Comment: Well-Conditioned Collinearity Indices

David A. Belsley

G. W. Stewart's paper, *Collinearity and Least Squares Regression*, is a substantive contribution to an important and often ignored problem of statistical analysis: assessing collinearity in least squares estimation. The summary of relevant results from numerical analysis given in Section 3 and the developments of Section 6 are alone worth the price of admission. Furthermore, Stewart, the numerical analyst, is to be commended for this foray into the world of statistics, for there is much these two disciplines (to which I'll add econometrics) have to teach each other. Indeed, the above two sections should become part of the basic material in all advanced courses in practical regression analysis. Sadly missing from the Stewart paper, however, is one of the more important notions that applied statistics has to teach the numerical analyst, namely, the necessity of a context for application: the fact that the data are not just a given set of numbers and the model is not just a linear combination of these data. Without this, elements are ignored that are vitally important for determining the meaning (or lack of meaning) of collinearity diagnostics in a statistical (as opposed to a numerical) application and that allow some conclusions to be drawn which cannot truly be supported. I discuss these issues here.

MODELS VERSUS DATA

Model and Data Confusion

I begin with a discussion of the relation between model and data, a confusion between which mars the Stewart paper and whose resolution motivates many of the comments that follow. Thus, for example, on numerous occasions throughout the paper, statements are made to the effect that "The diagnostics are large, and this should make one pause about the model," or "... should lead us to reject the model." In no such instance, however, are there proper grounds for such a conclusion, and indeed one should usually counsel the opposite. Let us see why.

The source of this confusion arises from the way Stewart defines and uses the term model. Initially, *model* is defined by (2.1) as $y = Xb + e$ where X is "simply a fixed array of numbers," and then, shortly thereafter, by "In other words, our model is specified by the matrix X alone." That is, a confusion occurs between the model and the data to which the model is applied. This is akin to confusing a random variable with a sample drawn from it. Perhaps such usage is current in some disciplines, but, as a rather exhaustive search of leading texts attests, it is certainly not in either statistics or econometrics, the disciplines toward which I presume the Stewart paper to be principally directed.

For those authors (on this, see Belsley (1986c)) who actually attempt a formal notion of an applied-statistical model (as opposed to a probabilistic model), modeling is an *a priori* description of the mechanism that generates the data. *It is not the data.* A model arises from the statistical investigator's (hopefully creative) imagination, and exists in a wholly different realm of discourse from the data associated with it. When applied to a specific context, it is assumed (not always validly) that the observed data are generated from the specific model, but, that they are only one set of data that could have occurred and for which that model is relevant. In fact, it is assumed that any of an infinity of other sets of data could have been generated by the same model, and the same model could have been applied conditionally to any of an infinity of other situations. That is, a model like (2.1) is assumed relevant to a class of X's (not just the observed ones), and given any one of those X's, any of a class of y's could have been generated.

Thus, the fact that there may be numerical problems with a given data set, in and of itself, says nothing about the validity of the model. A model can be rejected if it implies things inconsistent with the observed data, but a model cannot be adjudged invalid merely because some of the data to which it is applied are numerically funny. In so doing, one is putting the cart before the horse.

So, the strange diagnostic values given in the discussion surrounding Table 1 should not "give one pause about the model," rather they should give one

David A. Belsley is Professor of Economics, Boston College, and Principal Research Associate, Center for Computational Research in Economics and Management Science, Massachusetts Institute of Technology. His mailing address is Department of Economics, Boston College, Chestnut Hill, Massachusetts 02167.