefficients and partial correlation coefficients. Most of these texts start by coding the sums of squares and products by dividing each $\sum x_i x_j$ by $\sqrt{\sum x_i^2}\ \sqrt{\sum x_j^2}$, thus yielding a matrix of simple correlation coefficients, then solving for the matrix of c_{ij} values inverse to this coefficient matrix. The matrix of c_{ij} in this bulletin may be converted to exactly this matrix, since any c_{ij} of the matrix of correlation coefficients is given by

$$c_{ij}^{**} = c_{ij} m_i m_j \sqrt{\sum x_i^2}\ \sqrt{\sum x_j^2}. \qquad [13]$$

The mean square for error from such a matrix of correlation coefficients can be computed from the results of this analysis as $= s^2_{Y.123\ldots r}/g_g$.

LITERATURE CITED

(*1*) ANDERSON, R. L. AND BANCROFT, T. A. Statistical Theory in Research. McGraw-Hill Book Co., Inc., N. Y. 1952.

(*2*) BENNETT, C. A. AND FRANKLIN, N. L. Statistical Analysis in Chemistry and the Chemical Industry. John Wiley and Sons, Inc., N. Y. 1954.

(3) COX, G. M. AND MARTIN, W. P. Use of a Discriminant Function for Differentiating Soil with Different Azotabacter Population. Iowa State College Jour. Sci. VII: 323–31. 1937.

(*4*) CROXTON, F. E. AND COWDEN, D. J. Applied General Statistics. Prentice-Hall, Inc., Englewood Cliffs, N. J. 1955.

(*5*) DWYER, P. S. The Solution of Simultaneous Equations. Psychometrika 6: 101–129. 1941.

(*6*) ————. Linear Computations. John Wiley and Sons, Inc., N. Y. 1951.

(*7*) EZEKIEL, MORDECAI. Methods of Correlation Analysis, Revised. John Wiley and Sons, Inc., N. Y.1941.

(*8*) FISHER, R. A. Statistical Methods for Research Workers, 11th Ed. Oliver and Boyd, Edinburgh, Scotland. 1950.

(*9*) FRIEDMAN, JOAN AND FOOTE, R. J. Computational Methods for Handling Systems of Simultaneous Equations with Applications to Agriculture. U. S. Dept. Agr. Agricultural Handbook 94. 1955.

(*10*) GOGGANS, JAMES F. AND SCHULTZ, E. FRED, JR. Growth of Pine Plantations in Alabama's Coastal Plain. Ala. Agr. Expt. Sta. Bul. 313. 1958.

(*11*) GOULDEN, C. H. Methods of Statistical Analysis, 2nd Ed. John Wiley and Sons, Inc. N. Y. 1952.

(*12*) HENDRICKS, W. A. The Theory of Sampling with Special Reference to the Collection and Interpretation of Agricultural Statistics. N. C. Inst. Statis. Mimeo. Series No. 1. 1942.

(*13*) KRAMER, CLYDE Y. Simplified Computations for Multiple Regression. Ind. Quality Control Vol. 13. Feb. 1957.

(*13a*) LEONE, FRED C. Statistical Programs for High Speed Computers. Technometrics Vol. 3:123. 1961.

(*14*) MATHER, K. Statistical Analysis in Biology. Interscience Publishers, Inc., N. Y. 1946.

(15) MILLS, F. S. Statistical Methods. Henry Holt and Co., N. Y. 1955.

(*16*) PEACH, PAUL. Curve Fitting and Analysis of Variance. N. C. Inst. Statis. Mimeo. Series No. 3. 1947ca.

(*17*) QUENOUILLE, M. H. Associated Measurements. Academic Press, Inc., N. Y. 1952.

(*18*) RAO, C. R. Advanced Statistical Methods in Biometric Research. John Wiley and Sons, Inc., N. Y. 1952.

(*19*) SNEDECOR, G. W. Statistical Methods, 5th Ed. The Iowa State University Press, Ames, Ia. 1956.

(*20*) TIPPETT, L. H. C. Technological Application of Statistics. John Wiley and Sons, Inc., N. Y. 1950.

(*21*) ————. The Methods of Statistics, 4th Ed. John Wiley and Sons, Inc., N. Y. 1952.

(*22*) VILLARS, D. S. Statistical Design and Analysis of Experiments for Development Research. Wm. C. Brown Co., Dubuque, Ia. 1951.

(*23*) WERT, J. E., NEIDT, C. O., AND AHMANN, I. S. Statistical Methods in Education and Psychological Research. Appleton-Century-Crofts, Inc., N. Y. 1954.

TABLE 1. CALCULATING, CHECKING AND CODING THE SUMS OF SQUARES AND PRODUCTS OF DEVIATIONS FOR MULTIPLE REGRESSION WITH FOUR INDEPENDENT VARIABLES AND ONE DEPENDENT VARIABLE

			X_j					
X_i	Item	X_i Code	X_1	X_2	X_3	X_4	$X_5 = Y$	$X_6 =$ Check $\sum$
	$\sum X_j$		$\sum X_1$	$\sum X_2$	$\sum X_3$	$\sum X_4$	$\sum X_5 = \sum Y$	$\sum X_6$
	$\bar{X}_j$		$\bar{X}_1$	$\bar{X}_2$	$\bar{X}_3$	$\bar{X}_4$	$\bar{X}_5 = \bar{Y}$	. . .
	X_j Code		m_1	m_2	m_3	m_4	$m_5 = m_y$	. . .
X_1	$\sum X_1X_j$		$\sum X_1^2$	$\sum X_1X_2$	$\sum X_1X_3$	$\sum X_1X_4$	$\sum X_1X_5 = \sum X_1Y$	$\sum X_1X_6$
	$C_{1j} = (\sum X_1)(\sum X_j)/n$		C_{11}	C_{12}	C_{13}	C_{14}	$C_{15} = C_{1y}$	C_{16}
	$\sum x_1x_j = \sum X_1X_j - C_{1j}$	m_1	$\sum x_1^2$	$\sum x_1x_2$	$\sum x_1x_3$	$\sum x_1x_4$	$\sum x_1x_5 = \sum x_1y$	$\sum x_1x_6$
	$a_{1j} = (\sum x_1x_j)(m_1m_j)$		a_{11}	a_{12}	a_{13}	a_{14}	$a_{15} = g_1$	. . .
X_2	$\sum X_2X_j$			$\sum X_2^2$	$\sum X_2X_3$	$\sum X_2X_4$	$\sum X_2X_5 = \sum X_2Y$	$\sum X_2X_6$
	$C_{2j} = (\sum X_2)(\sum X_j)/n$			C_{22}	C_{23}	C_{24}	$C_{25} = C_{2y}$	C_{26}
	$\sum x_2x_j = \sum X_2X_j - C_{2j}$	m_2		$\sum x_2^2$	$\sum x_2x_3$	$\sum x_2x_4$	$\sum x_2x_5 = \sum x_2y$	$\sum x_2x_6$
	$a_{2j} = (\sum x_2x_j)(m_2m_j)$			a_{22}	a_{23}	a_{24}	$a_{25} = g_2$	. . .

X_3	$\sum X_3X_j$		$\sum X_3^2$	$\sum X_3X_4$	$\sum X_3X_5 = \sum X_3Y$	$\sum X_3X_6$
	$C_{3j} = (\sum X_3)(\sum X_j)/n$		C_{33}	C_{34}	$C_{35} = C_{3y}$	C_{36}
	$\sum x_3x_j = \sum X_3X_j - C_{3j}$	m_3	$\sum x_3^2$	$\sum x_3x_4$	$\sum x_3x_5 = \sum x_3y$	$\sum x_3x_6$
	$a_{3j} = (\sum x_3x_j)(m_3m_j)$		a_{33}	a_{34}	$a_{35} = g_3$	. . .
X_4	$\sum X_4X_j$			$\sum X_4^2$	$\sum X_4X_5 = \sum X_4Y$	$\sum X_4X_6$
	$C_{4j} = (\sum X_4)(\sum X_j)/n$			C_{44}	$C_{45} = C_{4y}$	C_{46}
	$\sum x_4x_j = \sum X_4X_j - C_{4j}$	m_4		$\sum x_4^2$	$\sum x_4x_5 = \sum x_4y$	$\sum x_4x_6$
	$a_{4j} = (\sum x_4x_j)(m_4m_j)$			a_{44}	$a_{45} = g_4$	. . .
$X_5 = Y$	$\sum X_5X_j$				$\sum X_5^2 = \sum Y^2$	$\sum X_5X_6$
	$C_{5j} = (\sum X_5)(\sum X_j)/n$				$C_{55} = C_{yy}$	C_{56}
	$\sum x_5x_j = \sum X_5X_j - C_{5j}$	$m_5 = m_y$			$\sum x_5^2 = \sum y^2$	$\sum x_5x_6$
	$a_{5j} = (\sum x_5x_j)(m_5m_j)$				$a_{55} = g_g$	. . .

TABLE 2. ABBREVIATED DOOLITTLE SOLUTION OF THE SIMULTANEOUS EQUATIONS FOR A MULTIPLE REGRESSION WITH FOUR INDEPENDENT VARIABLES AND ONE DEPENDENT VARIABLE

Row	Column					
	X_1	X_2	X_3	X_4	$X_5 = Y$	$X_6 = \text{Check} \sum$
X_1	a_{11}	a_{12}	a_{13}	a_{14}	$a_{15}=g_1$	$h_1=a_{15}+a_{14}+a_{13}+a_{12}+a_{11}$
X_2		a_{22}	a_{23}	a_{24}	$a_{25}=g_2$	$h_2=a_{25}+a_{24}+a_{23}+a_{22}+a_{12}$
X_3			a_{33}	a_{34}	$a_{35}=g_3$	$h_3=a_{35}+a_{34}+a_{33}+a_{23}+a_{13}$
X_4				a_{44}	$a_{45}=g_4$	$h_4=a_{45}+a_{44}+a_{34}+a_{24}+a_{14}$
$X_5=Y$					$a_{55}=g_g$	$h_5=g_g+g_4+g_3+g_2+g_1$
A_{1j}	$A_{11}=a_{11}$	$A_{12}=a_{12}$	$A_{13}=a_{13}$	$A_{14}=a_{14}$	$A_{1g}=g_1$	$A_{1h}=h_1$
B_{1j}	$B_{11}=A_{11}/A_{11}=1.0$	$B_{12}=A_{12}/A_{11}$	$B_{13}=A_{13}/A_{11}$	$B_{14}=A_{14}/A_{11}$	$B_{1g}=A_{1g}/A_{11}$	$B_{1h}=A_{1h}/A_{11}$
A_{2j}		$A_{22}=a_{22}-A_{12}B_{12}$	$A_{23}=a_{23}-A_{12}B_{13}$	$A_{24}=a_{24}-A_{12}B_{14}$	$A_{2g}=g_2-A_{12}B_{1g}$	$A_{2h}=h_2-A_{12}B_{1h}$
B_{2j}		$B_{22}=A_{22}/A_{22}=1.0$	$B_{23}=A_{23}/A_{22}$	$B_{24}=A_{24}/A_{22}$	$B_{2g}=A_{2g}/A_{22}$	$B_{2h}=A_{2h}/A_{22}$
A_{3j}			$A_{33}=a_{33}-A_{13}B_{13}-A_{23}B_{23}$	$A_{34}=a_{34}-A_{13}B_{14}-A_{23}B_{24}$	$A_{3g}=g_3-A_{13}B_{1g}-A_{23}B_{2g}$	$A_{3h}=h_3-A_{13}B_{1h}-A_{23}B_{2h}$
B_{3j}			$B_{33}=A_{33}/A_{33}=1.0$	$B_{34}=A_{34}/A_{33}$	$B_{3g}=A_{3g}/A_{33}$	$B_{3h}=A_{3h}/A_{33}$

A_{4j}				$A_{44}=a_{44}-A_{14}B_{14}$ $-A_{24}B_{24}-A_{34}B_{34}$	$A_{4g}=g_4-A_{14}B_{1g}$ $-A_{24}B_{2g}-A_{34}B_{3g}$	$A_{4h}=h_4-A_{14}B_{1h}$ $-A_{24}B_{2h}-A_{34}B_{3h}$
B_{4j}				$B_{44}=A_{44}/A_{44}=1.0$	$B_{4g}=A_{4g}/A_{44}$	$B_{4h}=A_{4h}/A_{44}$
A_{5j}					$A_{5g}=A_{gg}=g_g-A_{1g}B_{1g}$ $-A_{2g}B_{2g}-A_{3g}B_{3g}$ $-A_{4g}B_{4g}$	$A_{5h}=h_5-A_{1g}B_{1h}$ $-A_{2g}B_{2h}-A_{3g}B_{3h}$ $-A_{4g}B_{4h}$
B_{5j}					$B_{5g}=B_{gg}=A_{gg}/A_{gg}=1.0$	$B_{5h}=A_{5h}/A_{gg}=1.0$
c_{1j}	$c_{11}=1/A_{11}-B_{14}c_{14}$ $-B_{13}c_{13}-B_{12}c_{12}$	$c_{12}=-B_{14}c_{24}-B_{13}c_{23}$ $-B_{12}c_{22}$	$c_{13}=-B_{14}c_{34}-B_{13}c_{33}$ $-B_{12}c_{23}$	$c_{14}=-B_{14}c_{44}-B_{13}c_{34}$ $-B_{12}c_{24}$	$b_1=B_{1g}-b_4B_{14}$ $-b_3B_{13}-b_2B_{12}$	$h_{1b}=B_{1h}-h_{4b}B_{14}$ $-h_{3b}B_{13}-h_{2b}B_{12}$
c_{2j}		$c_{22}=1/A_{22}$ $-B_{24}c_{24}-B_{23}c_{23}$	$c_{23}=-B_{24}c_{34}-B_{23}c_{33}$	$c_{24}=-B_{24}c_{44}-B_{23}c_{34}$	$b_2=B_{2g}-b_4B_{24}-b_3B_{23}$	$h_{2b}=B_{2h}-h_{4b}B_{24}$ $-h_{3b}B_{23}$
c_{3j}			$c_{33}=1/A_{33}-B_{34}c_{34}$	$c_{34}=-B_{34}c_{44}$	$b_3=B_{3g}-b_4B_{34}$	$h_{3b}=B_{3h}-h_{4b}B_{34}$
c_{4j}				$c_{44}=1/A_{44}$	$b_4=B_{4g}$	$h_{4b}=B_{4h}$
b_j	$b_1=g_4c_{14}+g_3c_{13}$ $+g_2c_{12}+g_1c_{11}$	$b_2=g_4c_{24}+g_3c_{23}$ $+g_2c_{22}+g_1c_{12}$	$b_3=g_4c_{34}+g_3c_{33}$ $+g_2c_{23}+g_1c_{13}$	$b_4=g_4c_{44}+g_3c_{34}$ $+g_2c_{24}+g_1c_{14}$	Sum of squares due to regression $=\sum \hat{y}^2=\sum_i b_ig_i$ or $\sum_i A_{ig}B_{ig}$	
Check	$c_{14}a_{14}+c_{13}a_{13}$ $+c_{12}a_{12}+c_{11}a_{11}=1.0$	$c_{24}a_{24}+c_{23}a_{23}$ $+c_{22}a_{22}+c_{12}a_{12}=1.0$	$c_{34}a_{34}+c_{33}a_{33}$ $+c_{23}a_{23}+c_{13}a_{13}=1.0$	$c_{44}a_{44}+c_{34}a_{34}$ $+c_{24}a_{24}+c_{14}a_{14}=1.0$	Any predicted $Y=\hat{Y}=\bar{Y}$ $-\sum_i b_i\bar{X}_i+b_1X_1+b_2X_2+b_3X_3+b_4X_4$	

TABLE 3. FORM FOR CALCULATING SUMS OF SQUARES AND PRODUCTS REQUIRED IN SEARCH FOR POTENT VARIABLES IF THERE ARE 10 INDEPENDENT VARIABLES

X_i	Item	X_i Code	X_j											
			X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	$X_{11} = Y$	$X_{12} =$ Check $\sum$
	$\sum X_j$		(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)
	$\bar{X}_j$		(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	
	X_j Code		m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}	$m_{11} = m_y$	
X_1	$\sum X_1X_j$													
	$C_{1j} = (\sum X_1)(\sum X_j)/n$													
	$\sum x_1x_j = \sum X_1X_j - C_{1j}$	m_1	(1)		(3)			(2)	(4)				(1)	
	$a_{1j} = (\sum x_1x_j)(m_1m_j)$													
X_2	$\sum X_2X_j$													
	$C_{2j} = (\sum X_2)(\sum X_j)/n$													
	$\sum x_2x_j = \sum X_2X_j - C_{2j}$	m_2		(1)	(3)			(2)	(4)				(1)	
	$a_{2j} = (\sum x_2x_j)(m_2m_j)$													
X_3	Same 4 items as for X_1 and X_2	m_3			(1)	(3)	(3)	(2)	(3)	(3)	(3)	(3)	(1)	(3)
X_4	“	m_4				(1)		(2)	(4)				(1)	
X_5	“	m_5					(1)	(2)	(4)				(1)	
X_6	“	m_6						(1)	(2)	(2)	(2)	(2)	(1)	(2)
X_7	“	m_7							(1)	(4)	(4)	(4)	(1)	(4)
X_8	“	m_8								(1)			(1)	
X_9	“	m_9									(1)		(1)	
X_{10}	“	m_{10}										(1)	(1)	
$X_{11} = Y$	“	m_{11}											(1)	(1)

TABLE 4. ABBREVIATED DOOLITTLE SOLUTION SHOWING MASK AND PARTS THAT ARE IDENTICAL IN EACH OF THE EIGHT REGRESSIONS OF Y ON THREE INDEPENDENT VARIABLES (OTHER PARTS VARY AS X'_3 VARIES)[a]

Row	Column		$X'_3 = ?$		$X'_5 =$ Check $\sum$
	$X'_1 = X_6$	$X'_2 = X_3$		$X'_4 = Y$	
$X'_1 = X_6$	$a'_{11} = a_{66}$	$a'_{12} = a_{63}$	..	$a'_{14} = g'_1 = g_6$	..
$X'_2 = X_3$		$a'_{22} = a_{33}$	..	$a'_{24} = g'_2 = g_3$	..
$X'_3 = ?$			..	..	..
A_{1j}	A_{11}	A_{12}	..	A_{1g}	..
B_{1j}	1.0	B_{12}	..	B_{1g}	
A_{2j}		A_{22}		A_{2g}	..
B_{2j}		1.0	..	B_{2g}	..
A_{3j}			..	..	
B_{3j}			..	..	..
A_{4j}				..	..
B_{4j}				..	..

[a] "Primes" indicate that the subscripts of this table are not the original subscripts of Table 3.

TABLE 5. F TESTS OF SIGNIFICANCE OF ADDITIONAL VARIABLES IN MULTIPLE REGRESSION[a]

Source of variation	Degrees of freedom	Sum of squares[b]	Mean square	F	Chance probability
(1) Total	$n - 1$[c]	$\sum y^2$			
(2) Reduction due to most potent variable	1	$\sum \hat{y}^2$		↕	
(3) Residual = (1) − (2)	$n - 1 - 1$				
(4) Reduction due to most potent pair of variables	2	$\sum \hat{y}^2$			
(5) Second variable independent of first = (4) − (2)	1			↕	
(6) Residual = (1) − (4)	$n - 1 - 2$				
(7) Reduction due to most potent trio of variables	3	$\sum \hat{y}^2$			
(8) Third variable independent of first two = (7) − (4)	1			↕	
(9) Residual = (1) − (7)	$n - 1 - 3$				

[a] The testing process can be extended to as many variables as desired.

[b] All sums of squares are sums of squares of deviations in Y and must all be either coded or decoded, but not mixed. It would seem that decoded values would be better, since this would allow for changes in code as the problem proceeds, if desirable.

[c] n = the number of sets of observations of X_i and Y.

Table 6. Average Heights, Y, of Dominant Trees Together With Measurements of Six Possible Predictors, X_j, of Height for 40 Plantings of Longleaf Pine

Planting no.	X_1[a]	X_2[b]	X_3[c]	X_4[d]	$X_5=X_4^2$	$X_6=X_1X_4$	$X_7=Y$	$X_8=$ Check $\sum$
1	16.8	10.6	31.5	11	121	184.8	32.2	407.9
2	30.4	17.5	63.3	7	49	212.8	26.0	406.0
3	14.6	5.0	29.0	11	121	160.6	29.5	370.7
4	42.8	19.9	69.2	11	121	470.8	30.6	765.3
5	19.7	3.7	30.9	12	144	236.4	32.6	479.3
6	15.1	5.7	35.4	11	121	166.1	27.9	382.2
7	18.3	5.8	28.4	10	100	183.0	29.0	374.5
8	11.7	5.6	28.2	7	49	81.9	15.9	199.3
9	13.5	2.6	21.3	11	121	148.5	24.4	342.3
10	8.5	9.1	33.5	10	100	85.0	31.3	277.4
11	7.8	4.1	16.9	10	100	78.0	31.6	248.4
12	9.5	2.3	21.0	11	121	104.5	26.6	295.9
13	14.6	3.6	27.8	11	121	160.6	26.3	364.9
14	14.6	6.5	37.6	7	49	102.2	14.9	231.8
15	13.6	3.7	25.1	11	121	149.6	24.0	348.0
16	15.8	4.5	31.4	6	36	94.8	16.0	204.5
17	19.0	4.8	28.1	12	144	228.0	28.0	463.9
18	23.8	6.1	29.5	12	144	285.6	33.5	534.5
19	36.2	6.6	51.6	12	144	434.4	34.9	719.7
20	28.0	6.1	43.2	12	144	336.0	33.8	603.1
21	19.3	5.3	48.4	14	196	270.2	40.3	593.5
22	26.2	5.8	34.8	14	196	366.8	42.6	686.2
23	15.9	2.8	23.8	13	169	206.7	31.5	462.7
24	13.5	3.0	25.3	10	100	135.0	31.1	317.9
25	13.2	2.4	26.3	8	64	105.6	20.2	239.7
26	18.4	2.9	37.6	8	64	147.2	20.5	298.6
27	20.6	4.4	34.5	15	225	309.0	41.5	650.0
28	32.0	5.8	44.6	12	144	384.0	31.0	653.4
29	21.2	7.0	38.0	11	121	233.2	24.7	456.1
30	23.8	3.5	26.9	15	225	357.0	29.7	680.9
31	29.3	5.5	33.7	15	225	439.5	33.2	781.2
32	26.4	4.2	33.4	12	144	316.8	28.0	564.8
33	22.8	3.4	29.7	11	121	250.8	27.0	465.7
34	34.0	5.9	45.7	12	144	408.0	30.9	680.5
35	19.5	5.3	33.6	12	144	234.0	26.5	474.9
36	18.2	5.2	52.5	12	144	218.4	31.8	482.1
37	27.3	4.1	36.6	12	144	327.6	34.9	586.5
38	24.7	5.7	57.8	12	144	296.4	30.5	571.1
39	19.8	15.7	54.0	8	64	158.4	18.0	337.9
40	20.4	3.4	24.0	6	36	122.4	16.3	228.5
$\sum$	820.8	235.1	1424.1	437.0	4985.0	9190.6	1139.2	18231.8
$\bar{X}$	20.52	5.88	35.60	10.92	124.62	229.76	28.48	.....

[a] X_1 = silt plus clay content of topsoil in per cent, [b] X_2 = imbibitional water value of the most impervious soil horizon, [c] X_3 = silt plus clay content of B horizon in per cent, [d] X_4 = age of planting in years.

Table 7. Calculation of Coded Sums of Squares and Products for Multiple Regression of Height of Longleaf Pine on Six Other Variables

X_i	Item	X_i Code	X_j							
			X_1	X_2	X_3	X_4	X_5	X_6	$X_7 = Y$	$X_8 =$ Check $\sum$
	$\sum X_j$		820.8	235.1	1,424.1	437.0	4,985.0	9,190.6	1,139.2	18,231.8
	$\bar{X}_j$		20.5	5.9	35.6	10.9	124.6	229.8	28.5	. . .
	X_j Code		0.01	0.1	0.01	0.1	0.01	0.001	0.1	. . .
X_1	$\sum X_1X_j$		19,205.06	5,360.01		9,190.6			24,070.97	
	C_{1j}		16,842.82	4,824.25		8,967.2			23,376.38	
	$\sum x_1x_j$	0.01	2,362.24	535.76		223.4			694.59	
	a_{1j}		0.236,24	0.535,76		0.223,4			0.694,59	
X_2	$\sum X_2X_j$			1,956.37	9,677.79	2,499.2	27,806.4	57,903.39	6,680.69	111,883.85
	C_{2j}			1,381.80	8,370.15	2,568.5	29,299.4	54,017.75	6,695.65	107,157.40
	$\sum x_2x_j$	0.1		574.57	1,307.64	−69.3	−1,493.0	3,885.64	−14.96	4,726.45
	a_{2j}			5.745,7	1.307,64	−0.693	−1.493,0	0.388,564	−0.149,6	. . .
X_3	$\sum X_3X_j$				56,224.37	15,579.0			41,048.12	
	C_{3j}				50,701.52	15,558.3			40,558.37	
	$\sum x_3x_j$	0.01			5,522.85	20.7			489.75	
	a_{3j}				0.552,285	0.020,7			0.489,75	
X_4	$\sum X_4X_j$					4,985.	58,853.	107,160.4	12,949.1	211,216.30
	C_{4j}					4,774.	54,461.	100,407.3	12,445.8	199,182.42
	$\sum x_4x_j$	0.1				211.	4,392.	6,753.1	503.3	12,033.88
	a_{4j}					2.11	4.392	0.675,31	5.033	. . .
X_5	$\sum X_5X_j$						714.593.		152,323.5	
	C_{5j}						621,256.		141,972.8	
	$\sum x_5x_j$	0.01					93,337.		10,350.7	
	a_{5j}						9.333,7		10.350,7	
X_6	$\sum X_6X_j$							2,586,263.62	279,200.63	
	C_{6j}							2,111,678.21	261,748.29	
	$\sum x_6x_j$	0.001						474,585.41	17,452.34	
	a_{6j}							0.474,585,41	1.745,234	
X_7	$\sum X_7X_j$								34,171.52	550,444.53
	C_{7j}								32,444.42	519,241.66
	$\sum x_7x_j$	0.1							1,727.10	31,202.87
	a_{7j}								17.271,0	. . .

TABLE 8. TESTS OF SIGNIFICANCE OF ADDITIONAL VARIABLES IN PREDICTING HEIGHT OF LONGLEAF PINE

Source of variation	Degrees of freedom	Sum of squares[a]	Mean square	*F*	Chance probability
(1) Total	39	1,727.10			
(2) Reduction due to age	1	1,200.53	1,200.53	86.62	<0.001
(3) Residual = (1) − (2)	38	526.57	13.86		
(4) Reduction due to age and imbibitional water value	2	1,241.49			
(5) Reduction due to imbibitional water value independent of age = (4) − (2)	1	40.96	40.96	3.12	<0.1
(6) Residual = (1) − (4)	37	485.61	13.12		
(7) Reduction due to age, imbibitional water value and (age)^2	3	1,248.02			
(8) Reduction due to (age)^2 independent of others = (7) − (4)	1	6.53	6.53	0.49	>0.3
(9) Residual = (1) − (7)	36	479.08	13.31		

[a] Decoded sums of squares.

TABLE 9. "FORWARD" PORTION OF THE ABBREVIATED DOOLITTLE SOLUTION FOR MULTIPLE REGRESSION OF HEIGHT ON $X_1' = X_4$, AGE, AND $X_2' = X_1$, SILT AND CLAY OF TOPSOIL.[a]

Row	$X_1' = X_4$	$X_2' = X_1$	$X_3' = Y$	$X_4' = \text{Check} \sum$
$X_1' = X_4$	2.110,000,00	0.223,400,00	5.033,000,00	7.366,400,00
$X_2' = X_1$		0.236,240,00	0.694,590,00	1.154,230,00
$X_3' = Y$			17.271,000,00	22.998,590,00
A_{1j}	2.110,000,00	0.223,400,00	5.033,000,00	7.366,400,00
B_{1j}	1.0	0.105,876,78	2.385,308,06	3.491,184,83
A_{2j}		0.212,587,13	0.161,712,18	0.374,299,31
B_{2j}		1.0	0.760,686,59	1.760,686,59
A_{3j}			5.142,732,25	5.142,732,28
B_{3j}			1.0	1.0

$$\sum \hat{y}^2 = \sum_i A_{ig}B_{ig} = (5.033{,}000{,}00)(2.385{,}308{,}06) + (0.161{,}712{,}18)(0.760{,}686{,}59)$$
$$= 12.005{,}255{,}47 + 0.123{,}012{,}29 = 12.128{,}267{,}76$$

$$\sum \hat{y}^{2*} = \left(\sum \hat{y}^2\right)\left(\frac{1}{Y \text{ code}}\right)^2 = (12.128{,}267{,}76)\left(\frac{1}{0.1}\right)^2 = 1{,}212.83$$

[a] "Primes" indicate that the subscripts of this table are not the original subscripts of Tables 6 and 7.

TABLE 10. REDUCTIONS IN SUM OF SQUARES OF Y DUE TO THE SPECIFIED TWO VARIABLES

X_1'	X_2'	$\sum \hat{y}^2$	Decoded $\sum y^2$
X_4	X_2	12.414,865,59	1,241.49
X_4	X_3	12.356,524,65	1,235.65
X_4	X_1	12.128,267,75	1,212.83
X_4	X_5	12.087,520,95	1,208.75
X_4	X_6	12.075,158,38	1,207.52

TABLE 11. MASK TO AID IN SOLUTIONS OF REGRESSIONS OF HEIGHT ON $X_1'' = X_4 =$ AGE, $X_2'' = X_2 =$ IMBIBITIONAL WATER VALUE AND $X_3'' = ?$[a]

Row	Column		$X_3'' = ?$		$X_5'' =$ Check $\sum$
	$X_1'' = X_4$	$X_2'' = X_2$		$X_4'' = Y$	
$X_1'' = X_4$	2.110,000,00	−0.693,000,00		5.033,000,00	
$X_2'' = X_2$		5.745,700,00		−0.149,600,00	
$X_3'' = ?$					
$X_4'' = Y$				17.271,000,00	
A_{1j}	2.110,000,00	−0.693,000,00		5.033,000,00	
B_{1j}	1.0	−0.328,436,02		2.385,308,06	
A_{2j}		5.518,093,84		1.503,418,48	
B_{2j}		1.0		0.272,452,50	

Computational note: Since $\sum \hat{y}^2 = \sum_i A_{ig}B_{ig}$ from the $X_4'' = Y$ column, each and every $\sum \hat{y}^2$ will have (5.033,000,00)(2.385,308,06) + (1.503,418,48)(0.272,452,50) as part of the result; this quantity may be calculated once and entered on the mask to be used as needed.

[a] Double "primes" indicate possible changes in subscripts beyond those denoted by "primes".

TABLE 12. "FORWARD" PORTION OF THE ABBREVIATED DOOLITTLE SOLUTION FOR MULTIPLE REGRESSION OF HEIGHT ON $X_1'' = X_4$, AGE; $X_2'' = X_2$, IMBIBITIONAL WATER VALUE; $X_3'' = X_1$, SILT & CLAY CONTENT OF TOPSOIL, SHOWING MASK IN POSITION[a]

Row	Column		$X_3'' = ?$		$X_5'' =$ Check $\sum$
	$X_1'' = X_4$	$X_2'' = X_2$	$X_3'' = X_1$	$X_4'' = Y$	
$X_1'' = X_4$	2.110,000,00	−0.693,000,00	0.223,400,00	5.033,000,00	6.673,400,00
$X_2'' = X_2$		5.745,700,00	0.535,760,00	−0.149,600,00	5.438,860,00
$X_3'' = ?$			0.236,240,00	0.694,590,00	1.689,990,00
$X_4'' = Y$				17.271,000,00	22.847,990,00
A_{1j}	2.110,000,00	−0.693,000,00	0.223,400,00	5.033,000,00	6.673,400,00
B_{1j}	1.0	−0.328,436,02	0.105,876,78	2.385,308,06	3.162,748,82
A_{2j}		5.518,093,84	0.609,132,61	1.503,418,48	7.630,644,93
B_{2j}		1.0	0.110,388,23	0.272,452,50	1.382,840,73
A_{3j}			0.145,346,06	−0.004,247,52	0.141,098,53
B_{3j}			1.0	−0.029,223,52	0.970,776,44
A_{4j}				4.856,010,28	4.856,010,28
B_{4j}				1.0	1.0

$$\sum \hat{y}^2 = \sum_i A_{ig}B_{ig} = (5.033{,}000{,}00)(2.385{,}308{,}06) + (1.503{,}418{,}48)(0.272{,}452{,}50) + (-0.004{,}247.52)(-0.029{,}223{,}52)$$

$$= 12.005{,}255{,}47 + 0.409{,}610{,}12 + 0.000{,}124{,}13 = 12.414{,}989{,}72$$

$$\sum \hat{y}^{2*} = (\sum \hat{y}^2)\left(\frac{1}{Y \text{ Code}}\right)^2 = (12.414{,}989{,}72)\left(\frac{1}{0.1}\right)^2 = 1241.50$$

[a] Double "primes" indicate changes in subscripts from original

TABLE 13. REDUCTION IN SUM OF SQUARES OF Y DUE TO THE SPECIFIED THREE VARIABLES

X_1''	X_2''	X_3''	$\sum \hat{y}^2$	Decoded $\sum \hat{y}^2$
X_4	X_2	X_5	12.480,246,12	1,248.02
X_4	X_2	X_3	12.443,175,56	1,244.32
X_4	X_2	X_6	12.420,189,26	1,242.02
X_4	X_2	X_1	12.414,989,72	1,241.50

TABLE 14. MEASUREMENTS OF 3 POSSIBLE DISCRIMINATORS OF THE PRESENCE OF *Azotobacter* IN SOILS (FROM GOULDEN FROM COX AND MARTIN)

	Group I (n_{I} = 25) containing Azotobacter[a]				Group II (n_{II} = 27) without Azotobacter[1]			
	X_1	X_2	X_3	$\sum$	X_1	X_2	X_3	$\sum$
	6.0	46	24	76.0	6.2	49	30	85.2
	7.0	35	17	59.0	5.6	31	23	59.6
	8.4	115	28	151.4	5.8	42	22	69.8
	5.8	35	17	57.8	5.7	42	14	61.7
	6.9	55	25	86.9	6.2	40	23	69.2
	7.8	52	29	88.8	6.4	49	18	73.4
	7.8	52	29	88.8	5,8	31	17	53.8
	6.9	208	58	272.9	6.4	31	19	56.4
	7.0	70	13	90.0	5.4	62	26	93.4
	6.7	35	16	57.7	5.4	42	16	63.4
	6.2	27	44	77.2	5.7	35	22	62.7
	6.9	52	27	85.9	5.6	33	24	62.6
	8.0	60	58	126.0	5.8	24	15	44.8
	8.0	156	68	232.0	7.3	70	14	91.3
	8.0	90	37	135.0	6.1	21	21	48.1
	6.1	44	27	77.1	6.2	36	26	68.2
	7.4	207	31	245.4	6.7	35	26	67.7
	7.4	120	32	159.4	5.9	33	21	59.9
	8.4	65	43	116.4	5.6	25	32	62.6
	8.1	237	45	290.1	5.8	31	30	66.8
	8.3	57	60	125.3	6.1	30	24	60.1
	7.0	94	43	144.0	6.1	21	25	52.1
	8.5	86	40	134.5	5.7	35	22	62.7
	8.4	52	48	108.4	5.8	37	24	66.8
	7.9	146	52	205.9	5.8	28	19	52.8
					5.7	34	20	59.7
					5.8	16	19	40.8
$\sum$	184.9	2196	911	3,291.9	160.6	963	592	1,715.6
$\bar{X}$	7.3960	87.8400	36.4400		5.9481	35.6667	21.9259	

[a] X_1 = pH, X_2 = available phosphate content, X_3 = total nitrogen content.

TABLE 15. CALCULATION OF CODED SUMS OF SQUARES AND PRODUCTS FOR DISCRIMINATION OF PRESENCE OF *Azotobacter* IN SOIL

X_i	Item	X_i code	X_j			
			X_1	X_2	X_3	Check $\sum$
			Group I, containing *Azotobacter*			
	$\sum X_j$		184.9	2196.	911.	3,291.9
	$\bar{X}_j$		7.3960	87.8400	36.4400	. .
X_1	$\sum X_1X_j$		1,384.05	16,620.8	6,892.8	24,897.65
	C_{1j}		1,367.5204	16,241.6160	6,737.7560	24,346.8924
	$\sum x_1x_j$		16.5296	379.1840	155.0440	550.7576
X_2	$\sum X_2X_j$			278,162.	89,344.	384,126.8
	C_{2j}			192,896.6400	80,022.2400	289,160.4960
	$\sum x_2x_j$			85,265.3600	9,321.7600	94,966.3040
X_3	$\sum X_3X_j$				38,701.	134,937.8
	C_{3j}				33,196.8400	119,956.8360
	$\sum x_3x_j$				5,504.1600	14,980.9640
			Group II, without *Azotobacter*			
	$\sum X_j$		160.6	963.	592.	1,715.6
	$\bar{X}_j$		5.9481	35.6667	21.9259	. . .
X_1	$\sum X_1X_j$		959.70	5,770.8	3,514.5	10,245.00
	C_{1j}		955.2726	5,728.0667	3,521.3037	10,204.6430
	$\sum x_1x_j$		4.4274	42.7333	−6.8037	40.3570
X_2	$\sum X_2X_j$			37,979.	20,928.	64,677.8
	C_{2j}			34,347.0000	21,114.6667	61,189.7333
	$\sum x_2x_j$			3,632.0000	−186.6667	3,488.0667
X_3	$\sum X_3X_j$				13,566.	38,008.5
	C_{3j}				12,980.1481	37,616.1185
	$\sum x_3x_j$				585.8519	392.3815
			Both groups			
	X_j		X_1	X_2	X_3	$X_4 = d$
			0.1	0.01	0.01	1.0
X_1	$\sum x_1x_j$	0.1	20.9570	421.9173	148.2403	1.4479
	a_{1j}		0.209,570,00	0.421,917,30	0.148,240,30	0.144,790,00
X_2	$\sum x_2x_j$	0.01		88,897.3600	9,135.0933	52.1733
	a_{2j}			8.889,736,00	0.913,509,33	0.521,733,00
X_3	$\sum x_3x_j$	0.01			6,090.0119	14.5141
	a_{3j}				0.609,001,19	0.145,141,00

Table 16. Abbreviated Doolittle Solution for Discriminant Function for Presence of *Azotobacter* Based on X_1, pH; X_2, Available Phosphate Content; and X_3, Total Nitrogen Content

Row	Column				
	X_1	X_2	X_3	$X_4 = d$	X_5 = Check $\sum$
X_1	0.209,570,00	0.421,917,30	0.148,240,30	0.144,790,00	0.924,517,60
X_2		8.889,736,00	0.913,509,33	0.521,733,00	10.746,895,63
X_3			0.609,001,19	0.145,141,00	1.815,891,82
A_{1j}	0.209,570,00	0.421,917,30	0.148,240,30	0.144,790,00	0.924,517,60
B_{1j}	1.0	2.013,252,37	0.707,354,58	0.690,890,87	4.411,497,83
A_{2j}		8.040,310,00	0.615,064,20	0.230,234,19	8.885,608,38
B_{2j}		1.0	0.076,497,57	0.028,634,99	1.105,132,57
A_{3j}			0.457,091,82	0.025,110,77	0.482,202,58
B_{3j}			1.0	0.054,935,95	1.054,935,92
c_{1j}	c_{11} = 5.945,651,91	c_{12} = −0.157,788,52	c_{13} = −1.210,578,80	λ_1 = 0.602,842,85	h_1 = 1.602,842,84
c_{2j}		c_{22} = 0.137,175,72	c_{23} = −0.167,357,12	λ_2 = 0.024,432,52	h_2 = 1.024,432,54
c_{3j}			c_{33} = 2.187,744,25	λ_3 = 0.054,935,95	h_3 = 1.054,935,92
λ_j	λ_1 = 0.602,842,84	λ_2 = 0.024,432,52	λ_3 = 0.054,935,95	$D = 0.100{,}034{,}09 + 0.006{,}592{,}75$ $+ 0.001{,}379{,}48$ $= 0.108{,}006{,}32$	
Check	1.000,000,00	0.999,999,94	1.000,000,00	$Z^{a} = 0.060{,}28X_1 + 0.000{,}244{,}3X_2$ $+ 0.000{,}549{,}4X_3$ $Z^{b} = 246.7X_1 + X_2 + 2.248X_3$	

[a] The λ_i must be decoded; or the X_i must be coded.

[b] Divide each λ_i by the smallest λ-value.

TABLE 17. TESTS OF SIGNIFICANCE OF ADDITIONAL VARIABLES IN THE DISCRIMINANT FUNCTION FOR PRESENCE OF *Azotobacter* IN SOILS

Source of variation	Degrees of freedom	Sum of squares[a]	Mean square	*F*	Chance probability
(1) Total = (2) + (3)	51	0.229,930			
(2) *SS* due to most potent variable, X_1 = pH	1	0.129,896	0.129,896	64.92	<0.001
(3) Residual = *D*	50	0.100,034	0.002,001		
(4) *SS* due to 2 most potent variables X_1 & X_2 = pH and phosphate, (adjusted to *SS* for X_1)	2	0.133,487			
(5) Phosphate independent of pH = (4) – (2)	1	0.003,591	0.003,591	1.82	>0.10
(6) Residual = (1) – (4)	49	0.096,443	0.001,968		
(7) *SS* due to all 3 variables (adjusted to *SS* for X_1)	3	0.134,206			
(8) Nitrogen independent of other variables = (7) – (4)	1	0.000,719	0.000,719	0.36	>0.5
(9) Residual = (1) – (7)	48	0.095,724	0.001,994		

[a] All sums of squares were adjusted so that every discriminant would have the same total sum of squares as the most potent single discriminator, pH.

TABLE 18. "FORWARD" PORTION OF THE ABBREVIATED DOOLITTLE SOLUTION FOR THE DISCRIMINANT BASED ON X'_1 = pH AND X'_2 = SOIL PHOSPHATE CONTENT[1]

Row	$X'_1 = X_1$	$X'_2 = X_2$	$X'_3 = d$	$X'_4 =$ Check $\sum$
$X'_1 = X_1$	0.209,570,00	0.421,917,30	0.144,790,00	0.776,277,30
$X'_2 = X_2$		8.889,736,00	0.521,733,00	9.833,386,30
$X'_3 = d$			0.145,141,00	0.811,664,00
A_{1j}	0.209,570,00	0.421,917,30	0.144,790,00	0.776,277,30
B_{1j}	1.0	2.013,252,37	0.690,890,87	3.704,143,25
A_{2j}		8.040,310,00	0.230,234,19	8.270,544,19
B_{2j}		1.0	0.028,634,99	1.028,634,99

$$D = \sum_i A_{ig}B_{ig} = (0.144{,}790{,}00)(0.690{,}890{,}87) + (0.230{,}234{,}19)(0.028{,}634{,}99)$$

$$= 0.100{,}034{,}09 + 0.006{,}592{,}75 = 0.106{,}626{,}84$$

[1] "Primes" indicate that the subscripts of this table may not be the original subscripts of Tables 14, 15, and 16.