Use of the Singular Value Decomposition in Regression Analysis

JOHN MANDEL*

Principal component analysis, particularly in the form of singular value decomposition, is a useful technique for a number of applications, including the analysis of two-way tables, evaluation of experimental design, empirical fitting of functions, and regression. This paper is a discussion in expository form of the use of singular value decomposition in multiple linear regression, with special reference to the problems of collinearity and near collinearity.

KEY WORDS: Collinearity; Multiple linear regression; Principal component regression; Singular value decomposition.

INTRODUCTION

While multiple linear least squares regression has been in use for a long time as the major statistical technique for "fitting equations to data," the full implications, limitations, and inherent problems associated with it have been treated in the literature only recently. In addition to clarifying the issues, much recent work has also provided modifications of the technique aimed at increasing its reliability as a data-analytic tool.

Undoubtedly, the greatest source of difficulties in using least squares is the existence of "collinearity" in many sets of data, and most of the modifications of the ordinary least squares approach are attempts to deal with the problem of collinearity. Among these modifications one can cite principal components regression (Draper and Smith 1981; Hocking, Speed, and Lynn 1976), latent root regression (Webster, Gunst, and Mason 1974), shrinkage (Hocking, Speed, and Lynn 1976; Stein 1960), ridge regression (Chatterjee and Price 1977; Draper and Smith 1981; Hocking, Speed, and Lynn 1976; Hoerl and Kennard 1970; Marquardt 1970; Marquardt and Snee 1975; Snee 1973), and a number of variants of these techniques.

The present paper does not attempt to discuss all these techniques, or to compare their relative merits. Its purpose is rather to present the nature of the problems through a careful exposition of the pertinent mathematical and conceptual aspects. It is almost indispensable, in order to achieve this purpose, to use matrix notation and to resort to the method of principal components or to related techniques. We use the technique known as singular value decomposition (SVD) (Rao 1973, p. 42) of the design matrix, a technique closely related to the method of principal components, to elucidate the problem of collinearity, and we attempt to explain this technique mainly through graphical interpretations.

This paper is of an expository type, and we do not claim completeness in our treatment. For a more comprehensive and more advanced treatment, to which this paper may serve as a useful introduction, the reader is referred to a recent book by Belsley, Kuh, and Welsch (1980). Another general reference dealing with this topic is the new edition of Draper and Smith (1981).

*John Mandel is Statistical Consultant, National Measurement Laboratory, National Bureau of Standards, Washington, D.C. 20234.

THE MODEL

We assume that the model is known and of the form

$$Y = X\beta + e, \tag{1}$$

where Y and e are vectors of N elements each, X is an $N \times p$ matrix of elements x_{ij} and β a vector of p elements. The matrix X, consisting of nonstochastic elements, is given. The Y vector consists of the measurements y_i, each of which is the sum of two terms, the expected value

$$E(y_i) = \sum_j x_{ij}\beta_j,$$

and the error term e_i. The errors e_i are assumed to be uncorrelated, of zero mean and constant variance σ^2, the value of which is not known. The vector of e_i is represented by the term e in (1). The general ideas in this paper will be illustrated for the artificial data displayed in Table 1, in which $N = 8$ and $p = 3$. There are in this case three regressor variables, x_1, x_2, and x_3, of which the first is equal to unity for all i. The regression equation is

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i,$$

but since $x_{i1} \equiv 1$, the equation becomes

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i,$$

with an "independent term" β_1. Inclusion of such an independent term is a common practice in regression work. Its usefulness is apparent when one considers regressors that can be expressed in linearly related, but nonproportional units. For example, if the regressor x_2 is *temperature*, a conversion from Celsius to Fahrenheit units would be impossible within the assumed model if no allowance had been made for an independent term.

Many practitioners of regression analysis perform a "standardization" on all regressor variables other than the independent term, prior to analysis. The standardization of regressor x_j consists in replacing it in the regression equation by

$$x_j = \bar{x}_j + s_j t_j,$$

where $\bar{x}_j$ is the average, and s_j is the standard deviation, of the elements x_{ij} in column x_j. The regression is now

Table 1. Data Set A

Point	x_1	x_2	x_3	y
1	1	16.85	1.46	41.38
2	1	24.81	−4.61	31.01
3	1	18.85	−.21	37.41
4	1	12.63	4.93	50.05
5	1	21.38	−1.36	39.17
6	1	18.78	−.08	38.86
7	1	15.58	2.98	46.14
8	1	16.30	1.73	44.47

that of y on the regressors t_j, and the latter are such that for each j $\bar{t}_j = 0$ (centering) and $s_{t_j} = 1$ (scaling). The uses and usefulness of centering and scaling are discussed in some detail in Draper and Smith (1981). For simplicity of presentation we omit this step throughout this paper.

The object of the regression analysis is to estimate the coefficients $\beta_j (j = 1 \text{ to } p)$, as well as σ^2, to "predict" the value of y for any "future" vector of regressor variables $x = (x_1\ x_2\ \ldots\ x_p)$, and to estimate the error of such a predicted value, say $\hat{y}$. To avoid confusion, a set of values $(x_1, x_2, \ldots, x_p)$ for which a y-value is to be found are referred to henceforth as a "point in X space," or simply as a "point," rather than as a vector.

Additional aspects of the regression problem appear in the course of our discussion.

GEOMETRIC REPRESENTATION OF THE REGRESSION

Assume a situation in which there are only two regressor variables, x_1 and x_2. Then the "design points" (x_1, x_2) can be plotted in a plane D, such as that shown in Figure 1. At each of the design points a segment is erected, in a direction perpendicular to the D plane and of height y, where y is the value of the "response" variable at the point (x_1, x_2). According to the model equation (1), the end points of these segments should lie *close to* a plane. They would lie exactly *in* a plane if the response variable y were completely free of experimental error.

Let P designate the (true) response plane. Because of the errors in y, P cannot be exactly determined, but it can be *approximated* by a *fitted* plane, say P_f, as shown in Figure 1.

In the more general situation of p regressor variables, the D plane becomes a p-dimensional hyperplane, and so do the P and P_f planes.

SINGULAR VALUE DECOMPOSITION OF X

Given any $N \times p$ matrix X, it is possible to express each element x_{ij} of X in the following way:

$$x_{ij} = \theta_1 u_{1i} v_{1j} + \theta_2 u_{2i} v_{2j} + \ldots + \theta_r u_{ri} v_{rj} \qquad (2)$$

or, more compactly, by the relation

$$x_{ij} = \sum_{k=1}^{r} \theta_k u_{ki} v_{kj},$$

where $\theta_1 \geq \theta_2 \geq \ldots \geq \theta_r$. This is known as the *singular value decomposition* (SVD) of the matrix X. The number of terms in (2) is r, the rank of the matrix X; r cannot exceed N or p, whichever is smaller.

We will always assume that $N \geq p$, and it follows that $r \leq p$. The r vectors u are orthogonal to each other, as are the r vectors v. Furthermore, each of these vectors has unit length, so that

$$\sum_i u_{ki}^2 = \sum_j v_{kj}^2 = 1 \text{ for all } k. \qquad (4)$$

In matrix notation, we have

$$\underset{N\times p}{X} = \underset{N\times r}{U}\ \underset{r\times r}{\theta}\ \underset{r\times p}{V'}. \qquad (5)$$

The matrix θ is diagonal, and all θ_k are positive. The columns of the matrix U are the u vectors, and the rows of V' are the v vectors of (2). The orthogonality of the u and v and their unit length implies the conditions

$$U'U = I \qquad (6)$$

$$V'V = I, \qquad (7)$$

where a prime ($'$) indicates transpose of a matrix and I is an $r \times r$ identity matrix. The θ_k can be shown to be the square roots of the nonzero eigenvalues of the square matrix $X'X$ as well as of the square matrix XX'. The columns of U are the eigenvectors of XX' and the rows of V' are the eigenvectors of $X'X$. Excellent algorithms exist for obtaining the SVD of a matrix (see Chambers 1977). Table 2 shows the matrix X of data set A, along with its SVD. The U, θ, and V' matrices are displayed to show their dimensional relations to the X matrix. In this case the rank r is 3, that is, $r = p$. This is known as the *full-rank* case. Each element of X is easily reconstructed by multiplying the corresponding elements of the U, θ, and V' matrices and summing the

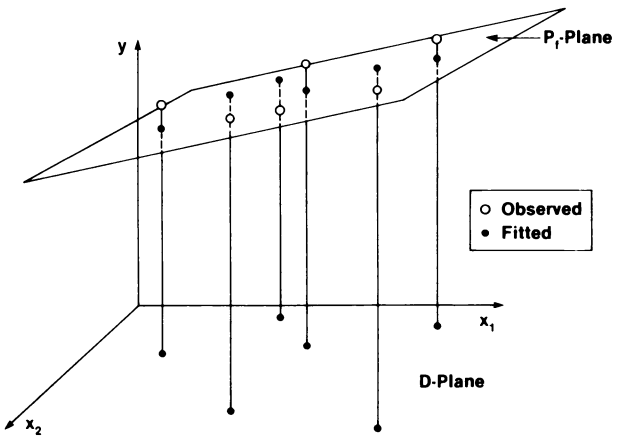

Figure 1. Geometric Representation of a Regression Surface

Table 2. SVD of the X Matrix of Data Set A

	X			U		
				u_1	u_2	u_3
	1	16.85	1.46	.322575	.176104	.193765
	1	24.81	−4.61	.473864	−.603455	.049884
	1	18.85	−.21	.360574	−.038169	.333727
	1	12.63	4.93	.242392	.621398	−.036214
	1	21.38	−1.36	.408731	−.186949	−.659192
	1	18.78	−.08	.359251	−.021580	.241838
	1	15.58	2.98	.298488	.370545	−.453637
	1	16.30	1.73	.312108	.210988	.385321
		V′			θ	
v_1	.053067	.998579	.004786	52.347807	0	0
v_2	.067340	−.008360	.997695	0	7.853868	0
v_3	.996317	−.052622	−.067688	0	0	.055690

Table 3. SVD of an X Matrix of Two Variables

	X-Matrix		U-Matrix	
	x_1	x_2	u_1	u_2
	1.3	1.2	.1202	.4037
	4.2	2.8	.2511	−.5113
	6.3	7.4	.4867	.6912
	8.0	7.1	.5391	−.1689
	9.4	8.2	.6285	−.2634
	V′		Θ-Matrix	
v_1	.7309	.6825	19.8360	0
v_2	−.6825	.7309	0	1.6040

terms. For example, the element 4.93 in the fourth row and third column is equal to:

$$[.004786 \times 52.347807 \times .242392]$$
$$+ [.997695 \times 7.853868 \times .621398]$$
$$+ [(-.067688) \times .055690 \times (-.036214)].$$

GEOMETRIC INTERPRETATION OF SVD

To simplify the exposition, we consider an example with only two regressor variables. Table 3 shows the X matrix, consisting of five points, as well as the two u-vectors, the two v-vectors, and the diagonal θ matrix. Each row of the X matrix consists of two numbers x_1, x_2. We can interpret these as a "point" in 2-dimensional space, with coordinates x_1 and x_2 (see Fig. 2). The vector joining the origin to that point can also be used to represent that point, and we will therefore not hesitate to refer to any such set of two numbers as a *vector* as well as a *point.* Thus the X-matrix is represented by five points, or five vectors.

Similarly, the rows labeled v_1 and v_2 also represent one point each or one vector each. The coordinates of the point v_1, for example, are the numbers .7309 and .6825. First, we note that the distance of the origin to that point is unity, and that the same holds for vector v_2. Thus, the vectors v_1 and v_2 have unit length. Second, it is easily verified that these two vectors are perpendicular (or "orthogonal") to each other (just as the original coordinate axes are perpendicular to each other). This follows from the fact that the sum of products of corresponding terms in v_1 and v_2 is zero.

Therefore, the two vectors v_1 and v_2 can be considered as an alternative set of orthogonal coordinate axes. If we now refer any one of the five points of X to these new axes, for example the second point (4.2, 2.8), the new "coordinates" of this point (i.e., the projections of the point on the v_1, v_2 axes) will be given by $\theta_1 u_1$ and $\theta_2 u_2$, in this case by (19.8360) (.2511) and (1.6040) (−.5113), or 4.9808 and −.8201.

The relative sizes of these two numbers are not coincidental; the projections of the points on the v_1-axis cover a wider range than those on the v_2-axis. In other words, the five points of the design matrix fall predominantly "along" the v_1-axis, and less along the v_2-axis. Note that if we had $\theta_2 = 0$, the coordinate of each of the five points on the v_2-axis would be zero; in that case, all five points would lie on the v_1-line (the line that is perpendicular to v_2 at the origin). We see that the purpose accomplished by the SVD is to reorient the coordinate axes in such a way as to make them follow more closely the pattern made by the points of the X matrix themselves. The SVD helps us understand the *structure* of the X matrix.

An entirely analogous interpretation holds for Table 2, but here the vector space of the design variables is three-dimensional. If in Table 2 θ_3 were exactly zero, all the points would lie in the v_1, v_2 plane, that is, in the plane that is perpendicular to v_3 at the origin. Since, for data set A, θ_3 is actually close to zero, the points lie *close to* the v_1, v_2 plane, rather than *in* this plane.

PRINCIPAL COMPONENTS REGRESSION

The main objective of this paper can now be stated in more precise terms. Having introduced the SVD of the matrix X, we now propose to demonstrate the advan-

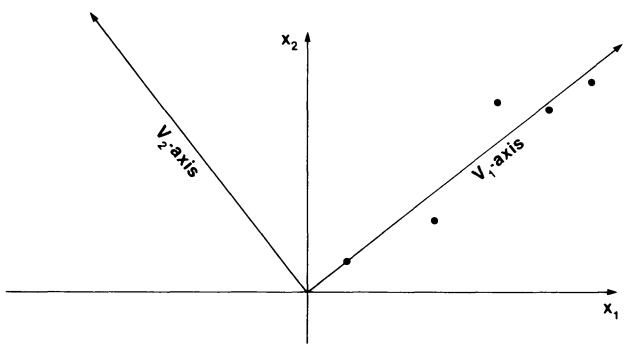

Figure 2. Geometric Interpretation of Singular Value Decomposition in the Case of Two Regressor Variables

tages of replacing X by its SVD in carrying out the regression of Y on X. This procedure is called *principal components regression*. We will see that while this technique can be used in every regression situation covered by (1) as defined above, it is particularly illuminating in the case of *collinearity* or *near collinearity*. These terms will be explained below.

Introducing (5) into (1), we obtain

$$Y = U\theta V'\beta + e. \tag{8}$$

Written in this form, the model is referred to as the *principal components regression model.*

Equation (8) can be written as

$$Y = U(\theta V'\beta) + e \tag{9}$$

where $\theta V'\beta$ is a $r \times 1$ matrix, that is, a vector of r elements. Let us denote this vector by α. Then

$$\alpha = \theta V'\beta \tag{10}$$

and

$$Y = U\alpha + e. \tag{11}$$

The vector Y and the matrix U are known. The least squares solution for the unknown coefficients α is obtained by the usual matrix equation,

$$\hat{\alpha} = (U'U)^{-1}U'Y,$$

which, as a result of (6), becomes

$$\hat{\alpha} = U'Y. \tag{12}$$

This equation is easily solved since $\hat{\alpha}_j$ is simply the inner product of the vector Y with the jth vector u_j.

Through application of (12) we obtain, for the data set A using the U matrix shown in Table 2,

$$\hat{\alpha} = U'Y = \begin{pmatrix} 111.285635 \\ 36.565303 \\ .018803 \end{pmatrix}.$$

It follows from (11) and (12) that

$$\hat{\alpha} = U'(U\alpha + e) = \alpha + U'e \text{ or}$$

$$\hat{\alpha} - \alpha = U'e. \tag{13}$$

It follows that

$$E(\hat{\alpha}_j - \alpha_j) = 0$$

or

$$E(\hat{\alpha}_j) = \alpha_j. \tag{14}$$

Thus, $\hat{\alpha}_j$ is unbiased. Furthermore, the variance of $\hat{\alpha}_j$ is

$$\begin{aligned} \text{var}(\hat{\alpha}_j) &= E(\hat{\alpha}_j - \alpha_j)^2 \\ &= E\Big[\sum_l \sum_t u_{lj} u_{tj} e_l e_t\Big] \\ &= \Big(\sum_l u_{lj}^2\Big)\sigma^2 = \sigma^2. \end{aligned}$$

Hence,

$$\text{var}(\hat{\alpha}_j) = \sigma^2. \tag{15}$$

It is also easily shown that the $\hat{\alpha}_j$ are mutually uncorrelated. Note that whereas the number of elements of β is p, that of α is r, which may be less than p. Now the relation between β and α is given by (10), which also holds for the least squares estimates of α and β,

$$\hat{\alpha} = \theta V'\hat{\beta}. \tag{16}$$

In (16), $\theta V'$ has the dimensions $r \times p$. Thus, given the r values of $\hat{\alpha}$, the matrix relation (16) represents r equations in the p unknown parameter estimates $\hat{\beta}$. If $r = p$, (the full-rank case), the solution is possible and unique.

In this case V' is a $p \times p$ orthogonal matrix (see (7)). Hence, the solution is given by

$$\hat{\beta} = V'^{-1}\theta^{-1}\hat{\alpha} = V\,\theta^{-1}\hat{\alpha}. \tag{17}$$

Note that $\theta^{-1}\hat{\alpha}$ is a $p \times 1$ vector, obtained by dividing each $\hat{\alpha}_j$ by the corresponding θ_j. Applying (17) to data set A, using the SVD of X of Table 2, we obtain at once

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = \begin{pmatrix} .053067 & .067340 & .996317 \\ .998579 & -.008360 & -.052622 \\ .004786 & .997695 & -.067688 \end{pmatrix} \times \begin{pmatrix} \hat{\alpha}_1/52.347807 \\ \hat{\alpha}_2/7.853868 \\ \hat{\alpha}_3/.055690 \end{pmatrix} \tag{18}$$

An important use of (17), apart from its supplying the estimates of β as functions of the $\hat{\alpha}$, is the ready calculation of the variances of the $\hat{\beta}_j$. We will illustrate this calculation in terms of the numerical relations (18). Thus, we obtain from (18)

$$\hat{\beta}_1 = (.053067)\frac{\hat{\alpha}_1}{52.347807} + (.067340)\frac{\hat{\alpha}_2}{7.853868} + (.996317)\frac{\hat{\alpha}_3}{.055690}.$$

In general notation this equation is written as

$$\hat{\beta}_j = \sum_{k=1}^{p} v_{jk}\frac{\hat{\alpha}_k}{\theta_k}. \tag{19}$$

Since the $\hat{\alpha}_j$ are mutually orthogonal and have all variance σ^2, we see that

$$\text{var}(\hat{\beta}_j) = \Big(\sum_{k=1}^{p}\frac{v_{jk}^2}{\theta_k^2}\Big)\sigma^2. \tag{20}$$

Applied to our data, for β_1, (20) becomes

$$\text{var}(\hat{\beta}_1) = \Big(\frac{(.053067)^2}{(52.347807)^2} + \frac{(.067340)^2}{(7.853868)^2} + \frac{(.996317)^2}{(.055690)^2}\Big) \times \sigma^2. \tag{20a}$$

The numerators in each term are the squares of the elements in the first row of the V matrix, that is, values between 0 and 1. But the denominators are the squares of θ_j. Now, we note that θ_3 is considerably smaller than θ_1 and θ_2, so that the third term in (20) contributes an unduly large portion to the variance of $\hat{\beta}_1$ (and also to the variances of $\hat{\beta}_2$ and $\hat{\beta}_3$). In fact, we find from (20a)

$$\text{var}(\hat{\beta}_1) = [(1.03 \times 10^{-6}) + (74.4 \times 10^{-6}) + (320)]\cdot\sigma^2. \tag{21}$$

The reason for this unfavorable state of affairs is the very small value of θ_3 (as compared to those of θ_1 and θ_2). We see that the use of the SVD technique allows us to pinpoint the cause or causes for the large variances found for some coefficients. In our example, the value

of θ_3 may be considered, for all practical purposes, to be equal to zero. But, a θ-value equal to zero has important consequences for the interpretation of regression results, as we will see shortly. We therefore interrupt the discussion of our numerical example to explore the consequences of a zero or near-zero value of θ.

COLLINEARITY AND ITS EFFECTS ON REGRESSION

Since the θ's are the square roots of the eigenvalues of the $X'X$ matrix, a zero θ value implies a zero eigenvalue. This in turn implies that a linear relationship exists between at least some of the p x-vectors of the X matrix. A *near-zero* θ-value consequently implies an *approximate* linear relationship between at least some of these p x-vectors.

Figure 3 shows that an approximate linear relation, namely

$$15x_{i1} - .75x_{i2} - x_{i3} \cong 0 \text{ for all } i, \qquad (22)$$

holds among the columns of X (recall $x_{i1} \equiv 1$).

We have therefore discovered the relation that causes the very small value for θ_3. This can, however, also be accomplished without making graphs.

First we observe that the meaning of (22) is more readily grasped if we consider a Euclidean space in three dimensions, with axes labeled x_1, x_2, and x_3. For each i, the triplet (x_1, x_2, x_3) represents a point in this space. (See Figure 2 for a two-dimensional analog.) Equation (22) simply means that all N points *lie in a single plane* (just as a linear relation between two x-variables indicates that all points lie on a straight line). They are therefore *coplanar,* an extension of the concept of *collinearity* (points on the same line). The existence of a linear relation such as (22) is however always referred to as collinearity (which is thus used as a generic term, including a generalization of the more limited case of points on the same line).

To study the effects of collinearity on the regression analysis, we introduce a second set of data, labeled data set B, which is shown in Table 4. Set B is merely a modified form of set A (Table 1): the relation between x_2 and x_3 is now an *exact* straight line.

The SVD of the X matrix of data set B is shown in Table 5. The value of θ_3 is now exactly zero, so that the rank of the matrix is 2, rather than 3. We now have the case $r < p$.

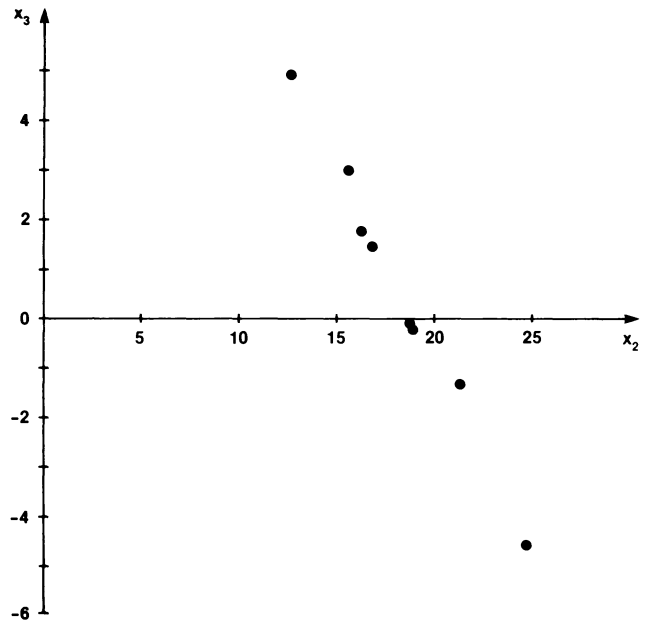

Figure 3. Near Collinearity for Three Regressor Variables x_1, x_2, and x_3 When $x_1 = 1$. The (x_2, x_3) Points Fall Close to a Straight Line

Using (12), we can calculate $\hat{\alpha}$ which now consists of only two values:

$$\hat{\alpha} = \begin{pmatrix} 111.524562 \\ 35.628363 \end{pmatrix}. \qquad (23)$$

Even though there are only two α-values, there still are three β-values, one for each of the three x-vectors. We try to obtain estimates for these β-values by using (10). Writing

$$(\theta V')\hat{\beta} = \hat{\alpha}, \qquad (24)$$

Table 4. Data Set B

Point	x_1	x_2	x_3	y
1	1	16.85	2.3625	41.38
2	1	24.81	−3.6075	31.01
3	1	18.85	.8625	37.41
4	1	12.63	5.5275	50.05
5	1	21.38	−1.0350	39.17
6	1	18.78	.9150	38.86
7	1	15.58	3.3150	46.14
8	1	16.30	2.7750	44.47

$x_3 = 15\,x_1 - .75\,x_2$

Table 5. SVD of X Matrix of Data Set B; X = UθV′

$$X = \begin{pmatrix} .323879 & .193143 \\ .469906 & -.598138 \\ .360569 & -.005670 \\ .246463 & .612642 \\ .406982 & -.257171 \\ .359285 & .001287 \\ .300581 & .319391 \\ .313789 & .247817 \end{pmatrix} \quad \begin{matrix} 52.406330 \\ 0 \end{matrix} \quad \begin{matrix} 0 \\ 8.036461 \end{matrix} \quad \begin{matrix} .053074 & .997433 & .048047 \\ .063871 & -.051408 & .996633 \end{matrix}$$

we obtain

$$\begin{pmatrix} 2.781414 & 52.271803 & 2.517967 \\ .513297 & -.413138 & 8.009402 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = \begin{pmatrix} 111.524562 \\ 35.628363 \end{pmatrix} \quad (25)$$

Equation (25) represents two equations in three unknowns, and does not yield a unique solution for the β. However, it allows us to express any two of the three $\hat{\beta}$-values as a function of the third.

To treat the general case, let us denote the $r \times p$ matrix $(\theta V')$ by Z,

$$Z \equiv \theta V', \quad (26)$$

and partition Z into Z_A and Z_B, the dimensions of which are respectively $r \times r$ and $r \times (p - r)$. Partitioning the vector β accordingly, into $r \times 1$ and $(p - r) \times 1$, we have

$$(Z_A Z_B) \begin{pmatrix} \hat{\beta}_A \\ \hat{\beta}_B \end{pmatrix} = (\hat{\alpha}) \quad , \quad (27)$$

which can be written

$$Z_A \hat{\beta}_A + Z_B \hat{\beta}_B = (\hat{\alpha}) \quad . \quad (28)$$

For the data in (25), this equation becomes

$$\begin{pmatrix} 2.781414 & 52.271803 \\ .513297 & -.413138 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} + \begin{pmatrix} 2.517967 \\ 8.009402 \end{pmatrix} (\hat{\beta}_3) = \begin{pmatrix} 111.524562 \\ 35.628363 \end{pmatrix} \quad . \quad (29)$$

Premultiplying both sides of (28) by Z_A^{-1} (the inverse of an $r \times r$ nonsingular matrix), we obtain

$$\hat{\beta}_A + Z_A^{-1} Z_B \hat{\beta}_B = Z_A^{-1} \hat{\alpha} \quad . \quad (30)$$

Equation (30) shows that, once a value for $\hat{\beta}_B$ has been arbitrarily selected, $\hat{\beta}_A$ is uniquely determined for this selection. Thus, in (29), $\hat{\beta}_1$ and $\hat{\beta}_2$ are uniquely determined for any arbitrarily selected value of $\hat{\beta}_3$. We could of course also have chosen either $\hat{\beta}_1$ or $\hat{\beta}_2$ as the arbitrarily selected parameter.

PREDICTION IN THE CASE OF COLLINEARITY

Continuing with our quest for the consequences of collinearity, let us consider a "new" point x, for which we wish to estimate $\hat{y}$. We have

$$\hat{y} = x \hat{\beta} \quad (31)$$

or, introducing the partitioning of $\hat{\beta}$,

$$\hat{y} = x \begin{pmatrix} \hat{\beta}_A \\ \hat{\beta}_B \end{pmatrix} \quad . \quad (32)$$

Since $\hat{\beta}_A$ is of dimensions $r \times 1$, and x has the dimensions $1 \times p$, we partition x, accordingly, into x_A (of dimensions $1 \times r$) and x_B (of dimensions $1 \times (p - r)$). Thus,

$$\hat{y} = (x_A x_B) \begin{pmatrix} \hat{\beta}_A \\ \hat{\beta}_B \end{pmatrix}$$

or

$$\hat{y} = x_A \hat{\beta}_A + x_B \hat{\beta}_B . \quad (33)$$

Introducing (30) into (33), we obtain

$$\hat{y} = x_A [Z_A^{-1} \hat{\alpha} - Z_A^{-1} Z_B \hat{\beta}_B] + x_B \hat{\beta}_B$$

or

$$\hat{y} = x_A (Z_A^{-1} \hat{\alpha}) + (x_B - x_A Z_A^{-1} Z_B) \hat{\beta}_B \quad . \quad (34)$$

Recall that $\hat{\beta}_B$ cannot be determined by the data, but is arbitrary. In order for (34) to make sense, then, we must insist that the value of $\hat{y}$ be unchanged for *any arbitrary value* of $\hat{\beta}$. This implies that

$$x_B - x_A Z_A^{-1} Z_B = 0 \quad . \quad (35)$$

It is easily seen that $Z_A^{-1} Z_B = (V_A')^{-1} V_B$ where V' is partitioned analogously to Z. (note that V_A is *not* an orthogonal matrix.) Thus, this product does not involve the θ matrix, and (35) can be written as

$$x_B = x_A [V_A'^{-1} V_B] \quad (35a)$$

If (35) is fulfilled, the solution is

$$\hat{y} = x_A (Z_A^{-1} \hat{\alpha}) \quad . \quad (36)$$

It is important to realize that (34) yields two relations: the condition—(35)—and the solution—(36). But the solution is valid only when the condition is fulfilled.

For data set B we have

$$Z_A^{-1} = \begin{pmatrix} 68.206904 \\ -1.495779 \end{pmatrix} \quad (37)$$

$$Z_A^{-1} Z_B = \begin{pmatrix} 15.000 \\ -.750 \end{pmatrix} \quad . \quad (38)$$

Thus, the *condition* becomes

$$x_3 = 15 x_1 - 0.75 x_2 \quad , \quad (39)$$

and the solution, subject to (39), is

$$\hat{y} = 68.207 x_1 - 1.496 x_2 \quad ,$$

which, as a result of (39), can be written as

$$\hat{y} = 1.9144 x_2 + 4.5471 x_3 \quad . \quad (40)$$

Since $x_1 \equiv 1$, (39) is the graph of a straight-line relation between x_2 and x_3. A schematic plot of x_3 versus x_2, using the data of Table 4, is shown in Figure 4 and exhibits this relationship. But our derivation of (39)

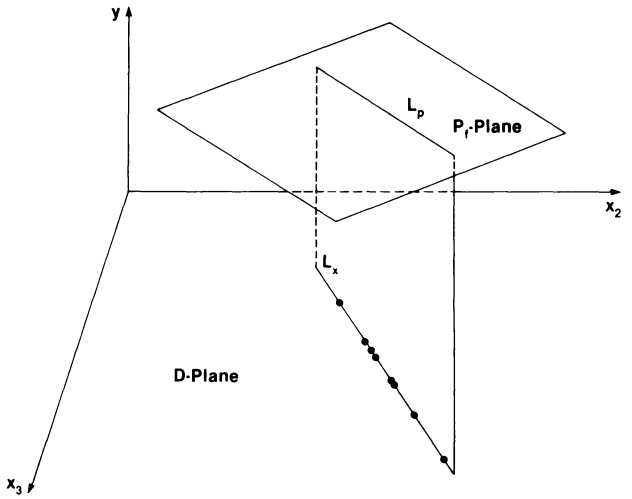

Figure 4. Effect of Collinearity on the Regression Surface

shows that this relation must also hold for any future point (x_1, x_2, x_3) for which an estimate $\hat{y}$ is desired.

Thus, we can obtain a valid estimate for $x = (1 \quad 15 \quad 3.75)$, namely $\hat{y} = 45.767$; but not for the point $x = (1 \quad 20 \quad 3.75)$. The reason is readily understood in terms of Figure 4. The X matrix on which our estimation process is based is such that an exact relationship exists between x_2, and x_3, represented by a straight line in the D plane of the (x_2, x_3) points. Call this line L_X. The regression problem is to find a plane P_f in the (x_2, x_3, y) space, possibly with a nonzero intercept, that best fits the observed y. But since all the y occur at points (x_2, x_3) that fall along the line L_x, the desired plane is defined only in its intersection with the vertical plane erected on L_X. This intersection is denoted by L_p in Figure 4, and it is apparent that the plane to be found can freely swivel around this intersecting line, yielding no unique value of $\hat{y}$ for any point that is not situated on L_X.

To summarize: if $r < p$, there is no unique solution for $\hat{y}$ for an x point, except in the case in which this x point fulfills the collinearity condition resulting from the zero θ-value. This condition is given by (35), and if it is fulfilled the solution is given by (36).

THE CASE OF NEAR COLLINEARITY

Regression and Prediction

We now return to our data set A (Table 1), in which the X matrix was full-rank but had one very small θ-value. Here, the line L_X discussed in the previous section does not strictly exist, but all eight points of the X matrix fall very close to such a line. Strictly speaking, the plane to be found does not swivel freely now, and estimates for $\hat{y}$ can be obtained for any point x. However, because of the unfortunate choice of the points defining the X matrix, the exact position of the desired plane is poorly known, except in the close vicinity of its intersection with the vertical plane erected on the approximate L_X line.

To prove this contention, consider a "new" point x $(1 \times p)$. We first express this point in equivalent u-coordinates, by using the basic SVD equation

$$X = U\theta V' \quad .$$

When applied to a single point x (e.g., a single row in X, but also any "new" point of p elements), this equation yields:

$$x = u\,\theta V' \quad , \tag{41}$$

where both x and u are vectors of p elements. From (41) we derive (since $V^{-1} = V'$ for the full-rank case)

$$u = x\,V\theta^{-1} \quad . \tag{42}$$

We will further discuss this relation below, but first deal with the prediction of $\hat{y}$ at the point x. Expressing the point in u coordinates, given by (42), we have

$$\hat{y} = u\hat{\alpha} \quad , \tag{43}$$

and consequently (see (15))

$$\text{var}(\hat{y}) = \sigma^2 \sum_{j=1}^{p} u_j^2 \quad . \tag{44}$$

This equation shows that even a single numerically large component of the u-vector can substantially increase the variance of $\hat{y}$.

We now investigate under what circumstances the vector u can have large components.

From the basic relation $X = U\theta V'$ we obtain

$$U\theta = XV. \tag{45}$$

Let us divide the θ's into two groups, θ_A and θ_B, such that the latter contains all the θ values that are considerably smaller than the others. For example, in Table 2 we would make

$$\theta_A = \begin{pmatrix} 52.347807 & 0 \\ 0 & 7.853868 \end{pmatrix} \tag{46}$$

and

$$\theta_B = .055690 \tag{47}$$

In this case θ_B consists of a single value, but in general it could contain, say, l values. Then θ_A is a diagonal matrix of $p - l$ values. Let

$$p - l = t. \tag{48}$$

Then θ_A is a square, diagonal $t \times t$ matrix and θ_B a square, diagonal $l \times l$ matrix. Partitioning both θ and V in (45), we have

$$U \begin{pmatrix} \theta_A & 0 \\ 0 & \theta_B \end{pmatrix} = X(V_A\, V_B) \quad , \tag{49}$$

where V_A is $p \times t$ and V_B is $p \times l$. This equation can be written

$$\left(U\begin{pmatrix} 0 \\ \theta_A \end{pmatrix} \; U\begin{pmatrix} 0 \\ \theta_B \end{pmatrix} \right) = (XV_A \quad XV_B) \quad . \tag{50}$$

Now, because all the θ-values in the B group are very small, the columns represented by $U\binom{0}{\theta_B}$ contain elements that are all very small. The same is then necessarily true of all the elements in XV_B. Thus the smallness of the θ-values in the B group implies that

$$XV_B \cong 0 \quad , \tag{51}$$

where $\cong$ represents approximate equality. Equation (51) is the *condition* imposed on the X matrix by the smallness of the θ's in the B group, and represents one or more linear relations between the columns of X. For example, for Table 2, we have

$$XV_B = X \begin{pmatrix} .996317 \\ -.052622 \\ -.067688 \end{pmatrix} \tag{52}$$

implying that for every row in X,

$$.996317x_1 - .052622x_2 - .067688x_3 \cong 0 \quad . \tag{53}$$

This can be verified to hold for matrix X, and it is essentially the equation of a line L_X that represents a linear fit to the points (x_2, x_3) (recall that $x_1 \equiv 1$). In fact, by dividing both sides of (53) by the coefficient of x_3, it is seen that (53) is essentially the same as (39).

Returning now to the problem of predicting y for a new point x, we write equation (42) in the form

$$u = (xV_A \;\; xV_B)\theta^{-1} \tag{54}$$

or

$$u = (xV_A \;\; xV_B) \begin{pmatrix} \theta_A^{-1} & 0 \\ 0 & \theta_B^{-1} \end{pmatrix} \quad . \tag{55}$$

The last l elements of u arise from the product

$$(xV_B)(\theta_B^{-1}) \quad ,$$

and because each θ in θ_B is very small and its reciprocal θ^{-1} consequently very large, these elements of u will be very large, causing the variance of $\hat{y}$ to be very large, unless $xV_B \cong 0$, that is, unless (cf. (52)) the new point x satisfies the near-collinearity condition of the X matrix. The further the new point x lies from the L_X line, the larger will be the variance of $\hat{y}$, that is, the poorer will be the precision of the predicted value. This is really the heart of the collinearity problem, when viewed from the standpoint of usability of the estimated regression for prediction purposes.

If the near-collinearity condition is *exactly* fulfilled for the point x, we have from (55) that

$$xV_B = 0, \text{ and } u = xV_A(\theta_A^{-1}) \quad , \tag{56}$$

which implies that the last l elements of u are zero, and

$$\hat{y} = u_1\hat{\alpha}_1 + \ldots + u_t\hat{\alpha}_t \quad . \tag{57}$$

In this equation, only the first t $\hat{\alpha}$-values are involved, the others having zero-multipliers.

For data set A (Table 1), the near-collinearity condition is given by Eq. (53), and the predicted $\hat{y}$ for x satisfying (53) (considered as an *equality*) is

$$\begin{aligned} \hat{y} = u\hat{\alpha} &= x(V_A)(\theta_A^{-1})(\hat{\alpha}) \\ &= x \begin{pmatrix} .053067 & .067340 \\ .998579 & -.008360 \\ .004786 & .997695 \end{pmatrix} \\ &\quad \begin{pmatrix} \frac{1}{52.347807} & 0 \\ 0 & \frac{1}{7.853868} \end{pmatrix} \begin{pmatrix} 111.524562 \\ 36.628363 \end{pmatrix} \\ &= x \begin{pmatrix} .418452 \\ 2.089422 \\ 4.536091 \end{pmatrix} \quad . \end{aligned} \tag{58}$$

To summarize: The near-collinear case is characterized by one or more very small (though nonzero) θ-values. The rank of X is p; hence predictions can in principle be made for any x-point for which the basic model is known (or assumed) to be valid. However, the precision of the predicted value becomes poorer and poorer as the x-point departs more and more from the near-collinearity condition given by the equation

$$xV_B \cong 0 \quad . \tag{59}$$

When the equality sign holds exactly, the predicted $\hat{y}$ for an x-point satisfying this condition is

$$\hat{y} = xV_A(\theta_A^{-1})\hat{\alpha} \quad . \tag{60}$$

The poorer the approximation in (59), the poorer is the precision of $\hat{y}$ at that point.

An important remark must be made at this point. If the θ-values in group B are very small, the corresponding portion of the V matrix, namely V_B, will be known with very poor numerical precision and (59) may then contain very large rounding errors. In that case, it is far better to act as though θ_B were exactly equal to zero and to express the near-collinearity condition through the use of (35). Using data set A as an illustration (though in this case (59) was adequate), we obtain, by ignoring θ_3 and the third row of V',

$$Z_A^{-1}Z_B = \begin{pmatrix} 14.7193 \\ -.7774 \end{pmatrix}$$

and consequently, applying (35), we obtain the collinearity condition

$$x_3 = 14.7193x_1 - .7774x_2 \quad ,$$

which is equivalent to (53).

Biased Estimation

We have seen that in the case of near-collinearity the variance of $\hat{y}$ increases drastically for x-points that are appreciably removed from the subspace of X in which the points of the design matrix essentially lie (the subspace defined by the collinearity condition). This condition has led to various attempts to obtain more precise estimates by sacrificing the condition of unbiasedness of the $\hat{y}$ estimator. The proposed procedures are therefore known as "biased estimation." We will deal with only one of the proposed methods: that directly associated with the *principal components regression technique.* Our discussion will however suffice to reveal the basic nature of the problem and of the proposed solutions.

Principal Components Regression

In the following, we assume that we have a near-collinear X matrix. According to (56) and (57), if a point x satisfies the condition $xV_B = 0$, the estimate of y corresponding to this point is given by the first t terms in $\hat{y} = \sum_j u_j\hat{\alpha}_j$, where

$$u = xV\theta^{-1} \quad .$$

While this estimate of y yields the correct least squares solution for points satisfying $xV_B = 0$, it is not the least squares solution for x-points that do not satisfy this condition, that is, for points that do not lie in the subspace defined by the collinearity condition.

Let us, however, denote by $\bar{y}$ the quantity

$$\bar{y} = \sum_{j=1}^{t} u_j\hat{\alpha}_j \quad , \tag{61}$$

regardless of whether the condition $xV_B = 0$ is or is not satisfied. Equation (61) is often called the biased principal component prediction equation. The reason for

considering such an estimate is of course the reduction of variance it accomplishes. We have

$$\text{var}(\hat{y}) = \text{var}(\sum_{j=1}^{p} u_j \hat{\alpha}_j) = \sigma^2 \sum_{j=1}^{p} u_j^2 \quad , \tag{62}$$

$$\text{var}(\tilde{y}) = \text{var}(\sum_{j=1}^{t} u_j \hat{\alpha}_j) = \sigma^2 \sum_{j=1}^{t} u_j^2 \quad . \tag{63}$$

For data set A, making $t = 2$, and considering as an example the point $x = (1 \quad 23 \quad -6)$, we have

$$u = x(V\theta^{-1}) = (.439 \quad -.778 \quad 3.450) \quad ,$$

and consequently

$$\text{var}(\tilde{y}) = [(.439)^2 + (-.778)^2 + (3.45)^2]\sigma^2 = 12.701\sigma^2 \quad ,$$

$$\text{var}(\tilde{y}) = [(.439)^2 + (-.778)^2]\sigma^2 = .7980\sigma^2 \quad .$$

In this case, then, the reduction in the variance of the estimate of y is very large.

On the other hand the estimator $\hat{y}$ is unbiased, while $\tilde{y}$ is a biased estimator. We have

$$E(\hat{y}) = E(y) \quad , \tag{64}$$

$$E(\tilde{y}) = E(y) - \sum_{j=t+1}^{p} u_j \alpha_j \quad . \tag{65}$$

Hence

$$\text{Bias}\ (\tilde{y}) = E(\tilde{y}) - E(y) = -\sum_{j=t+1}^{p} u_j \alpha_j \quad . \tag{66}$$

Denoting by MSE the mean squared error, defined by

$$\text{MSE} = \text{Variance} + (\text{Bias})^2 \quad , \tag{67}$$

we have

$$\text{MSE}(\hat{y}) = \sigma^2 \sum_{j=1}^{p} u_j^2 \quad , \tag{68}$$

$$\text{MSE}(\tilde{y}) = \sigma^2 \sum_{j=1}^{t} u_j^2 + (-\sum_{j=t+1}^{p} u_j \alpha_j)^2 \quad . \tag{69}$$

By using $\tilde{y}$ in the place of $\hat{y}$, we actually "trade" the reduction in variance, equal to $\sigma^2 \sum_{j=t+1}^{p} u_j^2$, for the introduction of a (bias)^2 equal to $(-\sum_{j=t+1}^{p} u_j \alpha_j)^2$.

Let us examine the case where $t = p - 1$ (occurring when only one θ-value is very small). Then the reduction in variance is $\sigma^2 u_p^2$ and the (bias)$^2 = u_p^2 \alpha_p^2$. Thus, $\tilde{y}$ will have a smaller MSE than $\hat{y}$ if, and only if, (bias)^2 < variance reduction, that is if

$$u_p^2 \alpha_p^2 < \sigma^2 u_p^2$$

or

$$\alpha_p^2 < \sigma^2 \quad . \tag{70}$$

Unfortunately, both α_p and σ are unknown parameters, for which only estimates are available. The estimate for α_p is of course $\hat{\alpha}_p$, and the estimate for σ^2 is the residual variance after fitting $\hat{\alpha}_1$, $\hat{\alpha}_2$, and $\hat{\alpha}_3$. We have, from least squares theory,

$$\hat{\sigma}^2 = \frac{\left(\sum_{i=1}^{N} (y_i - \hat{y}_i)^2\right)}{(N-p)} \quad , \tag{71}$$

which, when using the $\hat{\alpha}$ estimators, becomes

$$\hat{\sigma}^2 = \frac{\left(\sum_{i=1}^{N} (y_i - \hat{y}_i)^2\right)}{(N-p)} = \frac{\left(\sum_{i=1}^{N} y_i^2 - \sum_{j=1}^{p} \hat{a}_j^2\right)}{(N-p)} \quad . \tag{72}$$

For data set 1, we have ($p = 3$)

$$\hat{\alpha}_3 = .018803,$$

$$\hat{\sigma} = \sqrt{\frac{13729.3041 - 13721.5143}{8-3}} = 1.248 \quad .$$

Undoubtedly, in this case, $\hat{\alpha}_3 < \hat{\sigma}$, but this does not necessarily imply that $\alpha_3 < \sigma$. Indeed, the standard error of $\hat{\alpha}_3$ is σ (cf. (15)); if $\hat{\alpha}_3$ is of the same order of magnitude as σ, it will be virtually impossible to decide, on the basis of the data alone, whether condition (70) is satisfied. To decide whether $\tilde{y}$ is preferable to $\hat{y}$, that is whether MSE($\tilde{y}$) < MSE($\hat{y}$), it is therefore often necessary to make further assumptions. For example, if we *assume* that $\alpha_3 = 0$, then the $\tilde{y}$ estimator is obviously the one to use. The hypothesis $\alpha_3 = 0$ can be tested in principal component regression using Student's t test,

$$t = \frac{\hat{\alpha}_3 - \alpha_3}{S_{\hat{\alpha}_3}} = \frac{\hat{\alpha}_3 - \alpha_3}{\hat{\sigma}} \quad , \tag{73}$$

which, for the hypothesis $\alpha_3 = 0$, becomes

$$t = \frac{\hat{\alpha}_3}{\hat{\sigma}}, \tag{74}$$

For data set 1, we obtain

$$t = \frac{.088}{1.248} = .0151$$

The hypothesis $\alpha_3 = 0$ cannot be rejected, but this does not allow us to infer that condition (70) is satisfied.

From the viewpoint of the philosophy of experimentation, another important matter deserves consideration. Our problem involves both an assumption of a linear model, (1), and a set of data. In a "good" experiment, the data themselves generally provide some diagnostic means for testing the model. In the case of near collinearity, these diagnostic means are totally confined to the close vicinity of the subspace defined by the near collinearity. Thus, even if we knew that condition (70) is fulfilled, we would still be uncertain about the validity of the model for points with large u_3-values, that is points far removed from the (v_1, v_2) plane, in the v_3 direction.

Our analysis has led us to the conclusion that linear model inferences based on near-collinear X matrices should be made with great caution.

OTHER APPLICATIONS OF THE SVD TECHNIQUE

We have seen that the detection and treatment of collinearity and near collinearity is greatly facilitated through the use of the singular value decomposition. This interesting technique has other important uses in data analysis, such as in the study of the structure of two-way tables (Bradu and Gabriel 1978; Mandel 1971), the evaluation of experimental designs in regression (Hahn, Meeker, and Feder 1976), and the em-

pirical fitting of functions of two or more arguments (Mandel 1981).

SUMMARY AND CONCLUSION

We have presented algebraic and geometric aspects of multiple linear regression, based primarily on the singular value decomposition technique, and shown that difficulties of interpretation arise when constraints (such as exact or approximate linear relations) exist between the regressor variables. Linear constraints are known as collinearity. Near collinearity manifests itself in the form of one or more very small singular values.

Under linear constraints, the true coefficients of the relation between the response and the regressor variables cannot be estimated unambiguously without introducing additional assumptions. Nevertheless, it is possible to make valid predictions of the response, provided that the point for which the prediction is made lies in the same subspace as the points on which the regression calculations were based.

Even though the coefficients of the regression equation cannot be estimated precisely when applying the least squares technique to the case of collinearity, or near collinearity, certain linear combinations of the coefficients can be estimated with confidence.

[*Received August 1980. Revised July 1981.*]

REFERENCES

BELSLEY, D. A.; KUH, E.; and WELSCH, R. E. (1980), *Identifying Influential Data and Sources of Collinearity*, New York: John Wiley.

BRADU, D., and GABRIEL, K. R. (1978), "The Biplot as a Diagnostic Tool for Models of Two-Way Tables," *Technometrics*, 20, 47–68.

CHAMBERS, J. M. (1977), *Computational Methods for Data Analysis*, New York: John Wiley.

CHATTERJEE, S., and PRICE, B. (1977), *Regression Analysis by Example*, New York: John Wiley.

DRAPER, N. R., and SMITH, H. (1981), *Applied Regression Analysis* (2nd ed.), New York: John Wiley.

HAHN, G. J.; MEEKER, W. Q. Jr.; and FEDER, P. I. (1976), "The Evaluation and Comparison of Experimental Design for Fitting Regression Relationships," *Journal of Quality Technology*, 8, 140–157.

HOCKING, R. R.; SPEED, F. M.; and LYNN, M. J. (1976), "A Class of Biased Estimators in Linear Regression," *Technometrics*, 18, 425–437.

HOERL, A. E., and KENNARD, R. W. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55–67.

MANDEL, J. (1971), "A New Analysis of Variance Model for Non-Additive Data," *Technometrics*, 13, 1–18.

——— (1981), "Fitting Curves and Surfaces with Monotonic and Non-Monotonic Four Parameter Equations," *Journal of Research of the National Bureau of Standards*, 86(1), 1–25.

MARQUARDT, D. W. (1970), "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation," *Technometrics*, 12, 591–612.

MARQUARDT, D. W., and SNEE, R. D. (1975), "Ridge Regression in Practice," *The American Statistician*, 29, 3–20.

RAO, C. R. (1973), *Linear Statistical Inference and Its Applications*, New York: John Wiley.

SNEE, R. D. (1973), "Some Aspects of Nonorthogonal Data Analysis, Part 1. Developing Prediction Equations," *Journal of Quality Technology*, 5, 67–79.

STEIN, C. M. (1960), "Multiple Regression Contributions to Probability and Statistics," in *Essays in Honor of Harold Hotelling*, ed. I. Olkin, Stanford: Stanford University Press, 424–443.

WEBSTER, J. T.; GUNST, R. F.; and MASON, R. L. (1974), "Latent Root Regression Analysis," *Technometrics*, 16, 513–522.