Statistical Science
1987, Vol. 2, No. 1, 68–100

Collinearity and Least Squares Regression

G. W. Stewart

Abstract. In this paper we introduce certain numbers, called collinearity indices, which are useful in detecting near collinearities in regression problems. The coefficients enter adversely into formulas concerning significance testing and the effects of errors in the regression variables. Thus they provide simple regression diagnostics, suitable for incorporation in regression packages.

Key words and phrases: Collinearity, ill-conditioning, linear regression, errors in the variables, regression diagnostics.

1. INTRODUCTION

Statisticians and numerical analysts share a concern about the effects of near collinearities on regression models—and with good reason. For the statistician, near collinearities inflate the variances of regression coefficients and magnify the effects of errors in the regression variables. For the numerical analyst, they combine with rounding errors to introduce inaccuracies in computations. It is not surprising then that both groups have devoted a great deal of effort to issues related to collinearity. In spite of this the subject has a certain vagueness about it, and it is instructive to ask why.

In Section 3 we are going to survey some measures of collinearity that have appeared in the statistics and numerical analysis literature. It is significant that, with one exception, none of these measures was originally introduced to measure collinearity. For example, we shall see that large variance inflation factors imply near collinearity. Yet they were introduced by C. Daniel to show how the variance in a response vector is magnified in the regression coefficients (the name is due to D. W. Marquardt). Similarly, the numerical analysts' condition number was introduced by Turing (1948) to bound perturbations in the solutions of linear systems and was later extended by Golub and Wilkinson (1966) to least squares solutions. Its relation to collinearity usually appears as a curious incidental. While such a way of proceeding is likely to leave one with the (correct) impression that collinearity is troublesome, it is not the same as a systematic development of the subject (however, see Belsley, Kuh, and Welsch, 1980).

G. W. Stewart is Professor in the Computer Science Department and Research Professor in The Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742.

In this paper we will turn things around by starting with a set of "collinearity indices," one for each column of a regression matrix. When the regression matrix has been centered and scaled so that the cross-product matrix is a correlation matrix, the numbers are simply the diagonals of the inverse cross-product matrix—the variance inflation factors; however, they are defined in such a way that they are independent of column scaling and are applicable to models without a constant term. We will first show that the indices indicate the presence of near collinearity in a precisely quantifiable manner. We will then show that near collinearity is a bad thing by showing how the indices appear adversely in formulas concerning significance testing and the effects of errors in the variables. A bonus of this approach is that it provides simple diagnostics, suitable for incorporation into regression packages.

The paper is organized as follows. In Section 2 we will introduce the notation and conventions that will be observed throughout the paper. In Section 3 we will survey certain numbers associated with regression problems that have been found to be related to collinearity. This survey will lead us to our collinearity coefficients, whose definition and properties are the subject of Section 4. Since the justification for introducing these coefficients lies in their practical consequences, we will analyze the effects of near collinearity on significance testing in Section 5 and its interaction with errors in the variables in Section 6. The paper concludes with a summary and a discussion of further areas for research.

2. NOTATION AND CONVENTIONS

In this section we will introduce the notation that will be used throughout the paper. We will deal with least squares estimation in the linear model

$$y = Xb + e, \tag{2.1}$$

where X is an $n \times p$ matrix of rank p and e is a vector of uncorrelated random variables having mean zero and variance σ^2. Since it is sometimes a source of confusion, let us state at the outset that the matrix X is simply a fixed array of numbers; for example, X could be a design matrix from an unbalanced, fixed-effects analysis of variance or could consist of levels of controlled variables in an experiment. If the rows of X can be regarded as coming from some multivariate distribution or if the true value of y bears some underlying functional relationship to the columns of X, no account of it will be taken here. In other words, our model is specified by the matrix X alone.

To make our results as widely applicable as possible, we will not assume that the model (2.1) has a constant term. When there is one, we will assume that it is present in the regression matrix as a column of ones, unless it has been explicitly stated that X has been centered by subtracting column means. Similarly, we will not assume that the columns of X have been scaled so that the cross-product matrix $X^{\mathrm{T}}X$ has the form of a correlation matrix. Note that we do not preclude any of these things—we just do not assume them.

The jth column of X will be written x_j. The cross-product matrix will be written

$$A = X^{\mathrm{T}}X,$$

and the (i,j)-element of A^{-1} written $\alpha_{ij}^{(-1)}$. The pseudo-inverse of X will be written

$$X^{\dagger} = A^{-1}X^{\mathrm{T}},$$

and its jth row as $x_j^{(\dagger)}$.

Many of our results will be more easily derived from the QR decomposition of X (for details see Stewart, 1974). Specifically, there is an orthogonal matrix $Q = (Q_{\mathrm{X}}Q_{\perp})$ with Q_{X} an $n \times p$ matrix such that

$$\begin{pmatrix} Q_{\mathrm{X}}^{\mathrm{T}} \\ Q_{\perp}^{\mathrm{T}} \end{pmatrix} X = \begin{pmatrix} R \\ 0 \end{pmatrix}, \tag{2.2}$$

where R is an upper triangular matrix with positive diagonal elements. Multiplying this relation by Q, we get

$$X = Q_{\mathrm{X}}R, \tag{2.3}$$

from which it follows that the columns of Q_{X} form an orthonormal basis for the column space of X and the columns of $Q_{\perp}$ a basis for its orthogonal complement.

The QR decomposition is related to A and $X^{\dagger}$ as follows:

$$A = R^{\mathrm{T}}R \tag{2.4}$$

and

$$X^{\dagger} = R^{-1}Q_{\mathrm{X}}^{\mathrm{T}}. \tag{2.5}$$

From these relations and the triangularity of R it follows that

$$\rho_{pp}^{-2} = \alpha_{pp}^{(-1)} = \| x_p^{(\dagger)} \|^2, \tag{2.6}$$

where ρ_{pp} is the (p, p) element of R. We shall have occasion to refer to these relations later.

The norm in (2.6) is the usual Euclidean norm defined by $\| x \|^2 = x^{\mathrm{T}}x$. We shall also use two matrix norms: the spectral norm defined by

$$\| X \| = \max_{\| b \| = 1} \| Xb \| \tag{2.7}$$

and the Frobenius norm defined by

$$\| X \|_{\mathrm{F}}^2 = \sum_{i,j} x_{ij}^2 = \text{trace } X^{\mathrm{T}}X. \tag{2.8}$$

For more on these norms, see Golub and Van Loan (1983).

3. MEASURES OF NEAR COLLINEARITY

In the course of analyzing regression problems, numerical analysts and statisticians have introduced certain diagnostic numbers which turn out to be related to collinearity. Numerical analysts work with singular values and condition numbers. Statisticians work with correlations, both simple and multiple, and with variance inflation factors. The inexperienced in both groups sometimes suggest looking at the determinant of the scaled cross-product matrix. In this section we will discuss these numbers and their relations.

If a numerical analyst who is familiar with the art of matrix computations were asked for a reliable way of detecting near collinearity (or rank degeneracy as he might say), his first reply would probably be to compute the singular value decomposition and look at the smallest singular value. This is equivalent to looking at the number

$$\inf(X) \overset{\text{def}}{=} \min_{\| v \| = 1} \| Xv \|, \tag{3.1}$$

whose square is the smallest eigenvalue of the cross-product matrix A. The justification is the following result due to Eckart and Young (1936), as generalized by Mirsky (1960):

> $\inf(X)$ *is the spectral norm of the smallest matrix* E *such that* $X + E$ *is exactly collinear.*

Thus $\inf(X)$ measures the absolute distance of X from collinearity.

The fact that $\inf(X)$ is an absolute measure makes it difficult to interpret in the absence of information about the size of X. There are two solutions to this problem: scale X according to some fixed convention before computing $\inf(X)$, or scale $\inf(X)$ itself. Numerical analysts have tended to follow the latter

course, which leads directly to our second measure of collinearity—the condition number.

The condition number of a matrix X is defined by

$$\kappa(X) = \|X\|\,\|X^\dagger\|. \tag{3.2}$$

Since $\inf(X) = \|X^\dagger\|^{-1}$, it follows that

$$\kappa^{-1}(X) = \frac{\inf(X)}{\|X\|}.$$

Thus κ^{-1} is just $\inf(X)$ scaled by the norm of X. This means that the condition number is always greater than one (cf. (2.7) and (3.1)), and it does not change when X is multiplied by a nonzero constant.

In terms of the condition number, the Eckart-Young-Mirsky theorem reads as follows.

The smallest matrix E for which $X + E$ is collinear satisfies

$$\|E\|/\|X\| = \kappa^{-1}(X). \tag{3.3}$$

In other words κ^{-1} gives a lower bound on the *relative* distance to collinearity.

We shall give assessments of $\inf(X)$ and $\kappa(X)$ a little later. But first let us use the Eckart-Young-Mirsky result to dispose of the unhappy notion that $\det(A)$ bears a close relation to near collinearity. The rationale is that when X is exactly collinear, A is singular and $\det(A) = 0$. Consequently, a small value of $\det(A)$ ought to indicate near collinearity.

One difficulty in working with the determinant is its excessive sensitivity to scaling. This may be seen from the relation $\det(\alpha A) = \alpha^p \det(A)$, which implies that a 10-fold variation in the size of a 10×10 matrix A makes a 10 G-fold variation in the size of $\det(A)$. Anyone rash enough to make judgments about the size of such a sensitive number must expect difficulties.

Even when X has been scaled so that its columns have length unity, the determinant is unreliable. For example, consider any matrix X whose R factor in the QR factorization (2.3) has the form illustrated below for $p = 5$:

$$R_5 = \begin{pmatrix} 1 & 1/\sqrt{2} & 1/\sqrt{3} & 1/\sqrt{4} & 1/\sqrt{5} \\ 0 & 1/\sqrt{2} & 1/\sqrt{3} & 1/\sqrt{4} & 1/\sqrt{5} \\ 0 & 0 & 1/\sqrt{3} & 1/\sqrt{4} & 1/\sqrt{5} \\ 0 & 0 & 0 & 1/\sqrt{4} & 1/\sqrt{5} \\ 0 & 0 & 0 & 0 & 1/\sqrt{5} \end{pmatrix}.$$

(We leave it as an exercise to construct a seemingly uncontrived, centered regression matrix X with this property.) If A_p denotes the corresponding cross-product matrix, then from (2.4) $\det(A_p) = \det^2(R_p) = 1/p!$. Thus the determinant of A_p decreases factorially with p. However, it is easily verified that

$$R_5^{-1} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & \sqrt{2} & -\sqrt{2} & 0 & 0 \\ 0 & 0 & \sqrt{3} & -\sqrt{3} & 0 \\ 0 & 0 & 0 & \sqrt{4} & -\sqrt{4} \\ 0 & 0 & 0 & 0 & \sqrt{5} \end{pmatrix},$$

so that $\|R_p^{-1}\| \le p$ and $\inf(X) = \|R_p^{-1}\|^{-1} \ge p^{-1}$. Hence with increasing p the regression matrix suffers a gentle descent into collinearity, but not at the exaggerated rate suggested by the determinant.

Returning now to $\inf(X)$ and $\kappa(X)$, we note that these numbers have the virtue of simplicity. The condition number, in particular, carries its own scale with it. Thus, if the columns of X are roughly equal in size and $\kappa(X) = 10^5$, then we can attain collinearity by perturbing the elements of X in their fifth digits.

Moreover, a body of useful perturbation theory has been cast in terms of the condition number. For example, let $\hat{b} = X^\dagger y$ be the estimated vector of regression coefficients and let $\tilde{b} = \tilde{X}^\dagger y$ be the estimated regression coefficients for the perturbed regression matrix $\tilde{X} = X + E$. Then

$$\begin{aligned} \frac{\|\tilde{b} - \hat{b}\|}{\|\hat{b}\|} &\le \kappa(X)\frac{\|E\|}{\|X\|} \\ &\quad + \kappa^2(X)\frac{\|E\|}{\|X\|}\frac{\|\hat{e}\|}{\|X\|\,\|b\|} \\ &\quad + O(\|E\|^2), \end{aligned} \tag{3.4}$$

where $\hat{e} = y - X\hat{b}$ is the residual vector (see Stewart (1977) for this and other related inequalities). Thus the condition number can be used to predict the effects of errors in the regression variables on the regression coefficients.

However, the condition number has its defects. The statistician who attempts to use a bound like (3.4) will find that it is disappointingly pessimistic. The reason is that the bound is derived by repeated applications of the triangular and submultiplicative inequalities for matrix norms, and each application represents another backing off from sharpness. Numerical analysts are not overly concerned with this because their errors originate from rounding on a digital computer and are very small (see for example Wilkinson, 1963). However, the statistician must deal with measurement errors or errors made in recording data to a small number of figures, and here the lack of sharpness hurts.

Moreover, the condition number has its own scaling problems. For if we partition

$$X = (X_* x_p) \tag{3.5}$$

and write $X_\alpha = (X_* \alpha x_p)$, where α approaches zero,

then $\lim_{\alpha\to 0} \| X_\alpha \| = \| X_* \|$ and $X_\alpha^\dagger \cong \alpha^{-1} \| x_p^{(\dagger)} \|$. It follows that

$$\kappa(X_\alpha) \cong \alpha^{-1} \| X_* \| \| x_p^{(\dagger)} \| \to \infty.$$

Thus by scaling down any column of X, the condition number can be made arbitrarily large. This situation is known as *artificial ill-conditioning.*

The remedy for this problem is to adopt a standard scaling of the columns before computing the condition number; but what the standard should be is by no means clear. Belsley, Kuh, and Welsch (1980, pages 183–185) argue that the columns of X should all be scaled so that they are of equal length on the grounds that this scaling approximately minimizes the condition number, a result due to van der Sluis (1969). Moreover, if the quantity $\| X \|$ in (3.4) is to truly represent the size of the matrix X, all the columns x_j should be represented in equal measure—something that equal column scaling achieves.

However, this last heuristic argument cuts several ways. For example, if $\| E \|$ in (3.4) is to truly represent the size of E, then X should be scaled so that the columns of E have equal norms (additional arguments for this kind of scaling have been given by the author (Stewart, 1984), and it is recommended by the authors of LINPACK (Dongarra et al., 1979)). Furthermore, although the righthand side of (3.4) gives us precise information about the accuracy of the larger components of $\tilde{b}$, it is less precise about the smaller ones (consider, for example, the meaning of the inequality $\| \tilde{b} - \hat{b} \| / \| \hat{b} \| \leq 10^{-4}$ when $\hat{b} = (1\ 10^{-1}\ 10^{-2}\ 10^{-3}\ 10^{-4})^{\mathrm{T}}$). Thus the problem should also be scaled so that the components of $\hat{b}$ are roughly the same size. Needless to say, none of these three scalings have to be compatible.

To summarize, although the condition number is a useful indicator of collinearity, it is too crude for statistical applications. This is because it uses matrix norms to distill a large amount of information into a single number. What is needed is a *set* of numbers that can probe the effects of collinearities more delicately. Fortunately, two closely related sets of such numbers have been around for a long time, under the names of variance inflation factors and multiple correlation coefficients.

Our first order of business is to connect these numbers with the distance $\inf(X)$ of X collinearity. Since the definition of variance inflation factor presupposes that the cross-product matrix is a correlation matrix, for now we will assume that the matrix X has been centered and scaled so that $X^{\mathrm{T}}X$ is a correlation matrix.

Let us first suppose that X is nearly collinear in the sense that $\inf(X)$ is small. Let v be the vector for which the minimum in (3.1) is attained and write $v = (v_1 v_2 \cdots v_p)^{\mathrm{T}}$. Suppose without loss of generality that the largest component of v is v_p. Since $\| v \| = 1$, we must have $| v_p | \geq 1/\sqrt{p}$. If we set $\mu_j = v_j / v_p$, then

$$\| x_p - \mu_1 x_1 - \mu_2 x_2 - \cdots - \mu_{p-1} x_{p-1} \| = | v_p^{-1} | \inf(X) \leq \sqrt{p} \inf(X). \tag{3.6}$$

Now let $\hat{\mu}_1, \hat{\mu}_2, \cdots, \hat{\mu}_{p-1}$ be the regression coefficients obtained by fitting $x_1, x_2, \cdots, x_{p-1}$ to x_p, and let

$$\kappa_p^{-1} \overset{\text{def}}{=} \| x_p - \hat{\mu}_1 x_1 - \hat{\mu}_2 x_2 - \cdots - \hat{\mu}_{p-1} x_{p-1} \|. \tag{3.7}$$

By its very definition, κ_p^{-1} must be smaller than the righthand side of (3.6). Consequently we have the relation

$$\kappa_p^{-1} \leq \sqrt{p} \inf(X),$$

so that a near collinearity in X must make itself felt by at least one of the numbers $\kappa_1^{-1}, \kappa_2^{-1}, \cdots, \kappa_p^{-1}$ being small (here the other numbers κ_j^{-1} $(j < p)$ are defined in analogy with κ_p^{-1}).

Before we relate the numbers κ_j^{-1} to variance inflation factors and multiple correlation coefficients, let us derive the reciprocal relation between these numbers and near collinearity. Suppose that of all the κ_j^{-1}, the number κ_p^{-1} is the smallest. If we define v by

$$v^{\mathrm{T}} = \frac{(\hat{\mu}_1 \hat{\mu}_2 \cdots \hat{\mu}_{p-1} - 1)}{\sqrt{\hat{\mu}_1^2 + \hat{\mu}_2^2 + \cdots + \hat{\mu}_{p-1}^2 + 1}},$$

where the $\hat{\mu}_j$ are the coefficients appearing in (3.7), then $\| v \| = 1$, and it follows that

$$\inf(X) \leq \frac{\kappa_p^{-1}}{\sqrt{\hat{\mu}_1^2 + \hat{\mu}_2^2 + \cdots + \hat{\mu}_{p-1}^2 + 1}} \leq \kappa_p^{-1},$$

so that the smallest of the κ_j^{-1} is a bound on the distance to collinearity.

To establish the connection between κ_p^{-1} and other well known quantities, we shall first show that

$$\kappa_p^{-1} = \rho_{pp}, \tag{3.8}$$

where ρ_{pp} is the (p, p)-element of the R factor in (2.3). Let X be partitioned as in (3.5) and let $\hat{m} = (\hat{\mu}_1, \hat{\mu}_2, \cdots, \hat{\mu}_{p-1})^{\mathrm{T}}$. Then $\hat{m}$ is the unique vector satisfying

$$\| x_p - X_* \hat{m} \| = \min.$$

Since the norm is unchanged by orthogonal transformations, we may multiply $x_p - X_* \hat{m}$ by the orthogonal matrix Q from the QR decomposition to get

$$\| x_p - X_* \hat{m} \| = \| Q^{\mathrm{T}}(x_p - X_* \hat{m}) \| = \left\| \begin{pmatrix} r_{*p} \\ \rho_{pp} \end{pmatrix} - \begin{pmatrix} R_{**} \\ 0 \end{pmatrix} \hat{m} \right\|,$$

where we have partitioned

$$R = \begin{pmatrix} R_{**} & r_{*p} \\ 0 & \rho_{pp} \end{pmatrix}. \tag{3.9}$$

This norm is clearly minimized when $\hat{m} = R_{**}^{-1} r_p$, and its minimum value, which by definition is κ_p^{-1}, is ρ_{pp}. This establishes (3.8).

It now follows from (2.6) that κ_p^2 is the (p, p)-element of the inverse cross-product matrix A^{-1}. But since A is a correlation matrix, this is just the pth variance inflation factor, so called because the variance of the pth estimated regression coefficient is $a_{pp}^{(-1)}\sigma^2$ (Marquardt, 1970). Since reordering the columns of X simply reorders the diagonals of A^{-1}, it follows by interchanging the last column of X with the jth that

$$\kappa_j^2 = \alpha_{jj}^{(-1)} \quad (j = 1, 2, \cdots, p).$$

(Note, incidentally, that since we have assumed that $\| x_p \| = 1$, by (3.7) we must have $\kappa_p \geq 1$. This is an independent verification of the well known fact that the variance inflation factors are greater than one and really do inflate variances.)

The multiple correlation of x_p with $x_1, x_2, \cdots, x_{p-1}$ is by definition the simple correlation of x_p with the predicted value $X_*\hat{m}$. There is a wealth of expressions for these numbers. In particular, from Seber (1977, (4.30)) it follows that if R_j denotes the multiple correlation of x_j with the other columns of X then

$$R_j = \sqrt{1 - \kappa_j^{-2}}.$$

Thus multiple correlations near one are associated with near collinearities. This has been noted in the literature, simply on the basis of the definition of multiple correlation. The above development provides a precise, quantitative connection.

In defining κ_j^{-1}, we have presupposed correlation scaling. However, numerical analysts, who are equally affected by collinearity, seldom bother with such scaling in solving least squares problems. It is therefore desirable to produce a definition that is independent of scaling. We will do this in the next section, where we will define our collinearity indices and derive their properties.

4. COLLINEARITY INDICES AND THEIR PROPERTIES

In this section, building on the results of Section 3, we will introduce numbers, called collinearity indices, which are scale invariant measures of collinearity, and then describe their properties. With one exception, the results in this section are rather easy to establish and will be left as exercises (a useful technique is to use the QR decomposition to establish the result for the pth index and then generalize).

Definition

The numbers κ_j of the last section were defined under the restrictive assumption that the regression matrix X was centered and scaled. To remove this restriction, we note that from (3.8) to multiply x_j by a constant is to divide κ_j by the same constant. Hence, if we augment our original definition by a factor of $\| x_j \|$, we will always obtain the number κ_j however the columns of X have been scaled. This leads to the following definition.

For $j = 1, 2, \cdots, p$ the jth collinearity index *is the number*

$$\kappa_j \overset{\text{def}}{=} \| x_j \| \, \| x_j^{(\dagger)} \|. \tag{4.1}$$

The analogy in definition and notation between κ_j and the condition number (3.2) is deliberate. Like the condition number, the collinearity indices are invariant under scaling; however, whereas the condition number is invariant only under multiplication of X by a constant, the collinearity indices are invariant under any column scaling.

Since our collinearity indices (or rather their squares) are already present in the statistics literature as variance inflation factors, the introduction of new nomenclature requires some justification. There are four reasons why a change is desirable.

First, we have already noted that the scale invariance of the definition (4.1) makes it useful to people, like numerical analysts, who seldom bother with scaling. The notation κ_j also emphasizes the link with the condition number, which is widely used by numerical analysts and not unknown to statisticians.

Second, as we noted in the introduction, variance inflation factors and multiple correlations were not introduced to analyze collinearity in regression models, and their names show it. The nomenclature adopted here is more to the point.

Third, collinearity coefficients vary linearly with the relative distance to exact collinearity, whereas variance inflation factors vary as the square. Not only are the collinearity coefficients more readily interpreted, but their use removes unsightly square roots from formulas.

Finally, the use of collinearity indices represents a commitment to cast results in terms of relative errors. To see the utility of this, compare the statement *we are safe if the components of x_j are accurate to three figures* with the statement *we are safe if the errors in the components of x_j are less than* 10. The former makes sense in itself; the latter is hard to interpret unless the size of x_j is known.

Elementary Properties

Here we shall recapitulate the results of Section 3.

- *If X has correlation scaling, then*

$$\kappa_j^2 = \alpha_{jj}^{(-1)};$$

i.e., the square of the jth collinearity coefficient is the jth variance inflation factor.

- *If R_j denotes the multiple correlation between x_j and the other columns of X, then*

$$R_j = \sqrt{1 - \kappa_j^{-2}}.$$

- *If ρ_{pp} is the (p, p) element of the R factor of X, then*

$$\kappa_p = \frac{\| x_p \|}{\rho_{pp}}. \tag{4.2}$$

- *The index κ_j is greater than or equal to one, with equality if and only if x_j is orthogonal to the other columns of X.*
- *If X has unit column scaling, then*

$$\max\{\kappa_j\} \leq \inf^{-1}(X) \leq \sqrt{p} \max\{\kappa_j\},$$

and

$$\max\{\kappa_j\} \leq \kappa(X) \leq p \max\{\kappa_j\}.$$

Another Relation with Collinearity

An unsatisfactory aspect about the last item above is that the columns of X are required to have norm unity. However, there is a more direct relation between collinearity and collinearity indices.

The smallest perturbation e_j in x_j that will make X exactly collinear satisfies

$$\frac{\| e_j \|}{\| x_j \|} = \kappa_j^{-1}. \tag{4.3}$$

This result was stated by the author without proof in 1984. A proof of a more general theorem may be found in Golub, Hoffman, and Stewart (1984). Note the analogy between (4.3) and (3.3). Here, as there, everything carries its own scale: collinearity can be attained by a *relative* perturbation of size κ_j^{-1}.

A Single Variable Is Not Collinear

Collinearity is a group phenomenon. A single column cannot alone be a source of collinearity, since it must be collinear with other columns. The equivalent statement for *near* collinearity is that if one collinearity index is large then another must also be large. The following inequality quantifies this statement (a proof will be found in the Appendix).

For $j = 1, 2, \cdots, p$

$$\max_{i \neq j} \kappa_i \geq \sqrt{1 + \frac{\kappa_j^2 - 1}{(p-1)^2}}. \tag{4.4}$$

This result has been included to discourage the naive use of condition coefficients in selecting a variable to be thrown out of an unsatisfactory model. The temptation here is to choose the variable with the largest condition coefficient. However, (4.4) says that where there is one large coefficient there will also be others. Something as important as selecting or rejecting a variable should have a sounder basis than minor variations in the magnitudes of large collinearity indices.

Effects of Centering

Although the collinearity indices are invariant under column scaling, they shrink when the regression matrix is centered; however, they do it in an interesting way. To see this, let X be partitioned in the form $X = (x_1 X_2)$. In our application, x_1 will be a column of all ones representing the constant term, but that is not necessary for what is to follow.

Now centering amounts to computing the matrix $\bar{X}_2 = PX_2$, where P is the projection onto the space orthogonal to x_1 (i.e., $P = I - \| x_1 \|^{-2} x_1 x_1^T$). Consequently, for any column $\bar{x}_j$ of $\bar{X}_2$, we have

$$\| \bar{x}_j \| \leq \| x_j \|, \tag{4.5}$$

which certainly has potential for decreasing the collinearity indexes.

Let us now partition

$$X^\dagger = \begin{pmatrix} x_1^{(\dagger)} \\ X_2^{(\dagger)} \end{pmatrix}.$$

Then it can be shown that $\bar{X}_2^\dagger = X_2^{(\dagger)}$, whence

$$\| \bar{x}_j^{(\dagger)} \| = \| x_j^{(\dagger)} \|. \tag{4.6}$$

It now follows from (4.5) and (4.6) that

$$\bar{\kappa}_j = \| \bar{x}_j \| \, \| \bar{x}_j^{(\dagger)} \| \leq \| x_j \| \, \| x_j^{(\dagger)} \| = \kappa_j.$$

Thus the collinearity index does indeed decrease under centering, but all of the decrease comes from the decrease in the norm of x_j.

This phenomenon is a consequence of the fact that our definition of collinearity index compels us to work with relative errors. Consider, for example, the size (4.3) of the smallest perturbation of column j that makes X exactly collinear. The absolute size of this perturbation is not affected by centering, but the relative error $\| e_j \| / \| x_j \|$ becomes larger when x_j is replaced by the smaller $\bar{x}_j$. To compensate for this, κ_j must become smaller, precisely in proportion as x_j

becomes smaller. We will meet with this phenomenon again in §6.

5. SIGNIFICANCE

If $\|x_j\| = 1$ then κ_j^2 is the jth variance inflation factor, and the standard deviation of the jth regression coefficient is

$$\sigma_j = \kappa_j \sigma.$$

Thus near collinearity, as it is made manifest by large collinearity indices, is associated with large variances in the regression coefficients. This undesirable aspect of near collinearity has frequently been noted in the literature.

However, it is not a simple matter to state precisely what we have to fear from large variances. Informally, the problem is that a large variance could swamp an important regression coefficient. But if we attempt to replace the vague term "important" with the mathematically precise term "statistically significant," we become involved in a paradox; for a statistically significant regression coefficient is almost by definition one that is substantially greater than its standard deviation. This suggests that if we wish to use collinearity indices to assess the ill effects of near collinearity on regression coefficients, we must introduce concepts from outside the classical model. The following definition does just this.

In the model (2.1), *the* importance *of the variable* x_j *is the number*

$$\iota_j = \frac{|\beta_j| \, \|x_j\|}{\|y\|}. \tag{5.1}$$

In the expression

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e$$

the term $\beta_j x_j$ represents $100\iota_j\%$ of the total observed response, and a small value of ι_j therefore means that the contribution of x_j is unimportant. The point at which a variable becomes important must be determined by the application and by the judgment of the analyst. Most people will undoubtedly feel that a variable whose importance is greater than 0.5 is not one to ignore.

Let us suppose that we have chosen levels of importance λ_j above which the x_j would be considered important. Then the model (2.1) must be considered unsatisfactory if variables with importance above these levels are in danger of being declared insignificant. Since the λ_j must be fixed by extrastatistical consideration, there is no need to be overly precise about levels of significance. We shall therefore say that the model is unsatisfactory if the estimated standard deviation of a regression coefficient is half the size of the smallest value of the coefficient that would make $\iota_j > \lambda_j$. From (5.1) it is seen that this smallest value is

$$\lambda_j \frac{\|y\|}{\|x_j\|}.$$

The estimated standard deviation of β_j is

$$\|x_j^{(\dagger)}\| \, \hat{\sigma},$$

where $\hat{\sigma}$ is the usual estimate of σ. We thus require that

$$2\|x_j^{(\dagger)}\| \, \hat{\sigma} \geq \lambda_j \frac{\|y\|}{\|x_j\|}.$$

When this inequality is recast in terms of collinearity indices, it yields the following regression diagnostic.

Having chosen levels of importance λ_j *for the variables* x_j, *reject the model if for any* j

$$\text{IMP}_j \overset{\text{def}}{=} 2\kappa_j \frac{\hat{\sigma}}{\|y\|} > \lambda_j. \tag{5.2}$$

The criterion (5.2) has two particularly nice properties. First, it is scale invariant, not only with respect to the scaling of the columns of X but also the scaling of y. Second, it does not depend on estimates of the regression coefficients; only on the estimate $\hat{\sigma}$. For $\hat{\sigma}$ to be a good estimate all that is required is that the response vector be a linear combination of the columns of X. This will be true even if the model is overspecified, which is one of the most common sources of near collinearity.

However, the diagnostic (5.2) is not invariant under centering. The difficulty here is not only with the collinearity indices but with the definition (5.1), which is also not invariant under centering.

Now there is a sense in which we should be surprised if the notion of importance were to be invariant under centering. Centering amounts to removing x_1, which is a column of ones, from the other variables. If x_1 were anything but a column of ones, we would feel that we had defined a new set of variables—combinations of x_1 with the remaining variable—and the importance of the new variables would be open to reassessment. It is only the simplicity of the centering operation that makes us take exception to the lack of invariance in (5.1).

For example, consider the data in Table 1, which were introduced by Belsley (1984a) in an interesting discussion of the effects of centering on collinearity diagnostics. These data do not look promising, since their leading figures agree to three places, and the collinearity indices for the uncentered problem

κ_1	κ_2	κ_3
632	447	447

TABLE 1
Belsley's example

x_1	x_2	x_3	y
1	.996926	1.000060	2.69385
1	.997091	.998779	2.69402
1	.997300	1.000680	2.70052
1	.997813	1.002420	2.68559
1	.997898	1.000650	2.70720
1	.998140	1.000500	2.69550
1	.998556	.999596	2.70417
1	.998737	1.002620	2.69699
1	.999414	1.003210	2.69327
1	.999678	1.001300	2.68999
1	.999926	.997579	2.70003
1	.999995	.998597	2.70200
1	1.000630	.995316	2.70938
1	1.000950	.995966	2.70094
1	1.001180	.997125	2.70536
1	1.001770	.998951	2.70754
1	1.002310	1.001020	2.69519
1	1.003060	1.001860	2.70170
1	1.003940	1.003530	2.70451
1	1.004690	1.000210	2.69532

are rather large. The norm of the uncentered y is $\|y\| = 12.1$ and the estimated standard deviation is $\hat{\sigma} = 0.00555$. Thus the diagnostics numbers (5.2) are

	IMP$_j$	
0.581	0.411	0.411

which should give one pause about the model.

Centering does not much affect these results. The collinearity indices for the centered problem are both one (the columns of the centered regression matrix are orthogonal to working accuracy). The norm of the centered y is 0.0275, and hence the diagnostics are both 0.4029. The reason that centering has so little effect is that x_2, x_3, and y have the same number of constant leading figures, so that the decrease in the collinearity coefficients is almost exactly balanced by the increase in the relative error $\hat{\sigma}/\|y\|$.

Things become more complicated when the number of constant leading figures is different in the variables and the response vector. To see this, consider the diagnostics for the uncentered regression matrix and the centered y, which are

	IMP$_j$	
255	180	180

The reason for these large diagnostic numbers is that the constant part in x_1, x_2, and x_3 inflates their importance in relation to the comparatively small centered y.

I feel that when there is a constant term in the model, the model should be centered before the importance of the remaining variables is assessed and the test (5.2) applied. In order for centering to have a gross effect on the diagnostic, some variable x_j must have a large constant part, and in fact the larger the constant part the more "important" the variable becomes. Now a large constant part is usually an artifice of the way the data are collected, especially in the sciences where it is not uncommon to make very precise measurements over a narrow range. In these cases it is appropriate to regard the "importance" of such a variable as equally artificial. Otherwise put, the real variable is masked by the large constant part. Centering simply shows the variable for what it is.

6. ERRORS IN REGRESSION VARIABLES

In this section we shall use the collinearity indices to assess the effects of errors in the regression variables. This is a large subject, with a voluminous literature (Seber, 1977, pages 155–162, and Anderson, 1984, for surveys and further references), and it is important that we place the material in this section in context.

Approaches to errors in regression variables may be roughly divided into two classes. The first approach attempts to extract useful information from the regression model in spite of the presence of errors in the variables. Invariably, some precise information about the structure of the error is needed; for example, one may be required to furnish ratios of the variances of the errors in each column.

The second approach, to which our development belongs, attempts to determine when the errors are so small that they can be ignored or tolerated. Again information about the errors is required; but compared with the first approach it can be relatively imprecise—say the orders of magnitude of the variances of the errors. In many treatments (e.g., Davis and Hutton, 1975, or Beaton, Rubin, and Barone, 1976) the focus is on the well known asymptotic inconsistency of the estimates. When n is large, the errors tend to bias the estimates in a fixed way, which can be approximated from a rough knowledge of the size of the errors. The resulting diagnostic is to reject the model when the bias is unacceptably large.

The principle difficulty with the asymptotic approach is that the limits are attained so slowly (as $1/\sqrt{n}$) that it is hard to know what to make of the diagnostics when n is small. Accordingly, we shall attempt to ascertain the bias in regression coefficients due to errors in the variables for fixed n. However, in its full generality this problem is analytically intractable, at least in the sense of yielding realistic results. It is easy to use norms to get suggestive inequalities like (3.4), but they are too crude for practical work. Consequently, we will analyze the case where the model is exact ($y = Xb$) and only a single column is in error. In spite of the special nature of this case, it

gives a great deal of insight into the way errors in regression variables cause bias and other untoward effects in regression coefficients.

As we did in Section 3 we shall derive our results for the pth collinearity index in such a way that the generalization to the other indices is obvious. Although the following exposition is not technically very demanding, it is detailed, and to aid the reader we have broken it into short subsections.

The Problem

Let us suppose that we have the exact linear relation

$$y = Xb; \tag{6.1}$$

i.e., e is zero in (2.1). Let X be partitioned as (3.5) and let an error vector e_p be given. Define the perturbed set of regression $\tilde{b}$ as the solution of the least squares problem of minimizing

$$\left\| y - (X_* \; x_p + e_p)\begin{pmatrix}\tilde{b}_* \\ \tilde{\beta}_p\end{pmatrix} \right\|. \tag{6.2}$$

We wish to determine how $\tilde{b}$ compares with b.

Formulas from the QR Decomposition

The solution of the above problem is best cast in the form of the QR decomposition (2.2). Let Q be partitioned in the form

$$Q = (Q_* \; q_p \; Q_\perp).$$

Set

$$\begin{pmatrix}Q_*^{\mathrm{T}} \\ q_p^{\mathrm{T}} \\ Q_\perp^{\mathrm{T}}\end{pmatrix}(X_* \; x_p) = \begin{pmatrix}R_{**} & r_{*p} \\ 0 & \rho_{pp} \\ 0 & 0\end{pmatrix},$$

and

$$\begin{pmatrix}Q_*^{\mathrm{T}} \\ q_p^{\mathrm{T}} \\ Q_\perp^{\mathrm{T}}\end{pmatrix}y = \begin{pmatrix}z_* \\ \zeta_p \\ 0\end{pmatrix}.$$

If (6.1) is multiplied by Q^{T}, then

$$\begin{pmatrix}R_{**} & r_{*p} \\ 0 & \rho_{pp} \\ 0 & 0\end{pmatrix}\begin{pmatrix}b_* \\ \beta_p\end{pmatrix} = \begin{pmatrix}z_* \\ \zeta_p \\ 0\end{pmatrix},$$

from which we get the following formulas for the components of b:

$$\beta_p = \frac{\zeta_p}{\rho_{pp}} \tag{6.3}$$

and

$$b_* = R_{**}^{-1}z_* - \beta_p R_{**}^{-1}r_{*p}. \tag{6.4}$$

To derive corresponding formulas for the components of $\tilde{b}$, note that the norm of the vector in (6.2) is not changed when it is multiplied by Q. Hence if we set

$$\begin{pmatrix}Q_*^{\mathrm{T}} \\ q_p^{\mathrm{T}} \\ Q_\perp^{\mathrm{T}}\end{pmatrix}e_p = \begin{pmatrix}g_* \\ \gamma_p \\ h_p\end{pmatrix},$$

then $\tilde{b}$ is determined by minimizing

$$\left\| \begin{pmatrix}z_* \\ \zeta_p \\ 0\end{pmatrix} - \begin{pmatrix}R_{**} & r_{*p} + g_* \\ 0 & \rho_{pp} + \gamma_p \\ 0 & h_p\end{pmatrix}\begin{pmatrix}\tilde{b}_* \\ \tilde{\beta}_p\end{pmatrix} \right\|.$$

The solution of this problem is easily seen to be

$$\tilde{\beta}_p = \frac{(\rho_{pp} + \gamma_p)\zeta_p}{(\rho_{pp} + \gamma_p)^2 + h_p^{\mathrm{T}}h_p} \tag{6.5}$$

and

$$\tilde{b}_* = R_{**}^{-1}z_* - \tilde{\beta}_p R_{**}^{-1}(r_{*p} + g_*). \tag{6.6}$$

We are going to determine the effects of the error vector e on β_p by comparing (6.3) with (6.5), but first we must pause to consider the error itself.

The Error

As we indicated in the introduction to this section, to determine the effects of errors in regression variables, the analyst must supply independent information about the errors. To fix on something definite, we shall assume that the components of e_p are uncorrelated random variables with mean μ_p and variance σ_p^2. The analyst is then expected to provide a rough estimate of

$$\varepsilon_p = \sqrt{\mu_p^2 + \sigma_p^2}. \tag{6.7}$$

For later use, we will require an estimate of $h_p^{\mathrm{T}}h_p = \| Q_\perp^{\mathrm{T}}e_p \|^2$. It is easily verified that

$$\mathrm{E}(\| Q_\perp^{\mathrm{T}}e_p \|^2) = \mu_p^2 \| Q_\perp^{\mathrm{T}}\mathbf{1} \|^2 + (n - p)\sigma_p^2,$$

where $\mathbf{1}$ denotes the vector of ones. Since $Q_\perp$ has orthonormal columns, unless $\mathbf{1}$ bears some special relation to $Q_\perp$, the elements of $Q_\perp^{\mathrm{T}}\mathbf{1}$ should all be about unity in magnitude, and hence

$$\| Q_\perp^{\mathrm{T}}\mathbf{1} \|^2 \doteq (n - p). \tag{6.8}$$

Thus if (6.8) is valid, we may approximate

$$h_p^{\mathrm{T}}h_p \doteq (n - p)\varepsilon_p^2. \tag{6.9}$$

However, there is one important case where (6.8) does not hold: when the model has a constant term. For in this case the columns of $Q_\perp$ are orthogonal to $\mathbf{1}$, and hence $Q_\perp^{\mathrm{T}}\mathbf{1} = 0$. In this case we should replace ε_p in (6.9) with σ_p, or what is equivalent take $\mu_p = 0$ in (6.7).

In the same way we will approximate the size of γ_p by

$$|\gamma_p| \doteq \varepsilon_p, \tag{6.10}$$

where we take $\mu_p = 0$ in (6.7) if the model has a constant term.

We shall use (6.9) and (6.10) freely in deriving our results. However, since they are merely rough approximations, it is important to distinguish which results depend on them. We shall signal this by placing a dot over any relation involving them.

Bias in $\tilde{\beta}_p$

Comparing (6.3) and (6.5), we see that the error e_p affects the coefficient β_p through the numbers $h_p^T h_p$ and γ_p. It will be convenient to look at each of these effects separately, beginning with the former.

If we set $\gamma_p = 0$ in (6.5), we get

$$\tilde{\beta}_p = \frac{\rho_{pp}\zeta_p}{\rho_{pp}^2 + h_p^T h_p}.$$

Since

$$\frac{\rho_{pp}}{\rho_{pp}^2 + h_p^T h_p} \leq \frac{1}{\rho_{pp}},$$

we see that $\tilde{\beta}_p \leq \beta_p$; that is, the effect of $h_p^T h_p$ is to bias β_p downward. Let us agree to measure this bias by the relative error

$$\mathrm{RE}_{\mathrm{bias}} = \frac{\beta_p - \tilde{\beta}_p}{\beta_p} = \frac{h_p^T h_p}{\rho_{pp}^2 + h_p^T h_p}.$$

If we use the approximation (6.9) for $h_p^T h_p$, we get

$$\mathrm{RE}_{\mathrm{bias}} \doteq \frac{(n-p)\varepsilon_p^2}{\rho_{pp}^2 + (n-p)\varepsilon_p^2}.$$

Finally, if we set

$$\tau_p^2 = (n-p)\frac{\varepsilon_p^2}{\rho_{pp}^2} = (n-p)\kappa_p^2 \frac{\varepsilon_p^2}{\|x_p\|^2}, \tag{6.11}$$

then

$$\mathrm{RE}_{\mathrm{bias}} \doteq \frac{\tau_p^2}{1 + \tau_p^2}. \tag{6.12}$$

The right-hand side of (6.12) provides an approximation to the relative error due to h_p. When τ_p^2 is small, it is essentially the square of the product of a relative error $\sqrt{n-p}\varepsilon_p/\|x_p\|$ in x with the pth collinearity index. Thus the collinearity index serves as a factor showing how relative errors in the columns of X are amplified in the bias in the regression coefficient.

The Effects of γ_p

Turning now to the effect of γ_p, let us set $h_p = 0$ in (6.5) to get

$$\tilde{\beta}_p = \frac{\zeta_p}{\rho_{pp} + \gamma_p}. \tag{6.13}$$

If we define the relative error in β_p due to γ_p by

$$\mathrm{RE}_{\mathrm{lin}} = \left|\frac{\tilde{\beta}_p - \beta_p}{\tilde{\beta}_p}\right|,$$

then it is easily seen from (6.3) and (6.13) that

$$\mathrm{RE}_{\mathrm{lin}} = \frac{|\gamma_p|}{\rho_{pp}}.$$

Finally, if we use the approximation (6.10), then

$$\mathrm{RE}_{\mathrm{lin}} \doteq \frac{\varepsilon_p}{\rho_{pp}} = \kappa_p \frac{\varepsilon_p}{\|x_p\|}. \tag{6.14}$$

In this way we can approximate the relative error due to γ.

Relation between $\mathrm{RE}_{\mathrm{bias}}$ and $\mathrm{RE}_{\mathrm{lin}}$

From (6.12) and (6.14) it follows that

$$\mathrm{RE}_{\mathrm{bias}} \doteq \frac{(n-p)\mathrm{RE}_{\mathrm{lin}}^2}{1 + (n-p)\mathrm{RE}_{\mathrm{lin}}^2}. \tag{6.15}$$

Since $\mathrm{RE}_{\mathrm{lin}}$ depends linearly on ε_p (hence the subscript lin), we see that for fixed n as the error decreases, $\mathrm{RE}_{\mathrm{bias}}$ decreases quadratically and must ultimately become unimportant compared to $\mathrm{RE}_{\mathrm{lin}}$.

On the other hand, if we invert (6.15) then

$$\mathrm{RE}_{\mathrm{lin}} \doteq \sqrt{\frac{\mathrm{RE}_{\mathrm{bias}}}{(n-p)(1 - \mathrm{RE}_{\mathrm{bias}})}}.$$

Hence if with increasing n the number $\mathrm{RE}_{\mathrm{bias}}$ approaches a constant (as in the models of Davies and Hutton (1975) and Beaton, Rubin, and Barone (1976)), then $\mathrm{RE}_{\mathrm{lin}}$ ultimately becomes unimportant.

Stability of the Collinearity Index

One of the major problems with diagnostics for the effects of errors in the variables is that they must be computed from the perturbed regression matrix and are therefore contaminated by the very errors whose effects they are supposed to diagnose. We must therefore insure that this contamination is not great enough to invalidate the diagnostic. Although a formal analysis is possible, here we shall present an informal analysis based on our simplified model, which has the advantage that it shows clearly how the diagnostic is affected.

The number τ_p on which our approximation to $\mathrm{RE}_{\mathrm{bias}}$ depends must be calculated from ρ_{pp}. In practice we will be unable to compute ρ_{pp} directly, instead we must compute an approximation $\tilde{\rho}_{pp}$ from the perturbed regression matrix. If we ignore the effect of γ_p, we get

$$\tilde{\rho}_{pp}^2 = \rho_{pp}^2 + h_p^{\mathrm{T}} h_p,$$

and, estimating $h_p^{\mathrm{T}} h_p$ by (6.9),

$$\tilde{\rho}_{pp}^2 \doteq \rho_{pp}^2 + (n-p)\varepsilon_p^2. \tag{6.16}$$

Thus the number we actually compute is

$$\tilde{\tau}_p^2 \doteq (n-p)\frac{\varepsilon_p^2}{\tilde{\rho}_{pp}^2} \doteq (n-p)\frac{\varepsilon_p^2}{\rho_{pp}^2 + (n-p)\varepsilon_p^2}.$$

Dividing the numerator and denominator of the right hand side of this expresssion by ρ_{pp} we get

$$\tilde{\tau}_p^2 \doteq \frac{\tau_p^2}{1+\tau_p^2}. \tag{6.17}$$

Although the relation (6.17) is only approximate, it suggests that errors in x_p tend to depress the diagnostic, rendering any diagnostic based on (6.12) optimistic. However, if we invert the relation and write

$$\tau_p^2 \doteq \frac{\tilde{\tau}_p^2}{1-\tilde{\tau}_p^2},$$

we see that if $\tilde{\tau}_p^2 \leq 0.1$, then $\tau_p^2 \leq 0.112$ and the effect of the error on the diagnostic is small. On the other hand, if $\tilde{\tau}_p^2 \leq 0.5$, then τ_p^2 can be as large as one, and the diagnostic can be quite misleading. For this reason we recommend that $\tilde{\tau}_p^2$ not be used in a diagnostic unless it is less than 0.1.

Similarly, we see from (6.16) that the relative error in ρ_{pp} is

$$\left| \frac{\tilde{\rho}_{pp}^2 - \rho_{pp}^2}{\rho_{pp}^2} \right| \doteq \tau_p^2.$$

Consequently if $\tilde{\tau}_p^2 \leq 0.1$, we can use $\tilde{\rho}_{pp}$ in computing $\mathrm{RE}_{\mathrm{lin}}$ in (6.14).

A Diagnostic Procedure

In this subsection we will collect the results of the preceding subsections into a diagnostic procedure.

Let the errors in the jth column of X have mean μ_j and variance σ_j^2, and set

$$\varepsilon_j = \begin{cases} \sqrt{\mu_j^2 + \sigma_j^2} & \text{if there is no constant term,} \\ \sigma_j & \text{otherwise.} \end{cases}$$

Fix thresholds $\lambda_{\mathrm{bias}} \leq 0.1$ and $\lambda_{\mathrm{lin}} \leq 0.32$ for the relative errors $\mathrm{RE}_{\mathrm{bias}}$ and $\mathrm{RE}_{\mathrm{lin}}$. Letting $\tilde{\kappa}_j$ denote the jth collinearity coefficient computed from the perturbed regression matrix, set

$$\tau_j = (n-p)\tilde{\kappa}_j^2 \frac{\varepsilon_j^2}{\| x_p \|^2},$$

and reject the model if for any j

$$\mathrm{RE}_{\mathrm{bias},j} = \frac{\tau_j^2}{1+\tau_j^2} > \lambda_{\mathrm{bias}} \tag{6.18}$$

or

$$\mathrm{RE}_{\mathrm{lin},j} = \frac{\tau_j}{\sqrt{n-1}} = \tilde{\kappa}_j \frac{\varepsilon_j}{\| x_j \|} > \lambda_{\mathrm{lin}}. \tag{6.19}$$

Comments on the Diagnostic

In this subsection we will offer some observations on the diagnostic procedure just proposed.

Limitations. Since the diagnostics are based on the analysis of a simplified model, they cannot give absolute security. The analysis is straightforward enough that any problem failing the diagnostic is surely suspect (the trouble with a norm-based analysis of the general problem is it yields diagnostics that reject tractable models). However, a model that passes the diagnostic is not home free. For example, we have not considered the effects of e_p on coefficients other than β_p, although one can make an *a posteriori* assessment from (6.4) and (6.6). Again, we have not considered the effects of the error e in the observations. Although there is reason for believing that if $\mathrm{RE}_{\mathrm{bias}}$ is small enough, errors from these sources will be tolerable,[1] this must be left for future analyses or experiments to decide.

Invariance. It is obvious that scaling a column of a regression matrix will not affect the influence of errors in that column, since the error is scaled along with the column. Further reflection will show that centering also makes no difference, except to remove bias from the error. This corresponds to the fact that the diagnostic inequalities (6.18) and (6.19) are invariant under both scaling and centering.

The Diagnostics in Terms of τ_j. The numbers τ_j represent a combination of collinearity indices and the errors in the regression variables, and with a little practice can be interpreted by themselves. Table 2 is given as an aid. First for selected values of $\mathrm{RE}_{\mathrm{bias}}$ it gives values of τ such that $\tau^2/(1+\tau^2) = \mathrm{RE}_{\mathrm{bias}}$. Then for each value $\mathrm{RE}_{\mathrm{bias}}$ and for selected values of $\mathrm{RE}_{\mathrm{lin}}$, it gives the smallest values of $n-p$ such that $\tau/\sqrt{n-p} \leq \mathrm{RE}_{\mathrm{lin}}$.

Since the bias behaves quadratically with the error, we see that a small decrease in τ causes a big decrease

[1] For example, the inclusion of e in the model adds a term of the form $h_p^{\mathrm{T}} h$, where $h = Q_1^{\mathrm{T}} e$, to the numerator of (6.5). Not only does this term have mean zero, but its influence wanes as h_p decreases.

TABLE 2
Critical values of τ and $n - p$

RE_{bias}	τ	$n - p$ for values of RE_{lin}				
		0.1	0.05	0.01	0.005	0.001
0.1	0.34	12				
0.05	0.23	6	22			
0.01	0.10	2	5	102		
0.005	0.07	1	3	51	202	
0.001	0.03	1	1	11	41	1011

in RE_{bias}. On the other hand, once a level of bias has been chosen, $n - p$ must be very large (roughly RE_{bias}^{-1}) to bring RE_{lin} down to the same level. However, if the errors have mean zero or the model has a constant term, then it is not as critical to have a small value of RE_{lin}. To see this note that if $RE_{lin} \leq 0.1$, $\tilde{\beta}_p$ in (6.13) is well approximated by $\beta_p(1 + \gamma_p/\rho_{pp})$. Thus the contribution of γ_p is to add to β_p an approximately unbiased term whose standard deviation is less than $|\beta_p| RE_{lin}$; i.e., less one-tenth the size of β_p. In other words, the linear errors, provided they are not too large, increase the variability of the regression coefficients, but do not bias them. However, if RE_{lin} is too large, then it will introduce biases of its own, owing to the singularity of $(1 + \gamma_p/\rho_{pp})^{-1}$ at $\gamma_p/\rho_{pp} = 1$. For this reason we have recommended that $\lambda_{lin} \leq 0.32$ in our diagnostic.

Assumptions about the Errors. For convenience we have introduced stochastic assumptions about the error; however, they should not be taken too seriously. The main use of randomness was to justify the approximations (6.9) and (6.10). But clearly all that is required is that if e_p is written $e_p = \mu_p \mathbf{1} + w_p$, where $\mathbf{1}^T w_p = 0$, then the component of $Q^T w_p$ are all about σ_p in magnitude.

Our assumptions fail for polynomial regression. For example, suppose the rows of X are given by $(1\ \xi_i\ \xi_i^2)$ and we observe perturbed values $\xi_i + \eta_i$. Even if the η_i meet our assumptions, the errors in the third column will be $\xi_i\eta_i + \eta_i^2$, and their sizes will vary with ξ_i. Obviously, this and other models in which the columns are not independent will require a separate analysis.

Relation to Other Measures. Hodges and Moore (1972) develop a general expression for the expected bias that is closely related to ours. They assume that X is perturbed by a matrix E whose elements are independent with mean zero, the elements in e_j having variance σ_j^2. By expanding $\tilde{b} = (X + e)^{\dagger} y$ in a series and taking expectations in the second order terms, they obtain

$$E(\tilde{b}) \cong b - (n - m - 1)A^{-1}\Sigma b,$$

where $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2)$. The fact that their estimate of bias essentially reduces to ours when only one σ_j is nonzero, tends to confirm the validity of both estimates. However, it is not clear when we can use the perturbed values of A and b in their estimate.

Under the assumption that X is perturbed by a constant matrix U, Swindel and Bower (1972) derive the bound

$$\frac{\sqrt{b^T U^T U b}}{\sigma}$$

for the relative bias in an arbitrary linear combination of the components of b. However, here the bias is taken relative to the variance of the linear combination. Moreover, since U is regarded as fixed, the bias includes linear terms, which, as we observed above, contribute more to the variability of the regression coefficients than to their bias. Thus the measure is best suited to problems where the same regression matrix is to be used to analyze several different responses. Davies and Hutton (1975) discuss when $\hat{b}$ and $\hat{\sigma}$ can be used in place of b and σ.

Both Davies and Hutton (1975) and Beaton, Rubin, and Barone (1976) introduce diagnostics that amount to

$$\sum_{j=1}^{p} \tau_j^2, \tag{6.20}$$

i.e., the sum of our diagnostics for RE_{bias} when the τ_j are small (cf. (6.11)). Thus the diagnostics are closely related, with ours being less conservative. However, it must be kept in mind that (6.20) was derived to measure asymptotic (large n) bias, whereas ours measures the bias for fixed n.

An Example

Let us return to the example of Table 1 and assume that the only error is in the rounding of the data. Unfortunately, this makes the errors fall into two categories: numbers less than one will have errors of order 10^{-6} while those greater than one will have errors of order 10^{-5}. We will take a worst case approximation and set $\varepsilon_j = 10^{-5}$. With this, the diagnostics to be compared with RE_{bias} are both $0.17 \cdot 10^{-4}$, and those to be compared with RE_{lin} are 10^{-3}. Thus at this level of error, the problem is quite insensitive.

On the other hand if we set $\varepsilon_j = 0.01$, which corresponds to an example in Belsley (1984a), we get diagnostics 0.94 and 1.0. Thus, at this level of error the problem is completely intractable. Note that our model predicts a strong downward bias in the regression coefficients, which is exactly what Belsley observes.

Another Example

Belsley's example was concocted to illustrate some of the issues surrounding centering and is patently

TABLE 3
Woods, Steinour, and Starke problem

x_1	x_2	x_3	x_4	x_5	y
7	26	6	60	2.5	78.5
1	29	15	52	2.3	74.3
11	56	8	20	5.0	104.3
11	31	8	47	2.4	87.6
7	52	6	33	2.4	95.9
11	55	9	22	2.4	109.2
3	71	17	6	2.1	102.7
1	31	22	44	2.2	72.5
2	54	18	22	2.3	93.1
21	47	4	26	2.5	115.9
1	40	23	34	2.2	83.8
11	66	9	12	2.6	113.3
10	68	8	12	2.4	109.4

artificial. The data in Table 3 concerns the heat generated by cement during curing and were collected by Woods, Steinour, and Starke (1932). The independent variables are components of the cement, measured as percent of the whole, and the dependent variable is the heat generated. The data are actually one of a sequence of data sets taken at different times in the curing process, and the originators fit a linear model to each set. Daniel and Wood (1980, Ch. 9) (where the data were taken from) give a masterful treatment of the entire set of data using nonlinear least squares.

We first need to assess the size of the errors in the variables. Since the details of the experimental setup are not available, we shall assume that the data are accurate to all reported figures, subject only to rounding error. Following Beaton, Rubin, and Barone (1976), we will model as a uniformly distributed error over $[-0.5, 0.5] \cdot 10^t$, where t is the digit at which the rounding occurs (i.e., $t = 0$ for the first four columns and $t = -1$ for the last). The standard deviation of this error is $10^t/\sqrt{12}$. Hence we take for our values of ε_j

ε_1	ε_2	ε_3	ε_4	ε_5
0.289	0.289	0.289	0.289	0.029

This gives relative errors $\varepsilon_j/\|x_j\|$ of

$\varepsilon_j/\|x_j\|$				
.0086	.0016	.0060	.0024	.0030

There is reason to believe that this is a model which cannot support a constant term, since the rows in the regression matrix sum to about 100; i.e., the components measured make up the entire sample. Consequently, the inclusion of a constant term would make the model nearly collinear. This is confirmed by computing the collinearity indices of the model with a constant term appended:

κ_0	κ_1	κ_2	κ_3	κ_4	κ_5
148	14	75	21	50	6

From (4.3) we see that a relative error of $1/75 = 0.013$ in x_2 can make the problem collinear—a cause for concern since x_2 is reported to only two figures. Moreover, the $\mathrm{RE}_{\mathrm{lin}}$ diagnostics for the first four variables are all approximately 0.12, which is too large for comfort.

The collinearity indices for the problem without constant term are

κ_1	κ_2	κ_3	κ_4	κ_5
2.7	4.2	3.2	2.4	3.9

which are suitably small. We now get for our bias diagnostics

$\mathrm{RE}_{\mathrm{bias}}$				
.0041	.0004	.0029	.0003	.0011

and for the linear diagnostics

$\mathrm{RE}_{\mathrm{lin}}$				
.0227	.0067	.0191	.0057	.0116

From this we see that if the numbers reported are accurate, the inaccuracies due to their rounding have little effect. It is therefore appropriate to go on to compute the diagnostics for importance. Since $\hat{\sigma} = 2.6$ and $\hat{\sigma}/\|y\| = .0074$, we have

IMP_j				
.0391	.0622	.0466	.0356	.0569

which says that the model can detect variables of quite small importance. Thus the collinearity indices have alerted us to the dangers of inserting a constant term in the model and have certified the model without a constant term.

Another Diagnostic

Collinearity indices have the nice property that they can be computed without any knowledge of the importance of regression coefficients or the sizes of the errors in the regression variables. This means that the analyst is not forced to come up with importance or error estimates at the time the data is run through a computer. However, the price paid for this is that he must now work with p separate numbers. Things are different when error approximations can be furnished at the start of the analysis. In this subsection we shall show how to combine the errors with the regression matrix to produce a single, worst case diagnostic.

Let us suppose that the rows of the error matrix E are uncorrelated random vectors with mean zero and common variance Σ. For the moment, assume that Σ is nonsingular. If we replace X by $X_\Sigma = X\Sigma^{-1/2}$ and E by $E_\Sigma = E\Sigma^{-1/2}$, then we shall have transformed the problem into one with uncorrelated errors with mean zero and variance one. Let $b_\Sigma = \Sigma^{1/2}b$ denote the transformed regression coefficients.

We now ask: *Of all linear combination $v^T b_\Sigma$, which is most sensitive to the errors?* To answer this question, assume without loss of generality that $\| v \| = 1$ and let $(V_* \; v)$ be an orthogonal matrix. Set $Z = (Z_* \; z) = X_\Sigma (V_* \; v)$ and let

$$\cdot \begin{pmatrix} D_{**} & d_{*p} \\ 0 & \delta_{pp} \end{pmatrix}$$

be the partitioned R factor of Z. Now the pth regression coefficient for the matrix Z is $v^T b_\Sigma$. Consequently, if we agree to measure the sensitivity of $v^T b_\Sigma$ by the pth diagnostics (6.18) and (6.19) with $\epsilon_j = 1$, then the most sensitive linear combination is that one for which δ_{pp} is minimized.

In fact the minimum is attained for that vector v for which $\psi_p \overset{\text{def}}{=} \inf(X_\Sigma) = \| X_\Sigma v \|$.[2] To see this, note that for this choice of v we have $\psi_p = \sqrt{\delta_{pp}^2 + \| d_{*p} \|^2}$ and since for any choice of v we have $\psi_p \leq \delta_{pp}$, it follows that $d_{*p} = 0$ and hence $\delta_{pp} = \psi_p$.

Thus if we set

$$\tau = \frac{\sqrt{n - p}}{\inf(X\Sigma^{-1/2})},$$

and estimate the size of the error by one, the diagnostics for the transformed problem become

$$\frac{\tau^2}{1 + \tau^2} > \mathrm{RE}_{\mathrm{bias}} \tag{6.21}$$

and

$$\frac{\tau}{\sqrt{n - p}} > \mathrm{RE}_{\mathrm{lin}}. \tag{6.22}$$

The inconvenient restriction that Σ be nonsingular may be removed by observing that τ has the alternative definition

$$\tau = \sqrt{n - p} \, \| \Sigma^{1/2} X^\dagger \|. \tag{6.23}$$

Since this is continuous in Σ, the singular case can be treated as a limit of the nonsingular case.

For the Woods, Steinour, and Starke problem $\tau = 0.0802$, and the diagnostics (6.21) and (6.22) are, respectively, 0.0064 and 0.0284. This is in agreement with our previous analysis using collinearity coefficients.

The number τ as defined by (6.23) has the drawback that the spectral norm $\| \Sigma^{1/2} X^\dagger \|$ is difficult to compute. If in place of the spectral norm one uses the Frobenius norm (2.8) and if further Σ is diagonal, then τ reduces to (6.20), i.e., the diagnostic proposed by Davies and Hutton (1975) and Beaton, Rubin, and Barone (1976).

[2] Since $\inf(X)$ is the smallest singular value of X (Golub and Van Loan, 1983, Chapter 1), the letter ψ here stands for psingular value. Puns aside, ψ would not be a bad notation for both statisticians and numerical analysts to adopt (the latter use σ, which is impossible for the former).

7. CONCLUDING REMARKS

Although we have given a continuous mathematical exposition of our subject, the parts a person would want to use in practice are scattered in pieces through the last six sections, and it is desirable to have some sort of recapitulation. Perhaps the best way to do this to imagine ourselves documenting a regression package that uses collinearity indices. The relevant section might read as follows.

Collinearity Indices

Collinearity. *A regression problem is said to be collinear when there is a nontrivial linear combination of the variables that is zero. Collinear problems suffer from a number of difficulties. For example, the cross-product matrix $A = X^T X$ is singular, and there are an infinite number of regression coefficients that minimize the residual sum of squares. The cure for collinearity is to furnish additional information that makes the regression coefficients well determined.*

Fortunately, regression problems with exact collinearities usually arise in circumstances where it is clear how to fix them. Far more difficult to handle are near collinearities, in which a linear combination of the columns is merely small. The chief sources of near collinearities are overspecified models (the kitchen-sink approach to designing experiments) and poor choices of basis functions in problems like polynomial fitting. Near collinearities can affect a regression model adversely by inflating the variance of regression coefficients and magnifying the effects of errors in the regression variables. Here there is no easy fix, and a careful reexamination of the original problem is usually in order.

Collinearity Indices. *Although our package cannot tell you how to resolve near collinearities, it does print out numbers called* collinearity indices, *labeled* `kappa` *in the output, which can help you assess the effects of near collinearities. There is one such number κ_j for each variable x_j, and they are always greater than one. In other contexts the squares of the collinearity indices are known as variance inflation factors.*

Collinearity indices can tell you three things about your problem. First, they can tell you how near your regression matrix is to one that is exactly collinear. Second, they can estimate the ill effects of errors in the regression variables. Finally, they can tell you if you are in danger of declaring an important variable to be insignificant. Let us examine each in turn.

Distance to Collinearity. *It can be shown that if a problem is nearly collinear, it can be made exactly so by perturbing the values of one of the variables. Without going into details (for which see (4.3)), the rule of*

thumb is that if $\kappa_j^{-1} \cong 10^t$ *then perturbations in the t-th digits of the components of* x_j *can make the problem collinear. Another way of saying the same thing is that you should be troubled about your model if the number of digits in* κ_j *is not less than the number of accurate digits in the components of* x_j.

Errors in the Variables. *With the exception of specialized problems whose regression matrices have integer entries, most regression problems have errors in the components of the regression variables. These errors affect regression coefficients in two ways. First, if they are large enough, they can introduce bias into the coefficients. Second, even when the bias is small they can increase the variability of the coefficients. Let us see how collinearity coefficients can be used to measure both effects.*

The first thing you must do is provide an estimate of the size of the errors in the components of the carriers. To use the collinearity coefficients, you must first satisfy yourself that the errors in the components of a variable are all roughly of the same size, say they have mean μ_j *and standard deviation* σ_j *(if your model has a constant term take* $\mu_j = 0$*). If this is so estimate the error by* $\varepsilon_j = \sqrt{\mu_j^2 + \sigma_j^2}$. *For example, if your data has been rounded in the digit corresponding to* 10^t, *you might set* $\varepsilon_j = 0.3 \cdot 10^t$, *which amounts to approximating the rounding error by a uniform random variable.*

To use the collinearity coefficients to assess the effects of errors in the variables, carry out the following three step procedure:

1. *Compute*

$$\tau_j = \sqrt{n - p}\kappa_j \frac{\varepsilon_j}{\| x_j \|}.$$

   If $\tau_j > \frac{1}{3}$, *reject the model. The errors are so influential that the diagnostic procedure cannot be trusted.*
2. *Calculate* $\mathrm{RE}_{\mathrm{bias}} = \tau_j^2/(1 + t_j^2)$. *This is an estimate of the relative bias in* β_j *due to the error in the jth variable; that is, the errors in* x_j *can be expected to depress the value of* β_j *by* $100 \cdot \mathrm{RE}_{\mathrm{bias}}\%$.
3. *Compute* $\mathrm{RE}_{\mathrm{lin}} = \tau_j/\sqrt{(n - P)}$, *which is an estimate of the second source of error. If it is less than* $\frac{1}{3}$, *it will contribute an approximately unbiased error that is roughly* $\mathrm{RE}_{\mathrm{lin}}$ *percent of* β_j.

The above calculations require you to perform some simple calculations (to help you the package prints out the numbers $\| x_j \|$*). This price you must pay for not being required to enter estimates of the errors when the package is run. If you so desire, you may enter the errors and ask the package to print a single number* τ, *which corresponds to the most sensitive linear combination of the regression coefficients. The package will also print the corresponding values of* $\mathrm{RE}_{\mathrm{bias}}$ *and* $\mathrm{RE}_{\mathrm{lin}}$.

Two words of caution. First, do not apply the above procedure to models, like polynomial regression, where error propagates from variable to variable through functional relationships. Second, keep in mind that our current knowledge about errors in regression variables is far from complete, and the above procedure does not rigorously guarantee that all the procedures in this package are free from their influence. We feel that if all the numbers τ_j *are less than 0.1, then there is not much to fear; but this feeling is based more on intuition than analysis.*

Important Regression Coefficients. *Finally, the package prints out numbers* $\texttt{IMP}_j$ *to help decide whether an important variable may be declared insignificant. The importance of a variable is how much it contributes to y in the sum*

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e.$$

Specifically, the importance of x_j *is the number* $\iota_j = |\beta_j| \| x_j \| / \| y \|$. *Thus* x_j *explains* $100\iota_j\%$ *of y. The number* $\texttt{IMP}_j$ *is the level of importance at which* $|\beta_j|$ *is equal to twice its estimated standard deviation, and is therefore in danger of being judged insignificant. For example, one would almost certainly feel that a model which produced a* $\texttt{IMP}_j$ *of 0.5 was unsatisfactory, since* x_j *could account for 50% of the response and still be judged insignificant.*

If your model has a constant term, we recommend that you use the centering option of the package, which subtracts column means from the variables before computing the $\texttt{IMP}_j$. *This will not make much difference unless some variables have a large number of leading digits. Without the leading digits, it is easier to judge at what level the variables are truly important.*

Of course our regression package is imaginary, but most regression packages actually print out collinearity indices, usually as the diagonals of the inverse cross-product matrix (after X has been centered and scaled). If the package also prints out the norms of the centered x_j, then the above procedures can be carried out with a hand calculator.

We have indicated above that errors entering linearly into the regression coefficients may not be as harmful as the ones causing bias. Since this observation is a potential source of further research, I would like to conclude this paper by expanding on it.

The idea of using linear approximations to nonlinear functions of random variables in regression problems goes back at least to Gauss (1821, Art. 18–19), who used it to justify applying his theory of linear least squares to the nonlinear problems which were his chief concern. Hodges and Moore (1972) apply the same idea to the problem of errors in the regression variables, obtaining approximations to the variances of the regression coefficients.

One difficulty with this approach is that regression coefficients perturbed by errors in the variables may not have nice distributions. To take a simple example, if ε is distributed $N(1, \sigma^2)$ the approximation $(1 - \varepsilon)^{-1} = 1 + \varepsilon$ will with high probability be very accurate when σ is small. But $(1 - \varepsilon)^{-1}$ does not have a first moment whatever the value of σ. How then is a variance computed from the approximation $1 + \varepsilon$ to be interpreted?

An answer is provided by the following result (Stewart, 1983).

Let f: $R^m \to R^n$ be differentiable at x. Let e be a random m vector with mean zero and variance Σ. Then

$$\operatorname*{plim}_{\Sigma \to 0} \frac{f(x + e) - f(x) - f'(x)e}{\| \Sigma^{1/2} \|} = 0.$$

The key here is the denominator which renormalizes the collapsing distributions of $f(x + e)$ and its linear approximation $f(x) + f'(x)e$. These renormalized distributions converge to one another, so confidence intervals calculated from one apply approximately to the other. The author (1983) has used his to establish a linearized version of Gauss minimum variance theorem for regression coefficients computed by ordinary least squares when there are errors in the variables.

However, I feel that a more fruitful approach is to recognize that problems with errors in the variables can in many respects be regarded as a model with an altered error in the response. To see this, let $\tilde{X} = X + E$ and write the model (2.1) in the form

$$y = \tilde{X}b + (e - Eb). \tag{7.1}$$

Let $\tilde{X}^\dagger = X^\dagger + F^{\mathrm{T}}$, defining F. Then the vector of regression coefficients $\tilde{b} = \tilde{X}^\dagger y$ computed from $\tilde{X}$ is given by

$$\tilde{b} = b + X^\dagger(e - Eb) + F^{\mathrm{T}}(e - Eb).$$

Since F goes to zero with E, when E is sufficiently small, $\tilde{b}$ behaves as if it came from the model

$$y = Xb + (e - Eb).$$

It is important to stress here that E does not have to be so small that its effect is negligible. For example e could be zero so that *all* the variability in the model comes from E.

The residual vector $\hat{e}$ behaves in much the same way. Let $P_\perp$ be the projection onto the orthogonal complement of the columns space of X and let $\tilde{P}_\perp = P_\perp + G$ be the corresponding projection for $\tilde{X}$. Then from (7.1) we have

$$\tilde{e} = \tilde{P}_\perp y = P_\perp(e - Eb) + G(e - Eb).$$

Thus for small E the residual can be used to estimate the variance of $e - Eb$. In fact using this approach, David and Stewart (1982) have shown that the classical F tests of significance remain approximately valid for small E. Again it must be stressed that small does not mean negligible.

The chief problem with this approach—or any other approach through linearization—is how to determine from contaminated data when the approximations are sufficiently accurate to allow an analysis to proceed. There are two dangers here. The first is that an inept analysis of the general case might yield pessimistic diagnostics that reject good models. The second is the temptation to summarize complex issues in a few numbers. What progress we have been able to make in this paper came from applying rough approximations to special cases and recognizing that near collinearity has many adverse affects, each of which must be tested separately. I believe this will continue to be true for some time to come. Occasionally rigor must wait for insight and elegance give way to utility.

ACKNOWLEDGMENTS

I have written this paper as a numerical analyst, not a statistician, and I hope that the view from without has been interesting to those within. Over the years I have been fortunate in the statisticians I have known, particularly Steve Fienberg who introduced me to the world of statistics and statisticians. Ingram Olkin has encouraged me to get my ideas before the statistical community by preparing this contribution to *Statistical Science*. Donald Marquardt was kind enough to communicate his recollections of the term variance inflation factor. Finally, I am indebted to Paul Velleman for helping me to get to the heart of the matter by asking (insisting is more like it) that I tell him what a regression package should print out and how to explain it to users.

APPENDIX: PROOF OF (4.4)

As usual we will give the proof for $j = p$. Without loss of generality we may assume that $\| x_j \| = 1$, $j = 1, 2, \cdots, p$, so that by (2.5) $\kappa_j = \| x_j^{(\dagger)} \|$ is the norm of the jth row of R^{-1}. Since

$$\begin{pmatrix} R_{**} & r_{*p} \\ 0 & \rho_{pp} \end{pmatrix}^{-1} = \begin{pmatrix} R_{**}^{-1} & -\rho_{pp}^{-1} R_{**}^{-1} r_{*p} \\ 0 & \rho_{pp}^{-1} \end{pmatrix},$$

it follows that

$$\sum_{j \neq p} \kappa_i^2 = \| R_{**} \|_{\mathrm{F}}^2 + \kappa_p^2 \| R_{**}^{-1} r_{*p} \|^2. \tag{A.1}$$

Now since the columns of R_{**} have norm one, the diagonal elements of R_{**} are less than or equal to one, and the diagonal elements of R_{**}^{-1} are greater than or equal to one. It follows that

$$\| R_{**}^{-1} \|_{\mathrm{F}}^2 \geq p - 1. \tag{A.2}$$

To get a lower bound on the second term in (A.1), we use the fact that $\inf(R_{**}^{-1}) = \| R_{**} \|^{-1}$, from which it follows that

$$\begin{aligned} \| R_{**}^{-1} t_{*p} \| &\geq \inf(R_{**}^{-1}) \| r_{*p} \| \\ &= \| R_{**} \|^{-1} \| r_{*p} \| \\ &\geq \| R_{**} \|_F^{-1} \| r_{*p} \|. \end{aligned}$$

Since the columns of R_{**} have norm one, $\| R_{**} \|_F^2 = p - 1$ and $\kappa_p^{-2} = \rho_{pp}^2 = 1 - \| r_{*p} \|^2$. Hence

$$\| R_{**}^{-1} r_{*p} \|^2 \geq \frac{1 - \kappa_p^2}{p - 1}. \tag{A.3}$$

Combining (A.1), (A.2), and (A.3) we get

$$(p - 1) \max_{i \neq j} \kappa_i^2 \geq \sum_{i \neq j} \kappa_i^2 \geq p - 1 + \frac{\kappa_p^2 - 1}{p - 1},$$

which is equivalent to (4.4).

REFERENCES

ANDERSON, T. W. (1984). Estimating linear statistical relationships. *Ann. Statist.* **12** 1–45.

BEATON, A. E., RUBIN, D. B. and BARONE, J. L. (1976). The acceptability of regression solutions: another look at computational accuracy. *J. Amer. Statist. Assoc.* **71** 158–168.

BELSLEY, D. A. (1984a). Demeaning conditioning diagnostics through centering (with discussion). *Amer. Statist.* **38** 73–93.

BELSLEY, D. A., KUH, E. and WELSCH, R. E. (1980). *Regression Diagnostics.* Wiley, New York.

DANIEL, C. and WOOD F. S. (1980). *Fitting Equations to Data*, 2nd ed. Wiley, New York.

DAVID, N. A. and STEWART, G. W. (1982). Significance testing in a functional model. Technical Report 1204, Dept. Computer Science, Univ. Maryland.

DAVIES, R. B. and HUTTON, B. (1975). The effects of errors in the independent variables in linear regression. *Biometrika* **62** 383–391.

DONGARRA, J. J., BUNCH, J. R., MOLER, C. B. and STEWART, G. W. (1979). *The LINPACK User's Guide.* SIAM, Philadelphia.

ECKART, G. and YOUNG, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika* **1** 211–218.

GAUSS, C. F. (1821). Theoria Combinationis Observationum Erroribus Minims Obnoxiae: Pars Prior. In *Werke* 4. Königlichen Gesellschaft der Wissenschaften zu Göttingen, 1880.

GOLUB, G. H., HOFFMAN, A. and STEWART, G. W. (1984). A generalization of Eckhart-Young-Mirsky matrix approximation theorem. To appear in *Linear Algebra and Its Applications.*

GOLUB, G. H. and VAN LOAN, C. F. (1983). *Matrix Computations.* Johns Hopkins, Baltimore, Md.

GOLUB, G. H. and WILKINSON, J. H. (1966). Note on the iterative refinement of least squares solution. *Numer. Math.* **9** 139–148.

HODGES, S. D. and MOORE, P. G. (1972). Data uncertainties and least squares regression. *Appl. Statist.* **21** 185–195.

MARQUARDT, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear regression. *Technometrics* **12** 591–613.

MIRSKY, L. (1960). Symmetric gauge functions and unitarily invariant norms. *Quart. J. Math.* **11** 50–59.

SEBER, G. A. F. (1977). *Linear Regression Analysis.* Wiley, New York.

STEWART, G. W. (1974). *Introduction to Matrix Computations.* Academic, New York.

STEWART, G. W. (1977). On the perturbation of pseudo-inverses, projections, and linear least squares problems. *SIAM Rev.* **19** 634–666.

STEWART, G. W. (1983). A nonlinear version of Gauss's minimum variance theorem with applications to an errors-in-the variables model. Technical Report TR 1263, Dept. Computer Science, Univ. Maryland.

STEWART, G. W. (1984). Rank degeneracy. *SIAM J. Sci. Statist. Comput.* **5** 403–413.

SWINDEL, B. F. and BOWER, D. R. (1972). Rounding errors in the independent variables in a general model. *Technometrics* **14** 215–218.

TURING, A. M. (1948). Rounding-off errors in matrix processes. *Quart. J. Mech. Appl. Math.* **1** 287–308.

VAN DER SLUIS, A. (1969). Condition numbers and equilibration of matrices. *Numer. Math.* **14** 14–23.

WILKINSON, J. H. (1963). *Rounding Errors in Algebraic Processes.* Prentice-Hall, Englewood Cliffs, N.J.

WOODS, H., STEINOUR, H. H. and STARKE, H. R. (1932). Effect of composition of Portland cement on heat evolved during hardening. *Indust. Engrg. Chem.* **24** 1207–1214.

Comment

Donald W. Marquardt

Statisticians and numerical analysts owe a large debt of gratitude to Dr. Stewart for his demonstration and lucid exposition of the mathematical connection between the condition number and the parameter variance inflation factors. In doing so, he has also clarified the reasons why the condition number is not really helpful in the multiple regression context, nor in many other contexts. The insights he provides in this paper are important for all statisticians, because collinearity problems occur in many statistical contexts, including multiple linear regression, nonlinear regression, unbalanced analysis of variance, and estimation from inverse integral transform models. In this brief commentary I have selected three facets of Dr. Stewart's paper for discussion.

Donald W. Marquardt is Consultant Manager, Applied Statistics Group, Engineering Service Division, E. I. Du Pont De Nemours & Company, Wilmington, Delaware 19898.