NBER WORKING PAPER SERIES

MULTICOLLINEARITY: DIAGNOSING ITS PRESENCE AND ASSESSING THE POTENTIAL DAMAGE IT CAUSES LEAST-SQUARES ESTIMATION

David A. Belsley*

Working Paper No. 154

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

October 1976

Preliminary: Not for quotation

*NBER Computer Research Center and Boston College. Research supported in part by National Science Foundation Grants GJ-1154X3 and SOC-75-13626 to the National Bureau of Economic Research, Inc.

Abstract

This paper suggests and examines a straightforward diagnostic test procedure that 1) provides numerical indexes whose magnitudes signify the presence of one or more near dependencies among columns of a data matrix X, and 2) provides a means for determining, within the linear regression model, the extent to which each such near dependency is degrading the least-squares estimation of each regression coefficient. In most instances this latter information also enables the investigator to determine specifically which columns of the data matrix are involved in each near dependency.

The diagnostic test is based on an interrelation between two analytic devices, the singular-value decomposition (closely related to eigensystems) and a matching regression-variance decomposition. Both these devices are developed in full. The test is successfully given empirical content through a set of experiments that examine its behavior when applied to several different series of data matrices having one or more known near dependencies that are weak to begin with and are made to become systematically more nearly perfectly collinear. The general diagnostic properties of the test that result from these experiments and the steps required to carry out the test are summarized, and then exemplified by application to real economic data.

Acknowledgements

The author wishes to express gratitude to Professors Gene Golub of Stanford University, John Dennis of Cornell University, Edwin Kuh of MIT and the NBER, and Mrs. Virginia Klema of the NBER for their many helpful comments, and to Mr. David Jones for his very able research assistance. The author is additionally grateful to the Center for Advanced Studies in the Behavorial Sciences at Stanford for the opportunity to initiate research in this area during his fellowship there.

TABLE OF CONTENTS

Part 1: OVERVIEW AND PERSPECTIVE

1.0 Introduction

Multicollinear (ill-conditioned) data are a frequent, if often undetected, companion to econometric studies, and their presence, whether exposed or not, renders ordinary least-squares estimates less precise and less useful than would otherwise be the case. The ability to diagnose collinearity is therefore important to econometricians, and it consists of two related but separable elements: 1) detecting the presence of multicollinear relationships among the data series, and 2) assessing the extent to which these relationships have degraded[1] regression-estimated parameters. Such diagnostic information would aid the investigator in determining whether and where corrective action is necessary and worthwhile.[2] This paper suggests and examines a test procedure that treats both diagnostic elements. First, it provides numerical indexes whose magnitudes signify the presence of one or more near dependencies[3] among the columns of a data matrix X, and second it provides a means for determining, within the linear regression model, the extent to which each such near dependency is degrading the least-squares estimation of each regression coefficient. In most instances this latter information also enables the investigator to determine specifically which columns of the data matrix are involved in each near dependency, i.e., it isolates the variates involved (and therefore also those not involved) and the specific relationships in which they are included.

The remainder of this introductory part places the work reported here in its historical context. The next part, 2, provides an analytic background for the concepts to be employed and culminates in an empirical (as opposed to statistical) test procedure for diagnosing the presence of multicollinearity and assessing its potential harm to regression estimates. The techniques of Part 2 have long been known to numerical analysts, but only recently are they becoming part of the working vocabulary of econometrics. Part 3 provides a series of experiments designed to illuminate the empirical properties of the test procedure suggested in Part 2 and results in a set of interpretive tools. Finally, the process is summarized and exemplified in Part 4. There we learn, for example, that everything we thought we knew was bad about the data used to estimate the consumption function is true.

1.1 Historical Background

Two important approaches have been taken toward understanding the nature of multicollinearity[4] and diagnosing its severity and presence. For simplicity they will be referred to as the numerical analytic approach and the statistical approach.

The Statistical Approach

The single dominating work dealing with multicollinearity as a statistical problem is that of Farrar and Glauber (1967). Their widely-read article provides a prescribed set of steps for keying the presence of collinearity and for isolating the variates involved. The basis of their procedure is a statistical test against a null hypothesis of an orthogonal data matrix (the accepted standard of "clean" data). Both aspects of diagnosing collinearity are therefore treated by this technique, and Farrar and Glauber thereby presented the econometric practitioner with the first systematic means for determining just how bad his data was. The Farrar and Glauber technique has not, however, survived without its share of critical skepticism. Haitovsky (1969) criticized the use of a null hypothesis of orthogonality and proposed a widely accepted change of emphasis to tests against a null hypothesis of perfect singularity. This suggested change seemingly strengthened the Farrar and Glauber process. More recent criticism, however, has proved more damaging. Kumar (1975) highlights 1) the obvious fact that the Farrar and Glauber technique, in assuming the data matrix X to be stochastic, has no relevance to the standard regression model in which X is assumed fixed, and 2) the less obvious fact that even when X is stochastic, the tests employed by Farrar and Glauber depend upon an assumption (independence of the rows of X) that is without practical relevance to econometrics. In another vein, O'Hagan and McCabe (1975) quite directly question Farrar and Glauber's "statistical" interpretation of a measure of collinearity, concluding that their procedure misinterprets the use of a t-statistic as providing a cardinal measure of the severity of multicollinearity.

Indeed this latter criticism is correctly placed, for whether or not the statistical test employed by Farrar and Glauber is based on assumptions that are wholly relevant to econometrics, their interpretation of multicollinearity is a statistical phenomenon is fundamentally misleading. Farrar and Glauber justify their procedure as being one more in a long line of classical statistical tests of hypothesis - such as tests of significance, the Durban-Watson statistic, the Goldfeld-Quandt test - that test for the presence of a given problem through the use of a test statistic constructed under the assumed absence of that problem (the null hypothesis). To interpret their procedure as being a legitimate example of classical Neyman-Pearson hypothesis testing is, however, invalid, for such tests must be based on testable hypothesis; that is, the probability for the value of a relevant test statistic, calculated with actual sample data, is assessed in light of the distribution implied for it by the model under the null hypothesis. The Farrar and Glauber technique differs critically from this classical procedure exactly in the fact that the linear regression model makes no testable assumptions on the data matrix X.[5] That is, there are no distributional implications by the regression model for specific null hypotheses (such as orthogonality) on the nature of the data matrix against which tests can be made.

In short, multicollinearity can cause computational problems and reduce the precision of statistical estimates, but, in the context of the linear regression model, multicollinearity is not itself a statistical phenomenon subject to statistical test. A solution to the problem of diagnosing multicollinearity, then, must be sought elsewhere, in methods that deal directly with the numeric properties of the data matrix that cause calculations based on it to be unstable or sensitive in ways to be discussed.

The Numerical-Analytic Approach

Historical completeness requires one to begin this discussion with mention of Ragnar Frisch's (1934) bunch-map analysis. While not in the mainstream of numerical analysis, Frisch's technique of graphically investigating the possible relationships among a set of data series was the first major attempt in economics to uncover the sources of near linear dependencies in economic data series. Frisch's work addresses itself to the first of multicollinearity's diagnostic problems - the location of dependencies - but makes no attempt to determine the degree to which regression results are degraded by their presence. Bunch-map analysis has not become a major tool of the econometrician because its extension to dependencies among more than two variates is time consuming and quite subjective.

Recent efforts of the numerical analysts, however, have provided a very useful set of tools for a rigorous examination of multicollinearity. Their attention has, among other topics, been focused on an examination of the properties (conditioning) of a matrix A of a linear system of equations Ax = b that allow a solution for x to be obtained with numerical stability. The relevance of this to multicollinearity in econometrics is readily apparent, for the least-squares estimator is a solution to the linear system $(X'X)b = X'y$ with variance - covariance matrix $\sigma^2(X'X)^{-1}$. To the extent, then, that multicollinearity among the data series of X results in a matrix $A = X'X$ whose ill conditioning causes both the solution for b and its variance-covariance matrix to be numerically unstable, the techniques of the numerical analysts have direct bearing on understanding the econometrician's problems with multicollinearity. The important efforts of the numerical analysts relevant to this study are contained in Businger and Golub (1965), Golub (1968), Golub and Reinsch (1970), Lawson and Hansen (1974), Stewart (1973) and Wilkinson (1965).

Few of the techniques of the numerical analysts have been directly absorbed in econometrics, although a close cousin, eigensystems, has been an econometric staple for decades. This is in part because of a lack of communication between the two disciplines exacerbated by awkward differences in notation. Furthermore, the numerical analysts have placed much of their emphasis upon determining which columns of a data matrix can be discarded with least sacrifice to subsequent analysis,[6] a solution that is rarely open to the econometrician whose theory has already determined those variates that must always be present or those that may be deleted, but not solely on grounds of numerical stability. Nevertheless, with a change in emphasis, the numerical analysts' techniques have much to offer the econometrician in dealing with his twin concerns with multicollinearity.

For many years the eigensystem of the cross-products matrix X'X has been employed in dealing with multicollinearity. Kloeck and Mennes (1960), for example, depict several ways of using the principal components of X or related matrices to reduce the ill effects of collinearity. In a direction more useful for diagnostic purposes, Kendall (1957) and Silvey (1969) have suggested using the eigenvalues of the cross-products matrix X'X as a key to the presence of multicollinearity. In fact, the use of the eigenvalues of X'X is very closely related to one of the principal tools of numerical analysis, the singular-value decomposition (SVD) of the data matrix X. The intimate connection between these two notions, eigensystems and singular values, will be discussed below, but suffice to say that the best inroad numerical analysis has had into the econometric analysis of multicollinearity has been indirectly through the Kendall-Silvey line of research employing eigenvalues, and this paper draws heavily upon it.

Silvey (1969) correctly concludes that multicollinearity is easily discernable by the presence of a "small" eigenvalue of X'X, a fact first noted by Kendall. Silvey fails, however, to aid us in knowing when an eigenvalue is small, and it is here that numerical analysis adds important insights. Silvey also provides the basis for a mechanism for decomposing the estimated variance of each regression coefficient in a manner that can illuminate the degradation of each coefficient caused by collinear relationships, but fails to exploit this use.

This paper, then, 1) applies the relevant techniques of numerical analysis to Silvey's suggestion in order to provide a set of indexes that signal the presence of one or more near dependencies among the columns of X and 2) adapts the Silvey regression-variance decomposition in a manner that can be coordinated with the above indexes to uncover those variates that are involved in particular near dependencies and the degree to which their estimated ceofficients are being degraded by the presence of the near dependencies.

Part 2: TECHNICAL BACKGROUND

2.0 Introduction

In this section we develop the two principal tools of analysis employed in this paper, the singular-value decomposition (SVD) of a matrix X (and its associated notions of the conditioning of X), and the decomposition of the estimated regression variance in a manner corresponding to the SVD. As noted, none of these concepts is new; the innovation is their combination in a manner that helps the econometrician solve the two diagnostic problems of multicollinearity stated at the outset: detection and assessment of damage.

2.1 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 on K economic variates, may be decomposed as

$$X = U\Sigma V' \tag{2.1}$$

where $U'U = V'V = I_K$ and Σ is diagonal with non-negative diagonal elements σ_k, k=1 ... K,[2,3] called the singular values of X.

The singular-value decomposition is closely related to the familiar concepts of eigenvalues and eigenvectors, but has useful differences. Noting that $X'X = V\Sigma^2V'$, we see that V is an orthogonal matrix that diagonalizes X'X and hence the diagonal elements of Σ^2, the squares of the singular values, must be the eigenvalues of the real symmetric matrix X'X. Further, the orthogonal columns of V must be the eigenvectors of X'X (and, similarly demonstrated, the columns of U are the eigenvectors of XX').

The singular-value decomposition of the matrix X, therefore, provides information that encompasses that given by the eigensystem of X'X. As a practical matter, however, it is preferable to deal with the SVD of X rather than the eigensystem of X'X, because to the extent that X is ill conditioned, X'X is ill conditioned by the square.[4] Hence, calculations of the eigensystem based on X'X will be very much more unstable than will calculations of the singular values based directly on the matrix X and, of course, it is precisely on the case that X is ill conditioned that our interest is centered.[5]

Exact Linear Dependencies: Rank Deficiency

In the first instance let us assume X has exact linear dependencies among its columns, a case rarely encountered in econometric practice, so that rank X = r < K. Since, in the SVD of X, U and V are orthogonal (and hence are necessarily of full rank), we must have rank X = rank Σ. There will therefore be exactly as many zero elements along the diagonal of Σ as the nullity of X, and the SVD in (2.1) may be partitioned as

$$X = U\Sigma V' = U\begin{bmatrix}\Sigma_{11} & 0\\ 0 & 0\end{bmatrix}V' \tag{2.2}$$

where Σ_{11} is rxr and nonsingular. Postmultiplying by V and further partitioning we obtain

$$X\begin{bmatrix}V_1 & V_2\end{bmatrix} = \begin{bmatrix}U_1 & U_2\end{bmatrix}\begin{bmatrix}\Sigma_{11} & 0\\ 0 & 0\end{bmatrix} \tag{2.3}$$

where V_1 is K x r $\quad U_1$ is T x r

V_2 is K x (K-r) $\quad U_2$ is T x (K-r).

(2.3) results in the two matrix equations

$$X\,V_1 = U_1\,\Sigma_{11} \tag{2.4}$$

$$X\,V_2 = 0. \tag{2.5}$$

Interest centers on (2.5) 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.

If, then, X possesses K-r exact linear relation among its columns (and the computers possessed exact arithmetic), the number K-r of such dependencies would equal the number of zero singular values, and the variates involved in each of these dependencies would be determined by the non-zero elements of V_2 in (2.5).

Needless to say, in applied econometrics, the interrelations among the columns of X are not exact dependencies, and computers deal in finite, not exact, arithmetic. Exact zeros for the singular values or for the elements of V_2 will therefore rarely, if ever, occur. In general, then, it will be difficult to determine the nullity of X (as determined by zero σ's) or which columns of X do not enter into specific linear relationships (as determined by the zeros of V_2). There is nevertheless suggested in the foregoing the idea that each near linear dependency among the columns of X will result in a small singular value, a small σ. This corresponds to Silvey's notion that the presence of collinearity is revealed by the existence of a small eigenvalue. The question now is to determine what is small. Although what ultimately is to be judged as large or small must remain an empirical question, we are greatly aided in answering this question by the notion of a condition number of a matrix X.

The Condition Number

Intuition is pressed to define a notion of an ill-conditioned matrix. One is tempted to say a matrix is ill conditioned if it "almost isn't of full rank," or "if its inverse almost doesn't exist", two obviously absurd statements. Yet this is in effect what is meant when it is said that an ill-conditioned square matrix is one with a small determinant (or an ill-conditioned rectangular matrix is one with a small det X'X). A small determinant, of course, has nothing to do with invertability of a matrix, for the matrix αI_n has as its determinant the number α^n which can be made arbitrarily small; and yet it is clear that A^{-1} always exists for $\alpha \neq 0$ and is readily calculated as $\alpha^{-1} I_n$.

It is equally infeasible to obtain information on the invertability (conditioning) of a matrix from the smallness of some diagonal elements of a triangularization 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.[6]

A means for defining the conditioning of a matrix that accords somewhat with intuition and avoids the pitfalls of the above techniques is afforded by the singular-value decomposition. The motivation behind this technique derives from a more correct method of determining when 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 (a generalization of the well-known Euclidean norm of a vector)[7] in comparison with the spectral norm of the given matrix itself. In essence this measure, called the <u>condition number</u>, tells us how difficult it is to compute the inverse of a given matrix in the sense of specifying how sensitive the elements of A^{-1} are to small perturbations in the elements of A [Wilkinson (1965)]. The larger the condition number the more ill conditioned the given matrix.

Two examples aid our understanding of the condition number. Consider first the matrix $A = \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix}$. Clearly as $\alpha \to 1$, this matrix tends toward perfect singularity. The singular values of A are readily shown[8] to be $(1 \pm \alpha)$ and those of A^{-1} to be $(1 \pm \alpha)^{-1}$. Hence, as $\alpha \to 1$ the product $||A||\ ||A^{-1}|| = (1+\alpha)(1-\alpha)^{-1}$ explodes; the spectral norm of A^{-1} is large relative to that of A. A is ill conditioned for small α.

By way of contrast consider the admittedly well conditioned matrix introduced above, $B = \begin{bmatrix} \alpha & 0 \\ 0 & \alpha \end{bmatrix}$. As we have seen, the often held intuitive feeling that B becomes ill conditioned as $\alpha \to 0$ is incorrect, and this is correctly reflected in the condition number, for $\|B\| = \alpha$ and $\|B^{-1}\| = \alpha^{-1}$ and the product $||B|| \, ||B^{-1}|| = \alpha\alpha^{-1} = 1 \neq 0$ 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 $\alpha \neq 0$.

The conditioning of any square matrix A can be summarized, then, by a condition number $\kappa(A)$ defined as the product of the maximal singular value of A (its spectral norm) times the maximal singular value of A^{-1}. This concept is readily extended to a rectangular matrix and can be calculated without recourse to its inverse. From the SVD 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 non-zero diagonal elements inverted.[9] Hence the singular values of X^+ are merely the inverses of those of X, and the maximal singular value of X^+ is the reciprocal of the minimum (non-zero) singular value of X. We may therefore define the condition number of X as

$$\kappa(X) = \frac{\sigma_{max}}{\sigma_{min}} \geq 1 . \qquad (2.6)$$

It is readily shown that the condition number of any matrix with orthonormal columns is unity, and hence $\kappa(X)$ reaches its lower bound in this cleanest of all possible cases.

Near Linear Dependencies: How Small is Small

We have seen that for each exact linear dependency among the columns of X there is one zero singular value. Extending this property to near dependencies leads one to suggest, as did Kendall (1957) and Silvey (1969), that the presence of near dependencies (multicollinearity) will result in "small" singular values (or eigenvalues). This suggestion does not include a means for determining what small is. The preceding discussion of condition numbers, however, does provide such a measure. It was learned there that the degree of ill conditioning depends upon how small the minimum singular value is relative to the maximum singular value, i.e., σ_{max} provides a yardstick against which smallness can be measured.[10] In this connection, it proves useful to define

$$\eta(k) \equiv \frac{\sigma_{max}}{\sigma_k} \quad (k \neq \text{max index}) \tag{2.7}$$

to be the k-th condition index of the TxK data matrix X. There are, of course, K-1 such index values, the largest of which is also the condition number of the given matrix. A singular value that is small relative to its yardstick σ_{max}, then, has a high condition index.

We may therefore extend the Kendall-Silvey suggestion as follows: there are as many near dependencies among the columns of a data matrix X as there are high condition indexes (singular values small relative to σ_{max}). Two points regarding this extension must be emphasized.

First, we have not merely redirected the problem from one of determining when small is small to one of determining when large is large. As we saw above, taken alone, the singular values (or eigenvalues) shed no light on the conditioning of a data matrix. Equally well conditioned problems can have arbitrarily low singular values.[11] Determining when a singular value is small, then, has no relevance to determining the presence of a near dependency causing a data matrix to be ill conditioned. We did see however, in our discussion of the condition number, that determining when a singular value is small relative to σ_{max} (or, equivalently, determining when a condition index is high) is directly related to this problem. The meaningfulness of the condition index in this context is verified in the empirical studies of part 3.

Second, even if there is measurable meaning to the term "large" in connection with condition indexes, there is no a priori basis for determining how large a condition index must be before there is evidence of collinear data or, even more importantly, evidence of data so collinear that its presence is degrading or harming regression estimates. Just what is to be considered a large condition index is a matter to be empirically determined, and the experiments of part 3 are aimed at aiding such an understanding. There we learn that dependencies begin to be observable with condition indexes as low as 5 or 10, and, in comparison with other well-known standards, such as correlations and R^2's, become quite strong with values of 30 or 100.

The use of the condition index, then, extends the Kendall-Silvey suggestion in two ways. First, practical experience will allow an answer to the question of when small is small (or large is large) and second, the simultaneous occurrence of several large η's keys the simultaneous presence of more than one near dependency.

2.2 The Estimated Regression-Variance Decomposition

As we have seen, when any one singular value of a data matrix is small relative to σ_{max}, we interpret it as indicative of a near dependency among the columns of X associated with that singular value. In this section, reinterpreting and extending the work of Silvey (1969), we show how the estimated variance of each regression coefficient may be decomposed into a sum of terms each of which is associated with a singular value, thereby providing means for determining the extent to which near dependencies (having high condition indexes) degrade (become a dominant part of) each variance. This decomposition provides the link between the numerical analysis of a data matrix X as embodied in its singular-value decomposition, and the quality of the subsequent regression analysis using X as a data matrix as embodied in the estimated variance-covariance matrix of b.[12]

The variance-covariance matrix of the least-squares estimator $b = (X'X)^{-1}X'y$ is, of course, $\sigma^2(X'X)^{-1}$, where σ^2 is the common variance of the components of the T disturbances ε in the linear model $y = X\beta+\varepsilon$. Using the SVD of $X = U\Sigma V'$, we get

$$V(b) = \sigma^2(X'X)^{-1} = \sigma^2 V\Sigma^{-2}V' \tag{2.8}$$

or, for the k-th component of b

$$\mathrm{var}(b_k) = \sigma^2 \sum_j \frac{v_{kj}^2}{\sigma_j^2} \tag{2.9}$$

where the σ_j's are the singular values and $V \equiv (v_{ij})$.

(2.9), it is noted, decomposes var (b_k) into a sum of components, each associated with one and only one of the K singular values σ_k (or eigenvalues σ_k^2). Since these σ_k^2 appear in the denominator, other things equal, those components associated with near dependencies, i.e., with small σ_k, will be large relative to the other components. This suggests, then, that an unusually high proportion of the variance of two or more coefficients[13] concentrated in components associated with the same small singular values provides evidence that the corresponding near dependency is causing problems. Let us pursue this suggestion.

The variance-component proportions are readily displayed as follows. Let

$$\Phi_{kj} \equiv \frac{v_{kj}^2}{\sigma_j{}^2} \text{ and } \Phi_k \equiv \sum_{j=1}^{K} \Phi_{kj} \qquad k = 1 \ldots K. \tag{2.10}$$

Further, define the variance-component proportions as

$$\pi_{kj} \equiv \frac{\Phi_{kj}}{\Phi_k}\ , \quad k, j = 1 \ldots K. \tag{2.11}$$

The investigator seeking patterns of high variance-component proportions will be aided by a summary table (a Π matrix) of the form

Variance-Component Proportions

Components of

$$\begin{array}{cc|ccccc} & & \text{var}(b_1) & \text{var}(b_2) & \cdot\ \cdot & \cdot & \text{var}(b_k) \\ \hline & \sigma_1 & \pi_{11} & \pi_{12} & \cdot\ \cdot & \cdot & \pi_{1K} \\ & \sigma_2 & \pi_{21} & \pi_{22} & \cdot\ \cdot & \cdot & \pi_{2K} \\ & \cdot & \cdot & \cdot & & & \cdot \\ \text{associated with} & \cdot & \cdot & \cdot & & & \cdot \\ & \cdot & \cdot & \cdot & & & \cdot \\ & \sigma_K & \pi_{K1} & \pi_{2K} & \cdot\ \cdot & \cdot & \pi_{KK} \end{array} \tag{2.12}$$

Notice that the π_{kj} make use of the SVD information on near dependencies in a way that is directly applicable to examining their effects on regression estimates.

2.3 Two Interpretive Considerations

The next part will contain detailed experiments using the two tools developed here, the SVD and its associated Π-matrix of variance-component proportions. These experiments are designed to provide experience in the behavior of these tools when employed for analyzing multicollinearity, its detection and an assessment of the damage it has caused to regression estimates. Before proceeding to these experiments, however, it will be necessary to develop two important interpretive properties of the Π matrix of variance-component proportions. An example of these two properties completes the section.

Near Collinearity Nullified by Near Orthogonality

In the variance decomposition (2.9), small σ_j's, other things equal, lead to large components of var (b_k). However, not all var (b_k)'s need be adversely affected by a small σ_j, for the v_{kj}^2 in the numerator may be even smaller. In the extreme case where $v_{kj} = 0$, var (b_k) would be unaffected by any near dependency among the columns of X that would cause σ_j to become even very small. As is shown in Belsley and Klema (1975), the v_{kj} are equal to zero exactly as the kth and jth columns of X are orthogonal. The proof to this intuitively plausible statement is lengthy and need not be repeated here, for it reflects a fact well known to econometricians; namely, that the introduction into regression analysis of a variate orthogonal to all preceding variates will not change the regression estimates or the standard errors of the coefficients of the preceding variates. Thus, if two very collinear variates (near multiples of one another), that are also mutually orthogonal to all prior variates are added to a regression equation, the estimates of the prior coefficients and their variances must also be unaffected. In terms of the variance decomposition (2.9), this situation results in at least one σ_j (corresponding to the two closely collinear variates) which is very small, and which has no weight in determining any of the var (b_k) (for k corresponding to the initially included variates). Clearly, the only way this can occur is for the v_{kj} between the prior variates and the additional orthogonal variates to be zero. Hence we have the result that, in the SVD of $X = [X_1 X_2]$ with $X_1{}'X_2 = 0$, it is always possible to find[14] a V matrix with the form

$$V = \begin{bmatrix} V_{11} & 0 \\ 0 & V_{22} \end{bmatrix} .$$

Thus we see that the bad effects of collinearity, resulting in small σ's, may be mitigated for some coefficients by near orthogonality, resulting in small v_{kj}'s.

At Least Two Variates Must Be Involved

At first it would seem that the concentration of the variance of any one regression coefficient in any one of its components could signal that multicollinearity may be causing problems. However, since two or more variates are required to create a near dependency, it must be that two or more variances must be adversely affected by high variance components associated with a single singular value (i.e., a single near dependency).

To illuminate this, consider a data matrix X with mutually orthogonal columns - the best possible experimental data. Our previous result immediately implies that the V matrix of the singular-value decomposition of X is diagonal, since all $v_{ij} = 0$ for $i \neq j$. Hence the associated Π matrix of variance-component proportions must take the form

$$\begin{array}{ccc} & & \text{Proportions in} \\ & & \begin{array}{cccc} \mathrm{Var}(b_1) & \mathrm{Var}(b_2) & \cdots & \mathrm{Var}(b_k) \end{array} \\ \text{associated with} & \begin{array}{c} \sigma_1 \\ \sigma_2 \\ \\ \sigma_k \end{array} & \left[\begin{array}{cccc} 1 & & 0 & \\ & 1 & & \\ 0 & & \ddots & \\ & & & 1 \end{array} \right] \end{array} .$$

It is clear that a high proportion of any variance associated with a single singular value is hardly indicative of multicollinearity, for the variance proportions here are those for an ideally conditioned, orthogonal data matrix. Reflecting the fact that two or more columns of X must be involved in any near dependency, the degradation of a regression estimate due to collinearity can be observed only when a single singular value σ_j is associated with a large proportion of the variance of two or more coefficients. If, for example, in a case for K = 5, columns 4 and 5 of X are highly collinear and all other columns are mutually orthogonal, we would expect a variance-component Π matrix that has the form, say,

$$\begin{array}{ll|ccccc|}
 & & \multicolumn{5}{c}{\text{Proportions in}} \\
\text{associated with} & & \text{Var}(b_1) & \text{Var}(b_2) & \text{Var}(b_3) & \text{Var}(b_4) & \text{Var}(b_5) \\
 & \sigma_1 & 1.0 & 0 & 0 & 0 & 0 \\
 & \sigma_2 & 0 & 1.0 & 0 & 0 & 0 \\
 & \sigma_3 & 0 & 0 & 1.0 & 0 & 0 \\
 & \sigma_4 & 0 & 0 & 0 & 1.0 & 0.9 \\
 & \sigma_5 & 0 & 0 & 0 & 0 & 0.1
\end{array} .$$

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

An Example

An example of the preceding two interpretive considerations is useful at this point. Consider the 6x5 data matrix

$$X=[X_1X_2] = \left[\begin{array}{ccc|cc}
-74 & 80 & 18 & -56 & -112 \\
14 & -69 & 21 & 52 & 104 \\
66 & -72 & -5 & 764 & 1528 \\
-12 & 66 & -30 & 4096 & 8192 \\
3 & 8 & -7 & -13276 & -26552 \\
4 & -12 & 4 & 8421 & 16842
\end{array}\right] .$$

This matrix, essentially due to Bauer (1971), has the property that its fifth column is exactly twice its fourth, and both of these are in turn orthogonal to the first three columns (which are not, however, orthogonal to to each other). That is, X_2 is of rank 1 and $X_1'X_2 = 0$. We therefore know from the foregoing that, in the SVD of X, 1) one of the singular values associated wtih X_2 will be zero (i.e., within the machine tolerance of zero),[15] and 2) in $V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}$, $V_{12} = V_{21}' = 0$.

Indeed application of the program MINFIT[16] to obtain the singular-value decomposition of X gives

$$V = \left[\begin{array}{ccc|cc} .548 & -.625 & .556 & .15\text{x}10^{-18} & -.54\text{x}10^{-14} \\ -.836 & .383 & .393 & .22\text{x}10^{-19} & -.47\text{x}10^{-14} \\ .033 & .680 & .733 & .16\text{x}10^{-18} & -.73\text{x}10^{-14} \\ \hline -.64\text{x}10^{-15} & -.22\text{x}10^{-15} & .91\text{x}10^{-14} & .447 & .894 \\ .32\text{x}10^{-15} & .10\text{x}10^{-15} & -.46\text{x}10^{-14} & .894 & .447 \end{array}\right]$$

with singular values[17]

$$\begin{aligned} \sigma_1 &= 170.7 \\ \sigma_2 &= 60.5 \\ \sigma_3 &= 7.6 \\ \sigma_4 &= 36368.4 \\ \sigma_5 &= 1.3 \times 10^{-12} . \end{aligned}$$

A glance at V verifies the off-diagonal blocks are small - all of the order of 10^{-14} or smaller - and well within the effective zero tolerance of the computational precision. Only somewhat less obvious is that one of the σ_j is zero. σ_5 is of the order of 10^{-12} and would seem to be nonzero relative to the machine tolerance, but, as we have seen, the size of each σ_j has meaning only relative to σ_{max}, and in this case $\frac{\sigma_5}{\sigma_{max}} = \eta^{-1}(5) \leq 10^{-16}$, well within the machine zero.

The Π matrix of variance-component proportions for this data matrix is given in Table 0.

Table 0

Variance-Component 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

Several of its properties are noteworthy. First, we would expect that the small singular value σ_5 associated with the linear dependency $X_4 = .5X_5$ would dominate several variances - at least those of the two variates involved - and this is seen to be the case; the component associated with σ_5 accounts for virtually all the variances of both b_4 and b_5.

Second, we would expect that the orthogonality of the first three columns of X from the two involved in the linear dependency would isolate their estimated coefficients from collinearity's deleterious effects. Indeed, it is noted that the components of these three variances associated with σ_5 are very small, .003, .005 and .003 respectively.[18] This point serves also to exemplify that the analysis suggested here aids the user not only to determine which regression estimates are degraded by the presence of multicollinearity, but also which are not adversely affected and may therefore be salvaged.

Third, a somewhat unexpected result is apparent. The singular value σ_3 accounts for 97% or more of var(b_1), var(b_2) and var(b_3). This suggests that a second near dependency is present in X, one associated with σ_3, that involves the first three columns only. This, in fact, turns out to be the case, and we shall reexamine this example in Part 4, once we have gained further experience in interpreting the magnitudes of condition indexes and variance-component proportions.

Fourth, to the extent that there are two separate near dependencies in X (one among the first three columns, one between the last two), the Π matrix provides a means for determining which variates are involved in which near dependency. This property of the analytic framework being presented here is important, because it is not true of alternative means of analyzing near dependencies among the columns of X. One could hope, for example, to investigate such near dependencies by regressing selected columns of X on other columns or to employ partial correlations. But to do this in anything other than a shotgun manner would require prior knowledge of which columns of X would be best preselected to regress on the others, and to do so when there are several coexisting near dependencies would prove a terrible burden. Usually the econometrician, when presented with a specific data matrix, will have no rational means for such a preselection process, and can avoid the problem entirely, through the use of the Π matrix which displays all such near dependencies, treating all columns of X symmetrically, and requiring no prior information.

2.4 A Suggested Test Procedure

The foregoing discussion suggests a practical procedure for 1) testing for the presence of one or more near dependencies among the columns of a data matrix, and 2) assessing the degree to which each such dependency degrades the regression estimates based on that data matrix.

The Test

It is suggested that an appropriate means of detecting harmful collinearity is the double condition of

1) The presence of high variance-component proportions for two or more estimated regression variances associated with

2) A single singular value judged to have a high condition index.

The number of condition indexes deemed large in step 2 identifies the number of near dependencies among the columns of the data matrix X, and the magnitudes of these high condition numbers provides a measure of their relative "tightness". Furthermore the determination in Step 1 of large variance-component proportions associated with a high condition index identifies those variates that are involved in the corresponding near dependency, and the magnitude of these proportions in conjunction with the high condition index provides a measure of the degree to which the corresponding regression estimate has been degraded by the presence of multicollinearity.[19].

Examining the Near Dependencies

Once the variates involved in each near dependency have been identified by their high variance-component proportions, the near dependency itself can be examined, for example, by regressing one of the variates involved on the others. Another procedure is suggested by (2.5). Since V_2 in (2.5) has rank (K-r) we may partition X and V_2 to obtain

$$\begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} V_{21} \\ V_{22} \end{bmatrix} = X_1 V_{21} + X_2 V_{22} = 0 \qquad (2.13)$$

where V_{21} is chosen nonsingular and square.

Hence the dependencies among the columns of X are displayed as

$$X_1 = -X_2 V_{22} V_{21}^{-1} \equiv X_2 G \text{ where } G \equiv -V_{22} V_{21}^{-1}. \qquad (2.14)$$

The elements of G, calculated directly from those of V, provide alternative estimates of the linear relation between those variates in X_1 and those in X_2. Of course, (2.13) holds exactly only in the event that the linear dependencies of X are exact, and is not clearly interpretable otherwise. It is also straightforward to show that when the linear dependencies are exact, (2.13) provides identical estimates as those given by OLS. It seems reasonable therefore to employ OLS as the descriptive mechanism for displaying the linear dependencies once the variates involved are discerned in Step 2. It is important to reiterate the point made earlier that OLS does not and cannot substitute for the test suggested above, for OLS can be rationally applied only after it has first been determined how many dependencies there are among the columns of X and which variates are involved. The test suggested here requires no prior knowledge of the numbers of near dependencies involved or of the variates involved in each; it discovers this information - treating all columns of X symmetrically and requiring that none be choosen (as OLS requires) to become the "dependent" variable.

What is "large" or "high"

Just what constitutes a "large" condition index or a "high" variance-component proportion are matters that can only be decided empirically. We turn in Part 3 to a set of systemmatic experiments designed to shed light on this matter, but to provide a meaningful background against which to interpret those empirical results it is first useful to give a more specific idea of what it means for collinearity to harm or to degrade a regression estimate.

2.5 The Ill Effects of Collinearity

The ill effects that result from regression based on multicollinear data are two: one computational, one statistical.

Computational Problems

Computationally, it can be shown[20] that solutions to a set of least-squares normal equations, (or in general, a solution to a system of linear equations) contain a number of digits whose meaningfulness is limited by the conditioning of the data in a manner directly related to the condition number. Indeed the condition number gives a multiplication factor by which imprecision in the data works its way through to imprecision in the solution to a linear system of equations. Somewhat loosely, if data are known to n significant digits, and the condition number of the matrix A of a linear system Ax=b is of order of magnitude 10^r, then a small change in the data in its last place can (but need not) affect the solution $x=A^{-1}b$ in the (n-r)th place. Thus, if GNP data are trusted to 4 digits, and the condition number of (X'X) is 10^3, then a shift in the 5th place of GNP (which, since only the first four digits count, results in what must be considered an observationally equivalent data matrix), could affect the least-squares solution in its 2nd (5-3) significant digit. Only the first digit is therefore trustworthy, the others potentially being worthless, arbitrarily alterable by modifications in X that do not affect the degree of accuracy to which the data are known. Needless to say, had the condition number of X'X been 10^4 or 10^5 in this case, one could trust none of b's significant digits. This computational problem in the calculation of least-squares estimates may be minimized[21] but never removed. The econometrician's intuitive distrust of estimates based on ill-conditioned data is therefore justified.

Statistical Problems

Statistically, as is well known, the problem introduced by the presence of multicollinearity in a data matrix is the decreased precision with which statistical estimates conditional[22] upon those data may be known; that is, multicollinearity causes the conditional variances to be high. This problem reflects the fact that when data are ill conditioned some data series are nearly linear combinations of others and hence add very little new, independent information from which additional statistical information may be gleaned.

Needless to say, inflated variances are quite harmful to the use of regression as a basis for hypothesis testing, estimation and forecasting. All econometricians have had the suspicion that an important test of significance has been rendered inconclusive through a needlessly high error variance induced by collinear data, or that a confidence interval or forecast interval is uselessly large, reflecting the lack of properly conditioned data from which appropriately refined intervals could conceivably have been estimated.

Both of the above ill effects of collinear data are most directly removed through the introduction of new and well conditioned data.[23] In econometric analysis, however, new data are, more often than not, not available, and when they are, can be acquired only at great cost in time and effort. The usefulness of having diagnostic tools that key the presence of collinearity and even isolate the variates involved is therefore apparent, for with them the investigator can at least determine whether the effort to correct for collinearity (collect new data or apply

Bayesian techniques) is potentially worthwhile, and perhaps he can learn a great deal more. But just how much can be learned? To what extent can diagnostics tell the degree to which collinearity has caused harm?

Harmful vs. Degrading Collinearity

At the outset it should be noted that not all collinearity need be harmful. We have already seen, in the example of the Bauer matrix given in Section 2.3, that near orthogonality can isolate some regression estimates from the presence of even extreme collinearity. If by chance, the investigator's interest centers only on those unaffected parameter estimates, clearly no problem exists.[24] In a less extreme, and therefore a practically more useful example, we recall from (2.9) that the estimated variance of kth regression coefficient, var(b_k) is $s^2 \sum \frac{v_{kj}^2}{\sigma_j^2}$ where s^2 is the estimated regression error variance. If s^2 is sufficiently small, it may be that particular var(b_k) are small enough for specific testing purposes in spite of large components in the $\frac{v_{kj}^2}{\sigma_j^2}$ terms resulting from near dependencies. If, for example, an investigator is only interested in whether a given coefficient is significantly positive, and is able, even in the presence of collinearity, to accept that null hypothesis on the basis of the relevant t-test, then collinearity has caused no problem. Of course, the resulting forecasts or point estimates may have wider confidence intervals than would be needed to satisfy a more ambitious researcher, but for the limited purpose of the test of significance initially proposed, collinearity has caused no practical harm. These cases serve to exemplify the pleasantly pragmatic philosophy that collinearity doesn't hurt so long as it doesn't bite.

Providing evidence that collinearity has harmed estimation, however, is greatly more difficult. To do this one must show, for example, that a prediction interval that is too wide for a given purpose could be appropriately narrowed if made statistically conditional on better conditioned data (or that a confidence interval could be appropriately narrowed or the computational precision of a point estimator appropriately increased). To date there is no procedure that provides such information. If, however, the researcher were provided with information that 1) there are strong near dependencies among the data, so that collinearity is potentially a problem, and 2) that variances of parameters (or confidence intervals based on them) that are of interest to him have a large proportion of their magnitude associated with the presence of the collinear relation(s), so that collinearity is potentially harmful, then he would be a long way toward deciding whether the costs of corrective action were warranted. In addition such information would indicate when variances of interest were not being adversely affected and so could be relied upon without further action. The above information is, of course, precisely that provided by the condition indexes and high variance-component proportions used in the two-pronged test suggested earlier. And so we shall say that when this joint test has been met, the affected regression coefficients have been degraded (but not necessarily harmed) by the presence of multicollinearity, degraded in the sense that the magnitude of the estimated variance is being determined greatly by the presence of a collinear relation, and there is therefore a presumption that confidence intervals, prediction intervals and point estimates based on this estimate could be refined if need be by introducing better conditioned data.

At what point do estimates become degraded? Future experience may provide a better answer to this question, but for the experiments of the next section we will take as a beginning rule of thumb that estimates are degraded when two or more variances have at least half of their magnitude associated with a single singular value.

Part 3: EXPERIMENTAL EXPERIENCE

3.0 Introduction

The test for the presence of degrading collinearity suggested at the end of the prior section requires the joint occurrence of high variance-component proportions for two or more coefficients associated with a single singular value having a "high" condition index. Knowledge of what constitutes a high condition index must be empirically determined, and it is the purpose of this section to describe a set of experiments that have been designed to provide such experience.

3.1 The Experimental Procedure

Each of the three experiments reported below examines the behavior of the singular values and variance-component proportions of a series of data matrices that are made to become systematically more and more ill conditioned by the presence of one or more near dependencies constructed to become more nearly exact.

Each experiment begins with a "basic" data set X of T observations on K_1 variates. T, which is unimportant, is around 24-27, and K_1 is 3-5, depending upon the experiment. In each case the basic data series are chosen either as actual economic time series or as constructs that are generated randomly but having similar means and variances as actual economic time series.

These basic data series are used to construct additional collinear data series displaying increasingly tighter linear dependencies with the basic series as follows. Let c be a K_1-vector of constants and construct

$$x_i = Xc + e_i, \tag{3.1}$$

where e_i is generated randomly with mean zero and variance $\sigma_i^2 = 10^{-i}\,\sigma_{Xc}^2$,

$\sigma^2_{Xc} \equiv \frac{1}{T} c'X'Xc$, $i = 0 \ldots 4$. Each x_i, then, is, by construction, a known linear combination, Xc, of the basic data series plus a zero-mean random error term, e_i, whose variance becomes smaller and smaller (that is, the dependency becomes tighter and tighter) with increasing i. In the $i = 0$ case the variance in x_i due to the error term e_i is seen to be equal to variance in x_i due to the systematic part Xc. In this case, then, the imposed linear dependency is weak. The sample correlation between x_i and Xc in these cases tends toward .4 to .6. By the time $i = 4$, however, only 1/10,000 of x_i's variance is due to additive noise, and the dependency between x_i and Xc is tight, displaying correlations very close to unity. A set of systematically increasingly ill conditioned data matrices may therefore be constructed by augmenting the basic data matrix X with the x_i's, i.e., by constructing the set

$$X(i) = [X\ x_i] \qquad i = 0, . , 4. \tag{3.2}$$

The experiments are readily extended to the analysis of matrices possessing two or more simultaneous near dependencies by the addition of more data series similarly constructed from the basic series. Thus, for given K_1 vector b, let

$$z_j = Xb + u_j \tag{3.3}$$

where u_j is random with mean zero and variance $\sigma_j{}^2 = 10^{-j}\,\sigma^2_{Xb}$, $j = 0 \ldots 4$. Experimental matrices with two linear dependencies of varying strengths are constructed as

$$U(i,j) = [X\ x_i\ z_j] \qquad i, j = 0 .. 4 . \tag{3.4}$$

In the third experiment to follow, three simultaneous dependencies are examined.

The Choice of the X's

As mentioned, the data series chosen for the basic matrices X were either actual economic time series or variates constructed to have similar means and variances as actual economic time series. The principle of selection was to provide a basic data matrix that was reasonably well conditioned so that all significant ill conditioning could be controlled through the introduced dependencies such as (3.1).[1]

The various series of matrices that comprise any one experiment all have the same basic data matrix and differ only in the constructed near dependencies used to augment them. Within any one such series, the augmenting near dependencies become systematically tighter with increased i or j, and it is in this sense that we can speak meaningfully of what happens to condition indexes and variance-component proportions as the data matrix becomes "more ill conditioned," or "more nearly singular," or "the near dependencies get tighter," or "the degree of collinearity increases."

Experimental Shortcomings

The experiments given here, while not Monte Carlo experiments,[2] are nevertheless subject to a similar weakness; namely, the results depend upon the specific experimental matrices chosen and cannot be generalized to different situations with assurance. It has been attempted, therefore, within the necessarily small number of experiments reported here, to choose basic data matrices using data series and combinations of data series representing as wide a variety of economic circumstances as possible. Needless to say, not all meaningful economic cases can be considered, and the reader will no doubt think of cases he would rather have seen analyzed. However, the cases offered here are sufficiently varied that any systematic patterns that emerge from them are worthy of being reported and will certainly provide a good starting point for any refinements that subsequent experience will suggest.

The Need for Column Scaling

Data matrices that differ from one another only by the scale assigned the columns (matrices of the form XD, where D is a nonsingular diagonal matrix) represent equivalent economic structures; it doesn't matter for example, whether one specifies the model in dollars, cents, or billions of dollars. Such scale changes do, however, affect the numerical properties of the data matrix and result in very different singular-value decompositions and condition indexes.[3] Without further adjustment, then, we have a situation in which near dependencies among structurally equivalent economic variates (differing only in the units assigned them) can result in greatly differing condition indexes. Clearly the condition indexes can provide no stable information to the econometrician on the degree of collinear among the X variates in such a case. It is necessary, therefore, to standardize the data matrices corresponding to equivalent economic structures in a way that makes comparisons of condition indexes meaningful in an econometric application. A natural standardization process is to scale each column to have unit length.[4] This scaling is natural because it transforms a data matrix X with mutually orthogonal columns, the standard of ideal data, into a matrix whose singular values and condition indexes would all be unity, the smallest (and therefore most ideal) condition indexes possible. Any other scaling would fail to reflect this desirable property: the more ideal the data, the closer the condition indexes come to their lowest possible value, unity.[5]

In all the experiments that follow, then, the data are scaled to have unit column length before being subjected to an analysis of their condition indexes and variance-component proportions. In the event that the linear relations between the variates are displayed, the estimated coefficients have been rescaled to their original units.

The Experimental Report

Selected tables displaying variance-component proportions (Π-matrices) and condition indexes will be reported for each experiment, showing how these two principal pieces of information change as the near dependencies get tighter.

Additional statistics, such as the simple correlations of the contrived dependencies and their R^2's as measured from relevant regressions, will also be reported to provide a link between the magnitudes of condition indexes (with which we have little experience) and these more familiar notions. It cannot be stated too strongly, however, that these additional statistics cannot substitute for information provided by the variance-component proportions and the condition indexes. In the experiments that follow, we know a priori which variates are involved in which relations and what the generating constants (the c in (3.1)) are. It is therefore possible to construct simple correlations between x_i and Xc and run regressions of x_i on X. In practice, of course, c is unknown and one does not know which elements in the data matrix are involved in which dependencies. These auxiliary statistics are, therefore, not available to the investigator as independent analytic or diagnostic tools. However, one can learn from the variance-component proportions which variates are involved in which relationships, and regressions may then be run among these variates to display the dependency. Furthermore the t-statistics that result from these regressions can be used in the standard way for providing additional descriptive evidence of the "significance" of each variate in the specific linear dependency. Once the analysis by condition indexes and variance-component proportions has been conducted, then, it can suggest useful auxiliary regressions as a means of "exhibiting" the near dependencies; but regression by itself, particularly, if there are two or more simultaneous near dependencies, cannot provide similar information.[6]

3.2 The Individual Experiments

Three experiments are conducted, each using a separate series of data matrices designed to represent different types of economic data and different types of multicollinearity. Thus "levels" data (Manufacturing Sales), "trended" data (GNP), "rate of change" data (inventory investment), "rates" data (unemployment) are all represented. Similarly, the types of collinearity generated include simple relations between two variates, relations involving more than two variates, simultaneous near dependencies, and dependencies among variates with essential scaling problems. The different cases of relations involving more than two variates have been chosen to involve different mixes of the different types of economic variables listed above. In each case the dependencies are generated from the unscaled (natural) economic data, and hence the various test data sets represent as closely as possible economic data with natural near dependencies.

Experiment No. 1: The X Series

The basic data set employed here is

$$X \equiv [\text{MFGS}, \text{*IVM}, \text{MV}]^{7}$$

where MFGS is manufacturers' shipments, total
*IVM is manufacturers' inventories, total

MV is manufacturers' unfilled orders, total,

and each series is in millions of dollars, annual 1947-1970 (T=24).

This basic data set provides the type of series that would be relevant, for example, to an econometric study of inventory investment.[8]

Two sets of additional dependency series are generated from X as follows:

$$w_i = MV + v_i \qquad i = 0, \ldots, 4 \qquad (3.5a)$$

with v_i generated randomly with mean zero and variance $\sigma_i^2 = 10^{-i} s_{MV}^2$ (denoted $v_i \leftrightarrow f(0, 10^{-i} s_{MV}^2)$) s_{MV}^2 being the sample variance of the MV series,

and

$$z_j = .8MFGS + .2{*}IVM + v_j \qquad (3.5b)$$

$v_j \leftrightarrow f(0,\ 10^{-j} s_{\hat{z}}^2)$,

$s_{\hat{z}}^2$ being the sample variance of $.8MFGS + .2{*}IVM$.

The w_i and z_j series were used to augment the basic data set to produce three sequences of matrices

$$\begin{aligned} X1(i) &\equiv \begin{bmatrix} X & w_i \end{bmatrix} & i &= 0 \,..\, 4 \\ X2(j) &\equiv \begin{bmatrix} X & z_j \end{bmatrix} & j &= 0 \,..\, 4 \qquad (3.6) \\ X3(i,j) &\equiv \begin{bmatrix} X & w_i & z_j \end{bmatrix} & i, j &= 0 \,..\, 4, \end{aligned}$$

each of which is subjected to analysis.

The dependency (3.5a) is a commonly encountered simple relation between two variates. Unlike more complex relations, it is a dependency whose presence can be discovered through examination of the simple correlation matrix of the columns of the X1(i) or X3(i,j). Its inclusion, therefore, allows us to learn how condition indexes and simple correlations compare with one another.

The dependency (3.5b) involves three variates, and hence would not generally be discovered through an analysis of the simple correlation matrix. (3.5b) was designed to present no difficult scaling problems; that is, the two basic data series MFGS and *IVM have roughly similar magnitudes and variations, and the coefficients (.8 and .2) are of the same order of magnitude. No one variate, therefore, dominates the linear dependency, masking the effects of others. This situation should allow both the identification of the variates involved and the estimation of the relation among them to be accomplished with relative ease.

Experiment No. 2: The Y Series

The basic data set employed here is

$$Y \equiv [\text{*GNP58}, \text{*GAVM}, \text{*LHTUR}, \text{*GV58}] ,$$

where GNP58 is GNP in 1958 dollars
GV58 is annual change in total inventories, 1958 dollars
GAVM is net corporate dividend payments
LHTUR is total labor hours, unemployment rate.

Each basic series has been constructed from the above series to have similar means and variances, but chosen to produce a reasonably well conditioned Y matrix. Data are annual, 1948 to 1974 (T=27). The variates included here, then, represent "levels" variates, (*GNP58), rates of change (*GN58), and "rates" (LHTUR).

Three additional dependency series are constructed as

$$\begin{aligned} u_i &= \text{GNP58} + \text{GAVM} + v_i \\ v_i &\leftrightarrow f(0,10^{-i}\, s_{\hat{u}}^2) \\ s_{\hat{u}}^2 &= \text{var (GNP58 + GAVM)} \end{aligned} \qquad (3.7a)$$

$$v_j = 0.1{*}GNP58 + {*}GAVM + v_j$$

$$v_j \leftrightarrow f(0,\ 10^{-j}\ s_{\hat{v}}^{2}) \qquad (3.7b)$$

$$s_{\hat{v}}^{2} = \text{var}\ (.1{*}GNP58 + {*}GAVM)$$

$$z_k = {*}GV58 + v_k$$

$$v_k \leftrightarrow f(0,\ 10^{-k}\ s^{2}_{*GV58}). \qquad (3.7c)$$

These data were used to augment Y to produce four series of test matrices

$$\begin{aligned} Y1(i) &= [Y\ u_i] && i = 0\ ..\ 4 \\ Y2(j) &= [Y\ v_j] && j = 0\ ..\ 4 \\ Y3(k) &= [Y,\ z_k] && k = 0\ ..\ 4 \qquad (3.8) \\ Y4(i,k) &= [Y\ u_i\ z_k] && i, k = 0\ ..\ 4\ . \end{aligned}$$

Dependency (3.7a) presents a relation among three variates with an essential scaling problem; namely, in the units of the basic data, the variation introduced by GNP58 is less than one percent that introduced by GAVM. The inclusion of GNP58 is therefore dominated by GAVM, and its effects will be somewhat masked and difficult to discern. Dependency (3.7b) is of a similar nature except that the scaling problem has been made even more extreme. These are "essential" scaling problems in that their effects cannot be undone through simple column scaling. Dependency (3.7c) is a simple relation between two variates, except in this case the variate is a rate of change, exhibiting frequent shifts in sign.

Experiment 3: The Z Series

The basic data matrix here is an expanded version of that in the previous experiment

$$Z = [\text{*GNP58, *GAVM, *LHTUR, DUM1, DUM2}]$$

where DUM1 is generated similar to GV58 and DUM2 is similar to GNP58, except that DUM1 and DUM2 were generated to have very low intercorrelation with the first three variates. This configuration allows examination of the case described in Part 2 when some variates are isolated by near orthogonality from dependencies among others.

The additional dependency series are

$$u_i = \text{DUM1} + e_i$$
$$e_i \leftrightarrow f(0,\ 10^{-i}\ s^2_{\text{DUM1}}) \qquad i = 0\ ..\ 4 \tag{3.9a}$$

$$v_j = \text{DUM2} - \text{DUM1} + e_j$$
$$e_j \leftrightarrow f(0,\ 10^{-j}\ s^2_{\text{DUM2-DUM1}}) \qquad j = 0\ ..\ 4 \tag{3.9b}$$

$$w_k = 3\text{*GNP58} + 1.5\text{*LHTUR} + e_k$$
$$e_k \leftrightarrow f(0,\ 10^{-k}\ s^2_{3\text{*GNP}+1.5\text{*LHTUR}}) \qquad k = 0\ ..\ 4 \tag{3.9c}$$

$$x_m = \text{*GAUM} + .7\text{*DUM2} + e_m$$
$$e_m \leftrightarrow f(0,\ 10^{-m}\ s^2_{\text{*GAUM}+.7\text{*DUM2}}) \qquad m = 0\ ..\ 4. \tag{3.9d}$$

These data are used to augment Z to produce seven series of test matrices

$$\begin{array}{ll}
Z1(i) \equiv [Z\ u_i] & Z5(j,k) \equiv [Z\ v_j\ w_k] \\
Z2(j) \equiv [Z\ v_j] & Z6(i,m) \equiv [Z\ u_i\ x_m] \\
Z3(k) \equiv [Z\ w_k] & Z7(i,k,m) \equiv [Z\ u_i\ w_k\ x_m]. \\
Z4(m) \equiv [Z\ x_m] &
\end{array} \tag{3.10}$$

Each of the dependencies (3.9a, b and c) possesses essential scaling problems, with DUM2, 3*GNP58 and *GAVM, respectively, being the dominant terms. The problem is extreme in the relation of DUM2 and DUM1, where DUM1 introduces much less than .1% of the total variation, and difficult in the other cases. The relation defined by (3.9b) is isolated by near orthogonality from the one defined by (3.9c), and these relations occur separately in the Z2 and Z3 test series and together in the Z5 series. Relation (3.9d) bridges the two subseries.

3.3 The Results

Space limitations obviously prevent reporting the full set of experimental results. Fortunately, after reporting Experiment 1 in some detail, it is possible to select samples of output from Experiments 2 and 3 that convey what generalizations are possible.

Experiment 1: The X Matrices

X1: Let us begin with the simplest series of experimental matrices, the X1(i), i = 0, .., 4. Here the data series of column 4, C4, is related to that of column 3, C3, by (3.5a), i.e., C4 = C3 + e_i, i = 0 .. 4; and this is the only contrived dependency among the four columns of X1. We would therefore expect there to be one "high" condition index and a large proportion of var(b_3) and var (b_4) to be associated with it. Table 1.A presents the variance-component proportions and the condition indexes for this series as i goes from 0 to 4.

TABLE 1A*

Variance-Component Proportions
and Condition Indexes

<u>X1 Series</u>

1 constructed near dependency (3.5a)
$C4 = C3 + e_i$

X1(0)

	var(b_1)	var(b_2)	var(b_3)	var(b_4)	Condition index η
σ_1	.005	.012	.002	.003	
σ_2	.044	.799	.004	.032	5
σ_3	.906	.002	.041	.238	8
σ_4	.045	.187	.954	.727	14

X1(1)

	var(b_1)	var(b_2)	var(b_3)	var(b_4)	Condition index η
σ_1	.005	.011	.001	.001	
σ_2	.094	.834	.003	.002	5
σ_3	.899	.117	.048	.035	9
σ_4	.002	.038	.948	.962	27

X1(2)

	var(b_1)	var(b_2)	var(b_3)	var(b_4)	Condition index η
σ_1	.005	.012	.000	.000	
σ_2	.086	.889	.000	.000	5
σ_3	.901	.083	.003	.003	9
σ_4	.007	.016	.997	.997	95

X1(3)

	var(b_1)	var(b_2)	var(b_3)	var(b_4)	Condition index η
α_1	.005	.012	.000	.000	
α_2	.078	.903	.000	.000	5
α_3	.855	.079	.000	.000	9
α_4	.061	.006	.999	.999	461

X1(4)

σ_1	.005	.010	.000	.000	
σ_2	.084	.792	.000	.000	5
σ_3	.906	.070	.000	.000	9
σ_4	.004	.127	1.000	1.000	976

*Columns may not add to unity due to rounding error.

A glance at these results confirms our expectations. In each case there is a highest condition index that accounts for a high proportion of variance in two or more coefficients, and these are var(b_3) and var(b_4). Furthermore, the pattern is observable in the weakest case X1(0), and becomes clearer and clearer as the near dependency becomes tighter: all condition indexes save one remain virtually unchanged while the condition index corresponding to the imposed dependency increases strongly with each jump in i; the variance-component proportions of the two "involved" variates C3 and C4 become larger and larger, eventually becoming unity.

To help interpret the condition indexes in Table 1A we present in Table 1B the simple correlation between C3 and C4 for each of the X1(i) matrices and also regressions of C4 on C1, C2 and C3.

TABLE 1B*

	ρ(C3,C4)	Regression of C4 on C1, C2 and C3			
		C1	C2	C3	R^2
X1(0)	.766	.3905 [1.114]	-.1354 [.91]	.9380 [4.768]	.6229
X1(1)	.931	.1481 [.97]	.0925 [1.38]	.8852 [10.10]	.8765
X1(2)	.995	-.0076 [-.17]	.0142 [.72]	.9982 [38.80]	.9893
X1(3)	.999	.0111 [1.22]	.0015 [.37]	.9901 [188.96]	.9996
X1(4)	1.000	-.0012 [-.28]	.0033 [1.76]	.9976 [400.56]	.9999

*The figures in square brackets are t's. Since interest in these results centers on "significance", it seems proper to publish t's rather than estimated standard deviations.

In addition to observing the general pattern that was expected, the following points are noteworthy.

1. The relation between C3 and C4 of X1(0), having a simple correlation of .766 and a regression R^2 of .6229 (not very high in comparison with simple correlations present in most real-life economic data matrices), shows itself in a condition number of 14 and is sufficiently high to cause large proportions (.95 and .73) of the variance of the affected coefficients, var(b_3) and var(b_4).

2. Also at this lowest level (i=0), the diagnostic test proposed at the end of Part 2 would correctly indicate the existence of but one near dependency and correctly key the variates involved.

3. In light of 2) above, the regressions of Table 1B that were run for comparative purposes are also those that would be suggested by the results for displaying the near dependencies. Table 1B verifies that even in the X1(0) case with an η of 14, the proper relation among the columns of X1(0) is being clearly observed.

4. With each increase in i (corresponding to a ten-fold reduction in the noise in the near dependency), the simple correlations and R^2's increase one order, roughly adding another 9 in the series .9, .99, .999 etc., and the condition index increases in order of magnitude roughly along the progression 10, 30, 100, 300, 1000, a pattern we shall observe in further examples.[9] This relation suggests a means for comparing the order of magnitude of the "tightness" of a near dependency.

5. Also with each increase in i, the proportion of the variance components of the affected coefficients associated with the highest η increases markedly (again roughly adding one more 9 with each step).

6. As noted in Part 2, it is the joint condition of high variance-component proportions for two or more coefficients associated with a high condition index that signals the presence of degrading collinearity. In the case of X1(0), the second highest condition index, 8, is not too different from the highest, but it is a dominant component in only one variance, var (b_1). In this case, then, a condition index of 14 (roughly 10) is "high enough" for collinearity's presence to begin to be observed.

X2: The X2(i) series also possesses only one constructed near dependency, 3.5b, but involving three variates, columns 1, 2 and 4 in the form C4 = .8*C1 + .2*C2 + e_i. We expect, then, high variance-component proportions for these three variates to be associated with a single high condition index. Table 2A presents the Π-matrix of variance-component proportions and the condition indexes for the X2(i) data series, and Table 2B gives the corresponding simple correlations and regressions. In this case the correlations are between C4 in 3.5b and $\widehat{C4}$ = .8*C1 + .2*C2. The regressions are C4 regressed on C1, C2 and C3.

TABLE 2A

Variance-Component Proportions
and Condition Indexes

X2 Series

1 constructed near dependency (3.5b)

$C4 = .8*C1 + .2^{*}C2 + e_i$

X2(0)

	var(b_1)	var(b_2)	var(b_3)	var(b_4)	Condition index η
σ_1	.003	.012	.004	.005	
σ_2	.027	.735	.001	.068	4
σ_3	.009	.228	.636	.526	9
σ_4	.960	.030	.359	.401	11
			X2(1)		
σ_1	.001	.004	.003	.000	
σ_2	.021	.297	.011	.001	5
σ_3	.091	.026	.767	.006	10
σ_4	.887	.673	.219	.993	31
			X2(2)		
σ_1	.000	.001	.004	.000	
σ_2	.001	.039	.012	.000	5
σ_3	.004	.002	.983	.002	9
σ_4	.995	.958	.001	.998	102
			X2(3)		
σ_1	.000	.000	.004	.000	
σ_2	.000	.002	.014	.000	5
σ_3	.000	.000	.976	.000	9
σ_4	1.000	.997	.006	1.000	381

X2(4)

σ_1	.000	.000	.004	.000	
σ_2	.000	.000	.013	.000	5
σ_3	.000	.000	.938	.000	9
σ_4	1.000	1.000	.046	1.000	1003

TABLE 2B

	$\rho(C4, \hat{C4})$	Regression C4 on C1, C2 and C3			
	$\hat{C4}$ = .8C3 + .2C2	C_1	C_2	C_3	C_4
X2(0)	.477	.8268 [3.84]	-.0068 [-.07]	.1089 [.88]	.2864
X2(1)	.934	.6336 [10.14]	.1776 [6.49]	.1032 [2.87]	.8976
X2(2)	.995	.8186 [40.90]	.1879 [21.44]	.0007 [.06]	.9911
X2(3)	.999	.7944 [149.03]	.2023 [86.69]	.0012 [.40]	.9993
X2(4)	1.000	.7990 [393.94]	.1992 [224.32]	.0012 [1.02]	.9999

The following points are noteworthy.

1. Once again the expected results are clearly observed, at least for $i \geq 1$.

2. In the case of X2(0), the constructed dependency is weak, having a simple correlation of less than .5. The resulting condition index, 11, is effectively the same as the second highest condition index, 9, and hence this dependency is no tighter than the general background conditioning of the basic data matrix. We shall see as we proceed that in the case where several condition indexes are effectively the same, the procedure can have trouble distinguishing among them, and the variance-component proportions of the various variates involved can be arbitrarily distributed among the nearly equal condition indexes. In X2(0) the two condition indexes 9 and 11 account for over 90% of the variance in b_1, b_3 and b_4. b_3's presence in this group is

explained by the fact that the simple correlation between C1 and C3 is .58 (greater than the constructed correlation between C4 and $\widehat{C4}$). b_2's absence is explained by the fact that there is a minor scaling problem; C2 accounts for only half the variance of the constructed variate C4 = .8*C1 + .2*C2 + e and, in this case, its influence is being dominated by other correlations.

3. By the time the simple correlation between C4 and $\widehat{C4}$ becomes .934, in the case of X2(1), the above problem completely disappears. The contrived relation involving columns 1, 2 and 4 now dominates the other correlations among the columns of the basic data, and the variance-component proportions of these variates associated with the largest condition index, 31, are all greater than .5.

4. We again observe that, with each increase in i, the condition index corresponding to this ever-tightening relation jumps in the same progression noted before, namely 10, 30, 100, 300, 1000. In this regard it is of interest to observe that the contrived relation among columns 1, 2 and 4, becomes clearly distinguishable from the background in case X2(1) when its condition index becomes one step in this progression above the "noise," i.e., when it becomes 31 vs. the 10 associated with the background dependencies.

5. Once again, the presence of collinearity begins to be observable with condition indexes around 10. In this instance, however, an unintended relation (the .58 correlation between C1 and C3) also shows itself, confounding clear identification of the intended dependency among C4, C1 and C2.

6. In both this and the X1 case, a condition number of 30 signals clear evidence of the presence of the linear dependency and degraded regression estimates.

<u>X3(i,j):</u> This series of matrices combines the two dependencies (3.5a) and (3.5b) just examined into a single 5 column matrix with C4 = C3 + e and C5 = .8*C1 + .2*C2 + u. This, then, offers the first constructed example of simultaneous or coexisting dependencies.[10] We expect that there should be two high condition indexes, one associated with high variance-component proportions in var(b_3) and var(b_4) - due to dependency (3.5a) - and one associated with high variance-component proportions between var(b_1) and var(b_2) and var(b_5) - due to dependency (3.5b).

In an effort to reduce the increasing number of Π-matrices relevant to this case we shall concentrate our reported results in two ways. First, we shall report only representative information from among the 25 cases X3(i,j), i,j = 0 .. 4, and second, where possible, we will report only the rows of the variance-component proportions that correspond to the condition indexes of interest. We note in the previous two series that many of the rows, those corresponding to lower condition indexes, are effectively unchanging as i varies, and convey no useful additional information for the analysis at hand.

Let us begin by holding i constant at 2 and varying j = 0 .. 4; (3.5a) is therefore moderately tight, while (3.5b) varies. Table 3 presents the results for this case.

TABLE 3

Variance-Component Proportions
and Condition Indexes

X3 Series
2 constructed near dependencies (3.5a) and (3.5b)
$C4 = C3 + e_2$ (unchanging)
$C5 = .8*C1 + .2*C2 + u_i$ (i = 0 .. 4)

X3(2,0)

	$var(b_1)$	$var(b_2)$	$var(b_3)$	$var(b_4)$	$var(b_5)$	Condition index, η
σ_1	.002	.007	.000	.000	.003	
σ_2	.025	.734	.000	.000	.064	5
σ_3	.025	.240	.003	.003	.303	8
σ_4	.941	.002	.000	.001	.630	12
σ_5	.007	.016	.996	.996	.001	106

X3(2,1)*

σ_3	.076	.008	.003	.003	.009	9
σ_4	.765	.537	.000	.004	.792	34
σ_5	.147	.192	.997	.992	.199	118

X3(2,2)

σ_3	.004	.001	.003	.003	.002	8
σ_4	.746	.750	.459	.464	.758	127
σ_5	.249	.211	.537	.533	.241	99

X3(2,3)

σ_3	.002	.000	.003	.003	.000	8
σ_4	.999	.997	.173	.182	.999	469
σ_5	.000	.001	.824	.816	.001	83

X3(2,4)

σ_3	.000	.000	.003	.003	.000	8
σ_4	1.000	1.000	.003	.001	1.000	1124
σ_5	.000	.000	.994	.996	.000	106

*The unchanging and inconsequential rows corresponding to σ_1 and σ_2 have not been repeated. See text.

The auxiliary correlation and regression statistics need not, of course, be repeated, for these are the same as the relevant portions of Tables 1B and 2B. In particular, the correlations and regressions for the unchanging X3(2,j) relation for i = 2 between C4 and C3 are those for X1(2) in Table 1B and the regressions for column 5 on the basic columns 1, 2, and 3 for j = 0 ... 4 are those given in Table 2B for X2(j), j = 0 .. 4.

The following points are noteworthy.

1. The unchanging "tight" relation between columns 3 and 4 is observable throughout, having a correlation of .995 and a very large condition index in the neighborhood of 100.

2. The relation with varying intensity among C5, C1, and C2, begins weakly for the X3(2,0) case, and, as before, is somewhat lost in the background. Still, the involvement of var(b_1) and var (b_3) with the condition index 12 is observable even here, although it is being confounded with the other condition index, 8, of roughly equal value. The unchanging tight relation between C4 and C3 with index 106 is unobscured by these other relations.

3. When the relation between columns 1, 2 and 5 becomes somewhat tighter than the background, as in the case of X3(2,1), its effects become separable. This case clearly demonstrates the ability of the procedure to correctly identify two simultaneous dependencies and indicate the variates involved in each: the η of 34 is associated with the high variance-component proportions in var(b_1), var(b_2) and var(b_5) and the η of 118 is associated with those of var(b_3) and var(b_4).

4. When the two contrived dependencies become of roughly equal intensity, as in the case of X3(2,2), both having η's in the neighborhood of 100, the involvement of the variates in the two relations once again becomes confounded. However, it is only the information on the separate involvement of the variates that is lost through this confounding. It is still possible to determine that there are two near dependencies among the columns of X, and it is still possible to determine which variates are involved; in this case all of them, for the two condition indexes together account for well over 90% of the variance in b_1, b_2, b_3, b_4, b_5, indicating the involvement of each. The only information being lost here is which variates enter which dependency.

5. When the relation among columns 1, 2 and 5 again becomes strong relative to the unchanging relation between columns 3 and 4, as in the cases X3(2,3) and X3(2,4), their separate identities reemerge.

6. Once again, the order of magnitude of the relative tightness of a near dependency seems to increase with a progression in the condition index in the scale of 10, 30, 100, 300, 1000. Dependencies of roughly equal magnitude can be confounded; dependencies of differing magnitudes are able to be separately identified.

The preceding analysis examined X3(i,j) by varying the second dependency and holding the first constant at i = 2. Let us reverse this order and examine X3(i,1) for i = 0 .. 4, and j held constant at 1.

As i increases, we would expect there to be two high condition indexes. The one corresponding to the unchanging dependency between columns 1, 2 and 5 will not be too high, since j is held at 1. The relation between columns 3 and 4 will get tighter and more highly defined as i increases from 0 to 4.

Table 4 reports these results. Table 1B and the second row of Table 2B provide the relevant supplementary correlations and regression.

TABLE 4

Variance-Component Proportions
and Condition Indexes

X3 Series

2 constructed near dependencies (3.5a) and (3.5b)
$C4 = C3 + e_i$ (i = 0, .., 4)
$C5 = .8*C1 + .2*C2 + u_1$ (unchanging)

X3(0,1)

	$var(b_1)$	$var(b_2)$	$var(b_3)$	$var(b_4)$	$var(b_5)$	Condition Index, η
σ_1	.001	.002	.001	.002	.000	
σ_2	.008	.274	.003	.032	.000	5
σ_3	.092	.004	.044	.250	.011	9
σ_4	.885	.643	.171	.008	.988	35
σ_5	.015	.076	.781	.708	.001	15

X3(1,1)

	$var(b_1)$	$var(b_2)$	$var(b_3)$	$var(b_4)$	$var(b_5)$	Condition Index, η
σ_4	.821	.656	.226	.077	.900	35
σ_5	.073	.018	.720	.880	.089	30

X3(2,1)

σ_4	.765	.537	.000	.004	.792	34
σ_5	.147	.192	.997	.992	.199	118

X3(3,1)

σ_4	.871	.673	.000	.000	.985	34
σ_5	.028	.010	.999	.999	.005	519

X3(4,1)

σ_4	.842	.549	.000	.000	.901	34
σ_5	.062	.191	1.000	1.000	.089	1148

The following points are noteworthy.

1. Both relations are observable from the outset.

2. A condition number of 15 (greater than 10) corresponds to a dependency that is tight enough to be observed.

3. The confounding of the two dependencies is observable when the condition indexes are close in magnitude, the case of X3(1,1); but it is not as pronounced here as in the previous examples.

4. The rough progression of the condition indexes in the order of 10, 30, 100, 300, 1000 is observed again.

To complete the picture on the behavior of X3(i,j) we report the variance-component proportions tables for selected values of i and j increasing together.

TABLE 5

Variance-Component Proportions
and Condition Indexes

<u>X3 Series</u>

2 constructed near dependencies (3.5a) and (3.5b)
$C4 = C3 + e_i$ (selected values)
$C5 = .8*C1 + .2*C2 + u_i$ (selected values)

X3(0,0)

	$var(b_1)$	$var(b_2)$	$var(b_3)$	$var(b_4)$	$var(b_5)$	Condition Index, η
σ_1	.002	.007	.001	.002	.003	
σ_2	.016	.750	.000	.009	.041	5
σ_3	.028	.053	.044	.206	.323	7
σ_4	.953	.000	.018	.032	.610	12
σ_5	.001	.190	.937	.751	.023	15
			X3(1,0)			
σ_3	.015	.271	.041	.033	.341	8
σ_4	.952	.009	.012	.006	.585	12
σ_5	.006	.037	.946	.961	.004	30
			X3(1,2)			
σ_4	.995	.962	.094	.117	.998	123
σ_5	.005	.002	.860	.848	.001	30
			X3(3,2)			
σ_4	.967	.936	.000	.000	.979	115
σ_5	.028	.023	1.000	1.000	.020	523

X3(3,4)

σ_4	.999	.999	.015	.013	.999	1129
σ_5	.001	.001	.985	.986	.001	517

X3(4,4)

σ_4	.698	.705	.633	.631	.699	1357
σ_5	.302	.294	.369	.369	.301	960

The following points are noteworthy.

1. In the X3(0,0) case, three condition indexes are of close magnitude, 7, 12 and 15 and there is some confounding of variate involvement among all three of them.

2. The relation between C3 and C4 is freed from the background in the next case, X3(1,0), but there is still some confusion between the two similar condition indexes 8 and 12.

3. Quite clearly the two relations become more clearly identified as i and j increase, and are strongly separable so long as the condition indexes remain separated by at least one order of magnitude along the 10, 30, 100, 300, 1000 progression.

4. However, no matter how tight the individual relationship, they can be confused when of equal weight, as is seen in the case of X3(4,4).

Experiment 2: The Y Matrices

Our interest in examining these new experimental data series focuses on several issues. First, does a totally different set of data matrices result in similar generalizations on the behavior of the condition indexes and the variance-component proportions that were beginning to emerge from Experiment 1? Second, do "rate-of-change" data series and "rates" series behave differently from the "levels" and "trends" data of Experiment 1? The data series Y3 are relevant here. Third, do essential scale problems cause troubles? Data sets Y1 and Y2 examine this problem.

Y1 and Y2: The Y1(i) series, we recall, consists of a five column matrix in which C5 = C1 + C2 + e_i as in (3.7a). The variance in C5 introduced by C1 (GNP58) is relatively small, less than 1% of that introduced by C2. Its influence is therefore easily masked. The Y2(i) series is exactly the same except that C1's influence is made smaller yet. Here C5 = .1*C1 + C2 + e_i as is clear from (3.7b). These two series allow us to see how sensitive the diagnostic procedure for multicollinearity is to strong and even severe scaling problems.

For both experimental series, Y1 and Y2, we would expect one high condition index associated with high variance-component proportions in var(b_1), var(b_2) and var(b_5). Tables 6A and 7A present these results for Y1 and Y2 respectively as i = 0 .. 4. Tables 6B and 7B present the corresponding supplementary correlations and regressions.

TABLE 6A

Variance-Component Proportions
and Condition Indexes

Y1 Series

1 constructed near dependency (3.7a)
C5 = C1 + C2 + e_i (i = 0..4)

Y1(0)

	var(b_1)	var(b_2)	var(b_3)	var(b_4)	var(b_5)	Condition Index, η
σ_1	.005	.001	.003	.015	.002	
σ_2	.010	.001	.010	.898	.002	3
σ_3	.782	.036	.001	.010	.081	7
σ_4	.188	.045	.978	.071	.096	10
σ_5	.016	.916	.008	.007	.819	16

Y1(1)

σ_3	.607	.019	.031	.008	.003	8
σ_4	.024	.030	.941	.090	.012	10
σ_5	.360	.950	.014	.036	.985	40

Y1(2)

σ_3	.394	.001	.003	.010	.001	7
σ_4	.065	.001	.976	.068	.001	10
σ_5	.534	.998	.008	.020	.999	156

Y1(3)

σ_3	.099	.000	.003	.010	.000	7
σ_4	.016	.000	.959	.073	.000	10
σ_5	.883	.997	.024	.003	1.000	397

Y1(4)

σ_4	.001	.000	.965	.072	.000	10
σ_5	.993	1.000	.018	.000	1.000	1659

TABLE 6B

	$\rho(C5,\widehat{C5})$	Regression of C5 on C1, C2, C3, C4.				
	$\widehat{C5} = C1 + C2$	C1	C2	C3	C4	R^2
Y1(0)	.776	-.0264 [-.01]	.9262 [5.33]	97.8101 [1.03]	28.3210 [.34]	.6186
Y1(1)	.939	2.5629 [3.78]	.8238 [13.20]	35.7769 [1.05]	35.1060 [1.17]	.9116
Y1(2)	.997	.9505 [5.24]	1.0009 [59.80]	-2.7505 [-.30]	-4.9108 [-.61]	.9940
Y1(3)	.999	.9534 [13.36]	.9989 [151.80]	2.8747 [.80]	.9679 [.31]	.9991
Y1(4)	1.000	1.0178 [59.57]	.9975 [633.30]	.5632 [.65]	.0135 [.0178]	.9999

TABLE 7A

Variance-Component Proportions
and Condition Indexes

Y2 Series

1 constructed near dependency (3.7b)
$C5 = .1*C1 + C2 + e_i \ (i = 0 .. 4)$

Y2(0)

	var(b_1)	var(b_2)	var(b_3)	var(b_4)	var(b_5)	Condition Index
σ_1	.005	.003	.003	.014	.003	
σ_2	.009	.002	.009	.886	.005	3
σ_3	.804	.023	.007	.001	.205	7
σ_4	.157	.366	.201	.000	.781	10
σ_5	.025	.606	.781	.098	.006	11

Y2(1)

	var(b_1)	var(b_2)	var(b_3)	var(b_4)	var(b_5)	Condition Index
σ_3	.736	.009	.000	.022	.012	7
σ_4	.248	.005	.887	.072	.012	10
σ_5	.000	.985	.099	.016	.975	42

Y2(2)

	var(b_1)	var(b_2)	var(b_3)	var(b_4)	var(b_5)	Condition Index
σ_3	.747	.001	.000	.013	.000	7
σ_4	.190	.001	.880	.068	.001	10
σ_5	.049	.999	.108	.001	.998	153

Y2(3)

	var(b_1)	var(b_2)	var(b_3)	var(b_4)	var(b_5)	Condition Index
σ_3	.648	.000	.000	.012	.000	7
σ_4	.157	.000	.867	.069	.000	10
σ_5	.183	1.000	.120	.008	1.000	475

Y2(4)

σ_3	.512	.000	.000	.011	.000	7
σ_4	.121	.000	.984	.066	.000	10
σ_5	.357	1.000	.002	.063	1.000	1166

TABLE 7b

	$\rho(C5,\widehat{C5})$	Regression of C5 on C1, C2, C3 and C4.				
	$\widehat{C5} = .1*C1 + C2$	C1	C2	C3	C4	R^2
Y2(0)	.395	.3789 [.16]	.4585 [2.10]	260.8090 [2.18]	47.9654 [.46]	.1433
Y2(1)	.962	.0862 [.12]	1.0613 [16.64]	-35.0793 [-1.00]	27.0102 [.88]	.9294
Y2(2)	.997	.2077 [1.15]	1.0179 [60.94]	-13.8609 [-1.52]	-.4803 [-.06]	.9943
Y2(3)	.999	.1331 [2.287]	1.0078 [187.8]	-5.0668 [-1.73]	-1.0565 [-.41]	.9994
Y2(4)	1.000	.0845 [3.58]	1.0009 [460.06]	-.2182 [-.18]	1.3050 [1.25]	.9999

The following points are noteworthy.

1. The results for both data series are in basic accord with expectations.

2. However, the essential scale differences do cause problems in identifying the variates involved in generating the dependency. In the Y1 series, the involvement of the dominated column 1 is not observed at all in the weakest case Y1(0), having a condition index of 10 (and a correlation of .78). C1's involvement begins to be observed by Y1(1), but does not show itself completely until Y1(2) and Y1(3). By contrast, the involvement of the dominant column C2, along with the generated column 5 is observed from the outset. The same pattern occurs within the supplementary regressions. C1's regression parameter is insignificant in case Y1(0), becomes significant in Y1(1) and takes the proper order of magnitude (unity) in Y1(2) and Y1(3).

3. Aggravating the scale problem in the Y2 series (C1 now accounts for less than 1/100 of 1% of the variance in C5) has the expected effect. Now column 1's involvement is becoming apparent only by the tightest case Y2 (4) with a condition index of 1166.

4. The several other general patterns noted in Experiment 1 seem still to hold: a dependency's effects are beginning to be observed with condition indexes around 10; decreasing the variance of the generated dependency by successive factors of 10 causes the condition index roughly to progress as 10, 30, 100, 300, 1000, i.e., log η progresses in steps of 1/2.

Y3: The Y3(i) series consists of a 5 column data matrix in which the fifth column is in a simple relation 3.7c with the fourth column, $C5 = C4 + e_j$. C4 in this case is inventory investments, a rate-of-change variate. The results of this series are given in Tables 8A and 8B.

TABLE 8A

Variance-Component Proportions
and Condition Indexes

Y3 Series

1 constructed near dependency (3.7c)
$C5 = C4 + e_i \ (i = 0 \, .. \, 4)$

Y3(0)

	$\text{var}(b_1)$	$\text{var}(b_2)$	$\text{var}(b_3)$	$\text{var}(b_4)$	$\text{var}(b_5)$	Condition Index, η
σ_1	.005	.003	.003	.012	.012	
σ_2	.018	.007	.015	.094	.222	3
σ_3	.953	.180	.112	.003	.001	8
σ_4	.025	.810	.870	.013	.034	11
σ_5	.004	.001	.001	.877	.731	5
			Y3(1)			
σ_3	.706	.133	.118	.027	.019	8
σ_4	.000	.855	.693	.004	.022	11
σ_5	.270	.001	.171	.948	.939	15

Y3(2)

σ_3	.953	.171	.102	.000	.000	8
σ_4	.017	.757	.751	.002	.000	11
σ_5	.005	.063	.132	.996	.997	42

Y3(3)

σ_3	.956	.180	.115	.000	.000	8
σ_4	.018	.789	.859	.000	.000	11
σ_5	.001	.020	.008	.999	.999	147

Y3(4)

σ_3	.890	.179	.107	.000	.000	8
σ_4	.016	.790	.798	.000	.000	11
σ_5	.071	.020	.078	1.000	1.000	416

TABLE 8B

	ρ(C5,C4)	Regression of C5 on C1, C2, C3, C4				
		C1	C2	C3	C4	R^2
Y3(0)	.643	.0015 [.27]	.0004 [.81]	-.2014 [-.74]	.9180 [3.82]	.4231
Y3(1)	.950	-.0027 [1.93]	.0001 [.62]	.0845 [1.20]	.9491 [15.38]	.9188
Y3(2)	.995	.0002 [.36]	.0001 [1.23]	-.0450 [1.80]	1.0048 [45.67]	.9902
Y3(3)	.999	.0000 [.17]	-.0000 [-.65]	.0029 [.40]	1.0004 [161.11]	.9992
Y3(4)	1.000	.0000 [-1.3]	-.0000 [-.67]	.0035 [1.4]	1.0013 [455.66]	.9999

Points of Interest

1. An interesting phenomenon emerges in case Y3(0) that is in need of explanation and provides us the first opportunity to apply our diagnostic tools to a near dependency that arises naturally in the data and has not been artifically generated. First we note that the generated dependency between C5 and C4 is indeed observed - associated with the weak condition index 5. In addition, however, we note that over 80% of var(b_2) and var(b_3) is associated with the larger condition index 11, indicating at least their involvement in a low-level, unintended "background" dependency. The simple correlation between C2 and C3 is very low, .09,[11] so we must look further than a simple dependency between C2 and C3. Further examination of Y3(0) in Table 8A shows that there is really not one, but two unintended near dependencies of roughly equal intensity in the basic data matrix associated with the effectively equal condition indexes 11 and 8. Furthermore, these two background dependencies together account for over 95% of var(b_1), var(b_2), and var(b_3), and the three roughly equal condition indexes 5, 8 and 11 account for virtually all of each of the five variances. Applying what we have learned from Experiments 1 and 2, we conclude that there are three weak dependencies of roughly equal intensity whose individual effects cannot be separated, a problem we have seen arise when there are several condition indexes of the same order of magnitude. Since we know C5 and C4 are related, we would expect to find two additional near dependencies among the four columns C1, C2, C3, and C4.[12] Indeed, regressing C1 and C3 separately on C2 and C4 gives[13]

$$C1 = \underset{[5.86]}{.0540C_2} + \underset{[1.96]}{16.67C_4} \qquad R^2 = .8335$$

$$C3 = \underset{[8.43]}{.0015C_2} + \underset{[2.27]}{.0373C_4} \qquad R^2 = .9045 \quad .$$

These background dependencies are, of course, also present in the Y1 and Y2 series (the first four columns being the same in all Y series), but their effects there are overshadowed by the presence of the relatively stronger contrived dependency involving C1, C2 and C5. The experience we have attained from these experiments in the use of these diagnostic techniques, however, has clearly led us very much in the right direction.

2. The previously described phenomenon serves to emphasize the point that when there are two or more condition indexes of equal or close magnitude, care must be taken in applying the diagnostic test. In such cases the variance-component proportions can be arbitrarily distributed across the roughly equal condition indexes so as to obscure the involvement of a given variate in any of the competing (nearly equal) near dependencies. In Y3(0), for example, the fact that over 80% of var(b_2) and var(b_3) is associated with the single condition index of 11 need not imply only a simple relation between C2 and C3. Other variates (here C1 and C4), associated with competing condition indexes (8 and 5), can be involved as well. Furthermore, when there are competing condition indexes, the fact that a single condition index (like 8 in Y3(0)) is associated with only one high variance-component proportion (95% of var(b_1)), need not imply, as it otherwise could,[14] that the corresponding variate (C1) is free from involvement in any near dependency. Its interrelation with the variates involved in competing dependencies must be investigated.

In sum, when there are competing dependencies (condition indexes of similar value), they must be treated together in the application of the diagnostic test. That is, the variance-component proportions for each coefficient should be aggregated across the competing condition indexes, and high variance-component aggregate proportions for two or more variances associated with the set of competing high indexes is to be interpreted as evidence of degrading collinearity. The exact involvement of specific variates in specific dependencies cannot be learned in this case, but it is still possible to learn a) which variates are degraded (those with high aggregate component proportions) and b) the number of near dependencies present (the number of competing indexes).

3. Another, quite different, form of confounded involvement is also exemplified by the foregoing: the dominant dependency. C4 is apparently involved simultaneously in several near dependencies, weakly, and with scaling problems, in the dependencies associated with η's of 8 and 11, and without scaling problems in the contrived dependency between C4 and C5. In all cases, but particularly as it becomes tighter, this latter dependency dominates the determination of var(b_4), thereby obscuring C4's weak involvement in the other dependencies. Dominant dependencies (higher condition indexes), then, can mask the simultaneous involvement of a single variate in weaker dependencies. Thus, one cannot rule out the possibility that a variate whose variance is being greatly determined by a dependency with a high condition index is not also involved in dependencies with lower condition indexes, unless, of course, that variate is buffered from the other dependencies through near orthogonality.

4. From within the intricacies of the foregoing points, however, one must not loose sight of the fact that the test for potentially damaging collinearity requires the joint condition of 1) two or more variances with high component proportions associated with 2) a single high condition index[15]; condition 1) by itself is not enough. It is true in the Y3(0) case, for example, that the three condition indexes 5, 8, and 11 account for most of the variance of all five estimates, but by very rough standards, these condition indexes are not high, and the data matrix Y3(0), quite likely, could be suitable for many econometric applications. Let's examine this point further. In our prior examples we noted that contrived dependencies began to be observed when their "tightness" resulted in condition indexes of around 10. We were also able to calculate the correlations that correspond to these relations, so we can associate the magnitudes of condition

indexes with this more well known measure of tightness. A glance through Tables 1-8 shows that condition indexes of 10-11 result from underlying dependencies whose correlations are in the range of .4 to .6, relatively loose relations by much econometric experience. It is not until condition indexes climb to a level of 15-30 that the underlying relations have correlations of .9, a level that much experience suggests is high.[16] Further insight is afforded by examining the actual variances whose component proportions are given in Table 8A; these are presented in Table 9.

TABLE 9*

	var(b_1)	var(b_2)	var(b_3)	var(b_4)	var(b_5)
Y3(0)	9.87	16.94	16.32	3.71	3.04
Y3(1)	11.43	16.74	16.94	25.58	26.38
Y3(2)	9.90	17.56	18.18	207.76	204.17
Y3(3)	9.86	16.78	16.06	2559.53	2547.85
Y3(4)	10.59	16.79	17.28	20389.78	20457.85

*Figures reported here are diagonal elements of $(X'X)^{-1}$ - the ϕ_k's of (2.9) - and do not include the constant factor of s^2, the estimated regression variance. s^2 can, of course, only be calculated once a specific y has been regressed on X.

In the case of Y3(0), all variance magnitudes are relatively small, certainly in comparison to the size attained by var(b_4) and var(b_5) in the cases of Y3(2) and Y3(3) when the contrived dependency between them becomes tighter. In short, high variance-component proportions surely need not imply large component values. This merely restates the notion of Part 2 that degraded estimates (capable of being improved if calculated from better conditioned data), which apparently can result from even low-level dependencies, need not be harmful; harmfullness depends, in addition, upon the specific regression model employing the given data matrix, the estimated s^2 and statistical use to which the results are to be put.

5. The contrived dependency between C4 and C5, rates of change variates, seems to behave somewhat differently from previous experience, based on "levels" data; namely its condition index is lower for comparable tightness in the underlying relation. Perusal of Tables 1-7 indicates, that, quite roughly, the condition index jumps one step along the 10, 30, 100, 300, 1000 progression each time another "9" digit is added to the correlation of the underlying dependency. That is, an η of 10 has a corresponding correlation of about .5; an η of 30 with correlation .9; $\eta \simeq 100$, correlation = .99, $\eta \simeq 300$, correlation .999. Table 8 indicates the rate of change data to be one step lower with $\eta \leq 10$, correlation .6; $\eta \geq 10$, correlation .9; $\eta \simeq 30$, correlation .99, etc. It may be, therefore, that there is no simple pairing of the level of the strength of a relationship as measured by a condition index with that of the same relation as measured by a correlation. There does, however, seem to be stability in the relative magnitudes of these two measures along the progressions noted above.

6. In all of the foregoing, one should not lose sight of the fact that, basically, the diagnostic procedure works in accord with expectations. The contrived relation between C4 and C5 is observed from the outset and takes on unmistakable form once it is removed from the background, in case Y3(1) or Y3(2).

Y4: In the Y4(i,j) series the two dependencies of Y1 and Y3 occur simultaneously. Here $C5 = C1 + C2 + e_i$ according to (3.7a) and $C6 = C4 + u_i$ as in (3.7c). What is new to be learned from this experimental series can be seen from a very few selected variance-component proportion matrices. These are reported in Table 10.

TABLE 10

Variance-Component Proportions
and Condition Indexes

<u>Y4 Series</u>

2 constructed near dependencies (3.7a) and (3.7c)
$C5 = C1 + C2 + e_i$ (selected values)
$C6 = C4 + u_j$ (selected values).

Y4(0,0)

	var(b_1)	var(b_2)	var(b_3)	var(b_4)	var(b_5)	var(b_6)	Condition Index, η
σ_1	.004	.001	.002	.008	.001	.008	
σ_2	.010	.002	.009	.105	.002	.223	3
σ_3	.788	.036	.001	.001	.079	.004	7
σ_4	.181	.051	.981	.004	.100	.044	11
σ_5	.017	.910	.006	.008	.817	.001	17
σ_6	.001	.000	.000	.875	.001	.720	5
				Y4(0,2)			
σ_1	.003	.001	.002	.000	.001	.000	
σ_2	.011	.002	.008	.002	.003	.002	3
σ_3	.778	.036	.001	.000	.077	.000	8
σ_4	.186	.044	.822	.001	.092	.000	11
σ_5	.015	.916	.009	.000	.772	.000	17
σ_6	.006	.000	.159	.996	.055	.997	47
				Y4(1,1)			
σ_3	.477	.015	.045	.024	.002	.014	8
σ_4	.001	.035	.777	.006	.013	.025	11
σ_5	.346	.948	.010	.000	.984	.005	43
σ_6	.167	.006	.158	.948	.000	.936	17

Y4(1,2)

σ_3	.587	.018	.027	.000	.003	.000	8
σ_4	.023	.028	.847	.001	.011	.000	11
σ_5	.173	.453	.068	.234	.543	.250	39
σ_6	.209	.501	.048	.762	.442	.747	51

Y4(3,3)

σ_4	.016	.000	.954	.000	.000	.000	11
σ_5	.884	1.000	.025	.003	1.000	.003	428
σ_6	.000	.000	.007	.997	.000	.997	162

Points of Interest

1. Both relations are observable even in the weakest instance of Y4(0,0), one with condition index 17, the other with condition index 5. The presence of the contrived relation between C1, C2 and C5 has somewhat masked the background relation among C1, C2 and C3 that was observed in the Y3 series (although var(b_1) is still being distributed among these relations).

2. Dependencies with differing condition indexes tend to be separately identified, as in the cases of Y4(0,2), Y4(1,1) and Y4(3,3). When the condition indexes are nearly equal, however, as in the case Y4(1,2), the involvement of separate variates is confounded between the two. This fact, observed frequently before, is particularly important in this instance. In earlier experiments, roughly equal condition indexes corresponded to roughly equal underlying correlations. In this case, however, the relation between the "rate of change" variates C4 and C6 is .9 while that underlying the relation among C1, C2 and C5 is .99, one 9 stronger. Thus the problem of confounding of relations results from relations of nearly equal tightness as judged by condition indexes, not as judged by correlations.

3. In general, however, the two constructed relations behave together quite independently, and much as they did separately. This was true in Experiment 1 as well; the individual behavior of the dependencies in the X1 and X2 series was carried over to their simultaneous behavior in the X3 series. Thus, with the exception of the minor problem of confounded proportions that results from the presence of near dependencies with competing or dominating condition indexes, it seems fair to conclude that the simultaneous presence of several near dependencies causes the analysis no critical problems.

Experiment 3: The Z matrices

The purposes of Experiment 3 are 1) to analyze slightly larger data matrices (up to eight columns) to see if size has any noteable effect on the procedure; 2) to allow up to three coexisting near dependencies, again to see if new complications arise; 3) to recast some previous experimental series in a slightly different setting to see if their behavior remains stable; 4) to create cases where near orthogonality among data series exists in order to observe its buffering effect against dependencies within nearly orthogonal subgroups. Toward this last objective, columns 4 and 5 of the basic data matrix were generated having maximal correlations with columns 1-3 of roughly .18. Columns 1-3 here are the same as columns 1-3 in the previous Experiment 2. Four dependency relations are contrived according to (3.9a, b, c, and d). (3.9a, and b) generate dependencies among the two columns 4 and 5 which were constructed to have low intercorrelations with columns 1-3. (3.9c) generates a dependency using only C1, C2 and C3 and, hence, presumably buffered from C4 and C5. (3.9d) bridges these two data groups. There are scaling problems built into the generated dependencies. DUM1 is dominated (<.1% of the variance in (3.9b)). *LHTUR is dominated (<.1% of the variance in (3.9c)), and DUM2 is dominated (<.1% of the variance in (3.9d)).

Many of the Z-series experiments were designed to duplicate previous experiments with different data in order to observe whether the process exhibits some degree of stability. In those cases where such stability exists, such as Z1 below, and the experiment merely becomes repetitive it will be reported as such without additional and unnecessary tabulations.

Z1: In this series C6 = C4 + e_i. C4 (DUM1) in this basic data matrix Z is generated to have the same mean and variance as column 4, the rate-of-change variate (GV58), of the basic data matrix Y of Experiment 2. Hence the Z1 series is quite similar to the Y3 series of Experiment 2, and it is to be hoped that this experimental series would exhibit similar properties. This expectation was met in full.

Z2: In this series the dependency is generated by the two "isolated" columns, 4 and 5, by C6 = C5 - C4. It also mixes a rate-of-change variate, C4 and a levels variate, C5, and, as noted has a scaling problem. Table 11 presents two Π-matrices for the case Z2(2) and Z2(4).

TABLE 11

Variance-Component Proportions
and Condition Indexes

Z2 Series

1 constructed near dependency (3.9b)
C6 = C5 - C4 + e_i (selected values)

Z2(2)

	var(b_1)	var(b_2)	var(b_3)	var(b_4)	var(b_5)	var(b_6)	Condition index, η
σ_1	.004	.003	.002	.000	.000	.000	
σ_2	.000	.000	.000	.825	.000	.000	2
σ_3	.047	.114	.043	.019	.002	.002	6
σ_4	.944	.162	.085	.016	.000	.000	8
σ_5	.001	.721	.836	.000	.000	.000	11
σ_6	.004	.000	.034	.140	.998	.998	104

Z2(4)

σ_5	.001	.661	.839	.000	.000	.000	11
σ_6	.003	.088	.029	.817	1.000	1.000	1039

Points of Interest

1. The "background" relations with condition indexes 8 and 11 are still present (the first three columns here are the same as in Experiment 2).

2. The generated dependency is quite observable, but the scaling problem is evident. Even in Z3(2) with a condition index of 104, C4's involvement is not clearly observed, and does not become strongly evident until the condition index increases to the very high value of 1000.

3. The isolation of C1-C3 from C4-C6 is very evident. Even in the case of Z3(4), the high condition index of 1000 does not add any significant degradation to var(b_1), var(b_2) or var(b_3).

Z3: This series, in which C6 = 3*C1 + 1.5*C3 + e_i, is very similar to the Y1 series and shows effectively identical behaviour. The scaling problem here is severe and the involvement of the dominated variate C3 is not strong even in the Z3(4) case, as is seen by the one relevant row of the Π matrix.

Z3(4)

	var(b_1)	var(b_2)	var(b_3)	var(b_4)	var(b_5)	var(b_6)	Condition Index, η
σ_6	1.000	.027	.451	.072	.000	1.000	980

Z4: In this series there is the single contrived relation C6 = C2 + .7*C5 + e_i. The behavior is as according to expectation, paralleling that of the qualitatively similar Y1 and X2 series.

Z5: This series posseses two simultaneous dependencies each isolated from one another by low intercorrelations among the C1-C3 and C4-C5 columns of the basic data matrix Z. Here C6 = C5 - C4 + e_i and C7 = 3*C1 + 1.5*C3 + u_j. A typical Π-matrix for this series is given by

TABLE 12

Variance-Component Proportions
and Condition Indexes

Z5 Series

2 constructed near dependencies (3.9b) and (3.9c)
C6 = C5 - C4 + e_i (i = 2)
C7 = 3*C1 + 1.5*C3 + u_j (j = 3)

Z5(2,3)

	var(b_1)	var(b_2)	var(b_3)	var(b_4)	var(b_5)	var(b_6)	var(b_7)	Condition Index, η
σ_1	.000	.002	.001	.000	.000	.000	.000	
σ_2	.000	.000	.000	.742	.000	.000	.000	2
σ_3	.000	.029	.005	.029	.002	.002	.000	6
σ_4	.000	.243	.090	.005	.000	.000	.000	8
σ_5	.000	.711	.602	.000	.000	.000	.000	12
σ_6	.000	.000	.021	.121	.950	.949	.000	114
σ_7	.999	.015	.280	.102	.048	.049	.999	368

Points of Interest

1. The presence of the two relations is clear, and the scaling problems that beset the two relations are observed.

2. Of principal interest is the verification of the expected simultaneous isolation of the relation among C7, C1 and C3 from that among C4, C5 and C6. The low intercorrelations of these two sets of columns allows the variances within each group to be unaffected by the relation among the other group.

3. Although not shown, it should be noted that the usual confounding of relations occurs in this series, when the condition numbers are of equal magnitude.

Z6: This case has two contrived near dependencies C6 = C4 + e_i and C7 = C2 + .7*C5 + u_i. The results are fully in accord with expectations.

Z7: This case presents the first occurrence of three simultaneous relations C6 = C4 + e_i, C7 = 3*C1 + 1.5*C3 + u_i, and C8 = C2 + .7*C5 + v_k. Table 13 displays three selected cases.

TABLE 13

Z7 Series

3 near dependencies (3.9a), (3.9c) and (3.9d).
C6 = C4 + e_i (selected values)
C7 = 3*C1 + 1.5* + u_j (selected values)
C8 = C2 + .7*C5 + v_k (selected values)

Z7(2,2,3)

	var(b_1)	var(b_2)	var(b_3)	var(b_4)	var(b_5)	var(b_6)	var(b_7)	var(b_8)	Condition Index, η
σ_5	.000	.000	.657	.001	.020	.001	.003	.000	11
σ_6	.000	.000	.076	.835	.007	.852	.000	.000	35
σ_7	.970	.001	.071	.083	.000	.073	.973	.000	153
σ_8	.028	.999	.177	.078	.829	.072	.025	1.000	455
				Z7(2,3,3)					
σ_6	.000	.000	.071	.886	.007	.900	.000	.000	35
σ_7	.714	.105	.332	.104	.098	.089	.718	.108	345
σ_8	.286	.895	.003	.006	.734	.007	.282	.892	482
				Z7(3,2,2)					
σ_6	.401	.370	.017	.818	.065	.815	.380	.388	221
σ_7	.161	.113	.020	.179	.048	.182	.169	.094	116
σ_8	.436	.506	.068	.003	.124	.003	.451	.516	164

Points of Interest

1. The presence of three simultaneous near dependencies causes no special problems, each behaving essentially as it did separately.

2. The Z7(2,2,3) case illustrates the separate identifications of all three relationships, although the severe scaling problem of C3 is masking its influence in the relation associated with σ_7.

3. The other two cases exemplify the problem of separating the individual relationships when the condition indexes are of the same order of magnitude. In Z7(2,3,3) the two relations with similar condition indexes 345 and 482 are confounded; while in the Z7(3,2,2) case, the involved variates have the variance of their estimated regression parameter distributed over the three dependencies with roughly equal condition indexes 221, 116, 164.

One final conclusion may be drawn rather generally from this Experiment 3; namely, that those Z series that were qualitatively similar to previous X and Y series, resulted in quantitatively similar Π-matrices and condition indexes, attesting to a degree of stability in the diagnostic procedure.

Part 4: SUMMARY AND INTERPRETATION; AND EXAMPLES OF DIAGNOSING ACTUAL DATA FOR COLLINEARITY

4.0 Introduction

In Part 2 a test was suggested for diagnosing the presence of multicollinearity in econometric data matrices and for assessing the degree to which such near dependencies degrade ordinary least squares regression estimates. Part 3, recognizing the empirical element to this diagnostic procedure, reported a set of experiments designed to provide experience in its use and interpretation. This part summarizes and exemplifies the foregoing. In Section 1 the experimental evidence of Part 3 is distilled and summarized. Section 2 summarizes the steps to be followed in employing the diagnostic procedure on actual economic data sets, and Section 3 provides two examples of its use on actual data - analyzing naturally arising, uncontrived dependencies.

4.1 Interpreting the Diagnostic Results: A Summary of the Experimental Evidence

Before proceeding with a summary of the evidence, it is worth emphasizing its preliminary nature. The experiments of Part 3 are necessarily limited in scope and cannot hope to illuminate all that is to be known of the behavior of the proposed diagnostic procedure in econometric applications. Indeed, it is to be expected, as experience is gained from future application of these techniques to actual data, that the conclusions presented here will be refined and expanded. For the moment, however, the experimental evidence is gratifyingly stable and provides an excellent point of departure.

This summary begins with a presentation of the experience gained from experiments having a single contrived near dependencey. We then summarize the modifications and extensions that arise when analyzing data matrices in which two or more near dependencies coexist.

Experience With a Single Near Dependency

1. The Diagnostic Procedure Works. The diagnostic test suggested in Part 2 works well and in accord with expectations for a variety of data matrices with contrived dependencies. It is possible not only to determine the presence of the near dependency, but also, subject to the qualifications given below, to determine the variates involved in it.

2. The Progression of Tightness. The tighter the underlying dependency (as measured either by its correlation or relevant multiple correlation), the higher the condition index. Indeed, as the underlying correlations or R^2's increase along the progression <.9, .9, .99, .999, .9999 etc., the condition indexes increase roughly along the progression 3, 10, 30, 100, 300, 1000, 3000, etc. The correspondence between these two progressions, however, is not constant and depends upon the type of data. A given correlation, for example, among rates-of-change data appears to be translated into a lower condition index than for levels data. Some rough generalizations do, however, seem warranted, and these are given next.

3. Interpreting the Magnitude of the Condition Index. Most of the experimental evidence shows that weak dependencies (correlations of less than .9) begin to exhibit themselves with condition indexes around 10, and in some cases as low as 5. An index in the neighborhood of 15-30 tends to result from an underlying near dependency with an associated correlation of .9, usually considered to be the borderline of "tightness" in informal econometric practice. Condition indexes of 100 or more appear to be large indeed, causing substantial variance inflation and great potential harm to regression estimates.

4. Variance-Component Proportions. The rule of thumb proposed at the end of Part 2, that estimates shall be deemed degraded when more than 50% of the variance of two or more coefficients is associated with a single high condition index, still seems good. Future experience may suggest a more appropriate or a more sophisticated rule of thumb, but the 50% rule allows the involved variates to be identified in most instances even when the underlying dependency is reasonably weak (associated correlations of .4-.7). Indeed most evidence indicated proportions of over 80% were attained quite early.

5. Scaling Problems. Essential scaling imbalance causes the involvement of the dominated variates to be masked and more difficult to detect. Essential scaling imbalance occurs when several variates are interrelated so that the variance introduced by some is very much smaller than that introduced by others. Variates introducing less than 1% of the total variation are dominated and their involvement can be completely overlooked by this procedure until the condition index rises to 30 or more. Very strongly dominated variates (<.01%) can be masked even with condition indexes in excess of 300.

6. Data Type Matters. As already noted in 2 above, near dependencies among "rates-of-change" data seem to behave differently from those involving "levels" type data.

Experience With Coexisting Near Dependencies

7. Retention of Individuality. While some new problems of diagnosis and interpretation are introduced, in general it can be concluded that coexisting near dependencies cause the diagnostic procedure no critical problems. Subject to the modifications given in 11-13 below, the several underlying near dependencies behave together much as they did separately. In particular they remain countable (8 below) and to a great degree separable (9 below).

8. Countability. The number of coexisting near dependencies is correctly assessed in all cases by the number of high condition indexes. The presence of a very strong ($\eta > 300$) near dependency, for example, does not obliterate the presence of a much weaker near dependency.

9. Separability. The near dependencies remain greatly separable in the following two senses. First, near dependencies which, when existing alone, have a given condition index, retain roughly the same condition index when made to coexist with other near dependencies, regardless of their relative condition indexes, and second, subject to the qualifications given below, the individual involvement of specific variates in specific near dependencies remains observable.

10. Isolation Through Near Orthogonality. As the theory of Part 2 would have it, near orthogonality does indeed buffer the regression estimates of one set of variates from the deleterious effects of near dependencies among the nearly orthogonal variates.

11. Confounding of Effects with Competing Dependencies. When two or more near dependencies are competing, i.e., have condition indexes of the same order of magnitude, the high variance-component proportions of the variates involved in the separate competing dependencies can be arbitrarily distributed among them, thus confounding their true involvement. The number of coexisting dependencies is, however, not obscured by this situation, nor is the identification of the variates that are involved in at least one of the competing dependencies. It remains possible, therefore, to diagnose how many dependencies are present and which variates are being degraded by the joint presence of those dependencies. Only information on the separate involvement of specific variates in specific competing dependencies is lost. In this case the test procedure is trivially modified to examine those variates which have high variance-component proportions aggregated over the competing high condition indexes.

12. Dominating Dependencies. A dominating dependency, one with a condition index of higher order of magnitude, can become the prime determinant of the variance of a given coefficient and thus obscure information about its simultaneous involvement in a weaker dependency. Consider the example

TABLE 14

Variance-Component Proportions and Condition Indexes

	var(b_1)	var(b_2)	var(b_3)	var(b_4)	Condition Index, η
σ_1	.00	.01	.00	.00	
σ_2	.01	.99	.00	.00	3
σ_3	.99	.00	.01	.01	30
σ_4	.00	.00	.99	.99	300

Here there are two dependencies with high condition indexes, 30 and 300; and 300 dominates. The involvement of C3 and C4 in this dominant dependency is clear; however, equally clearly, we cannot rule out the potential involvement of C3 and C4 along with C1 in the dependency associated with η = 30. Thus, when there are dominated dependencies, such as η = 30 above, it is quite possible for only one high variance-component proportion to be associated with it and still give indication of degradation - the involvement of the other variate(s) being obscured by the dominant relation. In this case our test procedure must once again be qualified: "two or more high variance-component proportions associated with a single high condition number - unless that high condition index is dominated by an even larger one, in which case further investigations may be required." One reasonable procedure to adopt in such cases would be to run an auxiliary regression among the potentially involved variates (C1 on C3 and C4 in the above example) to verify their roles, if such information were required. In this example such additional information would be needed to demonstrate the degradation of var(b_1). There is no question but that var(b_3) and var(b_4) are degraded - not just by their presence in one, but possibly two, dependencies. var(b_1), however, cannot be said to be degraded unless C1 can be shown to be involved in a linear dependency with C3 and/or C4.

By way of contrast, had the last two rows of the above example read as in Table 15, the degradation of all variances

TABLE 15

Variance-Component Proportions
and Condition Indexes

	var(b_1)	var(b_2)	var(b_3)	var(b_4)	Condition Index, η
σ_3	.99	.99	.01	.01	30
σ_4	.00	.00	.99	.99	300

would be apparent without further analysis, and auxiliary regressions would not be required unless it was explicitly desired to know whether C3 and C4 entered along with C1 in the dependency with η = 30.

13. Non-degraded Estimates. On occasion it is also possible to identify those coefficients whose estimates show no evidence of being degraded by the presence of near dependencies. In the example given by Table 15, all four variances show degradation due to the two near dependencies with η's of 30 and 300. In the example given by Table 14, however, var(b_2) has virtually all of its variance determined in association with the relatively small condition index 3, and is not adversely affected by the two tighter dependencies with η's of 30 and 300. The same situation occurs closely in Part 3 , for example, in var(b_2) of the X1 or X2 series, Tables 1A and 2A.

Just where the dividing line between small and large is to be set is a matter that can be answered only with greater practical experience in the use of these techniques. The evidence of the experiments suggests $\eta = 10$, or a range of 7-11, to be a reasonable starting point.[1]

4.2 Employing the Diagnostic Procedure

Diagnosing any given data matrix for the presence of near dependencies and assessing the potential harm their presence may cause regression estimates is effected by a rather straightforward series of steps, the only problems of interpretation arising when there are competing or dominating near dependencies. Two thresholds must be determined at the outset, a condition-index cutoff, η^*, and component-proportion cutoff, π^*, as will be seen in Steps 3 and 5.

The Steps

Step 1. Scale the data matrix X to have unit column length.

Step 2. Obtain the Singular-Value Decomposition[2] of X, and from this calculate

a. The condition indexes, η_k, as in (2.7) and

b. The Π-matrix of variance-component proportions as in (2.10) - (2.12).

Step 3. Determine the number and relative strengths of the near dependencies by the condition indexes exceeding some chosen threshold, η^*, such as $\eta^* = 10$, or 15 or 30[3].

Step 4. Examine the condition indexes for the presence of competing dependencies (roughly equal condition indexes) and dominating dependencies (high condition indexes - exceeding the threshold determined for Step 3 - coexisting with even larger indexes.)

Step 5. Determine the involvement (and the resulting degradation to the regression estimates) of the variates in the near dependencies. Three cases are to be considered.

Case 1 - Only one near dependency present. A variate is involved in, and its estimated coefficient degraded by, the single near dependency if it is one of two or more variates with variance-component proportions in excess of some threshold value, such as 50%. Presumably, if there are not two high variance-component proportions associated with this single highest condition index, no degradation is exhibited.[4]

Case 2 - Competing dependencies. Here involvement is determined by aggregating the variance-component proportions over the competing condition indexes (see point 11 of Section 4.1). Those variates with aggregate proportions exceeding the threshold are involved in at least one of the competing dependencies, and therefore have degraded coefficient estimates. In this case, it is not possible exactly to determine in which of the competing near dependencies the variates are involved.

Case 3 - Dominating dependencies.[5] In this case 1) we cannot rule out the involvement of a given variate in a dominated dependency if its variance is being greatly determined by a dominating dependency, and 2) we cannot assume the noninvolvement of a variate even if it is the only one with a high proportion of the variances associated with the dominated condition index - other variates can well have their joint involvement obscured by the dominating near dependency. In this case additional analysis, such as auxiliary regressions, is warranted, directly to investigate the descriptive relations among all of the variates potentially involved. See point 13 of Section 4.1.

Step 6. Examine the underlying near dependencies. Once the number of near dependencies and the variates involved are determined (if the latter is possible), regressions among the indicated variates can be run to display the relations.

Step 7. Determine those variates that remain unaffected by the presence of the collinear relations. See point 13 of Section 4.1.

Once the X matrix has been analyzed and the potential harm to regression estimates has been assessed, it is possible to analyze the quality of an actual regression based on those data. In particular, one can often learn

1. How many near dependencies plague a given data set and what they are.
2. Which variates have estimates adversely affected by the presence of those dependencies.
3. Whether estimates of interest are included among those with inflated confidence intervals, and therefore whether corrective action (obtaining better conditional data or applying Bayesian techniques) is warranted.
4. Whether, rather generally, prediction intervals based on the estimated model are greatly inflated by the presence of ill-conditioned data.
5. Whether specific coefficient estimates of interest are relatively isolated from the ill effects of collinearity and therefore trustworthy in spite of ill-conditioned data.

Software. The computational foundation of the diagnostic procedure reported here is the singular-value decomposition of Step 2, a computational routine whose accessability would seem to be somewhat limited [6]. In point of fact, a routine called EISPACK - Release 2 contains a very efficient version of the SVD algorithm and already has been made available to over 200 university computer libraries.[7] Furthermore, an interactive routine has already been designed specifically to effect the computational steps 1, 2, and 6 and exists as part of the TROLL system at the National Bureau of Economic Research, Computer Research Center.[8]

4.3 Applications with Actual Data

With one very interesting exception,[8] we have, until now, employed the proposed diagnostic procedure just summarized only on data matrices with contrived near dependencies. We turn now to an analysis of two matrices of actual data to see how the procedure fares when dealing with naturally occurring, uncontrived near dependencies. The first example makes use of the Bauer matrix that was introduced in a different context in Part 2, and the second example makes use of data familiar to all econometricians, those relevant to an annual, aggregate consumption function.

The Bauer Matrix.

The Bauer matrix, we recall from Part 2, had an exact contrived dependency between its last two columns, C4 = .5*C5, which were in turn orthogonal to the first three. Its purpose there was to exemplify the isolation from collinearity that is afforded those variates that are orthogonal, or nearly so, from the variates involved in the offending near dependencies. In examining the Π-matrix, Table 0, of the Bauer matrix, the involvement of var(b_4) and var(b_5) in the exact, contrived dependency was clearly observed as well as the isolation of the first three variances from it. But, in addition, there appeared an unexpected occurrence: over 97% of var(b_1), var(b_2) and var(b_3) was associated with the condition index of σ_3. We were not prepared at that time to pursue this naturally arising phenomenon, but now we are.

First we note that the singular-values and variance-component proportions of Table 0 are based on data that have not been column scaled as required in Step 1. Since column scaling does not destroy the existence of dependencies, they can still be observed from Table 0, but there will be no standardized meaning to the singular values and the resulting conditon indexes. In Step 2, then, we compute the Π-matrix and condition indexes for a Bauer matrix on column-scaled data, resulting in Table 16.

TABLE 16

Scaled Bauer Matrix

	var(b_1)	var(b_2)	var(b_3)	var(b_4)	var(b_5)	Condition Index, η
σ_1	.000	.000	.000	.000	.000	
σ_2	.005	.005	.000	.000	.000	1.0
σ_3	.001	.001	.047	.000	.000	1.3
σ_4	.994	.994	.953	.000	.000	16
σ_5	.000	.000	.000	1.000	1.000	$2\text{x}10^{16}$

In analyzing Table 16, it will prove instructive to feign ignorance of any prior knowledge we have of the properties of the Bauer matrix to see how well the mechanism discovers all there is to know.

The first and obvious fact is that there are two near dependencies with condition indexes greater than 10. One is dominating; none is competing. The dominating dependency is clearly very tight, having the astronomically large condition number of 2x10^{16} and involving columns 4 and 5. It is safe to conclude that the involvement of C1, C2 and C3 in this dependency is minimal, if any.

The second dependency (and the one that we are really interested in) possesses the weak to moderate condition index of 16. Clearly, at least the first three columns, C1, C2, and C3, are involved in this dependency, but one cannot rule out the potential involvement of C4 and C5, their roles being usurped by their involvement in the dominate dependency.

We may display these two dependencies through auxiliary regressions; we need only to choose the two variates to act as dependent variates, the three remaining being independent. In this case, choosing one of C1, C2 or C3 and one of C4 or C5 is clearly appropriate and Table 17 presents the auxiliary regression results with C1 and C4 chosen as the two dependent variates to be regressed on C2, C3 and C5. The regressions are based on unscaled data, so that the dependencies are displayed in terms of the original data relationships.

TABLE 17

Auxiliary Regressions
Bauer Data-Unscaled

	Coefficient of			
	C2	C3	C5	R^2
C4	0.0000	0.0000	.5000	1.000
	[0.0]	[0.0]	[*]	
C5	-.7008	-1.2693	0.0000	.9820
	[-14.4]	[-7.5]	[0.0]	

*essentially infinite.

Both near dependencies are clearly displayed. In the first, we see the dominant, essentially perfect relation true of the Bauer data given in Part 2 in which C4 = .5C5, exactly. The noninvolvement of C2 and C3 in this relation is also discovered. In the second, we see a fairly strong (R^2=.98) relation involving C1, C2 and C3, but not C4. This is the naturally occurring dependency whose presence was first suggested in Part 2 and now verified. One can now conclude that all five regression estimates based on this matrix are degraded by the presence of two collinear relations. The variances for the coefficients of C4 and C5 are obviously very seriously degraded, while those for C1, C2 and C3 are less so.

It is fair to conclude that the diagnostic procedure, when applied to the Bauer matrix, has been very successful in uncovering all relevant properties of the near dependencies contained in it.

The Consumption Function

All economists are familiar with annual, aggregate consumption function data, and so we analyze the following data matrix.

$$[\iota\ C(T-1)\ DPI(T),\ r(T),\Delta\ DPI(T)],$$

annual series 1947-1976

where ι is a column of 1's (the constant term)

C is total consumption, 1958 dollars

DPI is Disposable Personal Income, 1958 dollars

and r is the interest rate (Moody's Aaa).

It must be emphasized that no attempt is being made here to analyze the consumption function. There are many well known, sophisticated alterations to basic consumption data involving, for example, per-capita weightings, disaggregations, wealth effects, and recognition of simultaneity. Our interest here necessarily centers on analysis of one fundamental variant without regard to additional econometric refinements; namely

$$C(T) = \beta_o + \beta_1 C(T-1) + \beta_2 DPI(T) + \beta_3 r(T) + \beta_4 \Delta DPI(T) + \varepsilon(T). \qquad (4.1)$$

Estimation of (4.1) with ordinary least squares results in

$$\begin{array}{llllll} C(T) = 6.7242 + & .2454\ C(T-1) + & .6984\ DPI(T) & -2.2097\ r(T) + & .1608\Delta DPI(T). & \\ \quad (3.83)^{9} & (.237) & (.208) & (1.838) & (.877) & (4.2) \\ & & & R^2 = .9991\ . & & \end{array}$$

Only one of these parameter estimates, that of DPI, is significant by a standard t-test; but few econometricians would be willing to reject the hypotheses that the other β's, either jointly or singly, are significantly different from zero. Furthermore, few econometricians would be happy with the prediction intervals that would result from such a regression. This dissatisfaction stems from the widely held belief that the consumption function data are highly ill conditioned and that estimates based on them are too noisy to prove conclusive or useful.[10] A mere glance at the simple correlation matrix for these data, Table 18, partially confirms this belief. But how ill conditioned are these

TABLE 18

Correlation Matrix

Consumption-Function Data

	ι	C(T-1)	DPI(TO	r(T)	ΔDPI(T)
ι	1.000				
C(T-1)	.000	1.000			
DPI(T)	.000	.997	1.000		
r(T)	.000	.975	.967	1.000	
ΔDPI(T)	.000	.314	.377	.229	1.000

data? How many near dependencies exist among them, and how strong are they? Which variates are involved in them, giving evidence of degradation? Which estimates might benefit most from obtaining better conditioned data or from the introduction of appropriate information through a Bayesian prior? Answers to these questions, of course, cannot be obtained from Table 18 along, but can be obtained from an analysis of the Π-matrix and condition indexes for the consumption function data. For this analysis, we will continue to set the condition-index threshold of $\eta^* = 10$, and the variance-component-proportion-threshold at $\pi^*=50\%$. Steps 1 and 2 of the diagnostic procedure applied to the consumption function data results in the Π-matrix given in Table 19.

TABLE 19

Variance-Component Proportions
and Condition Indexes

Consumption Function Data

	CONST	C(T-1)	DPI(T)	r(T)	ΔDPI(T)	Condition
	$var(b_1)$	$var(b_2)$	$var(b_3)$	$var(b_4$	$var(b_5)$	Index, η
σ_1	.001	.000	.000	.000	.001	1
σ_2	.003	.000	.000	.001	.135	4
σ_3	.301	.000	.000	.012	.000	8
σ_4	.263	.004	.004	.984	.048	39
σ_5	.420	.995	.995	.000	.813	376

Table 19 shows the existence of two near dependencies, one dominant with a large condition index of 376 and one strong with a condition index of 39. The dominant relation involves C(T-1), DPI(T) and ΔDPI(T). r(T) does not seem to be involved in this dependency, but it is likely that the constant term, CONST, is being shared in both. The weaker dependency definitely includes r(T); and all other variates are potentially involved, their effects clearly being dominated by their involvement in the stronger dependency with η=376.

Auxiliary regressions are required in this case to determine those variates involved in the weaker of the two dependencies. One possible choice for the two dependent variates of these auxiliary regressions would be DPI(T) and r(T). Table 20 reports these results.

TABLE 20

Auxiliary Regressions

Consumption-Function Data (Unscaled)

Coefficients of

	ι	C(T-1)	ΔDPI(T)	R^2	η
DPI(T)	-11.5472 [-4.9]	1.1384 [164.9]	.8044 [11.9]	.9999	376
r(T)	-1.0244 [-3.9]	.0174 [22.3]	-.0145 [-1.9]	.9945	39

We verify that the dominant relation does involve ι, C(T-1), DPI(T) and ΔDPI(T), and note that the weaker involves at least ι C(T-1) and r(T).

Quite generally, then, we may conclude that the data upon which the consumption function regression (4.2) was based possess two strong near dependencies (one very strong). Furthermore, each variate is involved in one or both of these near dependencies, and each is degraded to some degree by their presence. It would appear that the estimates of coefficients of C(T-1) and DPI(T) are most seriously affected, followed closely by that for ΔDPI(T), these variates being strongly involved in either the tighter of the two dependencies or both. The estimate of the coefficient of r(T) is adversely affected by its strong involvement in the weaker of the two dependencies, but, in our experimental experience, we found η's of 39 to be quite large, and the R^2 in Table 20 confirms this here. Thus we see that all parameter estimates in (4.2), and their estimated standard errors, show great potential for refinement through better conditioning of the estimation problem, either from more appropriate modeling or the introduction of better conditioned data or an appropriate Bayesian prior. One would be loath to reject, for example, the role of interest rates in the aggregate consumption function on the basis of the estimates of (4.2); and one would feel even more helpless in predicting the effects of a change in r on aggregate consumption from a regression equation like (4.2). Thus the econometrician's intuitive dissatisfaction with estimates of the aggregate consumption function, and his seemingly never-ending efforts to refine them, seems fully justified.

Several additional points of interest arise from this example, some of which suggest future directions for research. First, the fact that the estimated coefficient of DPI(T) appears to possess any degree of statistical significance at all reflects the fact that C(T) and DPI(T) are phenomenally highly correlated (.9999). In light of this, the seriousness of the degradation of the estimate of this parameter is seen in the fact that its standard error is still quite large, resulting in the very broad 95% confidence interval of [.28 - 1.11]. Second, as seen from Table 19, no one near dependency dominates the determination of the variance of the estimate of the constant term. This estimate is nevertheless degraded since nearly 70% of the variance is associated with the two near dependencies, as is verified by the auxiliary regressions in Table 20. This lack of dominance is to be contrasted with the estimates of the coefficients of C(T-1) and DPI(T), which also clearly enter both near dependencies but are greatly dominated by the stronger of the two. This situation suggests, in accord with intuition, that it is possible for a variate that is weakly involved in a strong near dependency to be confounded with one that is more strongly involved in a weaker near dependency. Similar results occur in the experiments of Part 3, but not in such a way that any definite conclusions can be drawn. Further experimentation will be needed directly to test this suggestion. Third, within a given near dependency, there appears to be a strong rank correlation between the relative size of the variance-component proportions of the variates involved and their t-statistics in the corresponding auxiliary regressions. Comparing the variance-component proportions for the near dependency with $\eta = 376$ in Table 19 and the corresponding t's for the DPI(T) regression in Table 20 exemplifies the point. Of course, allowance must be made for relations that are dominated (such as the one with $\eta = 39$) or are competing, but again there is considerable support, but no substantiation, for such an hypothesis from the experiments of Part 3, and further experiments aimed directly to this point are suggested. Fourth, even with this "real-world" data, the relative progression between correlations and condition indexes summarized in point 2 of Section 4.1 continues to hold. The near dependencies of the consumption data are of orders of magnitudes 30 and 300, two degrees apart along the 3, 10, 30, 100, 300, etc. progression. Similarly, the R^2's of the auxiliary regressions reported in Table 20 are .99 and .9999, two degrees apart along the 9's progression .9, .99, .999, .9999 etc. Fifth, we once again note the ability of these diagnostic tools to uncover complex relations among three or more variates that are overlooked by simple correlation analysis. The simple correlation matrix in Table 18 surely tells us that DPI(T) and C(T-1) are closely related; but the role of ΔDPI(T) (or, equivalently, the role of DPI(T-1)) is not at all observable from this information. The largest simple correlation with ΔDPI(T) is under .4. ΔDPI(T)'s role in a near dependency along with C(T-1) and DPI(T), however, is readily apparent from the variance-component proportions matrix of Table 19.

Part 1

Footnotes

[1] This term will be given meaning in Section 2.5.

[2] The emphasis of this paper is on diagnostics and does not deal with corrective mechanisms. The introduction of additional, well-conditioned data is, of course, one straightforward solution to the problem when such data are available. In other instances much interest attaches to the solutions offered through applications of Bayesian and mixed-estimation techniques. The reader is referred to Zellner (1971), Leamer (1973) and Theil (1971).

[3] This term will be defined in the next section.

[4] The terms multicollinearity, collinearity, dependencies, near dependencies, near collinearity, near singularity have all been used more or less formally in this context. The first five terms are used interchangably in this paper.

[5] X is, of course, assumed to have full rank, but this is not testable, for its absence, the null hypothesis, renders the regression model invalid.

[6] See, for example, Hawkins (1973) or Golub, Klema and Stewart (1976) or Webster, Gunst and Mason (1974) (1976).

Part 2

Footnotes

[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 (2.1) 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

$$\overset{\text{TxK}}{X} = \overset{\text{TxT}}{U}\ \overset{\text{TxK}}{\Sigma}\ \overset{\text{KxK}}{V'} \tag{2.1a}$$

or

$$\overset{\text{TxK}}{X} = \overset{\text{Txr}}{U}\ \overset{\text{rxr}}{\Sigma}\ \overset{\text{rxK}}{V'} \tag{2.1b}$$

where r = rank X. In this latter formulation Σ is always of full rank, even if X is not.

[4] See Golub (1969). These notions of conditioning and the reasons for the relation between the conditioning of X vs X'X will be explained subsequently.

[5] Furthermore, in operating directly in the TxK matrix X, the SVD avoids the additional computational burden of forming X'X, a TK^2 operation.

[6] Two examples from Golub and Reinsch (1970) and Wilkinson (1965) illustrate this point. Consider

$$\begin{bmatrix} .501 & -1 & & & 0 \\ & .502 & -1 & & \\ & & \ddots & \ddots & \\ & & & .599 & -1 \\ 0 & & & & .600 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & -1 & -1 & \cdots & -1 \\ & 1 & -1 & \cdots & -1 \\ & & & \ddots & \vdots \\ 0 & & & & 1 \end{bmatrix} .$$

Each of these matrices will be shown by the singular-value decomposition, in a way described later, to be quite ill conditioned even though neither possesses a small diagonal element.

[7] The spectral norm of the mxn matrix $A = (A_{ij})$, denoted $\|A\|$ is defined [Wilkinson (1965)] as $\sup\limits_{\|x\|_2=1} \|Ax\|_2$, where x is an n-vector and $\|\cdot\|_2$ denotes the Euclidean norm $\|y\|_2 = (\sum\limits_{i=1}^{n} y_i^2)^{\frac{1}{2}}$. It is readily shown that $\|A\| = \sigma_{max}$, i.e., the maximal singular value, a result that is equivalent to finding the first principal component of A.

Part 2 Footnotes (Continued)

[8] It is readily apparent from the application of the SVD (2.1) to a real symmetric matrix A that the singular values of A are also its eigenvalues.

[9] See Golub and Reinsch (1970) or Becker, et al. (1974).

[10] The reader who is interested in the analysis of the sensitivity of A^{-1} to perturbations in elements of A and in the determination of when a matrix is, as a practical matter, rank deficient (i.e., when σ_{min} may be considered equal to zero relative to σ_{max}) is referred to Wilkinson (1965) and to Golub, Klema and Stewart (1976). These studies give a greater appreciation of the interpretation of the condition number X than can be given here. See also Van der Sluis (1969) and (1970).

[11] The matrix $B = \alpha I$ employed above provides an excellent example here.

[12] This link is obviously of the utmost importance, for ill conditioning is a numeric property of a data matrix having, in itself nothing directly to do with least squares estimation. To have meaning in a regression context, then, there must be some means by which the numeric information on ill conditioning can be shown to directly affect the quality (variances) of the regression estimates. It is this link, for example, that is lacking in the Farrar and Glauber techniques described in Section 1.1.

[13] "Two or more" since there must be at least two columns of X involved in any dependency.

[14] The careful wording "it is always possible to find" is required here. As is shown in Belsley and Klema (1975), if there are multiple roots of X, there is a class of V's in the SVD of X, one, but not all, of which takes the partitioned form shown. Such multiplicities are therefore of theoretical importance but of little practical consequence since they will occur with probability zero in a "real life" economic data matrix.

[15] 10^{-14} on the IBM 370 67 in double precision.

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

[17] The reader is warned against interpreting the condition indexes from these singular values at this point. For reasons explained in Part 3 the data should be scaled first to have equal column lengths, and the resulting singular values subjected to analysis. For the analysis of this section, however, scaling is unnecessary.

[18] That these components are non zero at all is only due to the finite arithmetic of the machine. In theory these components are an undefined ratio of zeros that would be defined to be zero for this case.

[19] Part III is devoted to experiments that help us put meaning to "high" and "large", two terms whose meaning in this context can only be determined empirically and necessarily must be used loosely here.

[20] For a discussion of the theoretical underpinning to this topic see Wilkinson (1965). A more detailed discussion of its implications in econometrics is contained in Belsley and Klema (1975).

Part 2 Footnotes (Continued)

[21] The condition number of the moment matrix X'X is the square of that of X. This is seen from SVD of $X = U\Sigma V'$. Hence $X'X = V\Sigma^2V'$, and, by definition, this must also be SVD of X'X. Clearly, then, $\kappa(X'X) = \frac{\sigma^2_{max}}{\sigma^2_{min}} = \kappa^2(X)$. Hence, any ill conditioning of X is greatly compounded in its ill effects on a least-squares solution calculated as $b = (X'X)^{-1}X'y$. Procedures for calculating b that do not require forming X'X or its inversion exist, however. See Golub (1969) or Belsley (1974).

[22] To avoid any possible confusion it is worth highlighting that this is the statistical use of the word <u>conditional</u>, having nothing directly to do with (and thus to be contrasted with) the numeric-analytic notion of ill-conditioned data.

[23] In addition, this statistical problem, but not necessarily the computational problem can be alleviated by the introduction of Bayesian prior information. See Zellner (1971), Leamer (1973).

[24] Either statistical or computational, provided, of course a regression algorithm is used that does not blow up in the presence of highly collinear data. Standard routines based on solving $b = (X'X)^{-1}\ X'y$ are quite sensitive to ill-conditioned X. This problem is greatly overcome by regression routines based on SVD of X, or a QR decomposition [see Belsley (1974)].

Part 3
Footnotes

[1] As we shall see in Experiment #2, this objective was only partially achieved, leading to an unexpected set of dependencies that nevertheless provided a further successful test of the usefulness of this analytical procedure.

[2] No attempt is made here to infer any statistical properties through repeated samplings.

[3] Scale changes do not, however, affect the presence of linear dependencies among the columns of X since for any nonsingular matrix D there exists a non zero c such that $Xc = 0$ if and only if $[XD][D^{-1}c] \equiv \overline{Xc} = 0$ where $\overline{X} = XD$ and $\overline{c} = D^{-1}c$.

[4] This scaling is similar to that used to transform the cross-products matrix X'X into a correlation matrix, except that the "mean zero" properly is not needed, and, indeed, would cause unnecessary problems in the event that X contains a constant column.

[5] Furthermore an important converse is true with scaled data; namely, when all condition indexes of a scaled data matrix are equal to unity, the columns are mutually orthonormal. This is readily proved by noting that all condition indexes equal to 1 implies $\Sigma = I$. Hence, in the SVD of X, we have $X = U\Sigma V' = UV'$, or $X'X = V'U'UV = I$, due to the orthogonality of U and V. This result is important, because it rules out the possibility that several high variance-component proportions could be associated with a very low (near unit) condition index.

[6] Although, perhaps a costly and time-consuming set of tests based on partial correlations or block regressions on the columns of the data matrix encompassing all possible combinations could be of some value

[7] A star, *, before a series name indicates a dummy series was used having the same mean and variance as the given series, but generated to provide a well conditioned basic data matrix.

[8] See, for example, Belsley (1969).

[9] This progression corresponds closely to equal increments in log η_i of $\frac{1}{2}$, i.e., $\log \eta_i = 1 + \frac{i}{2}$.

[10] Although we have already seen something like it above in the case of X2(0).

[11] Indeed the highest simple correlation between the four basic columns is -.32.

[12] Even though var(b_4) is not greatly determined by the condition indexes of 8 and 11, C4's involvement in these two dependencies cannot be ruled out. The variance-proportions, as we have seen, can be arbitrarily distributed among η's of nearly equal magnitude.

Part 3 Footnotes (Continued)

[13] The choice of C1 and C3 on C2 and C4 is arbitrary. Any two of the four variates with a nonvanishing jacobian could be selected for this discriptive use of least squares. The figures in the square brackets are t's not standard deviations and the R^2's are the ratio of predicted sum of squares (not deviations about the mean) to actual sum of squares since there is no constant term.

[14] See, however, point 3 following.

[15] As we have just seen in points 2 and 3 above, point 1 requires some modifications when there are either competing or dominating dependencies. These modifications will be treated fully in Section 4.1.

[16] Indeed the experience so far indicates that the condition index goes one further step in the 10, 30, 100, 300, 1000 progression as successive "9's" are added to the underlying correlation. For example, .5→10, .9→30, .99→100, .999→300, etc.

Part 4
Footnotes

[1] cf., however, 6) above.

[2] Programs effecting this decomposition will be discussed in the text below.

[3] Choosing this threshold is akin to choosing a test size (α) in standard statistical hypothesis testing - and only practical experience will help determine a useful rule of thumb. $\eta^* = 10$ seems a good start. As a matter of practice it seems reasonable to ignore all condition indexes below the threshold as being too weak for further consideration, regardless of what patterns of variance-component proportions may be associated with them.

[4] This situation has, as yet, not occurred in practice, and as long as the data matrix has been properly scaled, as in Step 1, it doesn't seem likely that it will, cf footnote 5, Part 3.

[5] The joint occurrence of dominating and competing dependencies causes no additional difficulties. The competing dependencies, whether dominated or dominating, are merely treated as one in association with their aggregate variance-component proportions.

[6] As noted in Part 2, the eigenvectors of X'X and the positive square roots of its eigenvalues provide identical information as the SVD of X, but it is not recommended that the calculations be so obtained, for calculations based on X'X are computationally very much less stable than those based on X when X is ill conditioned - the case that is central to this analysis, of footnote 21, Part 2.

[7] Copies of EISPACK-Release 2 and further information on it may be obtained from Dr. Wayne Cowell, Argonne Code Center, Argonne National Laboratories, Argonne, Illinois, 60439.

[8] The unexpected weak, "background" dependency that was discovered when we examined the experimental series Y3.

[9] Numbers in parenthesis are standard errors.

[10] Indeed, few functions have received greater attention than the consumption function in efforts made to overcome the ill-conditioned data and refine its estimation.

REFERENCES

Bauer, F.L. (1971), "Elimination with Weighted Row Combinations for Solving Linear Equations and Least Squares Problems", pp. 119-133 in *Handbook for Automatic Computation, Vol. II: Linear Algebra*, Eds. Wilkinson, J.H. and Reinsch, C. New York: Springer-Verlag.

Becker, R., N. Kaden and V. Klema (1974), "The Singular Value Analysis in Matrix Computation", Working Paper No. 46, Computer Research Center, National Bureau of Economic Research.

Belsley, D.A. (1969), *Industry Production Behavior: The Order-Stock Distinction*, North-Holland: Amsterdam.

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

Belsley, D.A. and Virginia C. Klema (1974), "Detecting and Assessing The Problems Caused by Multicollinearity: A Use of The Singular-Value Decomposition", Working Paper #66, Computer Research Center, National Bureau of Economic Research, Cambridge.

Businger, P. and G.H. Golub (1965), "Linear Least Squares Solutions by Householder Transformations". *Numerische Mathematik*, Vol. 7, pp. 269-276.

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

Frisch, R. (1934), "Statistical Confluence Analysis by Means of Complete Regression Systems", University Institute of Economics, Oslo.

Golub, Gene H. (1969), "Matrix Decompositions and Statistical Calculations". *Statistical Computation*, Academic Press, New York, pp. 365-397.

Golub, G., V. Klema and G.W. Stewart (1976), "Rank Degeneracy and Least-Squares Problems", Computer Science Technical Report Series, #TR-456, University of Maryland.

Golub, G.H. and C. Reinsch (1970), "Singular Value Decomposition and Least Squares Solutions". *Numerische Mathematik*, Vol. 14, pp. 403-420.

Haitovsky, Yoel (1969), "Multicollinearity in Regression Analysis: Comment". *Review of Economics and Statistics*, November, pp. 486-489.

Hansen, Richard J. and Charles L. Lawsen (1969), "Extensions and Applications of The Householder Algorithm for Solving Linear Least Squares Problems". *Mathematics of Computation*, Vol. 23, No. 108, October, pp. 787-812.

Hawkins, D.M. (1973), "On the Investigation of Alternative Regressions by Principal Component Analysis", *Applied Statistics*, 22, pp. 275-286.

Kendall, M.G. (1957), *A Course in Multivariate Analysis*, London: Griffin.

Kloeck, T. and L.B.M Mennes (1960), "Simultaneous Equations Estimation Based on Principal Components of Predetermined Variables", *Econometrica*, Vol. 28, No. 1, pp. 45-61.

Kumar, T.K. (1975), "Multicollinearity in Regression Analysis", *The Review of Economics and Statistics*, Vol. LVII, August, pp. 365-366.

Lawson, C.R. and R.J. Hanson (1974), *Solving Least-Squares Problems*, Prentice-Hall: Englewood Cliffs, N.J.

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

O'Hagan, J. and B. McCabe (1975), "Tests for the Severity of Multicollinearity in Regression Analysis: A Comment". *The Review of Economics and Statistics*, Vol. LVII, August, pp. 369-370.

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

Stewart, G.W. (1973), *Introduction to Matrix Computations*, Academic Press, New York.

Theil, H. (1971), *Principles of Econometrics*, John Wiley & Sons: New York.

Van der Sluis, A. (1969), "Condition Numbers and Equilibration of Matrices". *Numeriche Mathematik*, 14, pp. 14-23.

Van der Sluis, A. (1970), "Condition Equilibration and Pivoting in Linear Algebraic Systems". *Numeriche Mathematik*, 15, pp. 74-86.

Webster, J.T., R.F. Gunst and R.L. Mason (1974), "Latent Root Regression Analysis", *Technometrics*, 16, pp. 513-522.

Webster, J.T., R.F. Gunst and R.L. Mason (1976), "A Comparison of Least Squares and Latent Root Regression Estimators", *Technometrics*, 18, pp. 75-83.

Wilkinson, J.H. (1965), *The Algebraic Eigenvalue Problem*, Oxford University Press.

Zellner, A. (1971), *An Introduction to Bayesean Inference in Econometrics*, John Wiley & Sons, New York.

National Bureau of Economic Research

Working Papers from

THE CENTER FOR ECONOMIC ANALYSIS OF HUMAN BEHAVIOR AND SOCIAL INSTITUTIONS

Number	Author	Title	Date
1	Finis Welch	Education, Information, and Efficiency	6/73
2	Barry R. Chiswick	Hospital Utilization: An Analysis of SMSA Differences in Hospital Admission Rates, Occupancy Rates and Bed Rates (Explorations in Economic Research, summer 1976)	6/73
4	L. A. Lillard	Human Capital Life Cycle of Earnings Models: A Specific Solution and Estimation	7/73
5	James P. Smith	A Life Cycle Family Model	7/73
7	Lewis C. Solmon	The Definition and Impact of College Quality and Its Impact on Earnings (Explorations in Economic Research, Vol. 2, No. 4, Fall 1975)	8/73
9	L. A. Lillard	From Age-Earnings Profiles to the Distribution of Earnings and Human Wealth	9/73
12	Melvin Reder	Citizen Rights and the Cost of Law Enforcement (J. Legal Studies, June 1974)	10/73
14	Lewis C. Solmon and Paul Wachtel	The Effects on Income of Type of College Attended (Sociology of Education, Vol. 48, Winter 1975)	10/73
17	Paul Taubman	Schooling, Ability, Non Pecuniary Rewards, Socioeconomic Background and the Lifetime Distribution of Earnings (NBER, Income and Wealth, #41, in press)	11/73
18	Isaac Ehrlich	The Deterrent Effect of Capital Punishment: A Question of Life and Death (American Economic Review, June 1975)	11/73
19	Edward F. X. Hughes, Eugene M. Lewit, Richard N. Watkins, Richard Handschin	Utilization of Surgical Manpower in a Prepaid Group Practice (New England Journal of Medicine, Vol. 291, pp. 759-763, October 10, 1974)	12/73
20	Victor R. Fuchs	Short-Run and Long-Run Prospects for Female Earnings (American Economic Review, May 1974)	12/73
21	Robert T. Michael and Robert J. Willis	Contraception and Fertility: Household Production under Uncertainty (NBER, Income and Wealth #40, 1976)	12/73
22	Michael Grossman	The Correlation Between Health and Schooling (NBER, Income and Wealth #40, 1976)	12/73
23	Warren Sanderson	What Happened During the Baby Boom? New Estimates of Age- and Parity-Specific Birth Probabilities for American Women	12/73
25	John C. Hause	The Covariance Structure of Earnings and the On-the-Job Training Hypothesis	12/73
27	Arleen Leibowitz	Production within the Household (American Economic Review, May 1974)	1/74
28	Yoram Weiss	The Wealth Effect of Occupational Choice (International Economic Review, June 1976)	1/74
29	Elisabeth M. Landes	Male-Female Differences in Wages and Employment: A Specific Human Capital Model	1/74
30	Sue Goetz Ross	The Timing and Spacing of Births and Women's Labor Force Participation: An Economic Analysis	1/74
34	James J. Heckman and Robert J. Willis	Estimation of a Stochastic Model of Reproduction: An Econometric Approach (NBER, Income and Wealth #40, 1976)	2/74
36	Warren C. Sanderson	Economic Theories of Fertility: What Do They Explain?	3/74
39	Jacob Mincer	Unemployment Effects of Minimum Wages (Journal of Political Economy, 1976)	5/74
40	William M. Landes	Legality and Reality: Some Evidence on Criminal Procedure (Journal of Legal Studies, June 1974)	5/74
41	Richard A. Posner	Theories of Economic Regulation (Bell Journal of Economics and Management Science, Autumn 1974)	5/74
42	Gary S. Becker	A Theory of Social Interactions (Journal of Political Economy, November/December 1974)	6/74
47	Lee A. Lillard	The Distribution of Earnings and Human Wealth In a Life Cycle Context (NBER, Income and Wealth, #41, in press)	7/74

*Copies, if available, may be obtained from authors. Citations are included for those which have appeared in print.

Number	Author	Title	Date
49	Arleen Leibowitz	Years and Intensity of Schooling Investment	8/74
50	Orley Ashenfelter and James Heckman	Measuring the Effect of an Antidiscrimination Program	8/74
51	Edward Lazear	Age, Experience, and Wage Growth (American Economic Review, September 1976)	8/74
53	Jacob Mincer	Progress in Human Capital Analyses of the Distribution of Earnings (Royal Economic Society, Personal Income Distributions, 1976)	8/74
55	Richard A. Posner	The Social Costs of Monopoly and Regulation (J. Political Economy, Vol. 83, August 1975)	9/74
59	Ann Bartel	An Analysis of Firm Demand for Protection Against Crime (Journal of Legal Studies, Vol. 4, June 1975)	10/74
62	William M. Landes and Richard Posner	The Private Enforcement of Law (Journal of Law and Economics, Vol. 4, January 1975)	11/74
67	Alan S. Blinder and Yoram Weiss	Human Capital and Labor Supply: A Synthesis (Journal of Political Economy, June 1976).	1/75
71	Isaac Ehrlich and Uri Ben-Zion	On the Theory of Productive Saving (Economic Inquiry, forthcoming in 1976)	1/75
73	Mark Pauly	The Role of Physicians in the Production of Hospital Output	2/75
74	Robert T. Michael	Variation Across Households in the Rate of Inflation	3/75
80	Lee A. Lillard	Inequality: Earnings vs. Human Wealth (American Economic Review, forthcoming)	4/75
81	Arleen Leibowitz	The Parental Bequest to Children	5/75
92	Edward Lazear	Schooling As a Wage Depressant (Journal of Human Resources, forthcoming)	6/75
97	Edward Lazear	Human Wealth and Human Capital	7/75
98	Barry R. Chiswick	The Demand for Nursing Home Care: An Analysis of the Substitution Between Institutional and Non-Institutional Care (Journal of Human Resources, Summer 1976)	7/75
99	Yoram Weiss	The Earnings of Scientists, 1960-1970: Experience, Age and Vintage Effects	7/75
102	Kenneth Wolpin	Education and Screening	8/75

Number	Author	Title	Date
104	Edward Lazear	Education: Consumption or Production	9/75
105	Richard N. Watkins, Edward F. X. Hughes, Eugene M. Lewit	Time-Utilization of a Population of General Surgeons in a Prepaid Group Practice	10/75
107	Michael J. Boskin	Social Security and Retirement Decisions	10/75
108	Victor R. Fuchs	Are Health Workers Underpaid? (Explorations in Economic Research, Summer 1976)	10/75
110	William M. Landes, Richard A. Posner	Independent Judiciary in an Interest-Group Perspective (J. Law and Economics, forthcoming)	10/75
112	James J. Heckman, Robert J. Willis	A Beta-Logistic Model for the Analysis of Sequential Labor Force Participation by Married Women	11/75
115	Douglas Coate	The Production of Health Services in Fee for Service, for Profit Health Practices: The Case of Optometrists	11/75
116	Michael J. Boskin	Notes on the Tax Treatment of Human Capital	11/75
117	Ann P. Bartel	Job Mobility and Earnings Growth	11/75
120	Victor R. Fuchs	From Bismarck to Woodcock: The "Irrational" Pursuit of National Health Insurance (Journal of Law and Economics, 1976)	1/76
121	Lee A. Lillard and Yoram Weiss	Analysis of Longitudinal Earnings Data: American Scientists 1960-70	1/76
122	Douglas Coate	The Market for Optometric Services in the United States	2/76
123	Gary S. Becker, Nigel Tomes	Child Endowments, and the Quantity and Quality of Children	2/76
129	Michael Grossman	A Survey of Recent Research in Health Economics	3/76
132	Donald O. Parsons	Health, Family Structure, and Labor Supply	4/76
133	Sam Peltzman	Toward a More General Theory of Regulation	4/76
134	Fred Goldman, Michael Grossman	The Demand for Pediatric Care: An Hedonic Approach	4/76

Number	Author	Title	Date
135	Michael J. Boskin	Taxation, Saving and the Rate of Interest	4/76
137	Reuben Gronau	Leisure, Home Production and Work--The Theory of the Allocation of Time Revisited	5/76
138	Yoram Weiss and Lee A. Lillard	Experience, Vintage and Time Effects in the Growth of Earnings: American Scientists, 1960-70	5/76
139	Kenneth E. Warner	Treatment Decision-Making in Catastrophic Illness (Medical Care, forthcoming 1977)	5/76
140	Yoram Weiss	A Multiplicative Model of Investment in Human Capital	6/76
141	Edward Lazear	Family Background and Optimal Schooling Decisions	7/76
143	George J. Borjas and Jacob Mincer	The Distribution of Earnings Profiles in Longitudinal Data	8/76
144	Michael Boskin and Eytan Sheshinski	Optimal Income Redistribution When Individual Welfare Depends upon Relative Income	8/76
145	Edward Lazear	Intergenerational Externalities	8/76
146	William M. Landes and Richard A. Posner	Legal Precedent: A Theoretical and Empirical Analysis	8/76
147	Robert T. Michael	Factors Affecting Divorce: A Study of the Terman Sample	8/76
148	Reuben Gronau	Who Is the Family's Main Breadwinner? --The Wife's Contribution to Full Income	9/76
150	Lee A. Lillard and Robert J. Willis	Dynamic Aspects of Earnings Mobility	9/76
151	Joseph E. Stiglitz and Michael J. Boskin	Some Lessons from the New Public Finance	10/76
152	Michael J. Boskin	Optimal Tax Theory, Econometric Evidence, and Tax Policy	10/76
153	Gary S. Becker, Elisabeth M. Landes, Robert T. Michael	Economics of Marital Instability	10/76
158	B. Peter Pashigian	The Market for Lawyers: The Determinants of the Demand for and Supply of Lawyers	12/76

Number	Author	Title	Date
160	J. Huston McCulloch	Misintermediation and Business Fluctuations	1/77
161	Ann P. Bartel and George J. Borjas	Middle-Age Job Mobility: Its Determinants and Consequences	1/77