NBER WORKING PAPER SERIES

DETECTING AND ASSESSING THE PROBLEMS CAUSED BY MULTICOLLINEARITY: A USE OF THE SINGULAR-VALUE DECOMPOSITION

David A. Belsley*
Virginia C. Klema**

Working Paper No. 66

COMPUTER RESEARCH CENTER FOR ECONOMICS AND MANAGEMENT SCIENCE
National Bureau of Economic Research, Inc.
575 Technology Square
Cambridge, Massachusetts 02139

December 1974

Preliminary: not for quotation

NBER working papers are distributed informally and in limited numbers for comments only. They should not be quoted without written permission.

This report has not undergone the review accorded official NBER publications; in particular, it has not yet been submitted for approval by the Board of Directors.

*NBER Computer Research Center and Boston College. Research supported in part by National Science Foundation Grant GJ-1154X3 to the National Bureau of Economic Research, Inc.

**NBER Computer Research Center. Research supported in part by National Science Foundation Grant GJ-1154X3 to the National Bureau of Economic Research, Inc.

Abstract

This paper presents a means for detecting the presence of multicollinearity and for assessing the damage that such collinearity may cause estimated coefficients in the standard linear regression model. The means of analysis is the singular value decomposition, a numerical analytic device that directly exposes both the conditioning of the data matrix X and the linear dependencies that may exist among its columns. The same information is employed in the second part of the paper to determine the extent to which each regression coefficient is being adversely affected by each linear relation among the columns of X that lead to its ill conditioning.

Acknowledgments

The authors wish to express their gratitude to Professor Gene Golub of Stanford University, Professor John Dennis of Cornell, and Edwin Kuh of the NBER for many helpful discussions. Moreover, the first author wishes to express his gratitude to the Center for Advanced Studies in the Behavorial Sciences at Stanford for the opportunity to initiate his research in this area during his fellowship there.

Contents

Addenda to bibliography, p. 49

Becker, R., Kaden, N., and Klema, V., [1974], "The Singular Value Analysis in Matrix Computation", NBER Working Paper 46.

Golub, G. H., [1969], "Matrix Decomposition and Statistical Calculation", in R. C. Milton and J. A. Nelder (eds.), Statistical Computation, Academic Press 365-397.

Golub, G. H., and Kahan, W., [1965], "Calculating the Singular Values and Pseudo-Inverse of a Matrix", J. SIAM Numer. Anal., Ser. B. Vol. 2, No. 2, 205-224.

Golub, G. H., and Reinsch, C., [1971], "Singular Value Decomposition and Least Squares Solutions," in J. H. Wilkinson and C. Reinsch (eds.), Handbook for Automatic Computation, Volume II: Linear Algebra, Springer Verlag, 134-151.

Hanson, R. and Lawson, C. L., [1969], "Extensions and Applications of the Householder Algorithm for Solving Linear Least Squares Problems," Mathematics of Computation, vol. 23, no. 1080, 782-812.

Errata

page 6. line 13 triangulation → triangularization

line 16 replace line 16 with from Golub and Kahan (1965) and Wilkinson (1965) p. 195 illustrate this point.

page 7. line 3 posses → possesses

INTRODUCTION

There are three major questions related to the problem of multicollinearity: when does it exist? how much damage has it caused? and what, if anything, can be done about it? Making use of a technique of numerical analysis, the singular-value decomposition, this paper suggests a means for answering the first two of these questions that is devoid of the ad hoc quality of previous attempts. Part 1 introduces the concept of the singular-value decomposition and applies it to the determination of the existence of linear dependencies among the columns of any given data matrix X. An Appendix to Part 1 deals with the problems caused by scaling of the data matrix. Part 2 addresses the question of assessing the damage caused by the presence of multicollinearity and applies the understanding gained from Part 1 toward an answer. Part 3 presents an assessment of several of the techniques previously advanced in the literature for diagnosing collinearity and, additionally, presents a fundamental critique against the use of non-Baysian "ridge regression" as a means of correcting the problems caused by collinear data. While some contrived examples are provided for illustration, a true study of the application of these techniques to economic data will be the subject of a future paper.

Part 1. The Singular-Value Decomposition and The Detection of Linear Dependencies

1.1 The Singular-Value Decomposition

We learn from the numerical analysts[1] that any TxK matrix X, considered here to be a matrix of T observations of K economic variates, may be decomposed as

$$X = U\Sigma V' \tag{1.1}$$

where $U'U = V'V = I_K$ and Σ is diagonal with non-negative diagonal elements σ_k, $k = 1 \ldots K$.[2,3]

---

[1] See, for example, Golub (1969), Golub and Reinsch (1970), Hanson and Lawson (1969), and Becker et al (1974).

[2] This decomposition is efficiently and stably effected by a program called MINFIT [Golub and Reinsch (1970)].

[3] In (1) U is TxK, Σ is KxK and V is KxK. Alternative formulations are also possible and may prove more suitable to other applications. Hence one may have

$$\underset{TxK}{X} = \underset{TxT}{U}\ \underset{TxK}{\Sigma}\ \underset{TxK}{V'} \tag{1.1a}$$

or

$$\underset{TxK}{X} = \underset{Txr}{U}\ \underset{rxr}{\Sigma}\ \underset{rxK}{V'} \tag{1.1b}$$

where $r = \rho(X)$. In this latter formulation Σ is always of full rank, even if X is not.

The singular-value decomposition is closely related to the familar concepts of eigenvalues and eigenvectors, but its difference from those concepts is important. The non-negative diagonal elements of Σ are called the singular values of X, and these are also the non-negative square roots of the eigenvalues of X'X. This is readily seen by noting

$$X'X = V\Sigma U'U\Sigma V' = V\Sigma^2 V'. \tag{1.2}$$

Recalling the orthonormality of V, we note that V diagonalizes X'X, and hence the diagonal elements of Σ^2 must be the eigenvalues of the real symmetric matrix X'X.

Equally clear, the orthonormal columns of V must be the eigenvectors of X'X, and, as is similarly demonstrated, the columns of U must be the eigenvectors of XX'.

The singular-value decomposition does not, however, merely duplicate knowledge of the eigensystem of X'X, for the singular value decomposition applies directly to the data matrix X, and not to the moment matrix X'X. The singular-value decomposition thus leads to a means of determining the linear dependencies, if any, among the columns of the data matrix X.

1.2 The Determination of the Linear Dependencies of X.

Assume that X is rank deficient, i.e., $\rho(X) = r < K$. Since U and V are orthonormal, and hence necessarily of full rank, we must have $\rho(X) = \rho(\Sigma)$. There will, therefore, be as many zero elements along the diagonal of Σ as the nullity of X, and hence we may partition the singular-value decomposition in

(1.1) as

$$X = U\Sigma V' = U \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & 0 \end{bmatrix} V' \tag{1.3}$$

where Σ_{11} is rxr and nonsingular.

After postmultiplying (1.3) by V and further partitioning we obtain

$$X [V_1 \; V_2] = [U_1 \; U_2] \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & 0 \end{bmatrix} \tag{1.4}$$

where V_1 is Kxr U_1 is Txr
V_2 is Kx(K-r) U_2 is Tx(K-r).

(1.4) results in the two matrix equations

$$X V_1 = U_1 \Sigma_{11} \tag{1.5}$$

and

$$X V_2 = 0. \tag{1.6}$$

Interest centers on (1.6), for it displays all of the linear dependencies of X: the Kx(K-r) matrix V_2 provides an orthonormal basis for the null space that is spanned by the columns of X.

Two problems arise in applying the exact algebra leading to (1.6) to real data. First, how does one determine the rank of X, r, i.e., how are the zeros of Σ discovered? And second, how are the zeros of V_2 discovered? Both of these problems arise because computers use finite arithmetic, and only in very special cases will "true" zeros be calculated as such. There are problems of both rounding error and error in the representation of the data[1].

---

1. Also sometimes called truncation error. However, this term also applies to the error introduced by truncating an infinite series after a finite number of steps, and hence will not be employed here.

The importance of the first problem is obvious: only through a correct determination of the zeros of Σ can we correctly assess how many linear dependencies exist among the columns of X. The importance of the second problem is less obvious. But, in general, all elements of V_2 will be calculated as non-zeros, however small some may be relative to others. Since scaling of X will alter these non-zero elements arbitrarily (a problem that is dealt with in length in the appendix to this section), we may arrive at the conclusion that many columns of X enter each linear dependency, whether or not this is true. The econometrician will rarely be satisfied with such an answer; he would like to identify the zeros of V_2 (or some manipulation of it) so that he can say which variates do and which variates do not enter into a specific linear relation. The next two sections deal with these two problems in turn.

1.3 Determination of $\rho(X) = r$

The singular value decomposition presents a means for determining the rank of the data matrix X. Referring to (1.1) and recalling that U and V are orthogonal we see that Σ has both the same norm and the same rank as X. Since Σ is diagonal, were there no problems of calculation introduced by the imprecision of the computer, one need only determine the number of nonzero elements of Σ to discover the rank of X. Unfortunately the task is not quite so simple, for the nonexact, finite arithmetic necessarily employed by computers and the problems of rounding error will result in nonzero elements of Σ when, under ideal conditions, they should be zero. It is necessary, therefore, to find a means for determining when an element of Σ is "small enough" to be considered zero, and hence evidence of X's being rank deficient.

Proposed Alternatives. The singular value decomposition is useful in this context of determining rank because it preserves the norm of X (i.e. column lengths). The singular values are in the same units as the columns of X, and hence are measurably interpretable. Other suggested means for determining rank fail on this and other counts .

The determinant of the matrix (if square - or X'X if not) clearly fails, for a small determinant has little to do with the invertability of a matrix. The matrix αI_n has determinant α^n which can be made arbitrarily small, yet it is clear that αI has orthogonal columns and is always invertable for $\alpha \neq 0$.

It is equally infeasible to obtain information on the invertability (conditioning) of a matrix from the smallness of some of the diagonal elements of a triangulation of the given matrix. This process is closely related to the use of the determinant, since the determinant will be the product of the diagonal elements of the triangular factorization. Two examples from Golub and Reinsch (1970) and Wilkinson (1965) illustrate this point. Consider

$$\begin{bmatrix} .501 & -1 & & & \\ & .502 & -1 & & 0 \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \\ 0 & & & .599 & -1 \\ & & & & .600 \end{bmatrix}$$

and

$$\begin{bmatrix} 1 & -1 & -1 & \cdots & -1 \\ & 1 & -1 & \cdots & -1 \\ & & \ddots & & \\ & & & \ddots & \\ & 0 & & & \\ & & & & 1 \end{bmatrix}.$$

Each of these matrices will be shown by the singular value decomposition to be quite ill-conditioned even though neither posses a small diagonal element.

The Condition Number. A means of determining the conditioning of a matrix that avoids the pitfalls mentioned above is afforded by the singular value decomposition. The motivation behind this technique derives from a more correct method of determining whether an inverse of a given matrix "blows up". As we shall see it is reasonable to consider a matrix to be ill-conditioned if its inverse is large in spectral norm[1] in comparison with the spectral norm of the given matrix itself. Two examples aid this point. Consider first the matrix

$$A = \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix}.$$

Clearly as $\alpha \to 1$, this matrix tends toward perfectly singularity. Also the singular values of A are easily shown to be $1 \pm \alpha$, and those of A^{-1} are $(1 \pm \alpha)^{-1}$. Now as $\alpha \to 1$, the product $||A||\ ||A^{-1}|| = \lim_{\alpha \to 1} (1+\alpha)\ (1-\alpha)^{-1}$ explodes, and hence we conclude the norm of A^{-1} is large relative to that of A. A is ill-conditioned for small α.

---

[1] The spectral norm of $A = (a_{ij})$, denoted $||A||$, is simply σ_{max}, the maximum singular value.

By way of contrast, consider the matrix, introduced above,

$$B = \begin{bmatrix} \alpha & 0 \\ 0 & \alpha \end{bmatrix} .$$

There is some feeling that B becomes ill-conditioned as $\alpha \to 0$. However, $||B|| = \alpha$ and $||B^{-1}|| = \alpha^{-1}$, and the product $||B||\ ||B^{-1}|| = \alpha\alpha^{-1} = 1$ is constant as $\alpha \to 0$. In this case, then, the norm of B^{-1} does not blow up relative to that of B, and B is well conditioned for all α.

The conditioning of any square matrix can be summarized, then, by a condition number $\kappa(A)$ defined as the product of the maximal singular value of A times the maximal singular value of A^{-1}. This concept is readily extended to a rectangular matrix and can be calculated without recourse to the inverse matrix. From the singular value decomposition of $X = U\Sigma V'$, it is easily shown that the generalized inverse X^{+} of X is $U\Sigma^{+}V'$, where Σ^{+} is the generalized inverse of Σ and is simply Σ with its nonzero diagonal elements inverted.[1] Hence the singular values of X^{+} are merely the inverses of those of X, and the maximal singular value of X^{-1} is the reciprocal of the minimum (nonzero) singular value of X. We may therefore define the condition number of X as $\kappa(X) = \frac{\sigma_{max}}{\sigma_{min}}$.

The Use of The Condition Number in Determining Rank. We will now discuss the sense in which the condition number has meaning as a measure of the ill-conditioning of a matrix. This will further result in a meaningful criterion for determining when a singular value is small enough (relative to σ_{max}) to provide evidence of a rank deficiency.

---

1. See Golub and Reinsch (1970) or Becker et al. (1974).

Consider the linear system Xb = a, and suppose the data are known exactly, but stored in finite precision. It is shown in Stewart (1973) or Hanson and Lawson (1969) that a change in the last digit of the elements of X can result in a change in $\kappa(X)$ times as great in the solution b. That is, if the machine zero is 10^{-10}, and $\kappa(X)$ is 10^{4}, then a change in X in the tenth decimal place can affect b in the $10^{-10} \times 10^{4}$ or 10^{-6} place. Clearly, then, a condition number sufficiently large can wipe out all significance to a solution to a linear system. Such would be the case if κ were larger than the word length of the machine.

In a least-squares problem, the solution to $X'X\ b=X'y$, a similar result holds, except that now a perturbation in X affects X'X as the square, and we must have the square of the condition number to be like the word length, or, equivalently, the condition number like the square root of the word length.

Rather generally, then, in the least-squares context, we would suppose that any singular value, σ_k, which, relative to the σ_{max}, was less than the square root of the machine zero (the reciprocal of the word length--about 2^{-26} for IBM 360/370 long precision) to be evidence of rank deficiency.

When there is Fuzziness in the Data. The determination of the rank of the data matrix X is less straightforward when the data are known imprecisely--with fuzziness. The analysis of the previous section is based on data known exactly, and from it we learn that a perturbation in the last digit of the data's word length can affect digits on the order of $\kappa(X)$ from the last in the solution for b of a linear system. Thus if the word length is 10^{8} and $\kappa(X)$ is 10^{3}, a change in the eighth digit of X can affect b in the 5th digit, and a $\kappa(X)$ of 10^{8} can remove all significance from b.

When the data are fuzzy, further problems are encountered, because relevant perturbations in the data now affect, not necessarily the last digit it the word length, but possibly much higher order digits. Suppose again a word length of 10^8 and a $\kappa(X) = 10^3$ but the data are known only up to 10^3. Now relevant perturbations of the data as stored in the computer are $10^8 \times 10^{-5} = 10^3$ times greater than perturbations of the last digit of the word length. Hence the solution to the linear system will be known with even less precision, and could[1] be affected in the digits on the order of $\kappa(X) \times 10^3$. In this case that would be 10^6, leaving only the first two digits to be known with any accuracy.

In the least-squares solutions—as contrasted to the solution to a linear system used in the explanation above—the treatment of data fuzziness is quite analogous. If the data in X are exact to, say, 10^3, then the data of X'X are exact to 10^6. A word length of 10^8 now implies that perturbations of the order of $10^8 \times 10^{-6} = 10^2$ are now relavant, and these can in turn be magnified in the least-squares solution by a factor $\kappa^2(X)$, the condition number of X'X. Here, this would be $(10^3)^2 \times 10^2 = 10^8$, and hence the solution b may have no definition at all with an 8 digit word length.

---

1. The word could is used because the figure is an upper bound telling the worst possible story. It could be better in any given case.

The preceding leads to the following suggestion for determining when a singular value is small enough to be considered evidence of rank deficiency when there is fuzziness in the data. Let w be the word length[1], and f be the fuzziness[2] in the data matrix X -- f^2 that of $X'X$.[3] Then the foregoing argues that we must have $wf^{-2}\kappa^2(X) \leq w$ if the least squares solution is to have any meaning (any stable digits) at all. That is we must have $\kappa(X) \leq f$. If the data are known up to 10^3, we can allow X to have $\kappa(X) = \frac{\sigma_{max}}{\sigma_{min}} \leq 10^3$.

Hence any σ_k such that $\frac{\sigma_{max}}{\sigma_k} \leq 10^3$ (=f) would indicate the possibility of rank deficiency.

---

1. w can be measured as 10^{ℓ}, where ℓ is the number of digits carried by the machine.
2. f can be measured as 10^h, where h is the number of places known with exactness.
3. Provided $X'X$ is accumulated in double precesion relative to that of X.

1.4 Determining the Structure of the Linear Dependencies of X.

1.4.1 Defining the Structure

In this subsection we assume we have already determined the rank of X as described in the previous subsection. Our interest here centers on determining which variates do and which do not enter any specific linear dependency. It is this information that is meant by the term structure of the linear dependency. It is not sufficient to examine the zero structure of V_2 in (1.6) to determine the structure of the linear dependencies, for clearly, for any $(k-r)^2$ nonsingular matrix A, (1.6) becomes

$$X V_2 A = 0, \tag{1.7}$$

and we can alter the zero structure of these linear dependencies (given by the zeros of the matrix V_2A) arbitrarily. Rather we must rework (1.6) into a form that is invariant to linear transformations. This is accomplished by partitioning (1.6) to produce a "reduced form" as follows:

$$X V_2 \equiv [X_1 X_2] \begin{bmatrix} V_{21} \\ V_{22} \end{bmatrix} = 0 , \tag{1.8}$$

where X_1 is T x (k-r) $\quad$ V_{21} is (k-r) x (k-r)
$\quad\quad X_2$ is T x r $\quad\quad$ V_{22} is r x (k-r)

and V_{21} is chosen to be nonsingular. Since V_2, having orthonormal columns, is of full rank, such a nonsingular submatrix must exist. From (1.8) we obtain

$$X_1 = - X_2 V_{22} V_{21}^{-1} \equiv X G \tag{1.9}$$

where $G \equiv - V_{22} V_{21}^{-1}$.

The structure of (1.9) is clearly invariant to linear transformations since $X\ V_2A = 0$ implies $[X_1X_2]\begin{bmatrix} V_{21} \\ V_{22} \end{bmatrix} A = 0$ or $X_1 = -X_2V_{22}A\ A^{-1}V_{21}^{-1} = -\ X_2G$. The determination of the structure of the linear dependencies of X therefore is precisely the determination of the zero structure of the matrix G. From it we learn which columns of X_2 are involved in linear relationships with the variates composing the columns of X_1.

Unfortunately we cannot simply calculate G and look for its zeros, for, as already mentioned, the finite arithmetic used in determining V_2— now further compounded by the calculations determing G as $-V_{22}V_{21}^{-1}$ —will not guarantee that the zeros of G will indeed be calculated as zero.

1.4.2 Determining the Zeros of G.

Two methods are suggested here for giving numerical specification to the zeros of G.[1] The first is a linear-programming approach, the second a least-squares approach. Both methods are based upon the following rationale. Linear dependencies are exact only in perfect algebra. The econometrician has always sought to extend this concept to one of "near dependency", a notion that has been more intuitive than rigorous. In the previous section, however, we saw how "nearness" could be given meaning in a realistic context both by the natural fuzziness given by a "machine zero", and by the more usually encountered fuzziness that results from data inaccuracies. This latter concept requires some discussion.

---

[1] The authors are greatly indebted to Gene Golub of Stanford University and John Dennis of Cornell University for their contributions to these techniques.

Observational Equivalence

A published GNP figure of 1.054 trillion dollars is clearly not exact. Indeed all additional information regarding digits beyond 10^{-3} have been suppressed. The datum 1.054 is therefore observationally indistinguishable from 1.0542 or 1.0539. That is,there is some region of fuzziness such that, given normal rounding procedures, any data point lying in that region is equally valid for an entry into X. This concept of truncated data reporting is quite distinct from errors in observation. The latter would argue that one might not know for sure the correctness of the data actually reported. Hence observations error introduces yet another element of fuzziness into the degree of accuracy with which one knows one's data.

In any event there is reason to suppose that there exists a matrix E, determined by the investigator, that puts limits on the accuracy to which he believes he knows his data. These limits may, for example, take the form that "column 6 of X is known only up to 10^{-3}". Hence, any data matrix $\tilde{X}$ such that $|\tilde{X} - X| \leq E$ is observationally equivalent to X.[1] This notion of observational equivalence (which could no doubt also be cast into a statistical framework) is a data-analytic analogue to the identification problem. Given the fuzziness in X, any results based on any $\tilde{X}$ observationally equivalent to X must also be indistinguishable within the degree of precision to which the data are known. Hence the investigator must consider as observationally indistinguishable any $\tilde{V}$ resulting from the singular value decomposition of any appropriate $\tilde{X} = \tilde{U}\tilde{\Sigma}\tilde{V}'$. It is this notion of observational equivalence that is exploited to determine the zeros of G.

---

[1] The notation $|X|$ here is used to mean absolute value of a matrix, not the determinant.

Zero Enrichment

Given the data matrix X, we have from (1.9) that

$$X_1 - X_2G = 0, \tag{1.10}$$

and we propose to determine the zero structure of G by determining whether any of its elements (or specific of its elements) are observationally indistinguishable from (equivalent to) zero. To do this we employ a numeric-analytical analogue to hypothesis testing.[1] It is proposed that the investigator examine the G determined by the singular value decomposition of X and specify which of its elements he has reason to believe to be zero. This may be based upon *a priori* considerations of which variates would not belong in certain linear dependencies (hence implying the corresponding elements of G to be zero) or it may be based on experience he has regarding which values of G that are calculated to be small numerically are in fact zero. In any event the matrix G has, as a matter of hypothesis, certain of its elements made to be zero. The resulting *zero enriched* matrix is denoted $\tilde{G}$. In both of the following procedures a method is presented to test the hypothesized zero enrichment by determining whether $\tilde{G}$ is observationally equivalent to G in the sense that $\tilde{G}$ could indeed by calculated as the G matrix for a data matrix $\tilde{X}$ that is observationally equivalent to X.

Method 1: A Linear-Programming Approach

Let $\Delta = (\delta_{ij})$ be a TxK matrix to be determined. X is the given TxK data matrix and E is the "limits" matrix defined above. G is the matrix defined in

[1]Again a statistical formulation of this procedure may well be possible, but is not examined here.

(1.9) by the singular value decomposition of X and for which (1.10) holds. Partion $\Delta = [\Delta_1 \Delta_2]$ to correspond to X_1 and X_2. $\tilde{G}$ is an hypothesised zero-enriched matrix subject to test. We will say that $\tilde{G}$ is observationally equivalent to G (and hence accept the hypothesised zero enrichment) if there exists a $\Delta = [\Delta_1 \Delta_2]$ such that $\tilde{G}$ satisfies

$$(X_1 + \Delta_1) - (X_2 + \Delta_2)\tilde{G} = 0 \tag{1.11}$$

and

$$|\Delta| \leq E, \tag{1.12}$$

i.e., if $\tilde{G}$ can result from the singular-value decomposition of a data matrix that is observationally equivalent to X.

The existence of such a Δ can be established from the feasibility of a linear program. From (1.11) we have

$$\Delta_1 - \Delta_2\tilde{G} = -(X_1 - X_2\tilde{G}) \tag{1.13}$$

or

$$\Delta H = -XH \tag{1.14}$$

where $H \equiv \begin{bmatrix} I \\ \tilde{G} \end{bmatrix}$.

Using the change of variable

$$\Phi = \Delta + E, \tag{1.15}$$

the problem of finding a Δ that satisfies (1.14) subject to the inequalities (1.12) is equivalent to finding the Φ that satisfies

$$\Phi H = (E-X)H \qquad \text{subject to} \tag{1.16}$$

$$\Phi \leq 2E \quad \text{and} \qquad (1.17)$$
$$\Phi \geq 0.$$

The existence of such a $\Phi = (\phi_{tk})$ is clearly established if there exists a feasible solution to the contrived linear program

$$\min \iota_T' \Phi \iota_K = \sum\sum_{tk} \phi_{tk} \quad [\iota_n, \text{ a vector of n ones}] \qquad (1.18)$$

subject to (1.16 and 1.17).

It is worth emphasizing that it is not necessary to solve the LP (1.18) to accept the hypothesis of the zero enriched $\tilde{G}$, rather it is only required to demonstrate the feasibility of the program.

Method 2: A Least-Squares (minimum norm) Approach

The LP given above will, even for moderate sized economic problems, be large. Even the demonstration of a feasible solution could prove costly, and, hence, a second method appears worthy of consideration.

Our problem is to find a Δ satisfying (1.14) also obeys the inequalities (1.12). Since H in (1.14) necessarily has full rank, we can find all Δ satisfying this relation without regard to (1.12) (in general there will be an infinity of them) by considering all

$$\Delta = - XHH^{-} \qquad (1.19)$$

where H^{-} is any pseudoinverse of H. Among all these solutions, however, is one with minimum norm (i.e., a Δ with minimum $\sum_{i,j} \delta_{ij}^2$), which is found by using the generalized inverse H^{+}, i.e.

$$\Delta^* = -XHH^{+}. \qquad (1.20)$$

There is, of course, no guarantee that Δ^* will satisfy (1.12) in all cases, but there is reason to hope that its property of minimum norm will indeed also

result in (1.12) as a practical matter. This second method of determining Δ, then, is sufficient but not necessary to accept the zero enrichment hypothesis. That is, a solution to (1.20) that also satisfies (1.12) accepts the observational equivalence of $\tilde{G}$ (the hypothesized zero enrichment), but a solution to (1.20) that does not also satisfy (1.12) does not mean that a solution to the LP (1.18) does not exist.[1] The advantage of this technique over the LP is that it is quick and cheap to employ. If it works, no further effort is required. If it doesn't, further investigation may be warranted. It will be a matter for experience to determine just how well this short cut works in practice.

---

1. We are indebted to our colleague, Paul Holland, for highlighting these points.

APPENDIX TO SECTION 1. SCALING

The seemingly elaborate test procedures given in the previous section are motivated by the fact that the elements of G are scale sensitive and can be made arbitrarily small simply by a choice of scale. Determination of the zero structure of G, therefore, requires some meaningful (not arbitrary) measure of small, and this measure is afforded by the procedures outlined.

The purpose of this appendix is to demonstrate this problemful scale sensitivity.

Let X be the data matrix in "original units", and let $D = \text{diag}(d_1 \ \ldots \ d_K)$ be a scaling matrix (all $d_i \neq 0$). Call the scaled data matrix $\hat{X} = XD$. Now (using the notation of the text) the SVD of X is

$$X = U \Sigma V', \text{ implying } X V_2 = 0 \tag{1.21}$$

and that of $\hat{X}$ is

$$\hat{X} = \hat{U} \hat{\Sigma} \hat{V}', \text{ implying } \hat{X} \hat{V}_2 = 0.$$

The reduced forms corresponding to the original and scaled data are therefore

$$X_1 = -X_2 V_{22} V_{21}^{-1} \equiv X_2 G, \qquad G = -V_{22} V_{21}^{-1} \tag{1.23a}$$

and

$$\hat{X}_1 = -\hat{X}_2 \hat{V}_{22} \hat{V}_{21}^{-1} = \hat{X}_2 \hat{G}, \qquad \hat{G} = -\hat{V}_{22} \hat{V}_{21}^{-1} \tag{1.23b}$$

and the econometrician must insist that the zero structure of G be the same as $\tilde{G}$, since arbitrary scaling cannot affect the real linear dependencies.

We will now show that with exact arithmetic, these zero structures are indeed the same, but that they can be made to appear different due to finite arithmetic, hence necessitating the test procedures of Section 1.4.2.

From $X\ V_2 = 0$ we may write

$$X\ D\ D^{-1}\ V_2 \equiv \hat{X}\ D^{-1}\ V_2 = 0. \tag{1.24}$$

Since $\rho(\hat{X}) = \rho(X)$, the null space of $\hat{X}$ must have the same dimension as X, and hence $D^{-1}\ V_2$ provides a basis (not orthonormal) for the null space of $\hat{X}$. Hence any orthonormal basis for this null space (such as $\hat{V}_2$) must be a non-singular transformation of $D^{-1}V_2$. Let this be

$$\hat{V}_2 \equiv D^{-1}\ V_2\ H \quad \text{or} \quad \begin{bmatrix} \hat{V}_{21} \\ \hat{V}_{22} \end{bmatrix} = \begin{bmatrix} D_1^{-1} & 0 \\ 0 & D_2^{-1} \end{bmatrix} \begin{bmatrix} V_{21} \\ V_{22} \end{bmatrix} H = \begin{bmatrix} D_1^{-1}V_{21}H \\ D_2^{-1}V_{22}H \end{bmatrix} \tag{1.25}$$

for H nonsingular.[1]

Putting (1.25) into (1.23b) gives

$$\begin{aligned} \hat{X}_1 &= -\hat{X}_2\ \hat{V}_{22}\ \hat{V}_{21}^{-1} = -\hat{X}_2\ D_2^{-1}\ V_{22}\ H\ H^{-1}\ V_{21}^{-1}\ D_1 \\ &= -\hat{X}_2\ D_2^{-1}\ V_{22}\ V_{21}^{-1}\ D_1 = -\hat{X}_2\ D_2^{-1}\ G\ D_1 \end{aligned} \tag{1.26}$$

Comparing (1.26) with (1.23b) shows

$$\hat{G} = D_2^{-1}\ G\ D_1. \tag{1.27}$$

---

1. It is readily seen from $X = [\hat{U}_1\ \hat{U}_2] \begin{bmatrix} \hat{\Sigma}_{11} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \hat{V}_1 \\ \hat{V}_2 \end{bmatrix}$ that $\hat{V}_2$ can be any orthonormal basis for the null space of X. One can therefore derive at least one such $\hat{V}_2$ from V_2 by taking the QR decomposition of $D^{-1}V_2 = Q\ R$ to produce $\hat{V}_2 \equiv Q = D^{-1}V_2R^{-1}$.

Since D_2 and D_1 are both diagonal, we have $\hat{g}_{ij} = 0$ if and only if $g_{ij} = 0$, where $\hat{G} = (\hat{g}_{ij})$ and $G = (g_{ij})$. Hence, in exact arithmetic scaling does not change the zero structure of G. However, in finite arithmetic it is clear that any nonzero element of G may be made as small as desired in $\hat{G}$ by appropriate scaling. A nonzero in G may therefore be a zero in $\hat{G}$ and vice versa within the limits of the machine's calculations.

The solution to this problem (that the determination of linear dependencies may be scale-affected) is one of numerical analysis. Since there would be no problem from scaling if we had exact calculations, we should analyze the data matrix X in units chosen to allow for the numerically most stable calculations in light of the finite arithmetic. Column equilibration (scaling to produce roughly equal column lengths) enjoys some usefulness in this context.[1] Conclusions regarding the zero structure of V_2 should be based on a data matrix so scaled. Then, should the user desire information on a differently scaled matrix, the above determined V_2 with the zero structure imposed should provide the basis of the transformed structure. That is, let X be the data scaled for numerical accuracy, and let $\hat{X} = X D$ be the data scaled in terms of the user's preferences. Then the zero structure of the G applicable to the data in X is determined by analysis of (1.23a). Let G be the calculated matrix, and denote G with its "zero" elements replaced by exact zeros by G*. Information on $\hat{G}^*$ can then be had by the analog to (1.27), namely

$$\hat{G}^* = D_2^{-1} G^* D_1 . \tag{1.28}$$

Clearly $\hat{G}^*$ will have the same zero structure, invariant to scale.

---

1. See Van der Sluis (1969) and (1970).

Part 2. An Assessment of the Damage Caused by Linear Dependencies

In this part we address the second major question set out in the opening paragraph, namely, how much damage is caused to the regression estimates due to the presence of linear dependencies (near dependencies) in the data matrix. It is well known that any such damage manifests itself in unstable regression coefficients and in inflated sampling variances. But it has not been possible quickly to determine whether the size of any specific sampling variance was large because of collinear data or because of inherent noise (arising, for example, because the given variate does not belong in the hypothesized relationship). The former problem is potentially correctable through additional information that might take the form of new noncollinear data, a prior distribution for the regression parameters, or outside estimates for specific coefficients. The analysis presented here helps to determine whether collinear data is in fact a cause of inflated sampling variance, and further it helps to highlight which regression estimates are being most adversely affected - thereby keying where corrective measures are most profitably employed.

In Section 1, the decomposition of the sampling variance that forms the basis of the analysis is presented. Section 2 presents a theoretical result that helps to interpret possible outcomes of the decomposition. Section 3 examines the procedures suggested in Section 1 for assessing the damage caused to regression estimates from the use of collinear[1] data.

---

[1] It should be highlighted that the term collinear here means rank deficient in the sense of Part 1 and does not mean the existence of an exact linear dependency; nor, obviously is it the common but loose usage in econometrics.

2.1 The Basic Decomposition of the Variance of b_b.

The singular value decomposition of a data matrix X, as we saw in Part 1 of this paper produces a set of singular values that can be associated with potential linear dependencies in the data. The word "potential" is used because (as per Section 1.3) it must first be determined, through machine and data considerations, which singular values are small, and for each of these there is a linear dependency to be identified. As any one singular value, then, gets small relative to σ_{max}, there is a near dependency to be associated with that singular value.

The basis for the analysis presented here is the decomposition of the variances of the regression coefficients into components that are associated with the singular values of X and hence are directly related to the specific linear dependencies possesed by X. A derivation of this variance decomposition using eigensystems of X'X due to Silvey (1969) is given in Johnston (1972), but we rederive the result here using the singular-value decompostion to highlight the correspondence of the components to the singular values, and hence the linear dependencies, of X (not of the moment matrix X'X).

The variance-covariance matrix of the least squares estimator $b = (X'X)^{-1}X'y$ is, of course,

$$\text{Var}(b) = \sigma^2(X'X)^{-1} \tag{2.1}$$

where σ^2 is the common variance of the components of the T disturbances ε in $y = x\beta+\varepsilon$. Making use of the singular value decomposition of X

$$\underset{TxK}{X} = \underset{TxK}{U}\ \underset{KxK}{\Sigma}\ \underset{KxK}{V'} \quad \text{with} \quad \Sigma = \text{diag}(\sigma_1 \,..\, \sigma_K), \text{ and } V=(v_{ij}) \tag{2.2}$$

we may rewrite (2.1) as (recalling $U'U = I$)

$$\text{Var}(b) = \sigma^2 \, V\Sigma^{-2}V' \tag{2.3}$$

or, for the k-th component of b,

$$\text{var}(b_k) = \sigma^2 \sum_j \frac{v_{kj}^2}{\sigma_j^2} \quad . \tag{2.4}$$

(2.4), it will be noticed, decomposes $\text{var}(b_k)$ into a sum of components each containing the square of one of the singular values, σ_j. We recall from Section 1.3 how, for each linear dependency of X, some σ_j becomes small. Since these σ_j are in the denominator in (2.4), other things equal, those components of $\text{var}(b_k)$ associated with a linear dependency (with small σ_j) will be large relative to the other components. This suggests, then, that an unusually high proportion of the variance of one or more coefficients concentrated in components associated with a specific singular value gives evidence that the corresponding linear dependency may be causing problems. This suggestion is pursued in Section 2.3 after some interpretive considerations are developed in Section 2.2.

It is a relatively easy matter to display these proportions for all $\text{var}(b_k)$ so that the investigator can tell at a glance where problems may be arising

Define

$$\eta_{kj} = \frac{v_{kj}^2}{\sigma_j^2} \quad , \quad \eta_k \equiv \sum_{j=1}^{K} \eta_{kj} \quad k = 1 \,..\, K \tag{2.5}$$

and

$$\phi_{kj} = \frac{\eta_{kj}}{\eta_k} \quad k, j = 1 \ldots K.$$

Then all information is summarized by the tables

Variance-Components Table
(all entries $\times \sigma^2$)
Components of

Associated with	$\text{var}(b_1)$	$\text{var}(b_2)$	. . .	$\text{var}(b_K)$
σ_1	η_{11}	η_{21}	. . .	η_{K1}
σ_2	η_{12}	η_{22}	. . .	η_{K2}
.	.	.		.
.	.	.		.
	.	.		.
σ_K	η_{1K}	η_{2K}	. . .	η_{KK}

(2.6a)

and

Variance-proportions table

Associated with	Components of $\text{var}(b_1)$	$\text{var}(b_2)$	...	$\text{var}(b_K)$
σ_1	ϕ_{11}	ϕ_{12}	...	ϕ_{1K}
σ_2	ϕ_{21}	ϕ_{22}	...	ϕ_{2K}
.	.	.		.
.	.	.		.
.	.	.		.
σ_K	ϕ_{K1}	ϕ_{K2}	...	ϕ_{KK}

(2.6b)

An example of these tables is given in Sections 2.2.4 and 2.3.3 below.

2.2 An Interpretive Consideration: Orthogonality and the Zero Structure of V.

It will be necessary to gain much practical experience with the decomposition (2.4) before reasonable guidelines can be established for its use as a diagnostic tool. There is, however, one immediate consideration that can be given a rigorous foundation, namely, that if in (2.4) some v_{kj}^2 are zero, then it makes no difference to $\text{var}(b_k)$ if the corresponding σ_j are very small, i.e., the coefficient will be immune from collinearity associated with those particular singular values. This section examines the conditions under which certain of the v_{ij} will be zero (or small relative to the corresponding σ_j) and hence develops conditions under which certain regression coefficients need not be adversely affected by the presence of multicollinear data. We can anticipate this result by recalling the well known fact that the addition to a regression equation of a variate that is orthogonal to all previous variates will not affect the regression calculations based only on the original variates. Clearly then, it should also not affect any regression calculations to add a set of variates

that are orthogonal to all previous variates - whether or not this additional set itself contains with it a perfectly collinear relationship.

Indeed, through a series of telescoping theorems of increasing generality, we arrive at sufficient condition on X (and its singular values) under which orthogonal partitions of X imply specific v_{ij}'s to be zero in the singular value decomposition of X. These are approximate conditions, then, under which regression estimates may possibly be salvaged even in the presence of strongly collinear data. Special computational algorithms are required to exploit this possibility, however, for most regression programs are incapable of dealing with collinear data no matter how it occurs, and hence can make no attempt to identify and salvage any coefficients that need not be adversely affected.[1]

In the rest of this section four theorems are proved that show the conditions under which orthogonal blocks in the data matrix X imply specific v_{ij}'s to be zero.[2] The reader not interested in the proofs to these theorems is advised to read Theorems 2 and 4 for gist and continue to the next section.

2.2.1 The Zero Structure of V when X has Orthogonal Parts

Let us begin with a TxK data matrix X partitioned into two orthogonal blocks X_1 (TxK_1) and X_2 (TxK_2) with $X_1'X_2 = 0$. In this case we can determine the singular values of X by determining them separately for X_1 and X_2. Indeed

---

[1] A set of calculations that proceed correctly in the presence of perfectly collinear data are given in Belsley (1974). These algorithms form the basis of the NBER Computer Research Center's GREMLIN system - a comprehensive package for estimating simultaneous systems available through the Center's time sharing network.

[2] It should be emphasized that these are sufficient, but not necessary conditions. Indeed there may well be other conditions leading to v_{ij}'s being zero - and these too would lead to coefficients isolated from colinear relationships.

the SVD of X is

$$X = U \Sigma V' \tag{2.7}$$

while those of X_1 and X_2 are

$$\begin{aligned} X_1 &= U_1 \Sigma_1 V_1' \qquad \text{where } U_1'U_1 = V_1'V_1 = I_{K_1} \quad \Sigma_1 = \text{diag. matrix} \\ X_2 &= U_2 \Sigma_2 V_2' \qquad\quad\; U_2'U_2 = V_2'V_2 = I_{K_2} \quad \Sigma_2 = \text{diag. matrix} \end{aligned} \tag{2.8}$$

It is clear that the matrix V derived from (2.8) as

$$\tilde{V} = \begin{bmatrix} V_1 & 0 \\ 0 & V_2 \end{bmatrix} \tag{2.9}$$

is orthogonal and has the property of diagonalizing X'X

$$\tilde{V}'(X'X)\tilde{V} = \begin{pmatrix} V_1' & 0 \\ 0 & V_2 \end{pmatrix} \begin{pmatrix} X_1'X_1 & 0 \\ 0 & X_2'X_2 \end{pmatrix} \begin{pmatrix} V_1 & 0 \\ 0 & V_2 \end{pmatrix} = \begin{pmatrix} \Sigma_1^2 & 0 \\ 0 & \Sigma_2^2 \end{pmatrix}. \tag{2.10}$$

Hence the matrix

$$\tilde{\Sigma} = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix} \tag{2.11}$$

must be the matrix of singular values of X.

Since these values are unique they must be the same elements as Σ in (2.7) - although the order is not unique. We have shown

Theorem 1.

Let $X = (X_1 X_2)$ with $X_1'X_2 = 0$. Then the singular values of X may be determined directly from the separate SVD of $X_i = U_i \Sigma_i V_i'$, $i=1,2$.

This result can be used to show that orthogonality among sets of columns of X implies a certain zero structure on the elements of V in (2.7), and hence on certain relevant v_{ij} in the numerator of the variance decomposition (2.4). We begin with

Theorem 2.

Let $X = [X_1 X_2]$ with $X_1'X_2 = 0$. Then, if the singular values of X are distinct, the matrix V in the SVD of $X = U\Sigma V'$ has the form $\begin{bmatrix} V_1 & 0 \\ 0 & V_2 \end{bmatrix}$, where V_i is $K_i x K_i$.

Proof: The SVD of X_i is as in (2.8), and because of Theorem 1, we can write Σ as

$$\begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} .$$

Now

$$(X'X) = \begin{pmatrix} X_1'X_1 & 0 \\ 0 & X_2'X_2 \end{pmatrix} = V\Sigma^2 V'$$

and one V that clearly works is $V = \begin{bmatrix} V_1 & 0 \\ 0 & V_2 \end{bmatrix}$. But since the columns of the V_i are the eigenvectors of $X_i'X_i$, the distinctness of the singular values guarantees the uniqueness of the V_i (up to permutations and a

multiplier of modulus 1). Hence V is unique up to permutations within its first K_1 columns and its last K_2 columns - which clearly will not alter the zero structure

QED

The condition in Theorem 2 that the singular values be distinct is overstrong for the purpose at hand. Problems in guaranteeing the desired zero structure occur only when there are multiple roots in common between Σ_1 and Σ_2, overlap of roots. The following example demonstrates this. Let

$$X = [X_1 X_2] = \left[\begin{array}{cc:cc} \sqrt{3} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \sqrt{2} \end{array}\right] \quad \text{so that } X'X = \left[\begin{array}{cc:cc} 3 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 2 \end{array}\right]$$

The matrix

$$V = \left[\begin{array}{cc:cc} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 0 & 1 \end{array}\right]$$

is easily shown to be orthogonal and diagonalize X'X, but it clearly does not possess the desired zero structure. Even here, however, there is a V matrix that does possess the desired structure, namely V=I, but such a structure is not guaranteed.

If, however, there are multiple roots that do not overlap X_1 and X_2 (are not in common to Σ_1 and Σ_2) the desired zero structure is assured. This is seen by assuming otherwise, i.e., assume

$$V^* = \begin{bmatrix} V^*_{11} & V^*_{12} \\ V^*_{21} & V^*_{22} \end{bmatrix}$$

in any other orthogonal V such that $X'X = V^*\Sigma^2 V^{*'}$. Since the Σ_1 and Σ_2 have no overlap, the non-uniqueness of V^* (beyond permutations of columns) can occur only up to linear combinations with its first K_1 columns and within its last K_2 columns. Linear combinations across these two sets of columns are not possible. But we already know that $\begin{bmatrix} V_1 \\ 0 \end{bmatrix}$ is a basis for the range space of the first K_1 columns, and $\begin{bmatrix} 0 \\ V_2 \end{bmatrix}$ a basis for the last K_2 columns. Hence any permissible linear combinations must preserve the zero structure. We have proved

Theorem 3.

If in Theorem 2 Σ_1 and Σ_2 have no values in common (however great the multiplicities within each), then V in the SVD of X retains the zero structure shown there.

The assumptions behind Theorem 3 are too strong, but they may be weakened to produce a useful result, namely.

Theorem 4.

Let $X = [X_1 X_2]$ with $X_1'X_2 = 0$ and let σ_{2k} be the kth singular value of X_2 (kth element of Σ_2). Then, if σ_{2k} is distinct from all other σ (in both Σ_1 and Σ_2), regardless of any other multiplicities or overlaps, $V = (v_{ij})$ in the SVD of X has the property that

$$v_{j,K_1+k} = 0 \quad \text{for } j=1, \ldots, K_1,$$

i.e., the first K_1 elements of the K_1+k column of V are zero.

Proof Beyond permutations, the K_1+kth column of V is uniquely determined up to a linear combination of the eigenvectors associated with the value $\sigma^2_{2_k}$. Since this value is assumed distinct, there is only a one dimensional space associated with it, and we know that this space is spanned by the K_1+k^{th} column of $V = \begin{bmatrix} V_1 & 0 \\ 0 & V_2 \end{bmatrix}$, which clearly has the required zero.

2.2.2 Nearcollinearity Nullified By Near Orthogonality

Theorem 4 has the generality required to analyze the variance decomposition (2.4). Let us assume, in the extreme, that X has two orthogonal parts X_1 and X_2 and that X_1 is well conditioned but X_2 is ill conditioned. This means that the elements of Σ_1 are roughly of the same magnitude but that there are some elements of Σ_2 that are relatively small. Break up the sum (2.4) into its first K_1 terms and it last K_2 terms as

$$\text{var}(b_k) = \sum_{j=1}^{K} \frac{v_{kj}^2}{\sigma_j^2} = \sum_{j=1}^{K_1} \frac{v_{kj}^2}{\sigma_{1j}^2} + \sum_{j=1}^{K_2} \frac{v_{k,K_1+j}^2}{\sigma_{2j}^2} \tag{2.12}$$

The ill conditioning of X means that some σ_{2_j} will be small - indeed zero if X_2 is perfectly collinear. Let this σ_{2_j} be σ_{2_p}. Now Theorem 4 guarantees that for $k = 1 \ldots K_1$, $v^2_{k,K_1+p} = 0$, and hence the term

$$\frac{v^2_{k\,K_1+p}}{\sigma^2_{2p}} = 0$$

for $k = 1 \ldots K_1$. That is, $\text{var}(b_k)$ is unaffected by near collinearity for $k = 1 \ldots K_.$. These estimates are salvaged in the presence of collinearity due to orthogonality of X_1 from X_2. Of greater generality, however, one clearly need not assume X_1 strictly orthogonal to X_2. Since the v_{ij}'s are continuous functions of the columns of X, as the blocks of X become more nearly orthogonal (their inner products get closer to zero) the relevant elements of V also go to zero in the limit. Hence some v_{ij} can be small if the data are pleasantly well behaved. That is, the adverse effects of near collinearity in one block of data, X_2 (as measured by some small σ_{2j}'s) can be mitigated in the estimates of the coefficients corresponding to another block of data, X_1, as these two blocks are the more nearly orthogonal (as measured by small v_{kj}^2's, $k = K_1+1 \ldots K$).

2.2.3 An Example

An example of the preceding result is useful here. We will consider the matrix

$$X = [X_1 X_2] \equiv \left[\begin{array}{rrr:rr} -74 & 80 & 18 & -56 & -112 \\ 14 & -69 & 21 & 52 & 104 \\ 66 & -72 & -5 & 764 & 1528 \\ -12 & 66 & -30 & 4096 & 8192 \\ 3 & 8 & -7 & -13276 & -26552 \\ 4 & -12 & 4 & 8421 & 16842 \end{array}\right] \tag{2.13}$$

This matrix, essentially due to Bauer (1971), has the property that its fifth column is exactly twice its fourth, and both of these are orthogonal to the first three columns. That is, X_2 is singular and $X_1'X_2 = 0$.

The preceding theorems tell us the following about the Σ and V matrices that result from the singular value decomposition of X: unless there are multiplicities of roots (which, as a practical matter will occur with

probability zero), 1) one of the singular values associated with X_2 will be zero (i.e., within the machine tolerance of zero), and 2) in $V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}$, $V_{12} = 0$ and $V_{21} = 0$.

Application of the program MINFIT[1] to obtain the singular value decomposition of X results in:

(2.14)

$$V = \left[\begin{array}{llllll|ll} 0.547864D & 00 & -.625347D & 00 & 0.5556850 & 00 & 0.148362D & -18 & -.543183D & -14 \\ -.835930D & 00 & 0.383313D & 00 & 0.392800D & 00 & 0.215618D & -19 & -.470435D & -14 \\ 0.326342D & -01 & 0.679715D & 00 & 0.732750D & 00 & 0.158113D & -18 & -.729449D & -14 \\ \hline -.642653D & -15 & -.216297D & -15 & 0.913326D & -14 & -.447214D & 00 & 0.894427D & 00 \\ 0.321423D & -15 & 0.108174D & -15 & -.456672D & -14 & -.894427D & 00 & -.447214D & 00 \end{array}\right]$$

and the following diagonal elements of Σ

$$\begin{aligned} \sigma_1 &= 0.170701D \quad 03 \\ \sigma_2 &= 0.605332D \quad 02 \\ \sigma_3 &= 0.760190D \quad 01 \\ \sigma_4 &= 0.363684D \quad 05 \\ \sigma_5 &= 0.131159D \quad -11 \quad . \end{aligned} \tag{2.15}$$

A glance at V verifies that the off-diagonal block partitions are indeed small - all of the magnitude of 10^{-14} or smaller - and well within the effective zero of the computational precision.[2] Only somewhat less obvious is that one of the σ_i associated with X_2 is zero. Actually σ_5 is of the order of 10^{-11}, and

---

[1] Golub and Reinsch (1970), and Becker, et al. (1974).

[2] 10^{-14} on the IBM 67 in double precision.

would seem to be non-zero, but the relevant comparison[1] is the order of magnitude of the scale-free value $\frac{\sigma_k}{\sigma_{max}}$, which, in this case, is 10^{-16}. The practical results are thus in full accord with theory, and we can now examine the effects of the perfectly collinear data matrix on the estimated variances of the regression parameters $b = (X'X)^{-1}X'y$.

It is clear that any problem in the calculation of $Var(b_k)$ in (2.4) for this particular case will arise because of the very small σ_5. However, σ_5, small as it is, is several orders of magnitude larger than its corresponding v_{ij} for i=1, 2, 3. Hence the contributions of the $\frac{v_{i5}^2}{\sigma_5^2}$ components to calculations of $Var(b_1)$, $Var(b_2)$ and $Var(b_3)$ in (2.4) will be small. That is, the presence of pure multicollinearity will not significantly upset the precision with which we can estimate the coefficients of other variates provided these other variates are reasonably isolated from the offending collinear variables through near orthogonality.

To demonstrate this point, we calculate the relative components of $var(b_1^*)$ by means of (2.4).

$$Var(b^*_1) = \sigma^2 \sum_{j=1}^{5} \frac{v_{1j}^2}{\sigma_j^2} =$$

$$\sigma^2 (.0010 + .0107 + .5343 + 0.0 + .0017)\, 10^{-2} = \sigma^2 (.5488 \times 10^{-2}). \quad (2.16)$$

It is clear from (2.4) that the component of $var(b_1^*)$ affected adversely by the collinearity, namely $\frac{v_{15}^2}{\sigma_5^2}$, is small ($.0017 \times 10^{-2}$) relative to the total

[1] Professor Golub shows any σ_k having the property that $\frac{\sigma_k}{\sigma_{max}} \leq \sqrt{\varepsilon}$, where ε is the effective machine zero, is considered evidence of rank deficiency. [Golub and Reinsch (1970)].

$(.5488 \times 10^{-2})$. Indeed, it is only through the finite arithmetic of the machine that this term has any definition, for it, in theory, is an undetermined ratio of zeros. In practice, there is reason to cast out this component in actual calculations of $var(b_1^*)$.

The preceding is in stark contrast to the calculation of $var(b_4^*)$ or $var(b_5^*)$, for these are the variances of coefficients that correspond to variables involved in the singularity of X. Indeed

$$var(b_5^*) = \sigma^2 \sum_{j=1}^{5} \frac{v_{5j}^2}{\sigma_j^2} = \sigma^2 \, (0.0 + 0.0 + 0.0 + .0000 + 1.1626 \times 10^{23})^1. \tag{2.17}$$

This variance is obviously huge and completely dominated by the last term and its role in causing the singularity of X.

2.3 Assessing the Damage Caused by Collinear Data.

2.3.1 At Least Two Variates Must Be Involved

The theorems and example of the preceding section help to put meaning to the variance components and proportions summarized in tables like (2.6 a and b). At first it might seem that the concentration of the variance of any one regression coefficient ($var(b_k)$) in any one of its compoents ϕ_{kj} (j = 1 ... k) signals the fact that multicollinearity may be causing problems. But it is clear from Theorem 4 that if collinearity (ill conditioning)

---

[1]The difference between 0.0 and .0000 in these expressions is designed to differentiate between a number within the machine's zero (0.0), and a nonzero number with highly negative exponent (.0000). The 0.0's in (2.17), for example, are of the order of 10^{-30}, while the .0000 is of the order 10^{-10}.

is causing problems, more than one variance must be adversely affected by variance components associated with a single singular value. This is seen from the following example.

Suppose the data matrix X consists of K mutually orthogonal columns, and the singular values satisfy the conditions of Theorem 4 (as they will with probability 1). Theorem 4 immediately implies that the V matrix of the singular value decomposition of X is of the form[1]

$$V = \begin{bmatrix} v_{11} & & & \\ & v_{22} & & 0 \\ & & \ddots & \\ 0 & & & v_{KK} \end{bmatrix}$$

Hence only the v_{jj} terms in (2.5) will be non-zero, and (2.6b) will take the form

Associated with singular value	Proportions in var (b_1)	...	var (b_k)
σ_1	1		0
.		1	
.			
.			
σ_k	0		1

.

[1] While V has been made diagonal here, Theorem 4 insists only that it have one non-zero element in each row and column. V is unique only up to column permutations and a multiplier of modules 1. This, of course, does not affect the calculations of (2.4) or (2.5) since the σ_j^2 permute in a compensating manner and since the v_{ij}'s are squared and unique despite the multiplier of modulus 1.

It is clear that a high proportion of each variance associated with a single singular value is hardly indicative of multicollinearity, for the variance proportions here are for an ideally conditioned, orthogonal data matrix. Indeed, problems can arise only when a single singular value σ_j is associated with a large proportion of the variance of two or more coefficients. This simply reflects the fact that there must be two or more columns of X involved in any linear dependency.

We know by Theorem 4 that each of the columns, k, of V involved in such a linear dependency must necessarily have a nonzero v_{kj} associated with the small singular value σ_j. The ratio of these v_{kj} to the small σ_j must, therefore, loom large in the calculation of the variances var (b_k) by (2.4) for those coefficients corresponding to the collinear (nearly collinear) variates. If, for example, in a case of K = 5, columns 4 and 5 are collinear and all other columns are mutually orthogonal we would expect a variance-proportions table like (2.6b) that has the form, say

		Proporations in var (b_1)	var (b_2)	var (b_3)	var (b_4)	var (b_5)
Associated with singular value	σ_1	1	0	0	0	0
	σ_2	0	1	0	0	0
	σ_3	0	0	1	0	0
	σ_4	0	0	0	1	.9
	σ_5	0	0	0	0	.1

Here σ_4 plays a large role in both var(b_4) and var(b_5).

2.3.2 Variance Proportions: Necessary but not Sufficient

We have learned from the foregoing that near collinearity (ill conditioning) will manifest itself as high proportions for two or more variances in components associated with a single singular value.[1] Unfortunately, for the purposes of testing, the converse does not hold; such a pattern of high proportions need not imply the existence of collinearity. Whereas several variances may have most of their weight in a component associated with the same singular value, the overall magnitude of the variance may be pleasantly low--near collinearity, if it exists at all, causes no problem. The variance proportions table, then, is merely a quick means of telling whether collinearity may be problemful, but once the pattern of high proportions is detected, one must turn to the actual variance components in Table (2.6a) to tell whether the overall levels are high. An example will serve to make this clear.

Let us return to the modified Bauer matrix of Section 2.2.3. This five column matrix, we recall, has the property that column 4 is exactly twice column 5, and these two columns are orthogonal to columns 1, 2 and 3. We would fully expect that the small singular value σ_5 (= .1312 x 10^{-11})associated with the linear dependency $X_4 = .5X_5$ would dominate several variances--at least var(b_4) and var(b_5). The variance proportions table (2.6b) for the modified Bauer matrix is given below in Table 1, and a glance at the bottom row verifies that σ_5 does indeed account for the entirety of these two variances (the first three variances are isolated from this relationship by the orthogonality of the first three columns of X from the last two).

---

1. It should be noted in passing that the existence of collinearity in X may not produce practically harmful problems in estimates of a linear model relating y to X, as in $y = X\beta+\varepsilon$. Such problems also depend upon the size of σ^2 (which also enters in Var (b)). This point is dealt with below in greater detail in section 2.3.4.

TABLE 1

Variance Proportions - Modified Bauer Matrix

	$Var(b_1)$	$Var(b_2)$	$Var(b_3)$	$Var(b_4)$	$Var(b_5)$
σ_1	.002	.009	.000	.000	.000
σ_2	.019	.015	.013	.000	.000
σ_3	.976	.972	.983	.000	.000
σ_4	.000	.000	.000	.000	.000
σ_5	.003	.005	.003	1.000	1.000

A somewhat unexpected pattern, however, is also apparent: The single singular value σ_3 accounts for 97% or more of $var(b_1)$, $var(b_2)$ and $var(b_3)$. It may well be the case that a second linear relationship among the columns of X, one associated with σ_3, is accounting for these high proportions. But two facts would tend to discount this possibility. First, the three columns X_1, X_2 and X_3 that could be involved in such a relationship[1] (X_4 and X_5 are orthogonal) are reasonably well conditioned; and second, in spite of the concentrated variance proportions, the overall magnitudes of $var(b_1)$, $var(b_2)$ and $var(b_3)$ are small. This latter fact is seen from the actual variance components for the modified Bauer matrix given in Table 2.

---

[1] From Theorem 1 we know that the singular values for the matrix X_1 which is comprised of the first three columns of the modified Bauer matrix X are precisely the same as σ_1, σ_2 and σ_3 for the modified Bauer matrix itself. Hence, the condition number of X_1 is $\kappa(X_1) = \frac{\sigma_{max}}{\sigma_{min}} = \frac{.171 \times 10^3}{.76 \times 10} = 22.5$, a number quite low relative to most matrices of economic data.

TABLE 2

Variance - Components

Modified Bauer Matrix

$\times \sigma^2$

	$Var(b_1)$	$Var(b_2)$	$Var(b_3)$	$Var(b_4)$	$Var(b_5)$
σ_1	$.103 \times 10^{-4}$	$.240 \times 10^{-4}$	$.366 \times 10^{-7}$	$.142 \times 10^{-34}$	$.354 \times 10^{-35}$
σ_2	$.107 \times 10^{-3}$	$.401 \times 10^{-4}$	$.126 \times 10^{-3}$	$.128 \times 10^{-34}$	$.319 \times 10^{-35}$
σ_3	$.534 \times 10^{-2}$	$.267 \times 10^{-2}$	$.929 \times 10^{-2}$	$.144 \times 10^{-29}$	$.361 \times 10^{-30}$
σ_4	$.166 \times 10^{-46}$	$.351 \times 10^{-48}$	$.189 \times 10^{-46}$	$.151 \times 10^{-9}$	$.604 \times 10^{-9}$
σ_5	$.172 \times 10^{-4}$	$.129 \times 10^{-4}$	$.309 \times 10^{-4}$	$.465 \times 10^{24}$	$.116 \times 10^{24}$
Sum =	$.548 \times 10^{-2}$	$.275 \times 10^{-2}$	$.945 \times 10^{-2}$	$.465 \times 10^{24}$	$.116 \times 10^{24}$

In order to get the actual variances and variance components, each of the figures of Table 2 must be multiplied by σ^2, the variance of the error term in the linear model $y = X\beta + \varepsilon$. But, at least on a relative basis, it is clear that the high proportions associated with σ_5 are reflecting massive sizes for $var(b_4)$ and $var(b_5)$-on the order of $\sigma^2 \times 10^{24}$, while those associated with σ_3 reflect smaller variances on the order of $\sigma^2 \times 10^{-2}$. Whether this latter figure is small in fact depends, of course, on the size of σ^2.

2.3.3 A Suggested Test for Harmful Collinearity

High variance proportions, then, in themselves are not sufficient to reveal the existence of harmful collinearity--for, as the preceding example shows, the high proportions may not be associated with a singular value that has been determined to be small enough (in the sense of Section 1.3) to indicate rank deficiency. Such is the case with the high proportions associated with σ_3.

σ^5, however, has been determined to be associated with a linear dependency, and its high variance proportions indicate collinearity to be harmful.

It is suggested here, then, that an appropriate means for detecting harmful collinearity is the double condition of

1) high variance proportions for two or more variances associated with
2) a single singular value determined by the methods of Section 1.3 to be small and hence evidence of rank deficiency.

2.3.4 Multicollinearity as a Practical Problem

Whether multicollinearity turns out to be a problem of practical consequence is a different question from that addressed above. It will be noted that the test for harmful collinearity suggested above wholly ignores the error variance σ^2 that also enters the relation $\text{Var}(b) = \sigma^2(X'X)^{-1}$. Indeed, the terms cancel from the variance proportions ϕ_{ij} of (2.6b), but they are a factor in each of the entries of (2.6a). It is possible, then, that collinearity resulting in high variance proportions ϕ_{ij}, and indeed high components η_{ij} can be mitigated by low σ^2, for, from (2.4) and (2.5), $\text{var}(b_k) = \sigma^2\eta_k$ where $\eta_k = \sum_{j=1}^{K}\eta_{jk}$. In such a case, the actual variances may be small enough to allow acceptance of all desired tests of hypothesis, in spite of the fact that the precision of the least squares estimates would be better in the absence of ill-conditioned data. In other words, the presence of multicollinearity as determined here, need not be problemful as a practical matter.[1] The test suggested

[1] Another view of this point is useful. It will be noted that the entire analysis of collinearity presented here is based on the data matrix X in the linear regression model $y = X\beta + \varepsilon$ and no where requires knowledge of y. This is because ill conditioning, and the instability of calculations and estimates that result from it, has only to do with X, and one would be better off with a nicely conditioned X matrix whether or not the ill conditioning is bad enough to cause practical problems. It is the latter point that depends upon y, for only through the introduction of y can σ^2 be estimated in order to determine if the overall levels of the estimated variances are too high for conducting desired hypothesis tests. If they are, and ill conditioning can be determined as a problem, then corrective action is worthwhile.

here, however, highlights when estimated variances are being adversely affected (whether to a point of being problemful or not), and hence indicates when and where such variances could be improved should the need arise through the introduction of additional information that "breaks up" the ill conditioning. This point will be discussed further in Part 3.

Part 3. Some General Considerations on Multicollinearity and Its Corrections

It is not the purpose of this paper to suggest an answer to the third question raised in its introduction: that dealing with corrective measures. However, some general remarks on multicollinearity and its correction seem called for. Section 1 of this third part examines other tests for multicollinearity that have been proposed. Section 2 discusses corrective procedures and presents a fundamental criticism of the use of non-Bayesian ridge regression as a means of correction.

3.1 Other Tests for Multicollinearity

3.1.1 Simple Correlations

The use of simple, pairwise correlations as a means of showing the presence of multicollinearity has been so basically discredited that it seems hardly necessary to mention it. However, the technique appears to flair up anew with some regularity, and seems to require constant care to keep it extinguished. In favor of the procedure,it must be said that the existence of two variates with correlation ± 1 is a clear indication of multicollinearity and therefore it would seem that "high" correlation would be problemful. But a correlation of .9 need not result in any real problem of estimation. The test is, therefore, without proper interpretation, for there is no well defined notion of "high". Conversely, low correlations are no indication of the absence of multicollinearity, for three or more variates may be perfectly collinear but have low pairwise correlations. Examination of the correlation matrix, therefore, offers, at worst, erroneous and, at best, misleading information.

3.1.2 The Determinant of X'X

Another discredited test for multicollinearity is the value of det X'X. Since X singular implies det X'X = 0, the motivation is clearly that low det X'X indicates near singularity. The problem with this notion comes from the fact that nonsingularity-singularity is not a continum. This is readily seen by considering the obviously nonsingular nxn matrix $A = \alpha I_n$ for $\alpha>0$. Clearly the determinant of A ($= \alpha^n$) may be made as small as desired by choosing α sufficiently small, but equally clearly A is always perfectly invertable.

3.1.3 Method of Farrar and Glauber

Farrar and Glauber (1967) suggest determining the presence of multicollinearity based upon a statistical test of the hypothesis that the columns of X are in fact orthogonal. A rejection of the hypothesis leads to the alternative hypothesis that the columns of X are nonorthogonal, and hence collinear. There are several weaknesses with this approach, both theoretical and applied.

1) The Farrar and Glauber approach is based on the assumption that the X data resulted from some stochastic process whose orthogonality is subject to test. If the X data are properly assumed as nonstochastic, however, (as they are in the classicial linear model) the Farrar-Glauber analysis is irrelevant.

2) If the X data are assumed stochastic, the previous consideration does not apply, but it is still doubtful that the Farrar-Glauber technique is proper. To see this one must realize that multicollinearity is a condition when some linear combination of the data are observationally indistinguishable from zero, and as such multicollinearity is seen to be a special case of the identification

problem. As is well known, identification is a problem logically preceding, and not a part of, the statistical problem of estimation. Multicollinearity, then, is not an estimation problem and is not properly treated as such.

3) As a practical matter the test against the null hypothesis of orthogonality seems to lack power; that is, it indicates nonorthogonality very often when there is no real problem (all coefficients are alive, well and with strong t's). This practical problem is not surprising in light of the general inappropriateness of the technique. Haitovsky (1968) attempts to overcome this practical problem of Farrar and Glauber by making the test against the null hypothesis of singularity. Haitovsky's procedure, however, falls prey to the same criticisms advanced above.

3.2 Corrective Measures

3.2.1 The Introduction of Identifying Information

The recognition above that multicollinearity is an identification problem has implications not only for the proper way to test for it, but also for the proper way to correct it. A multicollinear data set results in an unidentified equation. As is well known[1], it requires the addition of new, independent information to identify an unidentified equation. As we shall see below, the use of ridge regression as has been suggested by some fails to add identifying information and, indeed, fails to remove the estimation problem that results from collinear data. Two methods have been suggested, however, that can

[1] See Fisher (1966).

properly introduce additional information, and hence stand as appropriate corrective measures. These are the time-honored methods of using outside estimates (such as combining estimates of coefficients in a time-series equation previously estimated from cross-sectional data), and the method of using a Bayesian prior for the coefficients. The former method has the practical weakness that it is very difficult to find "outside" conditions that are appropriate to obtain estimates for the given situation. A marginal propensity to consume, for example, determined from cross-sectional budget studies has dubious relevance to a time-series estimated consumption function. The second method, proposed in Zellner (1971) and Leamer (1973), has much promise.

3.2.2 The Failure of Ridge

Attempts have been made recently to utilize ridge regression to mitigate the effects of multicollinearity.[1] Short of a means of combining this procedure with some method of bringing in legitimate identifying information,[2] however, this method is doomed to failure--merely substituting collinearity in the data for a degenerate distribution of the estimated coefficients.

We begin with the usual normal equations for least squares

$$X'X\,b = X'y \tag{3.1}$$

and we assume X to be rank deficient. The suggested ridge solution is to create an invertable matrix by constructing and solving the ridge equation

$$(X'X + kQ)b^* = X'y \tag{3.2}$$

where Q is some positive definite matrix--often taken as I, and b^* is the ridge

---

[1] See, for example, Bushnell and Huettner (1973), Hoerl and Kennard (1970).

[2] Such, for example, as is done by Holland (1973) in which he combines ridge with a Bayesian prior.

estimator. k and Q are taken so that $(X'X + kQ)^{-1}$ does exist--and the presumption is that b* is now solvable and uniquely so as

(3.3) $$b^* = (X'X + kQ)^{-1}X'y$$

Unfortunately, this trick does not solve the problem for it is readily shown that Var(b*) is singular, i.e., b* has a degenerate distribution and is no more amenable to proper hypothesis testing than is the nonuniquely defined OLS estimator b from (3.1).

To see this, note that, since X is rank deficient, there exists a non-trivial $\gamma \neq 0$ such that $X\gamma = 0$. Hence (3.2) becomes

(3.4) $$(X'X + kQ)b^* = X'y = 0$$

or

(3.5) $$C'b^* = 0$$

where $$C' = (X'X + kQ)'$$

Clearly C depends only on X (k fixed), and hence remains fixed in repeated samplings. (3.5) therefore implies a fixed linear restriction on the ridge estimates b*, and renders them degenerately distributed.[1]

This exercise serves to highlight the point made above regarding the need for identifying information. In multicollinearity, as strongly as anywhere else, you cannot get something for nothing. There is something about multicollinearity that brings out the alchemist in econometricians, but there is no way one can squeeze, stamp or club more out of the data than was there in the first place. If several variates are all giving the same information, you cannot make them speak differently simply by looking at them from a different angle. Only through the addition of new, independent identifying information can the confounded effects of collinear data be undone.

---

1. Again, combining ridge with a Bayesian prior as in Holland (1973) solves this problem.

REFERENCES

Belsley, D.A. [1974], "Estimation of Systems of Simultaneous Equations, and Computational Specifications of GREMLIN", *Annals of Economic and Social Measurement*, October.

Bushnell, R.C. and D.A. Huettner [1973], "Multicollinearity, Orthogonality and Ridge Regression Analysis", Unpublished mimeo, presented December 1973 Meetings of Econometric Society, N.Y.

Farrar, D.E. and R.R. Glauber [1967], "Multicollinearity in Regression Analysis: The Problem Revisited", *Review of Economonics and Statistics*, February, pp. 92-107.

Fisher, F.M. [1969], *The Identification Problem.*

Haitovsky, Yoek [1969], "Multicollinearity in Regression Analysis: Comment", *Review of Economics and Statistics*, November, pp. 486-489.

Hoerl, A.E. and R.W. Kennard [1970a], "Ridge Regression: Biased Estimation for Nonorthogonal Problems", *Technometrics*, No. 1, pp. 55-68.

Hoerl, A.E. and R.W. Kennard [1970b], "Ridge Regression: Applications to Nonorthogonal Problems", *Technometrics*, No. , pp. 69-82.

Holland, P.W. [1973], "Weighted Ridge Regression: Combining Ridge and Robust Regression Methods", NBER CRC Working Paper No. 11.

Leamer, E.E. [1973], "Multicollinearity: A Bayesian Interpretation", *Review of Economics and Statistics*, August, pp. 371-380.

Silvey, S.D. [1969], "Multicollinearity and Imprecise Estimation", *Journal of the Royal Statistical Society*, Series B, Vol. 31, pp. 539-552.

Stewart, G.W. [1973], *Introduction to Matrix Computations.*

Van der Sluis, A. [1969], "Condition, Equilibration and Pivoting in Linear Algebraic Systems", *Numerische Mathematik*, 15, pp. 74-88.

Wilkinson, J.H. [1965], *The Algebraic Eigenvalue Problem.*

Zellner, A. [1971], *An Introduction to Bayesian Inference in Econometrics.*