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.

pause about the ability of the given data to tell meaningful things about the model. This issue is especially clear here, for the data in Table 1 are known relevant to the model given in Belsley (1984a, 1986a). There is no issue here about the validity of the model; in this case "truth" is known. The y values given in the Belsley papers were indeed generated correctly by that model with these rottenly conditioned X's. Thus, the conditioning of the X's clearly cannot be used to invalidate the model, but this conditioning can lead one to question the ability of these data (once generated) meaningfully to estimate the model. I cannot overstate the importance of this issue. The conclusion quoted above simply is not correct.

The same holds for the statement made early in Section 6: "The resulting diagnostic is to reject the model when the bias is unacceptably large." Much preferred would be "The resulting diagnostic is to reject the ability of these data meaningfully to estimate the given model by ordinary least squares." It is, however, to be noted that these same data could be employed along with other suitably chosen data in, say, an instrumental-variables estimator to produce valid estimates of the model—the point being that the diagnostics here do not reject the model but rather the ability of the data meaningfully to estimate the model with OLS.

Importance

This issue comes to a head in Section 5 defining the term *importance.* Given a proper view of modeling, it is clear that the importance of a variate to the model is determined *a priori.* A variate can only be adjudged important when the investigator has cast it in an important role by knowing it, *a priori*, to be a variate measuring a concept that figures centrally in the workings of the real-life mechanism. *The data, in and of themselves, do not and cannot determine the importance of the variate.* Thus, those statements directed against the unsatisfactoriness of the model should instead be directed against the usefulness of the given data set for providing meaningful information about the model. Stewart almost seems to recognize this problem when he says "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." This is true, but then the wrong outside source is used, for one cannot dip again into the data but must go to the prior information about the model and its real-life context.

The idea that the data can be used to determine "importance" follows a growing and disturbing tendency by some statisticians to use the data for model building. This process is philosophically unsound, as I (and many others before me) try to show in Belsley (1986b, 1986c).

Thus, I strongly urge that the term "importance" for the concept defined in Section 5 be dropped and replaced perhaps with a term like "presence." And I strongly urge that the lack of presence be used as an indication of data weakness rather than model weakness.

A related matter arises in the statement: "But if we attempt to replace the vague term "important" with the mathematically precise term "statistically significant," we become involved in a paradox. . . ." Several problems arise here. "Important," as noted, refers to part of the *a priori* model-building process, not the statistics (or data) used to estimate that model. Thus, there is no possibility of replacing "important" with "statistically significant." Once again we have two terms that exist in two wholly different realms of discourse, and it is meaningless to try to use them together, much less to replace one by the other. Secondly, there is nothing mathematically precise about "statistical significance." Such significance depends upon a test level chosen, greatly arbitrarily, by the statistical practitioner.

Despite my objections to the use of the term "importance," I do feel the concept in Section 5 is indeed of interest. It makes an attempt to determine the degree to which a given data set in a specific estimation context is able to manifest itself in the estimation procedure. That is, it makes an attempt to measure the degree of signal available in the data to estimate a particular parameter. A closely related concept, placed on a more solid statistical footing, is to be found in Belsley (1982).

COLLINEARITY MEASURES

Let us now turn to collinearity measures and centering.

Measures of Collinearity

I quite agree with some of the objections raised by Stewart against the condition number taken by itself. Indeed it was for this reason in Belsley, Kuh, and Welsch (1980) that I moved to a more complete set of diagnostics growing out of the condition number, namely the full set of *condition indexes* and the *variance-decomposition proportions* (VDPs). It is the diagnostic value of these, not the condition number, that must be compared to the collinearity indices (VIFs) advocated by Stewart, and I am disappointed that no such comparison is made. Thus, in Section 3 it is indicated that condition numbers are too crude for statistical applications and that a *set* of numbers is needed "that can probe the effects of collinearities more delicately." But, of course, the full set of condition indexes given in Belsley, Kuh, and Welsch (1980) along with VDPs do exactly that, and Stewart's effort here by no means shows that the VIFs or his collinearity indices *as a practical matter* are superior.

I do, by the way, have some regard for VIFs (or *tolerances*, as Simon and Lesage (1986) call their inverses). They involve less computation than do the diagnostics of Belsley, Kuh, and Welsch (1980) and are made more easily available in statistical and econometric packages, and so have some practical advantage. But they fall short of the Belsley, Kuh, and Welsch diagnostics in several important respects, and so I must consider them second best. Specifically, the Belsley, Kuh, and Welsch diagnostics give information that the VIFs cannot. The VIFs are incapable, by themselves, of probing into the delicacy of how many dimensions of X are deficient, i.e., they cannot tell for a particular parameterization how many near dependencies exist or what variates are involved in each. The Belsley, Kuh, and Welsch diagnostics, however, can, and this information is vitally useful in providing proper corrective action (Belsley, 1982, 1984b).

Consider, then, a model with four variates x_i, $i = 1, \cdots, 4$. Suppose in case 1 that all four variates are highly jointly collinear, so that any one regressed on the others would produce an R^2 near 1, or a very large VIF (or collinearity index) indeed. Suppose in case 2 that x_1 and x_2 were highly collinear and x_3 and x_4 were also, but the two pairs were not collinear with each other. Again all VIFs would be astronomical, and there would be no way of distinguishing these two cases using VIFs or Stewart's collinearity indices. By contrast, the condition indexes of Belsley, Kuh, and Welsch and the VDPs would clearly distinguish these two cases and correctly indicate the type of corrective action that was needed in each. In case 1, a single piece of prior information on any one of the four variates has the chance of removing the problems of collinearity for all four variates, whereas in case 2 it will require prior information on two variates, on either of x_1 or x_2 and on either of x_3 or x_4. Using VIFs alone, one could not determine this, whereas the diagnostics of Belsley, Kuh, and Welsch would point directly to the solution.

The inability of VIFs or collinearity indices to determine the number of near dependencies in X and to help point out where corrective action is best placed is a serious weakness that is not mentioned in the Stewart paper. Nor is the ability of the diagnostics of Belsley, Kuh, and Welsch to help in this task. This issue, of course, is related to the modeling considerations examined above. If one believes that data weaknesses, in and of themselves, are indicative of model weaknesses (an indefensible position), then corrective action takes the form of model changes until one gets a model that fits the data (regardless of how silly it is otherwise—see the examples in Hendry (1980) or Belsley (1986c) if you wish to see how comical life can get here). If, however, data weaknesses are merely that, affecting estimation but not model specification, then the issue centers on how to correct for them, and the introduction of prior information from the investigator's understanding of the phenomenon being modeled becomes all important. VIFs leave a good deal to be desired here.

Stewart also quite correctly criticizes the condition number on the grounds that it is an upper bound that can be unnecessarily pessimistic. He does not, however, show that his collinearity indices provide any tighter bound in practice. To be sure, the VIFs are bounded from above by $\inf(X)$, but that inequality also has an equal sign. Thus I do not feel the impression given in the paper that the collinearity indices will provide a tighter bound in practice is justified. While this could be true, it must be more than stated. I have not done a systematic study of this phenomenon in my own work, but it certainly has not been my experience that it is true to any noticeable degree. In fact I can say that I have never learned anything about a data set from VIFs that I did not learn as well or better from the diagnostics of Belsley, Kuh, and Welsch (which do not figure in the Stewart paper at all), whereas the reverse has indeed been true.

Finally, Stewart criticizes condition numbers on the grounds that it is an open question whether the scaling should be done relative to X or to E. This too is a legitimate issue, but not one that is automatically debilitating to the use of condition numbers (or the set of condition indexes and VDPs). In Belsley, Kuh, and Welsch (1980) it is assumed (as is so marvelously typical of econometricians) that the data are measured without error. Here, then, the E's do not indicate measurement errors, but merely perturbations in the X's. The concern is not to determine the potential effects of measurement error on the OLS estimates, but to see the sensitivity of the estimates to new data that are also measured without error but which are within some relative shift from the first data set, each column of X being treated equally. This is indeed a special case, and it need not always be the most interesting one, but under these circumstances equilibration with respect to the columns of X, not E, is what is called for. This view of the problem is, to my mind, that which pairs up with what econometricians are often interested in, and is what motivated the development in Belsley, Kuh, and Welsch. There are, then, cases where this criticism of condition numbers can be answered.

Clearly, however, if different columns of X are to be treated differently, the convention of scaling X has its shortcomings. We recognize this problem and demonstrate one way of dealing with it in Belsley and Oldford (1986). The perturbations used in the example given there are variate-dependent and correspond to equalizing the columns of the E matrix.

Centering

For those familiar with Belsley (1984a, 1986a), it will come as no surprise to learn that I completely disagree with Stewart's stance on centering. To begin with, note that he never demonstrates the efficacy of centering and, in the final analysis, resorts strictly to a psychological argument in its favor—not a numerical argument, not a statistical argument, but a psychological one. Thus, in Section 5:

> 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 variable 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). [Italics mine]

Note the italicized words. These are strictly psychological in value and in no way demonstrate the relevance or necessity of centering. Not only that, the statement isn't so. Centering does not "remove x_1," it only removes a very special multiple of x_1, namely mx_1, where m is the mean of the elements of the vector being projected (orthogonally) on x_1. But there are numerous other ways one could "remove x_1." One could use any other "centercept" (to use Tukey's terminology), replacing m with such equally reasonable measures as the median or the geometric mean or even the max or the min. In each case, x_1 is "being removed," but the effects on any conditioning diagnostics will be very different. Furthermore, all of these "removals" can be mortally criticized on the ground that they are data-dependent. That is, the adjustment is made on the basis of the specific data set at hand, despite the fact that the meaning of the variate in the model is model-dependent and any meaningful adjustment must therefore be made on *a priori* grounds.

The issue of centering and conditioning cannot properly be dealt with without an understanding of this distinction between data- and model-dependent centerings. Since modeling is an *a priori*, non-data, phenomenon, any adjustments in a given data set for interpretation within the model must be made on the basis of model considerations, not data considerations. That is, data-dependent adjustments are almost always inappropriate for a conditioning analysis. This point is made in greater detail in Belsley (1984a, 1986a) and Belsley and Oldford (1986). There, however, it is shown that model-dependent "centerings" given *a priori* can indeed be appropriate. I put centerings in quotes because the adjustments need not (and usually will not) be mean-centerings, and the adjustments need not always be within the range of the data. This latter point is important because, whereas centerings that remove a constant factor within the range of the data will tend to reduce the condition number, constant adjustments outside the range of the data (which are quite possible on *a priori*, theoretical grounds) will tend to increase the condition number. That is, "removal of x_1" need not reduce the condition number.

Nor do theoretical considerations always suggest removal of a constant. One of the examples I give in the previously cited papers deals with an adjustment of the Dow-Jones index. The psychological base for this index could change over time. Several years back, a level of 800 had great importance, first for getting over it, then for falling below it. Later 1200 became the relevant base, and now it is 2500. An adjustment of such a time series for a conditioning analysis could well involve removals of different levels for different elements of the time-series vector. The blind adjustment of the given data series by its own mean is merely mechanical, having, in general, nothing to do with the real meaning of the data or its relation to the model. Using such adjusted data for assessing conditioning can do nothing but produce mechanical and arbitrary results.

This point is seen perhaps more strongly by noting the following: The same data series could figure in two separate models. The conditioning of the data must be assessed relative to the role those data play in the specific model, and that role is only assessable *a priori*. If, then, one always mean-centers the data, one is assuming that the relevant role of the data can mechanically be determined outside any context and is always the same. This clearly cannot be. For any conditioning analysis, the data must be put in a form that has meaning for the model at hand. For any data series, this meaning will change from context to context. No uniform and mechanical adjustment which affects the conditioning of the data (such as shift of origin) can therefore be justified.

The bottom line, however, is that mean-centering simply and plainly produces collinearity diagnostics (either those of Belsley, Kuh, and Welsch or Stewart's collinearity indices) that can overlook important diagnostic information. This point is irrefutably documented by the example given in Belsley (1984a, 1986a) and the study of Simon and Lesage (1986) and is not and cannot be answered by Stewart's paper. If I ran only Stewart's diagnostics (or mine on centered data), I would miss vital information about the weakness of these data for estimating *all parameters* of this model, not just the intercept, and for providing *a priori* corrective information. This point cannot be overemphasized, because many feel the intercept often to be a nuisance parameter and ignoring it a costless blessing (notice I say ignoring it, for centering does not get rid of it (Belsley, 1986a)). Thus both Stewart (last paragraph, Section 5) and Gunst (1983) claim that one should examine the conditioning of the uncentered data only if the estimate of the intercept is of interest. But this is simply wrong. Collinearity with the intercept can quite generally corrupt the estimates of *all parameters* in the model whether or not the intercept is itself of interest and whether or not the data have been centered, a result readily seen theoretically and demonstrated practically in Simon and Lesage (1986). Hence, diagnostics that ignore the presence of the intercept, such as ones based on centered data, are insidiously misleading about data problems for estimating all parameters.

In short, whereas some *a priori* justifiable origin shift (perhaps not constant) may indeed be appropriate to produce data amenable to a conditioning analysis, mechanical mean-centering is not in general of this class. Centering is to be avoided. I do not feel this; I know it. I have shown centered data generally to lack the information needed properly to assess their usefulness in estimating a linear model by least squares. There is nothing psychological in this view.

Short Data

A related issue to centering is that of short data. Both Stewart's and my collinearity diagnostics miss the boat here. But both his notion of "importance" (whose name I hope will be changed) and my notion of low signal-to-noise deal with this related phenomenon. Space prohibits examining this issue at length here; one can find it in Belsley (1982). Briefly, however, if one mean-centers a data series with a strong constant component, such as my data reproduced in Stewart's Table 1, the resulting series become short data (or lack "presence"). The effect of this on the standard error of the estimate of its regression coefficient is equally as devastating as if the centered variate were tightly involved in a collinear relation, and indeed the problem is of an integrally related character. Leamer (1978) even goes so far as to refer to this as collinearity among a single variate. This is because data that are collinear relative to one parameterization can become noncollinear, but necessarily "short," relative to another. That is, it is always possible to transform an ill conditioned data matrix into a well conditioned one through a nonsingular linear transformation. But such practices cannot buy the statistical practitioner anything, for they merely convert the source of a high standard error from collinearity to short data, but the standard errors remain high. This is, of course, as it should be; we would otherwise be buying something for nothing.

CONCLUSIONS

It might be argued that the thrust of the above criticisms rests on the ability of the investigator to provide a good and meaningful model but that what is needed in practice is a set of conditioning diagnostics or collinearity measures that can be applied to a data set without a context (model in my sense) in mind. To which, I can only answer that the latter does not exist; it's a will-o'-the-wisp, and one is kidding oneself to hope otherwise. Oh, to be sure, one can always devise algorithms that provide the investigator with diagnostic numbers, but will they mean anything? In an applied statistical analysis, meaningful conditioning (collinearity) diagnostics can only be obtained relative to a specific goal and a specific context. It is shown in Belsley and Oldford (1986) that changing the goal or the context changes the diagnostic numbers, either Stewart's or mine. Diagnostic numbers calculated without a goal and context, then, are just numbers, wholly without meaning. Thus a regression package that provides the information indicated in the last section of Stewart's paper without first making sure that the data are relevant to a particular goal and context is merely providing the user with the false security that somehow comes with firm numbers whether or not they mean anything. Furthermore, providing these numbers on the basis of centered data virtually guarantees their being potentially misleading regardless of goal, for such data almost always lack context.

Thus, I do not object to VIFs (or Stewart's related collinearity indices), but they should be based on uncentered R^2 values and applied to data known to be interpretable in the context of a relevant model. They otherwise become ill conditioned diagnostics. Furthermore, if additional information is needed regarding multiple dependencies and the best use of prior information, something more than such VIF-based diagnostics will be needed. In this case, try mine, you'll like 'em.

ACKNOWLEDGMENT

This research was sponsored in part by the National Science Foundation under Grant IST-8420614.

ADDITIONAL REFERENCES

Belsley, D. A. (1982). Assessing the presence of harmful collinearity and other forms of weak data through a test for signal-to-noise. *J. Econometrics* **20** 211–253.

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."

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

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

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

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