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

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

We do not know why these two measures use different denominators. We also think that the distinction between RE_{bias} and RE_{lin} is artificial.

From Stewart's equation (6.9) we know that $h_p^T h_p \doteq (n-p)(\mu_p^2 + \sigma_p^2)$. This is a nonnegligible value even for small values of μ_p^2 and σ_p^2. Therefore it does not seem reasonable to set $h_p = 0$. Rather than considering the effects of h_p and γ_p separately, we prefer to consider them jointly. We assess the relative error in the pth coefficient due to errors in the pth predictor by

$$\text{(1)} \qquad \begin{aligned} RE_j &= \frac{\beta_p - \tilde{\beta}_p}{\beta_p} \\ &= \frac{(\rho_{pp} + \gamma_p)^2 + h_p^T h_p - (\rho_{pp} + \gamma_{pp})\rho_{pp}}{(\rho_{pp} + \gamma_p)^2 + h_p^T h_p}. \end{aligned}$$

It is easily seen that $E(RE_j) \geq 0$ which is in agreement with the well known fact that errors in variables attenuate regression coefficients. Now suppose either that $\mu_p = 0$ or that the model contains a constant term. Then

$$\text{(2)} \qquad \begin{aligned} E(h_p^T h_p) &\doteq (n-p)\sigma_p^2, \\ E(\gamma_p) &= 0. \end{aligned}$$

Substituting (2) in (1), we get

$$RE_j = \frac{(n-p)\sigma_p^2}{\rho_{pp}^2 + (n-p)\sigma_p^2} = RE_{bias}.$$

Thus, RE_{bias} is the special case of RE_j corresponding to the most common regression situations. If we follow Stewart and set $h_p = 0$ in (1), we get

$$RE_j = \frac{\gamma_p}{\rho_{pp} + \gamma_p},$$

which says that if $\mu = 0$ or the model contains a constant term then $RE_j = 0$. This anomaly is another argument against setting $h_p = 0$.

2. COLLINEARITY INDICES AND LEVERAGE POINTS

It is important to note that Stewart's collinearity indices are not resistant to the effects of high leverage points (points that are separated from the bulk of other data points in one or more dimensions). A high leverage point can hide a collinearity (see Figure 1) or create one (see Figure 2). Small collinearity indices do not necessarily mean that the model is home free, and large collinearity indices may be due to only one or two data points. We refer to points that either hide or create near collinearities as collinearity-influential points. Collinearity-influential points are usually but not necessarily points with high leverage, but not all high leverage points are collinearity-influential. Mason and Gunst (1985) show that collinearity can

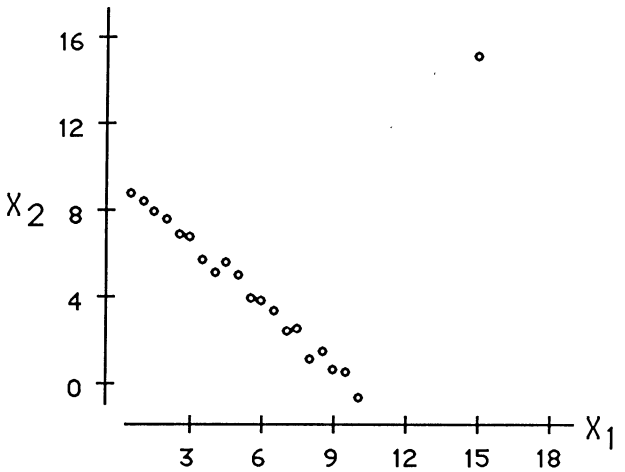

FIG. 1. *An example of a data point that hides a collinearity.*

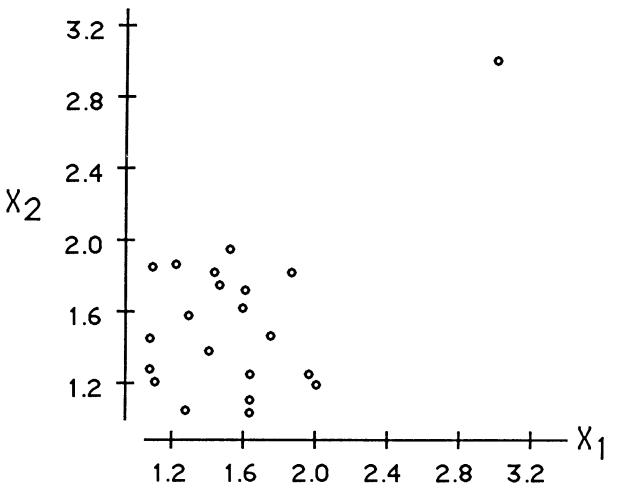

FIG. 2. *An example of a data point that creates a collinearity.*

be increased without bound by increasing the leverage of a point. They also show that a k variate leverage point can produce $k-1$ independent collinearities. Therefore, any careful diagnosis of collinearity must include diagnosis for collinearity-influential points. We suggest some diagnostic plots in the next section.

3. DIAGNOSING COLLINEARITY—INFLUENTIAL POINTS

As Stewart shows, the jth collinearity index is related to the residuals obtained from the regression of the jth column of X on all other columns. Let X_j and $X_{[j]}$ denote the jth column of X, and X without the jth column, respectively. The jth collinearity index can be expressed as

$$\text{(3)} \qquad \kappa_j = \| X_j \| / \| e_j \|, \quad j = 1, \cdots, p,$$

where e_j is the residual vector in the regression of X_j on $X_{[j]}$. Let P and $P_{[j]}$ be the projection matrices onto the spaces spanned by X and $X_{[j]}$, respectively. To

show the effect of the ith row of X on κ_j we write κ_j as

$$\kappa_j = \left[x_{ij}^2 + \sum_{r \neq i} x_{rj}^2\right]^{1/2} \left[\frac{p_{ii} - p_{ii[j]}}{e_{ij}^2}\right]^{1/2}, \tag{4}$$

where p_{ii} and $p_{ii[j]}$ are the ith diagonal elements of P and $P_{[j]}$, respectively, and x_{ij} is the ijth element of X.

A rash inspection of (4) may lead one to conclude incorrectly that if $e_{ij}^2 = 0$, then $\kappa_j = \infty$. That is, if X_{ij} happens to lie on the fitted equation when X_j is regressed on $X_{[j]}$, then $\kappa_j = \infty$. However

$$p_{ii} - p_{ii[j]} = \frac{e_{ij}^2}{e_j^{\mathrm{T}} e_j} \tag{5}$$

shows that as $e_{ij}^2 \to 0$, $(p_{ii} - p_{ii[j]}) \to 0$. In fact, the second expression on the right hand side of (4), $(p_{ii} - p_{ii[j]})/e_{ij}^2$, is constant for all i.

A careful inspection of (4) and (5) leads us to suggest the following graphical displays for the detection of collinearity-influential points:

(i) Pairwise scatterplots of the columns of X.
(ii) Plot e_j versus X_j, for each j.
(iii) Plot $e_{ij}^2/e_j^{\mathrm{T}} e_j$ versus $p_{ii[j]}$, for each j.

The pairwise scatterplots in (i) are common adjuncts to a careful regression analysis and can be useful in the detection of pairwise collinearities and pairwise collinearity-influential points. Of course, pairwise scatterplots may not show multivariate collinearities or multivariate collinearity-influential points. The plots in (ii) and (iii) perform better in that respect.

If X_j is orthogonal to $X_{[j]}$, the plot of $e_j = (I - P_{[j]})\, X_j$ versus X_j is a straight line through the origin with slope of one. Deviation from the 45° line indicates the existence of a collinearity between X_j and some columns of $X_{[j]}$. The pattern of points on this plot is also important. Collinearity-influential points are usually separated from the bulk of other points.

The plot of $e_{ij}^2/e_j^{\mathrm{T}} e_j$ versus $p_{ii[j]}$ must satisfy

$$0 \leq \frac{e_{ij}^2}{e_j^{\mathrm{T}} e_j} \leq 1$$

and

$$0 \leq p_{ii[j]} \leq \max(p_{ii}).$$

The scatter of points is governed by equation (5). In particular, for fixed p_{ii}, the larger $p_{ii[j]}$ the smaller $e_{ij}^2/e_j^{\mathrm{T}} e_j$. Collinearity-influential points appear in the lower right and upper left corners of this plot. A data point in the upper left corner will be a high leverage point only if X_j is included in the model, whereas a data point in the lower right corner will be a high leverage point even if X_j is not included in the model.

TABLE 1
Simple correlation coefficients for the data in Stewart's Table 3

	X_1	X_2	X_3	X_4	X_5
X_1	1				
X_2	0.23	1			
X_3	−0.82	−0.14	1		
X_4	−0.25	−0.97	0.03	1	
X_5	0.31	0.15	−0.32	−0.17	1

For those readers with access to three-dimensional scatterplots, we recommend a plot of e_i versus X_j versus $P_{ii[j]}$. This plot combines the merits of plots (ii) and (iii) in a single, effective diagnostic display. Three-dimensional scatterplots are not yet widespread, but their presence in statistics packages for inexpensive microcomputers (e.g., in the Data Desk package for the Macintosh, Velleman and Velleman (1986)) presages a growing availability.

Example. Accepting the legitimacy of fitting a five-predictor model to 13 data points, we follow Stewart's suggestion and fit the no-intercept model to the data in Stewart's Table 3. The simple correlation coefficients matrix $c = c_{ij}$ (shown in Table 1) indicates the existence of two pairwise collinearities; $c_{24} = 0.97$ and $c_{13} = 0.82$. The scatterplots in Figures 3 and 4 show that no data point seems to create or hide these collinearities. However, the inspection of the other pairwise scatterplots shows the existence of two other collinearities. They have been hidden by a collinearity-influential point. Figures 5 and 6 show that point 3 hides two collinearities; one between X_1 and X_5, and another between X_3 and X_5. In fact when the third row of X is omitted, c_{15} changes from 0.31 to 0.73 and c_{35} changes from −0.32 to −0.82.

For $j = 2, 3, 4$, the scatterplots of e_j versus X_j (not shown) do not show any discernible patterns that

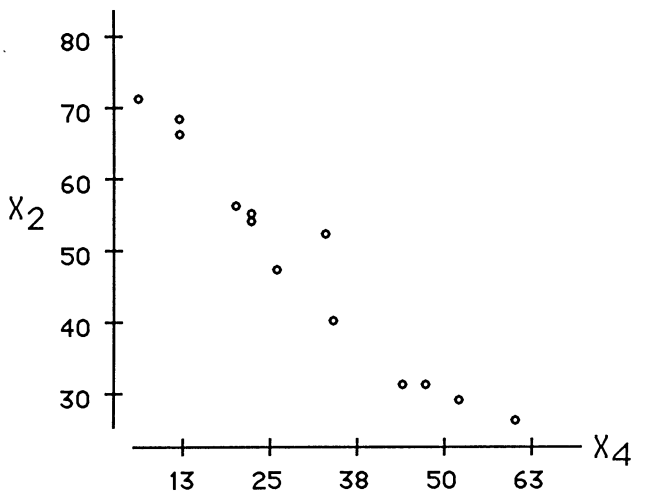

FIG. 3. *Scatterplot of X_2 versus X_4.*

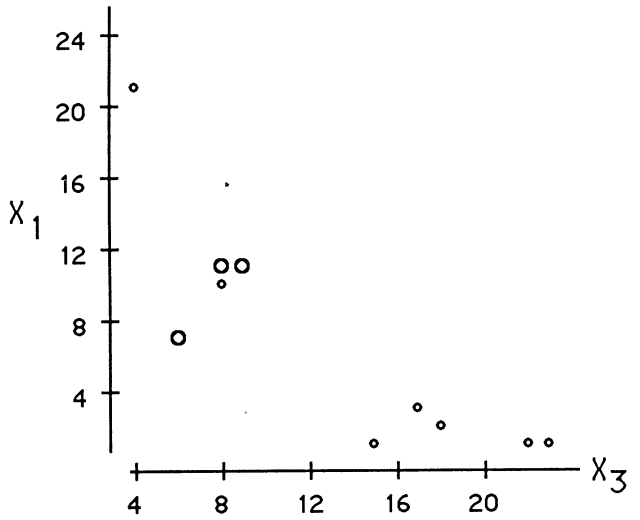

FIG. 4. *Scatterplot of X_1 versus X_3.*

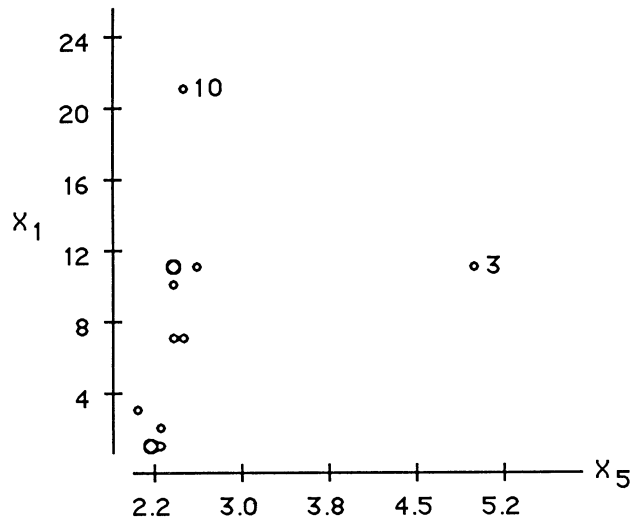

FIG. 5. *Scatterplot of X_1 versus X_5.*

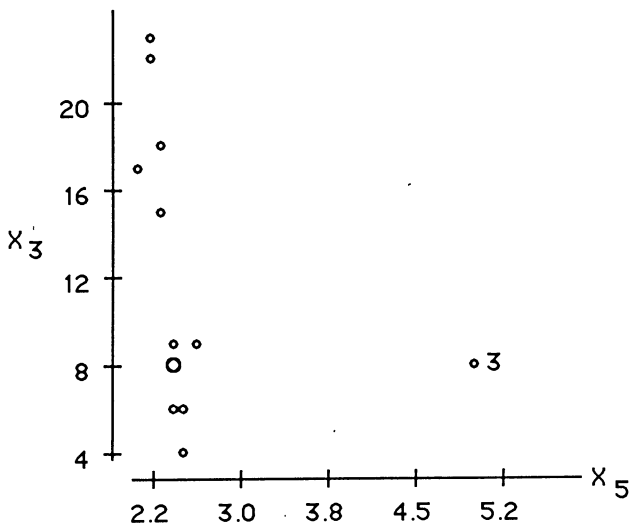

FIG. 6. *Scatterplot of X_3 versus X_5.*

deserve a comment. For $j = 1$ (Figure 7) point 10 is separated from the other data points and for $j = 5$ (Figure 8) point 3 lies far from other points. Thus points 3 and 10 have high potential for being collinearity-influential points. The scatter plots of $e_{ij}^2/e_j^T e_j$ versus $p_{ii[j]}$ explain why. Point 3 lies in the lower right corner of the plot in Figure 9 which means that it is a high leverage point even if X_1 is not included in the model. But because point 10 lies in the upper left corner, its influence is contingent upon the inclusion of X_1. When X_1 is deleted, the 10th diagonal element of the projection matrix changes from $p_{10,10} = 0.72$ to $p_{10,10[1]} = 0.16$. The position of points 3 and 10 are reversed in Figure 10. Thus point 3 is a high leverage point only if X_5 is included in the model. When X_5 is deleted, the third diagonal element of

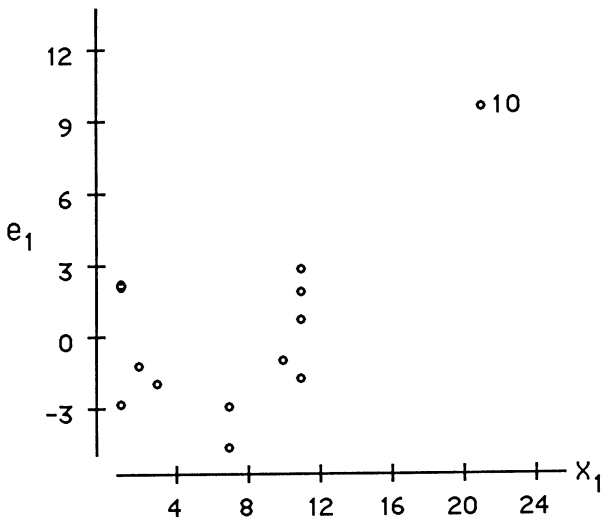

FIG. 7. *Scatterplot of $e_1 = (I - P_{[1]})X_1$ versus X_1.*

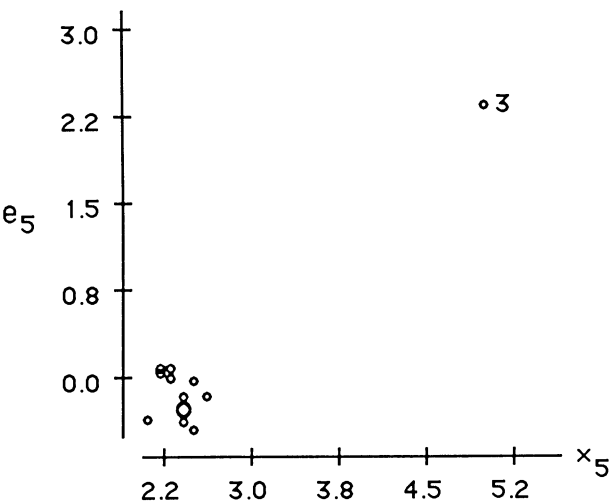

FIG. 8. *Scatterplot of $e_5 = (I - P_{[5]})X_5$ versus X_5.*

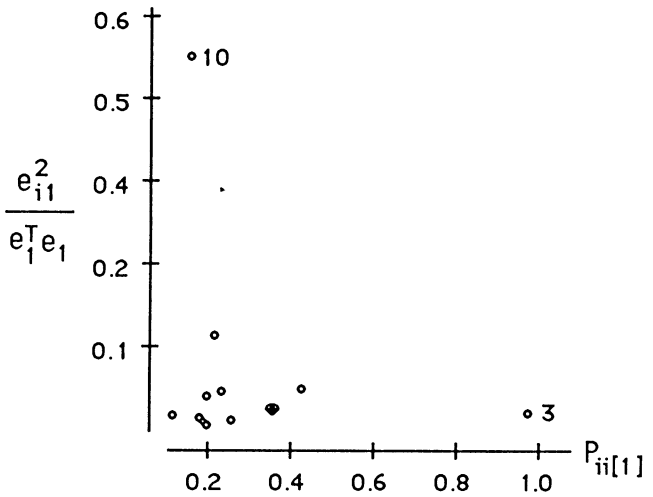

FIG. 9. *Scatterplot of normalized residuals when X_1 is regressed on $X_{[1]}$ versus the diagonal elements of $P_{[1]}$.*

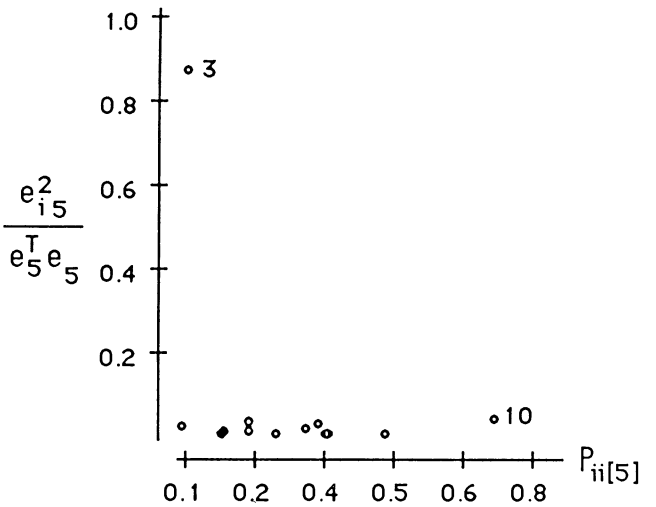

FIG. 10. *Scatterplot of normalized residuals when X_5 is regressed on $X_{[5]}$ versus the diagonal elements of $P_{[5]}$.*

the projection matrix changes from $p_{3,3} = 0.99$ to $p_{3,3[5]} = 0.12$.

Our analysis shows that points 3 and 10 are candidates for being collinearity-influential points. Let us now examine the effects of these points on collinearity measures. The condition numbers and the collinearity indices for the full and the reduced data are given in Table 2. The collinearity indices hardly change when point 10 is deleted but change substantially when point 3 is deleted. Thus we may conclude that point 3 is the only single point that influences collinearities. The impact of point 3 on the importance of the variables, and the estimated regression coefficients and their variances, is shown in Table 3. The results based on the full data are strikingly different from those obtained when point 3 is deleted. In this example collinearity indices support fitting the no-intercept model to the full data but not to the data without the third row.

TABLE 2
Condition numbers and collinearity indices for the data in Stewart's Table 3

	$\kappa(X)$	κ_1	κ_2	κ_3	κ_4	κ_5
Full data	85	2.7	4.2	3.2	2.4	3.9
3rd row deleted	779	4.5	17.0	3.8	12.1	31.4
10th row deleted	88	3.6	4.7	3.8	2.5	4.0

4. COLLINEARITY INDICES AND CENTERING

We agree with Stewart that when there is a constant term in the model, the data should be centered before computing the κ_j's and IMP_j's. In most multiple regressions, the origin is not a meaningful data point. A no-intercept model may represent a newborn patient with no systolic blood pressure and a body temperature of 0°F, or a house in the middle of a city with no floor space and no rooms, or some other nonsensical combination.

In some cases, however, it is more meaningful to fit a no-intercept model to the data. If the user chooses to fit a no-intercept model, we suggest that the regression package should: (i) advise the user to shorten any artificially long column by subtracting a constant. For example, if "year" (reported as a four digit number) is one of the columns of X, this column should be centered (or at least be given as $0, 1, \cdots, n-1$) even if the model does not contain a constant term and (ii) ask the user to nominate a "typical" data point; either one that is actually in the data set or one chosen for convenience, and then shift the data around this point. The package may suggest default values for a typical point, for example, the median of each column of X.

We commend Stewart for providing specific advice to developers of statistics packages and hope they adopt these methods. We think that they should take this opportunity and extend their packages to keep information about the precision of the data. A package that has this information could produce Stewart's diagnostics without requiring further computation from the user. The information would also be useful in drawing and labeling plots, in identifying data entry errors, and in deciding how many digits of a computed result to print. A package might reasonably use a natural default assumption that all of the entered digits are correct except for rounding errors and thus the errors follow a uniform distribution over the interval $(-.5, .5)$. This is equivalent to setting $\mu_p = 0$ and $\sigma_p^2 = 10^t/\sqrt{12}$, where t is the digit at which the rounding occurs. Of course, the user should be encouraged to override this default assumption with more accurate information.

TABLE 3
The impact of point 3 on some regression results for the data in Stewart's Table 3[a]

Var.	ι_j		IMP_j		$\tilde{\beta}$		Var $(\tilde{\beta})$	
	F	R	F	R	F	R	F	R
X_1	.213	.181	.04	.07	2.19	1.89	.041	.129
X_2	.602	.470	.06	.26	1.15	.90	.004	.063
X_3	.104	.090	.05	.06	.76	.63	.029	.043
X_4	.171	.083	.04	.19	.49	.23	.003	.066
X_5	.001	.247	.06	.48	.02	10.02	1.065	95.988

[a] F denotes full data and R denotes reduced data (point 3 deleted). Values of ι_j and IMP_j are from Stewart's equations (5.1) and (5.2), respectively.

ACKNOWLEDGMENT

This work was partly supported by the United States Army Research Office through the Mathematical Sciences Institute of Cornell University.

ADDITIONAL REFERENCES

MASON, R. L. and GUNST, R. F. (1985). Outlier-induced collinearities. *Technometrics* **27** 401–407.

VELLEMAN, P. F. and VELLEMAN, A. Y. (1986). *The Data Desk Handbook*. Data Description, Ithaca, N. Y.

Rejoinder

G. W. Stewart

I would like to begin by thanking the commentators for giving my paper a fair and careful reading. Since the following remarks must necessarily focus on our differences, let me stress at the outset that I find much to agree with in their comments.

I am happy to acknowledge that Donald Marquardt knew of the connection between variance inflation factors and collinearity. My only quibble is that one must read a rather small section of his 1970 paper very carefully in order to see it. Marquardt never uses the word collinearity and only asserts that the variance inflation factors depend on the partial correlations, without explicitly stating the nature of the dependency. From his comment one can deduce that he takes a partial correlation near one as a synonym for collinearity and means for the reader to infer that the dependency is the same as the one he writes down for two variables. However, the passage can also be read as a vague afterthought, which is how I interpreted it on first reading.

On nomenclature, the difficulty with the term variance inflation factor is that it draws attention to one effect of near collinearity to the exclusion of other, equally important effects. It seems more natural to me to give a simple characterization of near collinearity and then show how it affects statistical procedures. Taking the square root of the variance inflation factors not only simplifies the formulas but stresses a useful connection with the condition number.

David Belsley's comments are practically a paper in themselves, and a complete response would amount to another. Here I will only make a few observations and trust the reader to sort out the issues.

Belsley would make a distinction between data and models, and in a sense I heartily agree. Numerical and statistical tricks are no substitute for a knowledge of the science underlying a problem. However, on close inspection his distinction appears elusive. Is a constant term model or data? How do we classify the design matrix for an unbalanced analysis of variance? Moreover, the term model has come to mean many things. Belsley's "rather exhaustive" survey evidently did not include Draper and Smith (1981, page 86) or Seber (1977, pages 42 and 43), who use the term model in much the same sense as I do. Attempting to preempt the word model is like trying to tell the tide where to come in.

I will save my comments on importance for the end of this rejoinder. Regarding centering, I will simply restate that centering is a change of variables, and the new ones are not equivalent to the old. There is nothing vague or "psychological" about this observation, and it is ironic that Belsley quotes at length from a passage that describes the psychological biases in the opposing view.

Belsley points out that the collinearity indices do not tell the dimension of the approximate null space and provide little help in selecting an independent