NOTE ON COMPUTATION OF ORTHOGONAL PREDICTORS

By Bradford F. Kimball

New York State Department of Public Service

Summary. The present note calls attention to a simple algorithm for computation of the orthogonal matrix associated with the matrix of the normal equations of least squares. An application of the "forward" solution associated with the orthogonalization process is also pointed out.

1. Introduction. This note is supplementary to Dwyer's recent book *Linear Computations*, [1]. The author also acknowledges reference to extensive unpublished notes on matrix analysis of numerical methods kindly furnished him by A. S. Householder.

Orthogonalization of the predicting variables of the least square problem has been considered since the introduction of orthogonal polynomials into the least-squares problem by Tchebycheff 1853–73 and the significant doctoral dissertation of the Danish mathematician and actuary, J. P. Gram, in 1879, [3]. A discussion of the problem is also to be found in Poincaré's *Calcul des Probabilités*, 1912, Chapter XV. Various methods of simplifying the computations for purposes of mass application to statistical data have been studied since then. A complete bibliography is beyond the scope of this note.

The advantages of obtaining the solution of the least squares problem in terms of orthogonal predictors are numerous. Perhaps the most obvious are those associated with the resulting simplified expressions for the error formulae, correlations [2] and sampling variance of the fitted function [5]. Also the orthogonalization of the predicting variables is a starting point for the computation of principal components significant to structural relationships in psychology and economics. A further application of the associated "forward" solution is pointed out in connection with the lemma stated in Section 4 of the present paper.

Two solutions are obtained. One is in terms of a slight extension of the algorithm of the square root method of solving the normal equations [2], and the other is in terms of the algorithm of the Gauss-Doolittle method. The first method has the advantage that the coefficients of the orthogonal predictors have unit sampling variance weight. The second method is based on the more familiar Doolittle algorithm and does not involve square roots.

There seems to be some need for standardization of notation in matrix analysis. Since Householder has presented a consistent application of matrix analysis to a large body of material, we shall keep to his notation in the following respects. Capitals will be used for matrices and the transpose of a matrix will be indicated by the superscript T rather than by a prime. Single row or column matrices and vectors will be denoted by small letters.

Received 3/17/52.

2. Derivation of orthogonal matrix from matrix theory. Let

$$X = \| x_{1j}, x_{2j}, \cdots, x_{nj} \|, \qquad j = 1, 2, \cdots N, n \leqq N \tag{2.1}$$

denote the N rowed matrix of the N observations on the n predictor variables x_{ij}. With the subscript j suppressed, the observations on the n original predictor variables are treated as column vectors x_i, $i = 1$ to n.

Consider a matrix V of n rows and columns. The product matrix XV^T will consist of n columns, each of which is a linear transform of the n columns of the data matrix X. Hence, if the n linear functions defined in the n columns are to be orthonormal, the following matrix relation must obtain

$$(XV^T)^T(XV^T) = I, \tag{2.2}$$

where I denotes the identity matrix.

From general matrix theory $(XV^T)^T = VX^T$, and (2.2) can be written [4]

$$V(X^TX)V^T = I. \tag{2.3}$$

For purposes of setting up an algorithm for computation of V we introduce the triangular matrices of the Gauss-Doolittle and square root methods. Using Dwyer's notation, but replacing small letters by capitals where matrices are referred to, the upper triangular matrix of the summation rows of the Doolittle process is denoted by G^T, and the corresponding matrix (after division of the kth row by the square root of g_{kk}) for the square root method is denoted by S (see [1] p. 188). Let D denote the diagonal matrix composed of the diagonal elements of G. Then

$$S = D^{-\frac{1}{2}}G^T. \tag{2.4}$$

Denoting the matrix of the moments of the predictor variables by A, the fact that S factors A is expressed as

$$A = X^TX = S^TS. \tag{2.5}$$

Substituting in (2.3) we find $V(S^TS)V^T = I$ which can be written

$$(SV^T)^T(SV^T) = I. \tag{2.6}$$

A meaningful solution of this equation is given by

$$SV^T = I, \qquad V^T = S^{-1}, \qquad V = (S^T)^{-1} \tag{2.7}$$

and using (2.4) the solution can also be written as

$$V^T = (G^T)^{-1}D^{\frac{1}{2}}. \tag{2.8}$$

3. Computational algorithm for orthogonal multipliers. Dwyer has pointed out that the inverse of the matrix S can be very directly computed by using the identity matrix for the reference matrix of the dependent variable on the right in a computational schedule similar to the familiar Doolittle algorithm ([1] pp. 191 and 197, explicit in Table 13.6b and implicit in Table 13.8a).

For reference purposes we reproduce the schedule for $n = 3$ in terms of symbols, omitting the secondary subscripts:

$$A = \begin{matrix} a_{11} & a_{12} & a_{13} \\ * & a_{22} & a_{23} \\ * & * & a_{33} \end{matrix} \qquad \begin{matrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{matrix} = I$$

$$S = \begin{matrix} s_{11} & s_{12} & s_{13} \\ 0 & s_{22} & s_{23} \\ 0 & 0 & s_{33} \end{matrix} \qquad \begin{matrix} v_{11} & 0 & 0 \\ v_{21} & v_{22} & 0 \\ v_{31} & v_{32} & v_{33} \end{matrix} = V.$$

Because of the symmetry of the A matrix the computational algorithm assures that $SV^T = I$. Denote the n columns of the product matrix XV^T by the column vectors

$$\phi_1 = v_{11}x_1, \qquad \phi_2 = v_{21}x_1 + v_{22}x_2, \cdots \phi_n = \sum v_{ni}x_i. \tag{3.1}$$

These vectors constitute a set of n mutually orthogonal vectors which are linear transforms of the original set. They take the classical form, which one would expect, and are a "normal" set, since $\phi_k^T\phi_k = 1$.

For the uninitiated who are not familiar with the square-root method, or who prefer to use the Gauss-Doolittle method in the solution of the problem of least squares, recall from (2.8) that $V = D^{\frac{1}{2}}G^{-1}$ and note that G^{-1} is given by the triangular matrix represented by the lower of the doublet of rows extending under the identity reference matrix of the Doolittle algorithm ([1] p. 191, Table 13.6a). It follows that the triangular matrix from the upper of the doublet of rows, which we denote by R, satisfies the relation

$$R = D^{\frac{1}{2}}V. \tag{3.2}$$

Thus a set of mutually orthogonal vectors ψ_k is determined by the columns of the matrix product XR^T. These, although simpler to compute than the ϕ_k, have the disadvantage that $\psi_k^T\psi_k = g_{kk}$. One can of course compute ϕ_k from ψ_k by simply dividing through by $\sqrt{g_{kk}}$.

4. Accessory relations. Let y denote a column vector of N observations on the variable to be fitted in a least squares problem and let t denote the column vector of the coefficients of the orthogonal predictors ϕ_k. Recalling that $\Phi = XV^T$ represents an n column, N row matrix, conventional solution of the normal equations by matrix analysis leads to [4]

$$t = \Phi^T y, \qquad t_k = \phi_k^T y. \tag{4.1}$$

Furthermore it is easily seen from matrix analysis that

$$\Phi^T X = VX^T X = VS^T S = IS = S$$

and hence

$$s_{ki} = \phi_k^T x_i, \qquad i \neq 0. \tag{4.2}$$

Since the algorithm for finding s_{i0} in the y column is operationally the same as that used for finding s_{ik} in the x_k column, it follows that

$$s_{k0} = \phi_k^T y = t_k \tag{4.3}$$

and hence the fitted function u is given explicitly by

$$\begin{aligned} u &= \sum t_k \phi_k = \sum s_{k0} \phi_k , \\ \phi_k &= v_{k1} x_1 + v_{k2} x_2 + , \cdots , + v_{kk} x_k , \qquad k = 1, 2, \cdots n. \end{aligned} \tag{4.4}$$

Clearly if computations are based on the Doolittle algorithm and the predictors ψ_k are used, similar relations will hold:

$$g_{ki} = \psi_k^T x_i , \qquad i \neq 0; g_{k0} = \psi_k^T y, \tag{4.5}$$

and the coefficients of the predicting variables ψ_k will be given by

$$b_{k0} = g_{k0}/g_{kk} . \tag{4.6}$$

A forward solution is furnished by solving the above relations for the coefficients of x_i in the fitted function. Denoting these coefficients by the column vector h, we have $u = \Phi t = Xh$. Since X is not in general a square matrix, this equation is solved for h by multiplying it by what has been called the "pseudo-inverse" of X; namely, $(X^TX)^{-1}X^T$. The result is

$$h = (X^TX)^{-1}(X^T\Phi t) = (X^TX)^{-1}(X^TX)V^Tt = V^Tt.$$

In explicit, nonmatrix form

$$\begin{aligned} h_1 &= t_1v_{11} + t_2v_{21} + \cdots + t_nv_{n1} \\ h_2 &= t_2v_{22} + \cdots + t_nv_{n2} \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ h_{n-1} &= t_{n-1}v_{n-1,n-1} + t_nv_{n,n-1} \\ h_n &= t_nv_{nn} . \end{aligned} \tag{4.7}$$

A useful point about this forward solution which may easily be passed over without recognition is the following.

LEMMA. *If t_1, t_2 $\cdots$ t_k are determined as ϕ_1^Ty, ϕ_2^Ty, $\cdots$, ϕ_k^Ty, then h_1, h_2, $\cdots$, h_k derived from the above schedule* (4.7) *will satisfy the first k normal equations of the n predictor problem for arbitrary values of t_{k+1}, t_{k+2}, $\cdots$, t_n.*

The proof follows by applying the first k conditions for minimizing the sum of the squares of the deviations to the orthogonal form of the solution (4.4).

The writer has found application of this lemma to the following problem. Several linear functions are to be fitted to separate groups of data by least squares, where, say, one of the coefficients is to be determined so that it is optimum for *all* the groups *lumped together*, and the other coefficients are to be determined separately for *each separate group*. (One such application will be discussed by the author in an article which will appear shortly in the *Journal of the American Statistical Association*.)

REFERENCES

[1] P. S. DWYER, *Linear Computations*, John Wiley and Sons, 1951.
[2] P. S. DWYER, "The square root method and its use in correlation and regression," *J. Amer. Stat. Assn.*, Vol. 40 (1945), pp. 493–503.
[3] J. P. GRAM, "Ueber die entwicklung reeler functionen in reihen," Crelle's *J. für Math.*, Vol. 94 (1883), pp. 41–73. (This is a presentation in German of his 1879 dissertation.)
[4] A. S. HOUSEHOLDER, "Some numerical methods for solving systems of linear equations," *Amer. Math. Monthly*, Vol. 57 (1950), pp. 453–459.
[5] B. F. KIMBALL, "Orthogonal polynomials applied to least square fitting of weighted observations," *Ann. Math. Stat.*, Vol. 11 (1940), p. 352.

ON THE ASYMPTOTIC NORMALITY OF CERTAIN RANK ORDER STATISTICS[1]

BY MEYER DWASS

Northwestern University

1. Summary. Let $(R_1, \cdots, R_N)$ be a random vector which takes on each of the $N!$ permutations of the numbers $(1, \cdots, N)$ with equal probability, $1/N!$. Sufficient conditions are given for the asymptotic normality of $S_N = \sum_{i=1}^{N} a_{Ni} b_{NR_i}$, where $(a_{N1}, \cdots, a_{NN})$, $(b_{N1}, \cdots, b_{NN})$ are two sets of real numbers given for every N. These sufficient conditions are apparently quite different from those given by Wald and Wolfowitz [9] and extended by various writers [4, 7]. In some situations the conditions given here may be easier to apply than those given previously. The most general conditions available to date appear to be those of Hoeffding [4]. In the examples below, however, is given a case of an S_N which does not satisfy the conditions required by Hoeffding's theorem but which is asymptotically normal by our results.

2. Statement of theorem and its proof. We will assume hereafter that

$$\sum_{i=1}^{N} a_{Ni} = \sum_{i=1}^{N} b_{Ni} = 0, \qquad \sum_{i=1}^{N} a_{Ni}^2 = 1.$$

THEOREM. Suppose for an integer $k \geqq 1$ there is a random variable X satisfying the following conditions:

(a) X has a continuous cdf $F(x)$,

(b) if $X_1, \cdots, X_N$ are independent random variables each with the cdf $F(x)$ and $Z_{N1} \leqq \cdots \leqq Z_{NN}$ are the ordered values of $X_1, \cdots, X_N$ then

$$b_{Ni} = EZ_{Ni}^k - \sum_{j=1}^{N} EZ_{Nj}^k / N$$

for all N and i.

Received 7/21/52, revised 11/24/52.

[1] Work done under the sponsorship of the Office of Naval Research while the author was a graduate student at the University of North Carolina.