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 $\mathrm{IMP}_j = 0.63$. 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

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

and

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