BELSLEY, D. A. (1984b). Collinearity and forecasting. *Forecasting* **3** 183–196.
BELSLEY, D. A. (1986a). Centering, the constant, first-differencing, and assessing conditioning. In *Model Reliability* (E. Kuh and D. Belsley, eds.). M.I.T. Press, Cambridge.
BELSLEY, D. A. (1986b). Model selection in regression analysis, regression diagnostics and prior knowledge (with discussion). *Internat. J. Forecasting* **2** 41–46.
BELSLEY, D. A. (1986c). Modelling and forecasting reliability. Working paper, Center for Computational Research in Economics and Management Science, M.I.T.
BELSLEY, D. A. and OLDFORD, R. W. (1986). The general problem of ill-conditioning and its role in statistical analysis. *Comput. Statist. Data Anal.* **4** 103–120.
GUNST, R. F. (1983). Regression analysis with multicollinear predictor variables. *Comm. Statist. A—Theory Methods* **12** 2217–2260.
HENDRY, D. F. (1980). Econometrics—alchemy or science? *Economica* **47** 387–406.
LEAMER, E. E. (1978). *Specification Searches.* Wiley, New York.
SIMON, S. D. and LESAGE, J. P. (1986). The impact of collinearity involving the intercept term on the numerical accuracy of regression. Working paper, Dept. Applied Statistics and Operations Research, Bowling Green State Univ.

Comment

Ronald A. Thisted

The statistics profession is fortunate indeed to have such a friend as Professor Stewart. He has repeatedly taken the time and energy to inform statisticians about the relevance of numerical analysis to their day-to-day work, and he has also taken the trouble to understand and to explicate some of our problems from our own point of view. This paper is an example of what numerical analysis can have to say about statistical problems, and it shows that there is a lot that we statisticians can profit from. In particular, Professor Stewart greatly improves our understanding both of collinearity and of one indicator of collinearity—the variance inflation factor.

As is true of most important papers, this one raises as many questions as it answers. I would like to comment on three issues that Professor Stewart only touched on. First, although Stewart would relegate the condition number $\kappa = \|X\| \cdot \|X^\dagger\|$ to the dustbin for statistical purposes, there is an important statistical interpretation which rescues it. Second, Stewart's procedures for using collinearity diagnostics depend upon a measure ι_j of the importance of the jth regressor variable. The notion of relative importance of a regressor is an elusive one, however, particularly when collinearity is present. Finally, I discuss the question of whether statisticians should want collinearity diagnostics at all, and if so, what we should want from them. Where possible, I adopt Stewart's notation. References to equations in his paper are preceded by the letter "S."

Ronald A. Thisted is Associate Professor, Department of Statistics, The University of Chicago, 5734 University Avenue, Chicago, Illinois 60637.

1. THE CONDITION NUMBER

Stewart gives a clear description of the numerical relevance of the condition number κ. In numerical analysis, its primary significance is the inequality (S-3.4), the righthand side of which gives a good indication of the effect of *numerical* errors in the regressors on the regression coefficients themselves. Because the *statistical* errors represented by e in the regression model (S-2.1) are generally much larger in magnitude than the numerical errors resulting from rounding and truncation, the bound from (S-3.4) is often so pessimistic as to be useless. In addition, the condition number is not invariant with respect to rescaling columns of X, so that interpretation of κ is dependent on the way in which X has been scaled. Although Stewart discusses three alternatives for scaling X—equal column scaling of X, scaling X to produce equal column scaling of E, and implicitly, scaling X so that the components of β are roughly equal in size—he finds no single choice compelling.

The condition number of X has an important statistical interpretation in the regression problem which is generally overlooked. Consider an arbitrary linear combination of the estimated regression coefficients, say $\hat{\alpha} = v'\hat{\beta}$. The variance of $\hat{\alpha}$ is given by

$$\begin{aligned} \operatorname{Var}(\hat{\alpha}) &= \sigma^2 v'(X'X)^{-1}v \\ &= \sigma^2 \|v'X^\dagger\|^2. \end{aligned} \tag{1.1}$$

From this computation it is apparent that the linear combination with smallest variance (subject to the constraint, say, that $\|v\| = 1$) has variance $\sigma^2[\inf(X^\dagger)]^2$. The coefficients v_1 which achieve this minimum value explicitly give the linear combination $\hat{\alpha}_1 = v_1'\beta$ about which the regression data are most

informative. Similarly, it can be shown using the singular value decomposition of X that the variance of the linear combination with maximum variance is $\sigma^2[\inf(X)]^2$. The maximum variance coefficients v_p give the linear combination $\alpha_p = v_p'\beta$ about which the data at hand are least informative. Thus, κ^2 is the variance ratio between the most precisely estimated linear combination of β and the least precisely estimated. As a result, the condition number can indicate the range of relative precision with which linear combinations of the parameters can be estimated given the particular design matrix X.

The interpretation of the condition number in terms of the precision of estimating linear combinations of the regression coefficients also bears upon the matter of scaling X. When the columns of X all represent similar kinds of measurements—prices in local currency from Sweden, France, and Germany, for example—it is likely that linear combinations of regression coefficients will be most interpretable, and thus most interesting to the analyst, if expressed in terms of common units—dollars, for instance. Indeed, unless this is done, the variance of $v'\hat{\beta}$ is a number expressed in peculiar units indeed. This suggests, then, that the appropriate scaling for X is one in which the columns represent variables expressed in common units, *or* one from which the units have been removed.

2. IMPORTANCE OF VARIABLES

Collinearity is a problem because it produces regression coefficients with large standard errors which, as Stewart notes, may lead us to discount an important variable. But the notion of relative importance of variables in a regression is an elusive one, as a close inspection of Stewart's definition reveals. Small values of the "importance coefficient" $\iota_j \equiv |\beta_j| \, \| x_j \| / \| y \|$ are said to mean that x_j's contribution to y is unimportant, because "the term $\beta_j x_j$ represents $100\iota_j$% of the total observed response."

This definition is misleading, particularly in the presence of severe collinearity, since the ι_j's don't add to 100%. Indeed, their sum can and often does exceed 100%—by a considerable amount. The importance coefficients come closest to adding up when the x_j's are orthogonal, and the problem gets worse as collinearity increases. What is more, large importance coefficients can actually mask the role of the corresponding variable. Consider the following example.

The response variable y is the logarithm of the United States consumer price index (CPI), and x_1 and x_2 are, respectively, log Gross National Product (GNP) in nominal dollars and log of GNP expressed in constant (1972) dollars (CGNP). Let z_2 represent the difference GNP–CGNP. Now z_2 is simply the logarithm of the GNP deflator, which, like the CPI, is a measure of inflation. Neither GNP nor CGNP has much to do directly with the level of inflation as reflected in the CPI. The *important* variable, the GNP deflator, is not explicitly present in the regression model.

Table 1 contains data on these variables for seven nonconsecutive years. For the model

$$y = \beta_1 x_1 + \beta_2 x_2 + e, \tag{2.1}$$

the value of R^2 is 99.2%, so that CPI is essentially determined by GNP and CGNP together. For convenience and interpretability, let us suppose that y, x_1, and x_2 have all been centered. Then we have $\iota_1 = 1.086$ and $\iota_2 = 2.061$. Note that these importance coefficients add to more than 3000%. The reason for such large importance coefficients is the high correlation between x_1 and x_2, with the consequent high negative correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$. A naive interpretation of ι_1 and ι_2 would be that each variable is enormously important. This is true in only a very narrow sense.

Variable x_1 is "important" in this problem *only* if variable x_2 is also present in the model, and similarly for x_2. The model

$$y = \beta_1 x_1 + \gamma_2 z_2 + e, \tag{2.2}$$

where $z_2 = x_2 - x_1$, is equivalent to (2.1), but in this model the importance of x_1 is only $\iota_1 = 0.214$. The latter figure is a more realistic assessment of the substantive importance of x_1; its partial correlation with y after adjusting for z_2 is only -0.487. (It is easy to construct realistic examples in which both this partial correlation and the *raw* correlation are nearly zero.) These computations emphasize that "relative importance of variables" as measured by ι_j is highly conditional on the particular form of the model, and that greater collinearity leads to greater dependence on the particular way in which variables enter the model.

In short, the importance indices become less interpretable as collinearity increases. So it seems to be problematic to define a collinearity-diagnostic procedure such as (S-5.2) in terms of them.

TABLE 1
Economics data in logarithmic units

Year	CPI y	CGNP x_1	GNP x_2	DEFL z_2
1950	4.36691	6.28040	5.65599	−0.624404
1960	4.51632	6.60259	6.22654	−0.376051
1970	4.73181	6.98008	6.88959	−0.090485
1980	5.53576	7.30047	7.87322	0.572744
1982	5.57519	7.30317	8.03041	0.727240
1983	5.60396	7.33563	8.10289	0.767255
1984	5.63729	7.42833	8.23589	0.807558

3. WHY DO WE NEED COLLINEARITY DIAGNOSTICS?

In trivial problems such as the CPI regression, it is easy to understand the provenance of large variance inflation factors. (Actually, κ_1 is only a modest 7.5 for the centered data.) It is hard to imagine actually conducting a regression analysis with as little regard for the nature of the variables as I showed in the previous section, ignoring the clear *a priori* relationships between CPI, GNP, CGNP, and the GNP deflator. But in more complicated problems with many variables, relationships such as the one between GNP and CGNP can sneak into our regression models with the data analyst unaware.

The real value of collinearity diagnostics is to alert the statistician to the presence of a *potential* difficulty. Both the condition number and the collinearity indices can help to assess the magnitude of the potential problem. The κ_j's can also help to identify particular variables that are involved, so that they can indicate a starting point for further investigation. It is this latter property that makes diagnostics useful—they can be used to focus and to direct further efforts in refining the model. If they don't point a finger somewhere, they are not terribly useful.

In the economics data, the moderate value of κ_1 might lead us to question the role of x_1 in the model, as might the values $\text{IMP}_j = 0.63$. Yet, the model cannot be improved by removing either of the two variables. The problem is the GNP deflator, of course. How might the diagnostics lead us to discover the culprit?

There are two similar routes that can be followed to construct supplementary diagnostics. When κ_p (say) is large, by definition x_p is very nearly a linear combination of the other variables, and that linear combination is given by the coefficients $(\hat{\mu}_1, \cdots, \hat{\mu}_{p-1})$ from (S-3.7). These are simply the regression coefficients from the regression of x_p on the other variables. It is often the case when κ_p is large that the particular linear combination implied by $(\hat{\mu}_1, \cdots, \hat{\mu}_{p-1})$ is interpretable, and sometimes the linear combination $x_p - \sum \hat{\mu}_j x_j$ can be recognized as a more sensible "regressor" to have included in the first place than one or more of the x_j's.

A second route is to examine the $p \times 1$ vector v_p corresponding to the smallest singular value of X. This vector can be used to obtain the vector $u \equiv Xv$ which realizes $\inf(X)$; it is also the coefficient vector for $\alpha_p = v_p'\beta$, the linear combination of the regression coefficients about which the data are least informative. If one or more of the κ_j's is large, then $\inf(X)$ must be small, that is, the linear combination u is close to zero. The coefficients v_p point to the "worst collinearity." In practice, this linear combination is also often interpretable, and may suggest ways in which the original variables can be removed, rearranged, or reconstructed so as to avoid the near singularity.

ACKNOWLEDGMENT

This research was sponsored by National Science Foundation Grant DMS 84-12233. It was completed while the author was on leave at Stanford University.

Comment: Diagnosing Near Collinearities in Least Squares Regression

Ali S. Hadi and Paul F. Velleman

We congratulate Professor Stewart on a lucid presentation and a practical article. We will discuss several aspects of the proposed collinearity and relative error measures.

Ali S. Hadi is Assistant Professor of Economic and Social Statistics, and Paul F. Velleman is Associate Professor of Economic and Social Statistics, Cornell University, 358 Ives Hall, Ithaca, New York 14853.

1. COLLINEARITY AND ERRORS IN VARIABLES

Stewart gives simplified expressions for probing the effects of errors in regression variables by comparing his equations (6.3) and (6.5). Specifically, he defines

$$\text{RE}_{\text{bias}} = \frac{\beta_p - \tilde{\beta}_p}{\beta_p}$$

and

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