Computational Statistics & Data Analysis 4 (1986) 103–120
North-Holland

The general problem of ill conditioning and its role in statistical analysis *

David A. BELSLEY

Department of Economics, Boston College
and Center for Computational Research in Economics and Management Science,
MIT, Cambridge, MA, USA

R.W. OLDFORD

Center for Computational Research in Economics and Management Science, MIT,
Cambridge, MA, USA

Received 9 March 1986

Abstract: The notion of a conditioning analysis of a general, nonlinear set of relations is defined along with an associated definition of ill conditioning. From these, one may identify at least three different kinds of conditioning analyses of interest in statistics and econometrics: data, estimator, and criterion conditioning. While these three coincide in the OLS/linear case, they can and do diverge otherwise. The absence of a general mathematical solution for a conditioning analysis points to computer-intensive alternatives, one of which is suggested and illustrated.

Keywords: Collinearity, Conditioning analysis, Ill conditioning, Multicollinearity, Nonlinear estimation, Nonlinear systems, Perturbation analysis, Sensitivity analysis.

1. Introduction

Conditioning [1], initially a numeric-analytic concept pertaining to the sensitivity of solutions to linear systems, is finding increasingly useful applications in statistics and econometrics. Its main use to date has been to diagnose collinearity and its consequent ills, but, as this paper demonstrates, this is only the beginning of its value in statistics. Proceeding from a general definition, we distinguish three different kinds of conditioning analyses that are of statistical interest, namely

* All computation was accomplished using the TROLL system at MIT.

[1] The word 'conditioning' is sufficiently well-established in the numerical-analytic literature that its use here cannot reasonably be avoided. It should be clear from the context that no reference to conditional probability is intended.

data, estimator, and criterion conditioning. The three coincide in the ubiquitous case of least squares estimation of a linear model (OLS), but can and do diverge otherwise.

In what follows we first provide a general definition of ill conditioning and a conditioning analysis. We then emphasize a concept that is vitally important to a meaningful measure of conditioning but which is frequently overlooked in practice: contextual or structural interpretability. Section 4 introduces general considerations pertaining to three specific types of conditioning: data, estimator, and criterion. Section 5 exemplifies these three types of conditioning in the special case of OLS estimation of a linear model and demonstrates their coincidence there. Section 6 discusses the kinds of problems that arise when examining the conditioning of nonlinear models, and Section 7 introduces and demonstrates a computer-intensive means of perturbation analysis that would seem to allow the practical conditioning of any situation to be assessed.

2. A general notion of ill conditioning

Suppose rather generally any system of continuous equations

$$\lambda = f(\omega), \tag{2.1}$$

where λ, ω, and f are vectors and/or matrices. The elements of ω could be data, parameters, or random variables. Thus (2.1) might describe an estimator, a stochastic model, a system of data dependencies, or, in general, any system of interest in which elements λ are assumed to be dependent upon elements ω. In some applications this relationship might be defined implicitly as $f(\lambda, \omega) = \mathbf{0}$.

Frequently, it is of interest to know something about the sensitivity of λ with respect to particular changes in ω: changes that belong to some specified set Ω. Typically, concerns arise when, depending on circumstances, disproportionately large or small changes in λ can result from a given change in ω. In such a situation λ will be said to be *ill conditioned*.

For example, if ω represents data, then the set Ω might reasonably consist of changes corresponding to the measurement accuracy of the data. Further, if λ is an estimator based on these data, then λ's being ill conditioned with respect to Ω would indicate that the given estimate is not well-determined due to the inadequate quality of the observed data.

Mathematically, the problem can be decomposed as follows: Suppose the quantities ω and λ are related according to (2.1) so that an additive perturbation $\delta\omega$ in ω results in a perturbation in λ equal to $\delta\lambda = f(\omega + \delta\omega) - f(\omega)$. For fixed ω, a function $g(\delta\omega) \equiv f(\omega + \delta\omega) - f(\omega)$ may be defined which maps the elements of $\delta\omega$ of a given domain Ω to elements $\delta\lambda$ in the corresponding range set Λ. That is,

$$g\colon \delta\omega \mapsto \delta\lambda$$

or,

$$g\colon \Omega \to \Lambda.$$

Conceptually, there exists a third set Λ^* which consists of all those perturbations $\delta\lambda$ which are considered, a priori, to be reasonable given the set Ω. Concern arises when, corresponding to some $\delta\omega$ in Ω, there exists $\delta\lambda$ in Λ which is not in Λ^*. For example, in a sensitivity analysis, the set Ω consists of all 'reasonably small' perturbations $\delta\omega$, and concern arises when some such $\delta\omega$ can produce perturbations $\delta\lambda$ in λ which are not 'reasonably small'.

These considerations suggest the following general definitions:

Conditioning analysis: The specification of the *conditioning triple* $K = \{f, \Omega, \Lambda^*\}$ followed by a determination of whether λ is ill conditioned.

Ill conditioning: Given K and its implied Λ, λ is said to be ill conditioned with respect to ω (or Ω) if $\Lambda \not\subset \Lambda^*$. Equivalently, one can call the system f ill conditioned.

The relevance of a conditioning analysis depends critically on the determination of the sets Ω and Λ^* in K. As we see in the next section, these sets *must* be determined within the context of the problem at hand. To determine them otherwise is to render the definition meaningless.

For many practical applications, we have found the following specifications of Ω and Λ^* to be useful:

$$\Omega = \{\delta\omega: \|\delta\omega\|/\|\omega\| \leq m_1\} \tag{2.2}$$

and

$$\Lambda^* = \{\delta\lambda: \|\delta\lambda\|/\|\lambda\| \leq m_2\}, \tag{2.3}$$

where m_1, $m_2 \geq 0$ are real constants and $\|\cdot\|$ denotes some norm (Euclidean and spectral norms will be used here). Cases could also be considered in which either or both of the inequalities in (2.2) and (2.3) are replaced by '$\geq$' or '$=$', depending on the kind of conditioning being investigated.

In our experience, Λ^* often follows naturally once Ω has been specified. This being the case, it is important that the perturbations $\delta\omega$ be meaningfully interpretable within the context of the problem. To denote such perturbations, Belsley [2,3] has introduced the term *structural interpretability*, a concept we now briefly examine in greater detail.

3. Relevance of the conditioning analysis

A conditioning analysis has been defined formally above so that its parts may be recognized and understood. In particular, the conditioning triple K completely specifies the conditioning analysis: changing any of its elements produces a different conditioning analysis. Clear and explicit specification of K is therefore mandatory; anything less necessarily renders the analysis irrelevant. This section highlights two critical, but frequently overlooked, aspects in specifying K.

The first is the necessity of specifying clearly what sensitivity is to be addressed by the analysis, since a conditioning analysis intended for one purpose can be quite misleading for, and indeed often confused with, another. Belsley [2], for

example, shows how the condition number of centered data can inadvertently mask an intended analysis of the sensitivity of OLS estimates to perturbations in the basic, or uncentered, data.

The second is the necessity of having interpretable elements in K if the conditioning analysis itself is to be interpretable. Each element of K must be defined to be meaningfully interpretable within the context of some larger investigation. Only against such a backdrop (which typically comprises the information relevant to the study of some subject matter) can we hope to argue the meaningfulness of the results of a particular conditioning analysis.

Specifically, Ω represents a set of additive changes $\delta\omega$ in ω which are considered to have some special meaning or interpretation within an associated context. For example, in one analysis, knowledge from the underlying context may be used to show that the perturbations $\delta\omega$ may be considered as inconsequential, while in another that they are especially large. Clearly, care must be taken in the selection of ω and Ω. The changes $\delta\omega$ are necessarily defined relative to the ω included as arguments to f, and hence the $\delta\omega$ can only be argued to be interpretable if the ω are so.

Suppose, for example, Ω is defined as in (2.2). The value given to m_1 will depend upon both ω and $\delta\omega$. In the absence of any supporting context to provide a basis for interpretation, one might think that $m_1 = 0.01$ is an inconsequential proportionate change in any ω. For the particular ω under consideration, however, the context may in fact suggest values as large as 4.0 should be regarded as inconsequential. What constitutes the set of inconsequental changes Ω, then, depends upon ω and its interpretation within the overall context.

Similarly, $\lambda = f(\omega)$ must be interpretable within the underlying context before any defensible choice for Λ^* can be made.

If this discussion seems unnecessarily laborious, we stress it because we have found that overlooking its message can be the source of much needless confusion. To facilitate reference to this important concept, then, we describe those elements in K which can be meaningfully interpreted in terms of an underlying or associated context by the phrases *contextually* or *structurally interpretable*. We view the former term more broadly, the latter being reserved for those cases where the context is supplied by some underlying real-life situation that is the subject of some investigation, statistical or otherwise. [2] Thus, the relevance of a conditioning analysis must be argued on the basis of a contextually interpretable K, and it is incumbent upon the investigator to convince the reader of the relevance of this interpretation.

We turn now to three kinds of conditioning analyses which are often of interest

[2] This use of the term *structural* is borrowed from simultaneous-equations theory in Econometrics and may not be completely familiar outside that field. A particular set of values for all the parameters of a model of a real-life situation constitutes a *structure*. Structural elements, then, denote those parts of the model that can be paired up with particular elements of the real-life situation being modelled, and as such must be interpretable through one's a priori knowledge of that situation.

in statistical applications: data conditioning, estimator conditioning, and criterion conditioning. They are treated quite generally in the next section and discussed again for the particular case of least-squares linear regression in Section 5.

4. Data, estimator, and criterion conditioning: the general case

To fix ideas, we now examine three different kinds of conditioning analyses. The first, which we call data conditioning, is directly associated with the well-known concept of collinearity. The others, estimator and criterion conditioning, forge new directions.

Data conditioning (collinearity)

Given $z_1, \ldots, z_p$, observed n-vectors on p variables (endogenous, exogenous, or both), we wish to know if there exists an exact, or nearly exact, linear relation among them. If so, the variables $z_1, \ldots, z_p$, or the columns of the matrix $\boldsymbol{Z} = [z_1, \ldots, z_p]$, are said to be collinear. Among the various formalizations of this, that given in Gunst [6] is useful. Given $\| z_i \| = 1$ for all i, collinearity exists among the z_i's if, for a suitably small predetermined $\eta > 0$, there exist constants $\boldsymbol{c}^{\mathrm{T}} = [c_1, \ldots, c_p]$, not all zero, such that

$$\boldsymbol{Z}\boldsymbol{c} = \boldsymbol{\gamma} \tag{4.1}$$

and

$$\| \boldsymbol{\gamma} \| < \eta \| \boldsymbol{c} \|. \tag{4.2}$$

This definition of collinearity is equivalent to the following conditioning analysis. Let $\boldsymbol{\lambda}$ and $\boldsymbol{\omega}$ be defined by

$$\boldsymbol{\lambda} = \boldsymbol{f}(\boldsymbol{\omega}) \equiv \boldsymbol{Z}\boldsymbol{\omega}. \tag{4.3}$$

Consider perturbations

$$\delta\boldsymbol{\lambda} = \boldsymbol{g}(\delta\boldsymbol{\omega}) \equiv \boldsymbol{Z}\delta\boldsymbol{\omega} \tag{4.4}$$

having domain set

$$\Omega = \{ \delta\boldsymbol{\omega} : \| \delta\boldsymbol{\omega} \| / \| \boldsymbol{\omega} \| = m_1 \} \tag{4.5}$$

and acceptable response set

$$\Lambda^* = \{ \delta\boldsymbol{\lambda} : \| \delta\boldsymbol{\lambda} \| / \| \boldsymbol{\lambda} \| > m_2 \}. \tag{4.6}$$

That is, perturbations $\delta\boldsymbol{\omega}$ of fixed relative size m_1 are required to result in perturbations $\delta\boldsymbol{\lambda}$ whose length relative to $\boldsymbol{\lambda}$ is not less than m_2, some small number. If this cannot be the case, then the data of $\boldsymbol{Z}$ are ill conditioned with respect to Ω of (4.5). The degree of such data conditioning will depend upon the selection of the constants m_1 and m_2.

The equivalence of this notion of data conditioning to that of Gunst is shown as follows: Suppose first the data $\boldsymbol{Z}$ are ill conditioned as above; that is, for some

$\|\delta\omega\|/\|\omega\| = m_1$ we observe $\|\delta\lambda\|/\|\lambda\| < m_2$, which implies $\|\delta\lambda\|/\|\delta\omega\| < (m_2/m_1)\|\lambda\|/\|\omega\|$. Thus, taking $\eta \equiv (m_2/m_1)\|\lambda\|/\|\omega\|$, we find the data to be collinear according to Gunst's definition with (4.1) replaced by (4.4). Conversely, suppose we observe $\|\delta\lambda\| < \eta\|\delta\omega\|$ for some 'suitably small' $\eta > 0$ and $\delta\omega \neq \mathbf{0}$. Choose $\|\delta\omega\|$ so that $\|\delta\omega\| = m_1\|\omega\|$ for some $\|\omega\|$ and m_1, so we observe $\|\delta\lambda\| < \eta m_1\|\omega\|$, or, letting $m_2 = \eta m_1\|\omega\|/\|\lambda\|$, $\|\delta\lambda\|/\|\lambda\| < m_2$.

Estimator conditioning

Consider an estimator $\hat{\theta}$, of parameters θ

$$\hat{\theta} = f(X, Y), \tag{4.7}$$

where X is a matrix of observed exogenous variables and Y is a matrix of observed endogenous variables. This system (4.7) pairs up with (2.1) in obvious fashion with $\lambda = \hat{\theta}$ and $\omega \equiv [X, Y]$. Our interest here is to determine the potential sensitivity of $\hat{\theta}$ with respect to perturbations in X or Y or both, that is, the conditioning of $\hat{\theta}$ with respect to $\omega = [X, Y]$.

In the case when numerical problems are an issue, interest often attaches to the sets

$$\Omega = \{\delta\omega \equiv [\delta X, \delta Y] : \|\delta X\|/\|X\| < m_1,\ \|\delta Y\| = 0\}$$

and

$$\Lambda^* = \{\delta\lambda \equiv \delta\hat{\theta} : \|\delta\hat{\theta}\|/\|\hat{\theta}\| < m_2\},$$

where m_1 is usually chosen quite small, say 0.01, and m_2 usually larger, perhaps 0.05 or 0.1. To determine whether $\hat{\theta}$ is ill conditioned with respect to Ω we consider $m \equiv \sup_{\delta\omega \in \Omega}\|\delta\hat{\theta}\|/\|\hat{\theta}\|$. If $m > m_2$, then $\hat{\theta}$ is ill conditioned with respect to Ω.

With regard to shifts in Y, Ω might be chosen as

$$\Omega = \{\delta\omega \equiv [\delta X, \delta Y] : \|\delta Y\|/\|Y\| \leq m_1,\ \|\delta X\| = 0\},$$

where m_1 determines a region relative to the stochastically-generated δY so that Prob$(\Omega) = 0.95$, or some other such probabilistically defined region. This choice of Ω suggests a way of viewing the intimate relation known to exist between ill conditioning and high-variance estimates.

Criterion conditioning

Parameters θ are often estimated by minimizing some criterion function of the data and parameters. Let this function be $Q(X, Y, \theta)$, where X is exogenous, Y endogenous and $\theta \in \Theta$ is the vector of parameters to be estimated and Θ its domain. Suppose $\hat{\theta}$ is selected to satisfy

$$Q(X, Y, \hat{\theta}) = \inf_{\theta \in \Theta} Q(X, Y, \theta). \tag{4.8}$$

Frequently employed criteria $Q(X, Y, \theta)$ include

(i) $(y - f(X, y, \theta))^{\mathrm{T}}(y - f(X, y, \theta))$, for ordinary least squares estimates,

(ii) $-\log L(\boldsymbol{\theta})$, where $L(\boldsymbol{\theta})$ is the likelihood function of $\boldsymbol{\theta}$ for maximum-likelihood estimates, and

(iii) $\rho(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta})$, for M-estimation, where ρ is some function chosen for its robustness properties.

In each case, it is desirable that large changes $\delta\hat{\boldsymbol{\theta}} \equiv \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}$ from $\hat{\boldsymbol{\theta}}$ should be detectable by the selected criterion function, for failure to do so would indicate a poor determination of the parameter estimate. The following perturbation sets Ω and Λ^* are therefore of interest:

$$\Omega = \{\delta\boldsymbol{\theta} : \|\delta\hat{\boldsymbol{\theta}}\| = m_1 > 0\} \tag{4.9}$$

and

$$\Lambda^* = \left\{\delta Q : \|\delta Q\| / \left(\inf_{\delta\hat{\boldsymbol{\theta}} \in \Omega} \|\delta Q\|\right) \leq m_2\right\}. \tag{4.10}$$

If $\|\hat{\boldsymbol{\theta}}\| \neq 0$, then $\|\delta\hat{\boldsymbol{\theta}}\|$ in (4.9) might be more meaningfully replaced by $\|\delta\hat{\boldsymbol{\theta}}\| / \|\hat{\boldsymbol{\theta}}\|$. Note that the denominator in the definition of Λ^* should not be $\|Q(\boldsymbol{X}, \boldsymbol{Y}, \hat{\boldsymbol{\theta}})\|$, since $Q(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\theta})$ itself could be replaced by $Q(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\theta}) + c$, where $c > 0$ is an arbitrary constant, without changing $\hat{\boldsymbol{\theta}}$ or δQ. In this case the constant m_1 would be difficult to assign meaningfully since its value would depend on c. Of course the denominator could be set equal to one if absolute perturbations in the criterion were deemed important.

The appeal of Λ^* as defined in (4.10) lies in its assessing the effect of any perturbation only relative to the worst possible effect. That is, the worst case becomes a standard of acceptability. If, however, perturbations $\delta\hat{\boldsymbol{\theta}}$ in Ω produce δQ's which are not in Λ^*, then there must exist $\boldsymbol{\theta}^*$ that differ substantially from $\hat{\boldsymbol{\theta}}$ but which are relatively indistinguishable from $\hat{\boldsymbol{\theta}}$ by criterion $Q(\cdot)$. If $\inf_{\delta\hat{\boldsymbol{\theta}} \in \Omega} \|\delta Q\| = 0$, then $\boldsymbol{\theta}$ could be said to be *inestimable* with respect to this criterion and these data in those directions $\delta\hat{\boldsymbol{\theta}}$ for which $\|\delta Q\| = 0$.

As an example, consider

$$Q(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\theta}) = -\log L(\boldsymbol{\theta}).$$

Here $\delta Q = -\log L(\hat{\boldsymbol{\theta}} + \delta\hat{\boldsymbol{\theta}}) + \log L(\hat{\boldsymbol{\theta}})$, so that δQ represents the drop in log-likelihood due to setting $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}} + \delta\hat{\boldsymbol{\theta}}$ over $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, or the log of the likelihood ratio statistic of $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ versus $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}} + \delta\hat{\boldsymbol{\theta}}$. When ill conditioning of the maximum-likelihood estimate occurs according to Ω and Λ^* defined as above, there must be some perturbation $\delta\hat{\boldsymbol{\theta}}_1 \in \Omega$, say, for which the log-likelihood ratio can numerically distinguish the difference between $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}} + \delta\hat{\boldsymbol{\theta}}_1$ better than it can the difference between $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}} + \delta\hat{\boldsymbol{\theta}}_2$ for some other perturbation $\delta\hat{\boldsymbol{\theta}}_2 \in \Omega$.

More generally, to determine the criterion conditioning for the Λ^* defined as above, one must evaluate

$$\sup_{\delta\hat{\boldsymbol{\theta}} \in \Omega} \|\delta Q\| / \inf_{\delta\hat{\boldsymbol{\theta}} \in \Omega} \|\delta Q\|. \tag{4.11}$$

Since this is rarely easily done, we approximate this quantity to provide a rough guide to the possibility of ill conditioning. The first few terms of a Taylor

expansion of $Q(X, Y, \hat{\theta} + \delta\hat{\theta})$ about $\hat{\theta}$ (assuming that $\| \delta\hat{\theta} \|$ or m_1 is suitably small) yields

$$\begin{aligned} \delta Q &= Q(X, Y, \hat{\theta} + \delta\hat{\theta}) - Q(X, Y, \hat{\theta}) \\ &\simeq \delta\hat{\theta}^{\mathrm{T}}(\partial Q(X, Y, \theta)/\partial\theta) \,|_{\theta=\hat{\theta}} + \delta\hat{\theta}^{\mathrm{T}}(\partial^2 Q(X, Y, \theta)/\partial\theta\partial\theta^{\mathrm{T}}) \,|_{\theta=\hat{\theta}} \delta\hat{\theta}. \end{aligned} \tag{4.12}$$

For those many criterion functions for which the first-order term is zero at $\theta = \hat{\theta}$, (4.12) becomes

$$\delta Q \simeq \delta\hat{\theta}^{\mathrm{T}} A \delta\hat{\theta}, \tag{4.13}$$

where A is the Hessian matrix of Q with respect to θ at $\theta = \hat{\theta}$. [3] Using (4.13) in (4.11) and recalling the well-known (e.g. Rao [9]) extremum properties of eigenvalues of symmetric matrices, we see that (4.11) is equivalent to $\kappa(A)$, the condition number of the matrix A. In practice, then, $\kappa(A)$ may be used as a rough guide to assess the criterion conditioning.

5. Data, estimator, and criterion conditioning: the linear case

The three kinds of conditioning given above are easily developed for the linear model

$$y = X\beta + \varepsilon \tag{5.1}$$

with β estimated by ordinary least squares (OLS). Here y and ε are $n \times 1$, β is $p \times 1$, and X is an $n \times p$ matrix of rank p. A single number, $\kappa(X)$, is seen to be central to assessing all three forms of conditioning in this OLS/linear case — a coincidence that is not generally true. The quantity $\kappa(X)$ is the condition number of the matrix X and is defined to be the ratio of the largest to smallest singular values of X.

We first present an inequality important to the conditioning analyses that follow. Let $b = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$ be the OLS estimator of β, $\hat{y}$ be the OLS fitted values, and R be the uncentered multiple correlation coefficient of y regressed on X. Further, let δX, δb, and δy represent perturbations in X, b, and y, respectively. Both X and $X + \delta X$ are assumed of full rank, this latter assured if $\| \delta X \| / \| X \| < \kappa^{-1}(X)$ (Hanson and Lawson [7]). Then, from Golub and Van Loan [5],

$$\| \delta b \| / \| b \| \leq \kappa(X) R^{-1} \left[2 + (1 - R^2)^{1/2} \kappa(X) \right] \nu + \mathrm{O}(\nu^2), \tag{5.2}$$

where $\nu = \max(\| \delta y \| / \| y \|, \| \delta X \| / \| X \| < \kappa^{-1}(X))$. Exact equality is possible and does not depend upon $\kappa(X)$.

We now examine in this OLS/linear context the three forms of conditioning introduced in Section 2.

[3] This is, of course, the sample information matrix if $Q = -\log L(\hat{\theta})$.

Data conditioning

In (5.1), the data conditioning of interest is that of the data matrix X. Let $X^+ \equiv (X^TX)^{-1}X^T$ denote the Moore–Penrose inverse of X. Take Ω and Λ^* as in (4.5) and (4.6) and, for any ω, define λ by $\lambda \equiv X\omega$, so that ω and λ also obey $\omega = X^+\lambda$. Applying inequality (5.2) to this latter equation (with only λ perturbed) yields $\|\delta\lambda\|/\|\lambda\| \geq \frac{1}{2}\kappa^{-1}(X)\|\delta\omega\|/\|\omega\|$, where we note in this case that R, the uncentered multiple correlation of λ regressed on X, in (5.2) equals 1. From this inequality, which may be an equality for certain X and λ, it is readily seen that, if $\kappa(X)$ is large, the X data will be ill conditioned with respect to Ω.

Estimator conditioning

The estimator to be examined is the OLS $b = X^+y$. Two ω's are of immediate interest to perturb, namely $\omega = X$ and $\omega = y$. A third, $\omega = E(y) = X\beta$, will also be considered in this section.

$\omega = X$: As usual in a conditioning analysis, three items must be specified: the sets Ω and Λ^*, and the relation $\lambda = f(\omega)$ (from which we get $\delta\lambda \equiv g(\delta\omega)$). When X is perturbed, these quantities are taken to be

$$\Omega = \{\delta X : \|\delta X\|/\|X\| \leq m_1\}, \tag{5.3}$$

$$\Lambda^* = \{\delta b : \|\delta b\|/\|b\| \leq m_2\}, \tag{5.4}$$

and

$$\delta b = g(\delta X) = (X + \delta X)^+ y - X^+ y. \tag{5.5}$$

In practice, choices of 0.01 for m_1 often prove reasonable, as do values for m_2 of approximately $20m_1$. If the range Λ of δb given by (5.5) based on the Ω of (5.3) contains any element not in Λ^*, the OLS estimate is ill conditioned with respect to Ω. Thus, if small relative changes in the X matrix can produce large relative changes in the estimate, the estimate is said to be ill conditioned. To determine whether the OLS estimate is ill conditioned in any particular instance, we must calculate

$$\sup_{\delta X \in \Omega} \|\delta b\|/\|b\|. \tag{5.6}$$

Should this quantity be larger than m_2, then $\Lambda \not\subset \Lambda^*$, and b is ill conditioned. In practice (5.6) is not easily evaluated, but, by (5.2) with only X perturbed, it is known to be bounded from above by

$$m_1\kappa(X)R^{-1}\left[2 + (1 - R^2)^{1/2}\kappa(X)\right], \tag{5.7}$$

and may in fact be equal to (5.7) for some X, δX, and y. Thus, as a rough guide to the conditioning of b with respect to Ω, the quantity $2m_1\kappa(X)R^{-1}$ could be compared to m_2. If it is much larger than m_2, then b will be said to be ill conditioned. This guide is particularly good (and could become simply $2m_1\kappa(X)$) when the fit is good (i.e., when R is near unity), but could understate the extent

of ill conditioning when the fit is poor and the $\kappa^2(X)$ term dominates. In any event, we note that $\kappa(X)$ is an important multiplicative factor, it being possible, for example, that a 1 percent ($m_1 = 0.01$) relative change in X could produce a $\kappa(X)$ percent change in $\|\delta b\|/\|b\|$. For this reason the condition number has been given much attention in Belsley, Kuh and Welsch [4].

$\omega = y$: The second quantity to be perturbed is y, where Ω, Λ^* and $g(\delta y)$ are as follows:

$$\Omega = \{\delta y : \|\delta y\|/\|y\| \leq m_1\}, \tag{5.8}$$

$$\Lambda^* = \{\delta b : \|\delta b\|/\|b\| \leq m_2\}, \tag{5.9}$$

$$\delta b = g(\delta y) = X^+ \delta y. \tag{5.10}$$

By proceeding in a manner entirely analogous to that above, the relevant bound again becomes $2m_1\kappa(X)R^{-1}$, in which $\kappa(X)$ remains an important factor.

$\omega = E(y)$: Consider now a third perturbation for this estimator, again involving y. This time, however, we do not perturb y about its observed value but rather about its theoretical, or expected, value, namely $E(y) = X\beta$. Since the perturbations are taken about $X\beta$, it is reasonable, when constructing the set Ω, to take into account elements of the stochastic mechanism that generates y. That is, δy could be taken to be equal to a possible ε of (5.1). In this case

$$\Omega = \{\delta y : \|\delta y\|/\|X\beta\| \leq m_1\} \tag{5.11}$$

and m_1^{-1} could be chosen to be the minimum 'signal-to-noise' ratio expected to be encountered in the model (5.1). The set Λ^* is now taken to be

$$\Lambda^* = \{\delta b : \|\delta b\|/\|\beta\| \leq m_2\}. \tag{5.12}$$

To distinguish this basis for a conditioning analysis from the others, we call it a *stochastically-based* conditioning analysis.

Since y has been taken to be $X\beta$ and $y + \delta y = X\beta + \varepsilon$, we have

$$\delta b = X^+(y + \delta y) - X^+ y = X^+ \delta y = \hat{\beta} - \beta, \tag{5.13}$$

where $\hat{\beta}$ is the least squares estimate based on the realization $X\beta + \varepsilon$. Therefore, $\|\delta b\|^2 = (\hat{\beta} - \beta)^{\mathrm{T}}(\hat{\beta} - \beta)$ is a squared error, and determining whether b is ill conditioned in this situation also determines whether, for probable realizations of ε, the maximal squared error of the resulting OLS estimator is less than some amount $m_2\|\beta\|$. A guide to such an occurrence again results from applying (5.2) to (5.13), yielding ($R = 1$ here)

$$\begin{aligned}\|\delta b\|/\|\beta\| &= \|\hat{\beta} - \beta\|/\|\beta\| \leq 2\kappa(X)\|\delta y\|/\|X\beta\| \\ &= 2\kappa(X)\|\varepsilon\|/\|X\beta\|.\end{aligned} \tag{5.14}$$

The last quantity in (5.14), for $\delta y \in \Omega$, is less than $2m_1\kappa(X)$. Again $\kappa(X)$ is the dominant consideration in determining the conditioning.

Criterion conditioning

In OLS, the criterion to be minimized is

$$Q(X, Y, \beta) = (y - X\beta)^{\mathrm{T}}(y - X\beta). \tag{5.15}$$

Here the Taylor series expansion of (4.12) is exact, so that (4.11) can easily be evaluated. Taking Ω and Λ^* as in (4.9) and (4.10), where now $\theta = \beta$ and $\hat{\theta} = \hat{\beta} = X^+ y$, and $\delta Q = g(\delta\hat{\beta}) = 2\delta\hat{\beta}^T(X^TX)\delta\hat{\beta}$, we note that the criterion (5.15) is ill conditioned with respect to Ω if $\kappa^2(X)$ is greater than m_2. Once again, it is $\kappa(X)$ that provides information on the ill conditioning in the OLS/linear case.

6. Nonlinearities

So far we have given a general definition of conditioning, exemplified it quite generally in three forms (data, estimator, and criterion conditioning), and shown that the condition number coincidentally provides important information for assessing all three types of conditioning in the special OLS/linear case. The simple expedient of a condition number is not, however, always available. While some forms of nonlinearities (see, for example, Belsley [1]) can be normalized so that their conditioning admits of a similar analysis, this is readily seen not to be true for all forms of nonlinearities.

The several simple examples that follow well illustrate the types of problems that can arise and the divergences among the different types of conditioning that can occur when nonlinearities (in variables and/or parameters) are allowed.

Consider first the two orthogonal vectors $x_1 = (1, 1, 1, 1)^T$ and $x_2 = (-1, 1, -1, 1)^T$ which, for the analysis at hand, are considered to be the structurally interpretable, and therefore basic, data. These basic data would be very suitable for estimating the linear (in both parameters and variables) model

$$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon, \tag{6.1}$$

but would be useless if the model were

$$y = \beta_1 x_1 + \beta_2 x_2^2 + \varepsilon; \tag{6.2}$$

they would again become suitable for estimating

$$y = \beta_1 x_1^2 + \beta_2 x_2^3 + \varepsilon, \tag{6.3}$$

but not for

$$y = \beta_1 x_2 + \beta_2 x_1 x_2 + \varepsilon. \tag{6.4}$$

From this we see that, in the assessment of estimator conditioning when there are nonlinearities, both the data and the nature of the model must be considered. The divergence between the perfect basic-data conditioning and the ill conditioning of the OLS estimators of models (6.2) and (6.4) is clear.

Further problems can arise when there are nonlinearities in the parameters. Consider

$$y = \beta_1 x_1 + \beta_2 x_2^\alpha + \varepsilon. \tag{6.5}$$

For the basic data x_1 and x_2 as above, it is clear that these data might be suitable for estimation if $\alpha = 1$, but could be problemful if $\alpha = 2$. Unfortunately, α must

be estimated, and hence any measure of the suitability of the x series must depend on an estimate, a, of α. This is further complicated by the fact that $\alpha = 1$ does not guarantee that a will be near 1. Indeed, in most cases there will still be a non-zero probability that a will be arbitrarily close to 2, leaving one with little hope for correctly assessing the conditioning, say, by using the condition number of $X = [x_1, x_2^a]$. In such cases, assessment of conditioning may require the introduction of prior information on α.

7. An example of a computational alternative

In the OLS/linear case of Section 5, it is possible to provide a mathematical solution to the conditioning problem – $\kappa(X)$, for example, figures prominently in a theoretically derived bound on the sensitivity of $\|\delta b\|/\|b\|$ to relative changes in X or y. Such solutions are not generally available: either the mathematics becomes too cumbersome or the breadth of perturbations deemed relevant invalidates simple approximations. In this section we suggest and exemplify a more generally applicable computational alternative.

A computational alternative

We recall from Section 2 that any conditioning analysis consists in determining the triple $K = \{f, \Omega, \Lambda^*\}$, where from $\lambda = f(\omega)$ we derive $\delta\lambda = g(\delta\omega)$ which relates responses $\delta\lambda$ to perturbations $\delta\omega$, Ω is the set of a priori 'reasonable' perturbations, and Λ^* is the set of a priori 'reasonable' responses. The objective is to determine whether reasonable perturbations $\delta\omega \in \Omega$ can result in unreasonable responses $\delta\lambda \notin \Lambda^*$. A straightforward, albeit computer intensive, way of conducting such an analysis is randomly to select elements $\delta\omega$ from Ω, calculate their corresponding $\delta\lambda = g(\delta\omega)$ and check if any fall outside Λ^*.

This procedure has many advantages. First, it is universally applicable. So long as the conditioning triple K can be defined, the method can be employed. Second, for highly ill conditioned problems, a few random picks for $\delta\omega$ from Ω should suffice to find a response $\delta\lambda$ lying outside Λ^*. Our experience supports this. Third, it allows for complete generality in the way perturbations are defined. This last point is especially relevant in practice. In the mathematical solution to the conditioning of the OLS/linear case given in Section 5, it was necessary to assume that all perturbations are in terms of relative shifts $\|\delta\omega\|/\|\omega\|$ and that the appropriate measure is a vector norm. Such perturbations will often lack structural interpretability. More likely, the appropriate structurally interpretable perturbations will differ for each element of ω, as we see in the example that follows. Fourth, in analyzing nonlinear models g, this computational alternative does not rely on any linear approximations. Thus it can properly accommodate perturbations $\delta\omega$ that are considered reasonable relative to the problem at hand ($\delta\omega \in \Omega$) but which would be unreasonably large for a Taylor approximation to hold. Finally, the procedure measures more than conditioning as formally defined

in numerical analysis (i.e., the sensitivity of the exact solutions to a system of equations), for the observed sensitivities also contain elements of algorithmic stability. From the practical perspective of assessing the reliability of a given solution, this is as it should be.

The method, of course, has drawbacks. Like any computationally intensive procedure, it can be expensive. Of greater import is the fact that, whereas the method seems readily to show the presence of ill conditioning, it cannot (without a complete examination of Ω) demonstrate the absence of ill conditioning. At best, after many draws from Ω, a reasonable presumption may be allowed to the statement that ill conditioning is absent.

The consumption function

As an illustration of the method suggested above, consider a conditioning analysis of a nonlinear formulation of the U.S. personal-consumption function. The model f is given as

$$C_t = a \cdot C_{t-1}^{b_1} \cdot \mathrm{DPI}_t^{b_2} \cdot r_t^{b_3} \cdot (\mathrm{DPI}_t/\mathrm{DPI}_{t-1})^{b_4} + \varepsilon_t, \tag{7.1}$$

where all observations are annual 1948 to 1974 and C_t is U.S. consumption in billions of 1958 dollars, DPI_t is U.S. disposable personal income in billions of 1958 dollars, r_t is the interest rate in percent (Moody's Aaa).

Here the consumption function used in [4] is given a Cobb–Douglas form with an additive error. This function differs from a 'linear in the logs' formulation commonly employed in econometric analysis only in the specification of an additive rather than a multiplicative error. This alteration, however, makes (7.1) an essentially nonlinear function, one incapable of simple transformation into a form amenable to linear estimation and linear conditioning analysis. [4] Its lack of use in econometric studies is less on the grounds of economic plausibility than the added complications introduced through requiring nonlinear estimation.

To complete the specification of a conditioning analysis, we must give Ω and Λ^*. This is readily done here, since each of these economic time series is in a structurally interpretable form. That is, their magnitudes, and in particular, changes in their magnitudes, can be meaningfully assessed as being large or small through our knowledge of the underlying economic phenomena they measure.

Thus, we determine that perturbations δC_t and $\delta \mathrm{DPI}_t$ that are within ± 0.1 percent (± 0.001) of C_t and DPI_t, respectively, are small. Not only are such magnitudes of little macroeconomic consequence, but they would be perceived by most economists as lying within the bounds of measurement error. Relative perturbations make sense in this context. By contrast, we assume perturbations δr_t that lie in an interval of one basis point (± 0.05 of a percentage point) are reasonably considered small in measuring interest rates. Here we are assuming

[4] Had the error been multiplicative, the standard OLS/linear methods of [4] could be used with the modifications given in [1].

Table 1
Extreme responses to random perturbations in Ω of nonlinear consumption function (7.1)

Coefficient (i.e. λ)	Range of $\delta\lambda$	Largest percent increase	Largest percent decrease
$a = 0.975$	0.064	3.8	2.8
$b_1 = 0.130$	0.081	24.6	37.8
$b_2 = 0.867$	0.080	5.2	4.0
$b_3 = -0.022$	0.012	32.1	22.9
$b_4 = 0.097$	0.081	44.4	39.3

additive perturbations make sense. Thus, Ω is chosen as

$$\Omega = \{(\delta C_t, \delta \mathrm{DPI}_t, \delta r_t) \forall t: \delta C_t \in \pm 0.1\% \text{ of } C_t, \delta \mathrm{DPI}_t \in \pm 0.1\% \text{ of } \mathrm{DPI}_t, \delta r_t \in \pm 0.05\}. \quad (7.2)$$

To pick Λ^*, we merely state that a relative response to such perturbations by any coefficient estimate in excess of 10% is too large (e.g., would yield a substantively different policy analysis.)

The sensitivity analysis is now straightforward. First estimate (7.1) with nonlinear least squares (NLS) using the basic data $\omega = [C, \mathrm{DPI}, r]$ to obtain base estimates, $\boldsymbol{b}$. Repeatedly re-estimate (7.1) with perturbed data $\omega + \delta\omega$ determined by random draws $\delta\omega$ from Ω given in (7.2). This can be accomplished through uniform selections from

$$\delta C_t \sim U(0.999C_t, 1.001C_t),$$

$$\delta \mathrm{DPI}_t \sim U(0.999\mathrm{DPI}_t, 1.001\mathrm{DPI}_t),$$

$$\delta r_t \sim U(-0.05, 0.05).$$

Each re-estimation produces a new estimate $\boldsymbol{b}^*$, and a resulting $\delta\boldsymbol{b} = \boldsymbol{b}^* - \boldsymbol{b}$. Our interest centers on whether any of the $\delta b_i / b_i$ fall outside the 10 percent level chosen for Λ^*. Table 1 shows the extreme results over 30 replications for each of the parameters of (7.1).

The base estimates $\boldsymbol{b}$ are shown in column 1. These estimates are completely compatible with the estimates of the analogous linear model analyzed in [4]. Column 2 shows the range of the perturbed estimates over the 30 replications, and columns 3 and 4 show, respectively, the largest percent increase and the largest percent decrease for the particular coefficient. It is clear that the 10 percent target level of Λ^* has been exceeded in both directions for b_1, b_3, and b_4, and is almost met on an overall basis for b_2. Only the constant (or scale factor) a seems relatively stably determined.

These results are wholly consonant with the conditioning analysis given in [4] to the analogous linear model of the consumption function. Indeed the same patterns of instability are exhibited. Thus, for example, we can plot scatter diagrams showing how the instability in the estimate of one coefficient relates to that of another over the different perturbations. Such scatter plots are given in

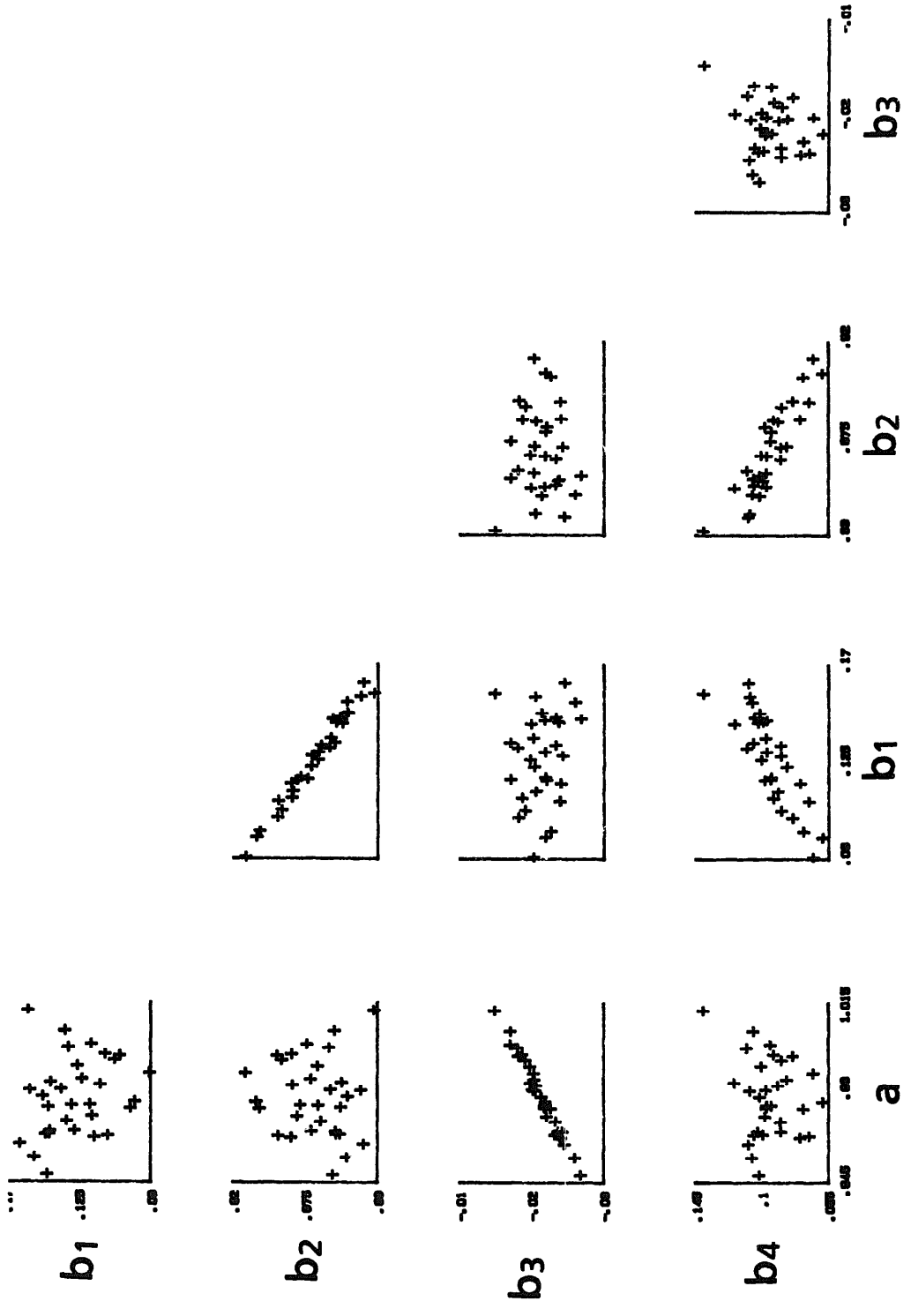

Fig. 1. Scatterplot matrix of the coefficients.

Table 2
Variance-decomposition proportions and condition indexes for consumption-function data, with linear model taken from [4]

Condition index	Const. var(a)	C_{t-1} var(b_1)	DPI_t var(b_2)	r_t var(b_3)	$\Delta\mathrm{DPI}_t$ var(b_4)
1	0.001	0.000	0.000	0.000	0.001
4	0.004	0.000	0.000	0.002	0.136
8	0.310	0.000	0.000	0.013	0.000
39	0.264	0.004	0.004	0.984	0.048
376	0.420	0.995	0.995	0.000	0.814

Figure 1. The tight dependency pairs (b_1 and b_2), (a and b_3), (b_1 and b_4), and (b_2 and b_4) draw immediate attention, each showing that instability in the estimate of one of the pair tends to be accompanied by covariant instability in the other. These scatter plots provide useful auxiliary information to a conditioning analysis and are, in this day and age, quickly acquired. They tell similar information for two-dimensional relations that one gets from the variance-decomposition proportion matrices of the linear analysis of [4]. Indeed, for comparison, this matrix for the linear consumption function of [4] is given in Table 2.

Some further thoughts on analyzing the Jacobian matrix

Unlike the variance-decomposition proportions matrices, two-dimensional scatter plots can (but need not) overlook joint dependencies involving three or more parameter estimates. We found it of interest, therefore, to pursue the suggestion motivated in [4] that the variance-decomposition matrix derived from the Jacobian matrix of the particular nonlinear model with respect to its parameters (here, the Jacobian of (7.1) with respect to $\boldsymbol{b}$) be used to analyze the composition of more involved dependencies.

It is unnecessary to reproduce the matrix that results from analyzing the Jacobian here because it is virtually identical to Table 2. The analysis of the Jacobian, then, holds promise, but it is premature to claim too much for it. As indicated in Section 5, the proper use of this Jacobian requires knowledge of the actual, not the estimated, parameters. In practice, of course, this cannot be. Use of the estimated Jacobian (one whose derivatives are based on estimated coefficients) runs the very real danger of basing a diagnostic of the extent to which the estimates of particular coefficients are ill conditioned on those possibly ill conditioned estimates themselves.

Furthermore, the condition indexes one obtains for this Jacobian have questionable value. In general they depend upon the units in which the basic variables are measured, and there is to date no uniform normalization that allows for a stable interpretation of these indexes outside the linear case of [4] and that dealt with in [1]. Still, if one feels the estimates are reasonable, the information from the variance-decomposition proportions matrix of the Jacobian can offer valuable

complementary hints as to the nature of the ill conditioning, hints that could then be tested for directly, using variations on the computer-oriented technique described above.

Fortunately, one need not bank on the value of the Jacobian here, for scatter plots like those of Table 1 offer much relevant information that is relatively easily and directly obtained. In addition, the scatter plots can provide visual indications of nonlinear dependencies (as possibly between b_3 and b_4) and bifurcated dependencies (as between b_1 and b_4 or between b_2 and b_4) that could never be seen in a table of variance-decomposition proportions.

8. Conclusion

A conditioning analysis is a sensitivity analysis carefully constructed to guarantee its results relate meaningfully to the problem at hand (through the selection of Ω and Λ^*). Such a conditioning analysis can be directed at many interesting elements of a given statistical analysis, as exemplified by data, estimator, and criterion conditioning.

In some circumstances, conditioning can be assessed mathematically, through the derivation of some measure that bounds the potential sensitivity. This is seen to be the case for the OLS/linear problem, for which the condition number $\kappa(X)$ applied to structurally interpretable data conveniently provides the needed measure for all three: data, estimator, and criterion conditioning.

In more general (for example, nonlinear) contexts, however, such a mathematically derived bound need not be forthcoming. But, in these cases it should always be possible to investigate any form of conditioning empirically, and a method for doing so is illustrated in the context of analyzing the estimator conditioning of a nonlinear version of the consumption function.

One way or another, then, it should always be possible to examine whether, for example, trivial changes in the inputs of a statistical analysis can produce substantive alterations in important outputs of the analysis. To us, such conditioning analyses and their resultant information are an important adjunct in interpreting the reliability of a statistical study.

References

[1] D.A. Belsley, Conditioning in models with logs, Technical Report, TR-41, Center for Computational Research in Economics and Management Science, Massachusetts Institute of Technology, Cambridge, MA (1983).

[2] D.A. Belsley, Demeaning conditioning diagnostics through centering, and Reply, *Amer. Statist.* **38** (1984) 73–77 and 90–93.

[3] D.A. Belsley, Centering, the constant, first differencing and diagnosing collinearity, in: D.A. Belsley and E. Kuh (Eds.), *Model Reliability* (MIT Press, Cambridge, MA, 1986).

[4] D.A. Belsley, E. Kuh, and R.E. Welsch, *Regression Diagnostics: Identifying Sources of Influential Observations and Collinearity* (John Wiley and Sons, New York, 1980).

[5] G. Golub and C.F. Van Loan, *Matrix Computations* (Johns Hopkins University Press, Baltimore, MD, 1983).
[6] R. Gunst, Comment: Toward a balanced assessment of collinearity diagnostics, *Amer. Statist.* **38** (1984) 79–82.
[7] R.J. Hanson and C.L. Lawson, Extensions and applications of the Householder algorithm for solving linear least squares problems, *Math. Comput.* **23** (1969) 787–812.
[8] P.J. Huber, *Robust Statistics* (John Wiley and Sons, New York, 1981).
[9] C.R. Rao, *Linear Statistical Inference and Its Applications, 2nd Edition* (John Wiley and Sons, New York, 1973).
[10] R.A. Thisted, Comment, *J. Amer. Statist. Assoc.* **75** (1980) 81–86.