THE PSEUDOINVERSE OF A RECTANGULAR MATRIX AND ITS STATISTICAL APPLICATIONS

By: T. N. E. Greville, Office of the Quartermaster General

SUMMARY

The connection between inversion of matrices and solution of nonsingular systems of linear equations is well known. The statistician, however, is often concerned with systems of equations in which the number of equations exceeds the number of unknowns and there is no exact solution. In such cases the least squares solution of the system is usually sought, and the classical matrix theory is of little avail.

In 1920 E. H. Moore announced a generalization of the notion of inverse of a matrix, which provides a generalized inverse or "pseudoinverse" for rectangular matrices, as well as for singular square matrices. Bjerhammar in 1951 and Penrose in 1956 have shown that this pseudoinverse is related to the least squares solution of an inconsistent system of linear equations in a way analogous to the relationship of the classical inverse to the solution of a nonsingular system.

In the present article two possible applications of this concept to statistical procedures are suggested. These relate to (1) the computation of multilinear regression coefficients and (2) least squares curve fitting, with particular reference to the fitting of polynomials. In the latter application the procedure suggested here has an advantage over the use of orthogonal polynomials in that unequal spacing of the arguments does not increase the amount of calculation required. Among other possible uses of the pseudoinverse not discussed here is its application to bivariate interpolation.*

A recursive algorithm is described by which one can derive from the pseudoinverse of a given matrix that of a second matrix obtained by the addition of a single column. Thus one computes first the pseudoinverse of the first column of the coefficient matrix, then that of the first two columns, and so on until the pseudoinverse of the entire coefficient matrix is obtained. In the regression application, this makes it possible to arrange the variables in decreasing order of their probable importance in the regression equation, and to stop the process when it appears that the introduction of further variables will not have a significant effect. Similarly, in fitting a polynomial, the process is arranged so as to fit polynomials of successively higher degree, and one can stop when it appears that the most suitable degree has been reached. In either case the residual variance is easily obtained as a by-product.

1. INTRODUCTION

It is of course well known that any square matrix A with nonzero determinant has a unique inverse A^{-1} such that

$$AA^{-1} = A^{-1}A = I , \tag{1}$$

where I denotes the unit matrix or identity matrix having 1's along its principal diagonal and 0's elsewhere. It seems to be not so well known that in 1920 [1,2] the eminent American mathematician E. H. Moore announced a generalization of the inverse concept to include rectangular matrices and those with vanishing determinant. Little notice was taken of Moore's discovery for about 30 years, but during the past decade the properties of this generalized inverse or pseudoinverse have been vigorously explored by Bjerhammar [3,4,5], Penrose [6,7] and Hestenes [8].1/ It is found to have some useful applications in numerical analysis and statistics. Of the latter, probably the most obvious are in connection with multiple regression and least squares curve fitting.

It is the purpose of the present expository article to indicate these statistical applications and to describe a simple numerical algorithm for computing the pseudoinverse of a given matrix. No claim to originality is made; everything in this article is explicit or implicit in the work of Bjerhammar, Penrose and Hestenes.

A knowledge of the elementary properties of matrices is assumed.2/ In particular, the reader should keep in mind that a vector can be thought of as a matrix of one column (or row), and that the row-by-column rule for multiplying two matrices together implies that, in a matrix product

$$AB = C ,$$

each column of C is a linear combination of columns of A and each row of C is a linear combination of rows of B. Extensive use will be made of the notions of vector spaces and orthogonality. A fuller version of this article, including a brief exposition of these concepts and proofs of certain important properties of the pseudoinverse, can be obtained from the author in mimeographed form.

2. SOLUTION OF SYSTEMS OF LINEAR EQUATIONS

It is well known that a system of linear equations

$$\sum_{j=1}^{n} a_{ij} x_j = b_i \qquad (i = 1, 2, \ldots, m) \tag{2}$$

can be written compactly as a matrix equation

$$Ax = b , \tag{2}'$$

where A denotes the matrix (a_{ij}), x is the vector (i.e., single-column matrix) whose elements are the values of the variables x_j which consti-

tute a solution of the system, and b is the vector whose ith element is b_i. If A is nonsingular (i.e., if m = n and its columns are linearly independent), it has a unique inverse A^{-1} satisfying equation (1), and moreover the system (2) or (2)' has a unique solution given by

$$x = A^{-1} b . \tag{3}$$

The statistician is often concerned with systems of equations in which m > n and there is no exact solution. In such a case, Bjerhammar [3,4] and Penrose [7] have shown that the "best" solution in the sense of least squares is given by

$$x = A^{\dagger} b ,$$

where $A^{\dagger}$ is the pseudoinverse of the rectangular matrix A. The definition and computation of the pseudoinverse will now be taken up.

3. DEFINITION AND PROPERTIES OF THE PSEUDOINVERSE

Any nonzero real matrix A of rank r can be expressed as a product

$$A = BC , \tag{4}$$

where the r-column matrix B and the r-rowed matrix C are both of rank r. To show this, let B be any matrix whose columns form a basis for the column-space of A. Then a matrix C exists such that A = BC, and C is of rank r, since the rank of a product cannot exceed the rank of any factor. Since the columns of B and the rows of C are linearly independent, the matrices $B^T B$ and CC^T (where the superscript T denotes the transpose) are positive definite, and therefore nonsingular.

We now define the pseudoinverse for the matrices B and C (and generally for rectangular matrices of maximal rank) as follows:

$$B^{\dagger} = (B^T B)^{-1} B^T \qquad C^{\dagger} = C^T (CC^T)^{-1} . \tag{5}$$

It will be noted that, for a nonsingular square matrix, these expressions reduce to the classical inverse. We have also

$$B^{\dagger} B = CC^{\dagger} = I , \tag{6}$$

and moreover $B^{\dagger} A = C$, which, together with (4), shows that the row-spaces of A and C are identical. It follows that the rows of C form a basis for the row-space of A.

For the nonzero real matrix A, we now define the pseudoinverse[3] as

$$A^{\dagger} = C^{\dagger} B^{\dagger} . \tag{7}$$

Finally, for completeness, we define the pseudoinverse of an m x n zero matrix as an n x m zero matrix.

In consequence of this definition, the row-space and column-space of $A^{\dagger}$ are the transposes of the column-space and row-space, respectively, of A. If A is m x n, evidently $A^{\dagger}$ is n x m; thus both the products $AA^{\dagger}$ and $A^{\dagger} A$ can be formed. Moreover, these products have interesting properties. First, we note that each is symmetric and idempotent (i.e., equal to its own square). It is easily verified that $I - AA^{\dagger}$ and $I - A^{\dagger} A$ are also symmetric and idempotent. We can show that the product $AA^{\dagger}$ is the same for any two matrices A_1 and A_2 having the same column-space. For, equations (4), (6) and (7) give $AA^{\dagger} = BB^{\dagger}$. Since the same matrix B can serve for both A_1 and A_2, the result follows.

Let u denote any vector in Euclidean m-space and consider the vectors:

$$x = AA^{\dagger} u , \qquad y = u - x = (I - AA^{\dagger})u .$$

It is evident that x is a vector in the space S_c, the column-space of A. Moreover, y is orthogonal to this space. For, if z is any vector of S_c, there exists a vector w, such that z = Aw; and, in view of the symmetry and idempotency of $AA^{\dagger}$,

$$y^T z = u^T (I - AA^{\dagger}) Aw = 0 ,$$

since (4), (6) and (7) give $AA^{\dagger} A = A$. This decomposition of u into a vector of S_c and a vector orthogonal to S_c is unique. To show this, let u = $x_1 + y_1$, where x_1 is in S_c and y_1 orthogonal to S_c. Then there exists a vector w_1, such that $x_1 = Aw_1$, and $AA^{\dagger} u = AA^{\dagger} Aw_1 = Aw_1 = x_1$, showing that $x_1 = x$.

We shall call the vector x the projection of u on S_c, and the matrix $AA^{\dagger}$, which has been shown to be characteristic of the space (since it is the same for all matrices A having this column-space), will be called the projector on S_c, and will be denoted by P_c.

Similar remarks apply to the matrix $A^{\dagger} A$ with regard to left multiplication by row-vectors, and it will be called the projector on the row-space S_r of A, and will be denoted by P_r.

4. APPROXIMATE SOLUTION OF INCONSISTENT SYSTEMS OF LINEAR EQUATIONS[4]

Consider the problem of approximating an arbitrary vector u by a vector v which is restricted to the extent that it must belong to a given vector space[5] S. We shall take the differences between corresponding components of u and v, and shall say that the approximation is "best" when the sum of the squares of these differences is a minimum. This sum of squares, which we shall

denote by q, is given by

$$q = (u - v)^T (u - v) . \tag{8}$$

Now, we have

$$u = Pu + y ,$$

where P is the projector on S and y is orthogonal to S. Therefore,

$$u - v = (Pu - v) + y .$$

Since $Pu - v$ belongs to S, it is orthogonal to y, and therefore equation (8) reduces to

$$q = (Pu - v)^T (Pu - v) + y^T y .$$

This is the sum of two positive terms, the second of which is independent of v, and is clearly a minimum when the first term is made to vanish by choosing $v = Pu$. In other words, the projection of u on S is the best approximation in S to u.

Now, let us return to the consideration of the system of equations represented by (2) or (2)'. An exact solution is possible if and only if b belongs to the column-space of A. When this is the case, the solution is unique when A is nonsingular, and is given by (3). If A is singular, the general solution is

$$x = A^\dagger b + y , \tag{9}$$

where y is any vector orthogonal to the row-space of A. It is clear that this is a solution, since

$$Ax = P_c b + Ay = b ,$$

since, under the hypotheses, $P_c b = b$ and $Ay = 0$. To show that this includes all solutions, we first note that any solution x can certainly be expressed in the form (9) if y is unrestricted. Substitution of this expression in (2)' then gives $b + Ay = b$, or $Ay = 0$, showing that y is indeed orthogonal to the row-space of A.

Since the two vectors in the right member of (9) are orthogonal, the length of the vector x is a minimum when we take $y = 0$. Thus, the solution of minimum vector length is given by

$$x_m = A^\dagger b . \tag{10}$$

Of greater interest to us, however, is the case in which b is not in the column-space of A, so that no exact solution is possible. In this case, we consider as the "best" solution the vector x for which the length of the vector $Ax - b$ is a minimum. This is tantamount to the usual least squares criterion, and implies, as we have already seen, that

$$Ax = P_c b ,$$

where P_c is the projector on S_c, the column-space of A. Since $P_c b$ is in S, the solution is given by (9) with b replaced by $P_c b$, and is therefore

$$x = A^\dagger P_c b + y = A^\dagger b + y ,$$

since $A^\dagger A A^\dagger = A^\dagger$ by (4), (6) and (7). Thus we have shown that when there is no solution, (9) and (10) give the "best" solution in the sense indicated.

5. STATISTICAL APPLICATIONS

Perhaps the most obvious statistical application is to multiple regression. Let a variate y depend on n variates $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$, and let it be required to determine the coefficients a_j in the regression equation

$$y = \sum_{j=1}^{n} a_j x^{(j)} .$$

It is assumed that corresponding numerical values $y_i, x_i^{(j)}$ are given for $i = 1, 2, \ldots, m$. If y denotes the column-vector whose ith component is y_i, a the column-vector whose jth component is a_j, and X the matrix $(x_i^{(j)})$, the regression coefficients are given by

$$a = X^\dagger y . \tag{11}$$

If the columns of X are linearly independent, as will usually be the case, the least squares regression equation is unique. Otherwise, there will be many solutions which yield the minimum value for the sum of the squared residuals. Of these possible solutions, (11) then gives the one for which the sum of the squares of the coefficients a_j is smallest.

Let (x_i, y_i), $i = 1, 2, \ldots, m$, be a set of points to which a curve $y = f(x)$ is to be fitted. It is stipulated that $f(x)$ is to be a linear combination of n given functions $g_1(x), g_2(x), \ldots, g_n(x)$: thus

$$f(x) = \sum_{j=1}^{n} a_j g_j(x) .$$

The coefficients a_j are to be determined so as to minimize the quantity

$$S = \sum_{i=1}^{m} \left[y_i - f(x_i) \right]^2 .$$

This covers many, but not all least squares curve fitting situations. The simplest and most usual case is that of fitting a polynomial of degree $n - 1$, for which $g_j(x) = x^{j-1}$.

If y and a are defined as before and Q denotes a matrix such that the element in the ith row and the jth column is $g_j(x_i)$, then

$$a = Q^\dagger y .$$

If u_i denotes the fitted ordinate corresponding to the given ordinate y_i and u is the vector whose ith component is u_i, we have

$$u = Qa = QQ^\dagger y = P_c y ,$$

where P_c is the projector on the column-space of Q.

6. RECURSIVE ALGORITHM FOR OBTAINING THE PSEUDOINVERSE OF A MATRIX

Equations (5) and (7) are not very practical for computational purposes. The writer has given elsewhere [13] the following algorithm[6] for obtaining the pseudoinverse of a matrix. Let a_k denote the kth column of a given matrix A, and let A_k denote the submatrix consisting of the first k columns. Then the pseudoinverse of A_k is of the form

$$A_k^\dagger = \begin{bmatrix} A_{k-1}^\dagger - \gamma_k c_k \\ c_k \end{bmatrix}, \tag{12}$$

where

$$\gamma_k = A_{k-1}^\dagger a_k, \tag{13}$$

and the last row c_k remains to be determined. In its determination two distinct cases arise, according to whether or not a_k belongs to the column-space S_{k-1} of A_{k-1}. In other words, it must be ascertained whether or not the space spanned by the first $k-1$ columns of A is enlarged by the addition of the kth column. Now a_k belongs to S_{k-1} if and only if the projection of a_k on S_{k-1} is equal to a_k itself: in other words, if

$$A_{k-1} \gamma_k = a_k . \tag{14}$$

If (14) is not satisfied

$$c_k = (a_k - A_{k-1} \gamma_k)^\dagger , \tag{15}$$

while if (14) is satisfied

$$c_k = (1 + \gamma_k^T \gamma_k)^{-1} \gamma_k^T A_{k-1}^\dagger . \tag{16}$$

Formulas (12) to (16) constitute a recursive procedure for obtaining successively $A_2^\dagger$, $A_3^\dagger$, ..., starting with $A_1^\dagger$. Both for the initial determination of $A_1^\dagger$ and for the evaluation of the right member of (15) a formula is required for the pseudoinverse of a single-column matrix a. It follows from (5) that this is

$$a^\dagger = \begin{cases} a^T & (a = 0) \\ (a^T a)^{-1} a^T & (a \neq 0). \end{cases} \tag{17}$$

For the purpose of statistical applications, some "streamlining" of the algorithm can be effected by noting that in these situations it is unnecessary to obtain the pseudoinverse explicitly. Rather, what is wanted is the "best" solution $x = A^\dagger b$ of an inconsistent system $Ax = b$. The algorithm can be modified to give $A_k^\dagger b$ for $k = 1, 2, \ldots$ successively. To this end it is convenient to define a matrix A' obtained by enlarging A through the addition of two columns on the right: (i) the vector b and (ii) a total column, which is the sum of all the preceding column vectors. Then (12) gives

$$A_k^\dagger A' = \begin{bmatrix} A_{k-1}^\dagger A' - \gamma_k (c_k A') \\ c_k A' \end{bmatrix} . \tag{18}$$

The penultimate column of this matrix is $A_k^\dagger b$, while the final column should be the sum of the preceding column vectors if the arithmetic has been correctly performed. Moreover, (13) shows that γ_k is the kth column of $A_{k-1}^\dagger A'$.

In order to obtain $c_k A'$ for use in (18) we must first compute

$$a_k - A_{k-1} \gamma_k . \tag{19}$$

If this vector vanishes, (16) shows that

$$c_k A' = (1 + \gamma_k^T \gamma_k)^{-1} \gamma_k^T A_{k-1}^\dagger A' .$$

If (19) does not vanish, it equals $c_k^\dagger$ by (15). By (17) we then have

$$c_k A' = (c_k^{\dagger T} c_k^\dagger)^{-1} c_k^{\dagger T} A' . \tag{20}$$

If we first compute the vector $c_k^{\dagger T} A'$, we note that its kth element is $c_k^{\dagger T} a_k$. By (13), $A_{k-1} \gamma_k = A_{k-1} A_{k-1}^\dagger a_k$ is the projection of a_k on the column-space of A_{k-1}. Therefore, $c_k^\dagger$ as given by (19) is orthogonal to this space. Thus $c_k^{\dagger T} A_{k-1} \gamma_k = 0$, and consequently, $c_k^{\dagger T} c_k^\dagger = c_k^{\dagger T} a_k$. It follows from (20) that c_k is obtained from the computed vector $c_k^{\dagger T} A'$ upon "normalizing" it by dividing by its kth element. With these explanations, (18), (19) and (20) constitute the recursive procedure desired.[7]

For example, if m sets of corresponding values of n statistical variables $x_1, x_2, \ldots, x_n$ are given, and it is required to compute regression coefficients of x_1 against $x_2, x_3, \ldots, x_n$, the matrix A' is formed so that its first column consists of all 1's, and the 2nd to nth columns exhibit the successive values of the variables $x_2, \ldots, x_n$, respectively. The $(n+1)$th column is the vector x whose components are the corresponding values of x_1, and this is followed by the total column. Then $A_1^\dagger x$ is the mean value of x_1, while at the kth stage of the process $A_k^\dagger x$ is a vector whose components are the coefficients in the regression equation

$$\hat{x}_1 = b_k + b_{12\cdot 3\ldots k} x_2 + b_{13\cdot 24\ldots k} x_3 + \ldots + b_{1k\cdot 2\ldots k-1} x_k .$$

The components of $A_k A_k^\dagger x$ are the values of x_1 predicted by the regression equation, while $x^T(x - A_k A_k^\dagger x)$ is the sum of the squares of the errors of estimate. Thus, if there is doubt as to how many of the variables should be included in the regression equation (and if one is fortunate in choosing the order in which the variables are introduced), this method shows at a glance how much the coefficients change as the less significant variables are brought into the equation, and, if desired, the reduction at each step in the standard error of estimate. It will be noted also that the vector γ_k given by (13) exhibits the coefficients in the regression equation of x_k against $x_2, x_3, \ldots, x_{k-1}$.8/

Consider now the problem of least squares fitting of a polynomial. Let (x_i, y_i), $i = 1, 2, \ldots, m$, be the set of points to which a polynomial is to be fitted, let A be the n-column matrix (x_i^{j-1}), and let y denote the vector whose ith component is y_i. Then, if the algorithm is applied, the vector $A_k^\dagger y$ exhibits the coefficients of successive powers of x in the least squares polynomial of degree $k - 1$, while $A_k A_k^\dagger y$ exhibits the values of the fitted polynomial corresponding to the given abscissas x_k, and $y^T(y - A_k A_k^\dagger y)$ is the sum of the squared residuals. The latter quantity, of course, may be used in testing to see what degree of polynomial is most suitable.9/

Extensive tables are available [16] to facilitate the use of orthogonal polynomials in fitting a least squares polynomial to data with equally spaced arguments, but they are of no avail when the abscissas are irregularly spaced. It is to be noted that the procedure described here makes no assumption about the spacing of the arguments. Further, the recursive nature of the process obviates the need to make use of orthogonal polynomials directly. If desired, however, they can easily be obtained as a by-product.10/ If $\gamma_k^{(i)}$ denotes the ith component of γ_k, the set of polynomials

$$(21)\qquad \begin{aligned} p_0(x) &= 1 \\ p_k(x) &= x^k - \sum_{i=1}^{k} \gamma_k^{(i)} x^{i-1} \\ &\qquad (k = 1, 2, \ldots, m - 1) \end{aligned}$$

constitutes an orthogonal set over the discrete domain $(x_1, x_2, \ldots, x_m)$. In other words,

$$(22)\qquad \sum_{i=1}^{m} p_j(x_i)\, p_k(x_i) = 0 \qquad (j \neq k).$$

To show this, we first note that j and k are interchangeable in (22), so that we can assume without loss of generality that $j < k$. Thus, (22) will be established if we can show that

$$\sum_{i=1}^{m} x_i^j\, p_k(x_i) = 0 \qquad (j = 0, 1, \ldots, k - 1).$$

But, in view of (21) and of the definition of A, this follows from the fact previously noted that the vector $a_k - A_{k-1}\gamma_k$ is orthogonal to S_{k-1}.

FOOTNOTES

*This possibility was suggested to the writer by William Hodgkinson, Jr., of the American Telephone and Telegraph Co.

1/The first two writers mentioned were unaware of Moore's work until it was brought to their attention by Rado [9].

2/An excellent brief treatment of the subject, more than adequate for the understanding of this article, is given in Chapter 1 of [10].

3/The pseudoinverse can be defined in several other ways [2,6,8,12, and the fuller, mimeographed version of this paper]. The present approach, which is probably the simplest, was suggested to the writer by A. S. Householder, whose assistance is gratefully acknowledged.

4/See also [3,4,7,11].

5/See also [12].

6/In essence, this algorithm is an abbreviated form of a particular case of the method of matrix inversion by biorthogonalization proposed by Hestenes [8]. See also [13,14].

7/A further "streamlining" is possible by working with the symmetric matrix $A'^T A'$, which, in essence, merely exhibits the usual "normal" equations. This kind of procedure is easily explained without reference to the pseudoinverse, and is probably the simplest approach for small-sized calculations. In large-scale calculations, it has the disadvantage that, if the "recursive" feature is retained, certain key quantities in the computations (e.g., those which we have called "normalizing" factors) are obtained as differences between large, and almost equal, numbers, and accuracy is rapidly lost.

8/A numerical example of the application to multiple regression is given in the fuller, mimeographed version of this article. It is not reproduced here because the actual arithmetic is essentially the same as that involved in other methods of calculating multiple regression coefficients.

9/See [15], pp. 461-465.

10/See also [17].

REFERENCES

[1] Moore, E. H., Bulletin of the American Mathematical Society, 26 (1920), 394-395.

[2] Moore, E. H., "General Analysis, Part I," Memoirs of the American Philosophical Society, 1 (1935), especially pp. 197-209.

[3] Bjerhammar, A., "Application of calculus of matrices to method of least squares; with special reference to geodetic calculations," Transactions of the Royal Institute of Technology (Stockholm), No. 49 (1951), 1-86.

[4] Bjerhammar, A., "Rectangular reciprocal matrices with special reference to geodetic calculations," Bulletin Géodésique, 1951, 188-220.

[5] Bjerhammar, A., A Generalized Matrix Algebra, National Research Council of Canada, Division of Applied Physics, Ottawa, 1957. Pp. 24.

[6] Penrose, R., "A generalized inverse for matrices," Proceedings of the Cambridge Philosophical Society, 51 (1955), 406-413.

[7] Penrose, R., "On best approximate solutions of linear matrix equations," Ibid., 52 (1956), 17-19.

[8] Hestenes, M. R., "Inversion of matrices by biorthogonalization and related results," Journal of the Society for Industrial and Applied Mathematics, 6 (1958), 51-90.

[9] Rado, R., "Note on generalized inverses of matrices," Proceedings of the Cambridge Philosophical Society, 52 (1956), 600-601.

[10] Rao, C. R., Advanced Statistical Methods in Biometric Research, New York, John Wiley and Sons, 1952.

[11] Greville, T. N. E., "The pseudoinverse of a rectangular or singular matrix and its application to the solution of systems of linear equations," SIAM Review, 1 (1959), 38-43.

[12] Greville, T. N. E., "On smoothing a finite table: a matrix approach," Journal of the Society for Industrial and Applied Mathematics, 5 (1957), 137-154.

[13] Greville, T. N. E., "Some applications of the pseudoinverse of a matrix," to be published in SIAM Review, 1 (1959).

[14] Dent, B. A., and Newhouse, A., "Polynomials orthogonal over discrete domains," SIAM Review, 1 (1959), 55-59.

[15] Snedecor, G. W., Statistical Methods, 5th Edition, Ames, Iowa, Iowa State College Press, 1956.

[16] Anderson, R. L., and Houseman, E. E., Tables of Orthogonal Polynomial Values Extended to N = 104, Research Bulletin 297, Agricultural Experiment Station, Iowa State College of Agriculture and Mechanic Arts, Ames, 1942.

[17] Forsythe, G. E., "Generation and use of orthogonal polynomials for data fitting with a digital computer," Journal of the Society for Industrial and Applied Mathematics, 5 (1957), 74-88.