Interval Least Squares — a Diagnostic Tool

David M. Gay

AT&T Bell Laboratories
Murray Hill, New Jersey 07974

ABSTRACT

Linear least-squares models are often used to describe how an endogenous variable depends on some exogenous variables and to make forecasts based on that description. In using such models, one should consider various sources of uncertainty in the parameter estimates and forecasts. Often there are uncertainties in the exogenous variables. When these uncertainties are confined to small intervals and symmetrically distributed, interval least-squares estimates of the model's parameters can furnish either bounds on the component of parameter estimation or forecast error contributed by errors in the exogenous variables, along with an assurance that there is little bias, or else a warning that linear least-squares estimates and forecasts may be subject to significant bias. Indeed, one byproduct is an index of nonlinearity that can warn of possible bias; a similar measure is available from singular value analysis. On problems where bias is insignificant, interval least-squares solutions provide a more detailed collinearity diagnostic than the scaled condition number espoused in the book of Belsley, Kuh, and Welsch.

Introduction

Often one observes, say, n instances $y_1, \cdots, y_n$ of an endogenous (dependent) variable y and corresponding values $X_{i,j}$, $1 \le i \le n$, $1 \le j \le p$ of some exogenous (independent) variables $X_1, \cdots, X_p$. Sometimes it is convenient to assume these variables are related by

$$y = X\beta + e,$$

where the components of $e = y - X\beta$ are independently distributed with mean 0 and common variance σ. As many texts explain, when X has full rank, there is a unique b that minimizes $\|y - Xb\|_2$, i.e.,

$$b = X^{\dagger}y = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y, \tag{1}$$

which is the maximum likelihood estimate of β when e is normally distributed. (Here $\|\cdot\|_2$ is the Euclidean norm: $\|z\|_2 = \left(\sum_i z_i^2\right)^{1/2}$, and superscript T means "tran-

ISBN 0-12-505630-3

spose''.) Even without the assumption of normality, b is an unbiased estimate, i.e., the expected value of b is β.

Often one is interested in some linear forms (or functions of linear forms) involving β. In particular, models are often used to make forecasts: one predicts that if the independent variables had the value $x^{new} = (x_1^{new}, \cdots, x_p^{new})$, then the dependent variable $\beta^T x^{new} + e^{new}$ would have mean value $b^T x^{new}$. In general, under the above assumptions, if $c \in \mathbb{R}^p$ is an arbitrary p-vector, then $c^T b$ is an unbiased estimate of $c^T \beta$.

Unfortunately, the independent variables themselves are sometimes subject to error, e.g. measurement error. For example, we may only know $n \times p$ matrices $\underline{X}$ and $\overline{X}$ such that

$$\underline{X} \le X \le \overline{X}, \tag{2a}$$

where the inequalities are understood componentwise. Similarly, we may only know lower and upper bounds $\underline{y}$ and $\overline{y}$ (in $\mathbb{R}^n$) on the dependent variable y:

$$\underline{y} \le y \le \overline{y}. \tag{2b}$$

In this case (2) it is reasonable to consider estimating β by using $X := \frac{1}{2}(\underline{X} + \overline{X})$ and $y := \frac{1}{2}(\underline{y} + \overline{y})$ in (1). As many authors have noted (e.g., [8, 10, 18, 30, 31, 33]), this gives a biased estimate of β, and one should assess whether this bias may be large enough to be worrisome, e.g. to affect decisions about whether the current linear model is appropriate or to affect decisions based on forecasts derived from the model.

Linear models like (1), (2) that account for errors in the independent variables (X) are often called errors-in-variables models; I shall refer to the parameter estimation problem for such models as an errors-in-variables problem. The present work emphasizes an interval least-squares approach to this problem, but several other approaches are available. Hodges and Moore [18], for example, mention instrumental variables (which approach relies on additional data), maximum likelihood (which requires a distributional assumption), and grouping of observations (which requires enough data to group). Approaches of more recent interest include Golub and Van Loan's total least squares method [16, 17, 19] and the orthogonal distance regression treatment of Boggs, Byrd, and Schnabel [7], both of which I now discuss.

Figure 1 illustrates a least-squares fit to six data points:

$$X = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 5 & 1 \\ 6 & 1 \\ 9 & 1 \\ 10 & 1 \end{bmatrix}, \quad y = \begin{bmatrix} 2.5 \\ 1.5 \\ 3.5 \\ 4.5 \\ 7.5 \\ 6.5 \end{bmatrix}; \tag{3}$$

the dashed arrows emphasize that distance is measured vertically. Sometimes it is more appropriate to measure distance to the nearest point on the fitted line, as in Figure 2; this is a simple instance of orthogonal distance regression (ODR) [7], the idea of which is to find the smallest perturbations $\xi_{i,j}$ to $X_{i,j}$ and η_i to y_i such that $\sum_{j=1}^{p}(X_{i,j}+\xi_{i,j})b_j = y_i + \eta_i \quad (1 \le i \le n)$. In measuring these perturbations, one can specify separate scale factors $\delta_{i,0}$ and $\delta_{i,j}$ for each η_i and $\xi_{i,j}$; ODR then minimizes $\sum_i(\delta_{i,0}\eta_i)^2 + \sum_{i,j}(\delta_{i,j}\xi_{i,j})^2$.

For linear problems in which one is content with a rank 1 matrix of scale factors δ for (ξ, η), orthogonal distance regression reduces to total least squares (TLS), which has the advantage that b, ξ, and η can be computed from a single singular value decomposition. (See, e.g., [29] or [17] for discussion of the singular value decomposition and of QR factorizations, which come into play below.) The ODR algorithm in [7] is sufficiently general that it readily accommodates independent variables (columns of X) that are not subject to error. As hinted in [16] (in the note added in proof) and as Stewart's argument in the appendix of [30] shows, one can handle error-free variables in the TLS setting by "regressing out" the corresponding constant columns of X, i.e., projecting y and the error-prone columns of X onto the orthogonal complement of the column space of the constant columns. If X is partitioned as $X = [K,W]$, where K and W encompass, respectively, the constant ("*k*nown") and error-prone ("*w*iggly") columns, and if we factor

$$X = [K,\ W] = QR = [Q_K,\ Q_W] \begin{bmatrix} R_{KK} & R_{KW} \\ \mathbf{0} & R_{WW} \end{bmatrix}, \tag{4}$$

where Q is an orthogonal matrix ($Q^TQ = I$, the identity matrix) and Q and R are partitioned conformably with X, then, as indicated in [16], the "regressing out" amounts to applying the TLS procedure to matrix R_{WW} and right-hand side Q_W^Ty to get b_W (and, if desired, ξ and η), then solving $R_{KK}b_K = Q_K^Ty - R_{KW}b_W$. Figure 3 shows what

happens if we apply TLS to (3) directly, allowing perturbations in the column of ones. Figure 2 results from treating the column of ones as constant. (All scale factors $\delta_{i,j}$ are 1 in Figures 2 and 3. The above discussion assumes X to be of full rank and the relevant TLS problems to have solutions.)

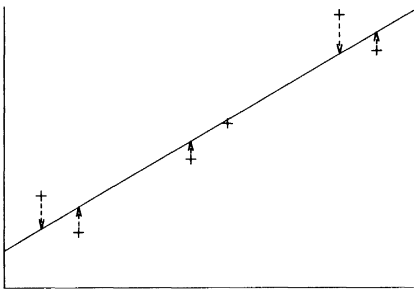

Fig. 1. OLS: $y = .603x + 1.017$

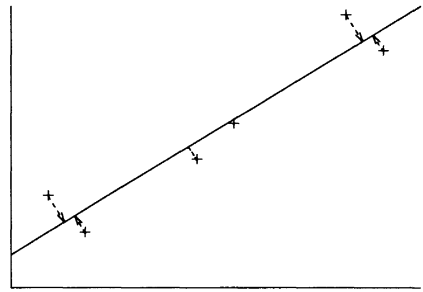

Fig. 2. ODR: $y = .624x + .902$

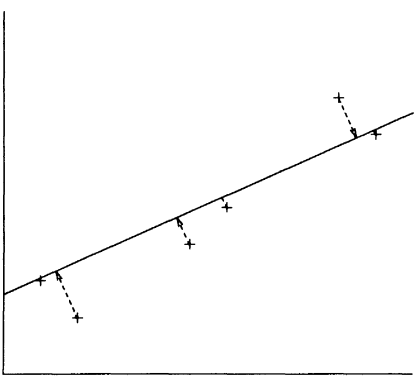

Fig. 3. TLS: $y = .452x + 2.164$

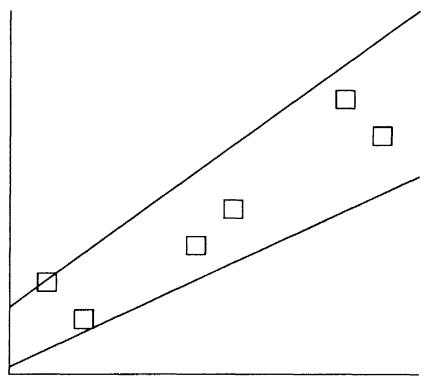

Fig. 4. ILS: $y = (.603 \pm .133)x + (1.017 \pm .812)$

Linearity

This paper is concerned with detecting cases when the errors (2) are large enough that decisions based on ordinary least squares parameter estimates may be unreliable. The diagnostics proposed below may be worth examining when one uses other parameter estimates, but this is a topic for further study.

By differentiating (1), it is straightforward to obtain a first-order estimate of the effects that the errors (2) in X and y can have on our linear least squares estimate b.

As shown in [18],

(5a) $$\frac{\partial b}{\partial X_{i,j}} = (X^{\mathrm{T}}X)^{-1}\left[r_i I_j - b_j X^{\mathrm{T}}_{i,*}\right]$$

and

(5b) $$\frac{\partial b}{\partial y_i} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}_{i,*},$$

where $X_{i,*}$ is the ith row of X,

(5c) $$r = y - Xb$$

is the residual vector, and I_j is the jth column of the identity matrix. Left-multiplying (5a) and (5b) by c^{T}, we obtain $\frac{\partial(c^{\mathrm{T}}b)}{\partial X_{i,j}}$ and $\frac{\partial(c^{\mathrm{T}}b)}{\partial y_i}$ and hence can compute first-order estimates of the effects on $c^{\mathrm{T}}b$ of a particular perturbation to X.

Using (5), it is straightforward to compute worst-case bounds on the effect that any perturbation (2) can have on $c^{\mathrm{T}}b$. Moreover, one of the diagnostics suggested by this paper arises from trying to compute rigorous worst-case bounds that are near the best possible ones. (The results of this attempt, applied to (3) with perturbations of $\pm.25$ to each component of y and of the first column of X are shown in Figure 4.) But such bounds are often irrelevant, as Inman [20] has pointed out and the following example illustrates. Suppose we have n real values f_i that satisfy

$$-1 \le f_i \le 1.$$

Then clearly

$$-n \le \sum_{i=1}^{n} f_i \le n.$$

If $n = 100$ and the f_i are independently and uniformly distributed over $[-1, 1]$, then

$$\Pr(\sum_{i=1}^{n} f_i \ge n/2) \approx 2.8\times10^{-7};$$

the standard deviation $\text{s.d.}(\sum_{i=1}^{n} f_i) = n^{1/2}/\sqrt{3} = 10/\sqrt{3}$ is likely to be much more relevant. And so it often happens that standard deviations are of greater interest than worst-case bounds.

In the linear case it is easy to relate worst-case bounds to standard deviations. If

f_i, $(1 \le i \le n)$ are independently and identically distributed over $[\underline{f}, \bar{f}]$ with standard deviation $\text{s.d.}(f_i) = \sigma$, then $s := \sum_{i=1}^{n} (d_i + a_i f_i) \in [\underline{s}, \bar{s}]$, where $\underline{s} = \sum_{i=1}^{n} (d_i + \min\{\underline{f} a_i, \bar{f} a_i\})$ and $\bar{s} = \sum_{i=1}^{n} (d_i + \max\{\underline{f} a_i, \bar{f} a_i\})$ are the best possible bounds on s. Then $\bar{s} - \underline{s} = (\bar{f} - \underline{f}) \sum_{i=1}^{n} |a_i|$ and $\text{s.d.}(s) = (\bar{s} - \underline{s}) \left(\frac{\sigma}{\bar{f} - \underline{f}} \right) \frac{\sqrt{\sum_i a_i^2}}{\sum_i |a_i|}$.

In nonlinear cases that are close to linear, we can similarly compute a good estimate of the standard deviation. Suppose $F: \mathbf{R}^n \to \mathbf{R}$ is mildly nonlinear. Then $\text{s.d.}(F(f)) \approx \Delta F \left(\frac{\sigma}{\bar{f} - \underline{f}} \right) \frac{\sqrt{\sum_i a_i^2}}{\sum_i |a_i|}$, where $a_i = \frac{\partial F}{\partial f_i}(f^\circ)$, $f^\circ := (\underline{f} + \bar{f})/2$, and $\Delta F := (\bar{f} - \underline{f}) \sum_{i=1}^{n} \left| \frac{\partial F}{\partial f_i}(f^\circ) \right|$. In such cases, if we can compute good bounds on $F(f)$, then their difference times $\left(\frac{\sigma}{\bar{f} - \underline{f}} \right) \frac{\sqrt{\sum_i a_i^2}}{\sum_i |a_i|}$ will often give a good upper bound on $\text{s.d.}(F(f))$. Thus, to some extent we can relate standard deviations to worst-case bounds.

Interval Notation

It is now convenient to introduce some interval notation. As above, inequalities involving vectors or matrices are understood componentwise. $\mathbb{I}(\cdot)$ denotes the interval hull of $(\cdot)$, i.e., the smallest Cartesian product containing $(\cdot)$, where $(\cdot)$ can be a scalar, vector, or matrix expression. $\tilde{\mathbb{I}}(\cdot)$ denotes a Cartesian product containing $\mathbb{I}(\cdot)$, computed by interval arithmetic (see, e.g., [24, 1]). Boldface letters denote interval vectors or matrices, as in

$$\mathbf{X} \equiv [\underline{X}, \bar{X}] = \{X : \underline{X} \le X \le \bar{X}\}$$

and

$$\mathbf{y} \equiv [\underline{y}, \bar{y}] = \{y : \underline{y} \le y \le \bar{y}\}.$$

It is convenient to use the natural extension of functions to the power sets of their domains: if $f: A \to B$ and $S \subset A$, then $f(S) \equiv \{f(s): s \in S\}$.

Effects of Nonlinearity

Despite the word "linear" in "linear least squares", the linear least-squares estimate (1) is a nonlinear function of X. The present work began as an attempt to specialize [15] to the linear least-squares problem, which would provide an enclosure of

$$B^* := \mathbf{X}^\dagger \mathbf{y} = \{X^\dagger y : X \in [\underline{X}, \overline{X}],\ y \in [\underline{y}, \overline{y}]\}. \tag{6}$$

As mentioned above, $c^T b$ is often of greater interest than b itself, where c is an arbitrary linear form, so it may be more relevant to compute an enclosure of

$$C^* := c^T \mathbf{X}^\dagger \mathbf{y} = \{c^T X^\dagger y : X \in [\underline{X}, \overline{X}],\ y \in [\underline{y}, \overline{y}]\}. \tag{7}$$

Since we can take columns of the identity matrix for c, there is no loss of generality in considering (7) rather than (6). Of course, we could compute an enclosure $\mathbf{C}$ of C^* by first computing an interval enclosure (i.e., Cartesian product of intervals) $\mathbf{B}$ of B^*, then enclosing $\{c^T b : b \in \mathbf{B}\}$. But this would deliver unduly pessimistic bounds because of *simultaneity*, i.e., because it would also bound $c^T b$ for many $b \notin B^*$.

We can get tighter bounds on $c^T b = c^T b^\circ + c^T(b - b^\circ)$ by bounding $c^T(b - b^\circ)$ for some nominal solution

$$b^\circ = X^{\circ\dagger} y^\circ, \tag{8a}$$

say from the midpoints of $\mathbf{X}$ and $\mathbf{y}$:

$$X^\circ = \tfrac{1}{2}(\underline{X} + \overline{X}) \tag{8b}$$

$$y^\circ = \tfrac{1}{2}(\underline{y} + \overline{y}). \tag{8c}$$

Interval Linear Equations

To assess the effects of nonlinearity on $c^T b$, we are often led to considering sets of interval linear equations, i.e., to computing (outer) approximations to sets of the form

$$\mathbb{I}(\mathbf{A}^{-1}\mathbf{v}) = \mathbb{I}(\{A^{-1}v : A \in [\underline{A}, \overline{A}],\ v \in [\underline{v}, \overline{v}]\}).$$

The computations reported below proceed as in [14]; for fuller discussions and extensive references, see Neumaier's papers [25, 26, 27]. For general interval linear equations, it is often essential to use a preconditioner M, i.e., to work with $\mathbb{I}(M\mathbf{A})$ and $\mathbb{I}(M\mathbf{v})$. An important, computationally useful measure of possible nearness to singularity is given by

(9) $$\rho := \sup \| \mathbb{I}(MA) - I \|.$$

Neumaier [25] has shown that if $\rho < 1$ is possible, then choosing $M := 2(\underline{A} + \bar{A})^{-1}$ gives the smallest ρ in (9) when $\|\cdot\|$ is a scaled max-norm, i.e.,

(10) $$\|u\| := \|Du\|_\infty = \max\{ |D_{ii}u_i| : 1 \le i \le p \}$$

for some nonsingular diagonal matrix $D \in \mathbf{R}^{p\times p}$ and all $u \in \mathbf{R}^p$. The choice of D can strongly affect computational results; in the computing reported below, D is chosen as described in §5 of [14].

Normal Equations

The most straightforward way to enclose C^* (see (7)) is by attacking the normal equations (1) directly: since

$$C^* = \{c^T(X^TX)^{-1}X^T(y - Xb^\circ): \underline{X} \le X \le \bar{X}, \underline{y} \le y \le \bar{y}\}$$

(11a) $$= \{c^TM(X^TXM)^{-1}X^T(y - Xb^\circ): \underline{X} \le X \le \bar{X}, \underline{y} \le y \le \bar{y}\},$$

it is tempting to work with (X^TXM) for some preconditioner M, say

(11b) $$M := (X^{\circ T}X^\circ)^{-1}.$$

However, the computational experience reported below suggests that an approximate QR factorization generally gives better bounds and succeeds in delivering good bounds in some cases where the normal-equations approach fails, i.e., where the latter approach finds $\rho \ge 1$ in (9).

***QR* Approach**

Suppose we compute the QR factorization

(12a) $$X^\circ = Q^\circ R,$$

where Q° is an orthogonal matrix and R is nonsingular. Let

(12b) $$Z := R^{-1}.$$

Then XZ is close to orthogonal when X is not too far from X°, and

$$(X^TX)^{-1}(y - Xb^\circ) = Z(Z^TX^TXZ)^{-1}(XZ)^T(y - Xb^\circ),$$

so we can express (7) as

(12c) $$C^* = \{c^T Z(Z^T X^T XZ)^{-1}(XZ)^T(y - Xb^\circ): X \in [\underline{X}, \overline{X}],\ y \in [\underline{y}, \overline{y}]\}.$$

We could approximate $\square(\{(Z^T X^T XZ)^{-1}(XZ)^T(y - Xb^\circ): \underline{X} \le X \le \overline{X},\ \underline{y} \le y \le \overline{y}\})$, then multiply the result by $c^T Z$, but unless $Z^T c$ is a scaled unit vector, i.e., a scaled column of the identity matrix, this would yield unduly pessimistic bounds because of the simultaneity problem described above. At least two cures are possible. We could replace Z by ZP, where P is an orthogonal matrix chosen to make $c^T ZP$ a multiple of a standard unit vector. But if we were interested in $c^T b$ for several values of c, this would be inefficient.

Perhaps a better way to circumvent the simultaneity problem (to first order) is as follows. Let $X \in [\underline{X}, \overline{X}]$ and $y \in [\underline{y}, \overline{y}]$ be fixed for a moment, and let $H := I - (XZ)^T(XZ)$, so that $b := X^\dagger y = \overline{Z}(I - H)^{-1}(XZ)^T y$. Since

$$(I - H)^{-1} = I + H(I - H)^{-1} = I + (I - H)^{-1}H,$$

we have

$$c^T b = c^T b^\circ + c^T Z(I - H)^{-1}(XZ)^T(y - Xb^\circ)$$

(13a) $$= c^T b^\circ + c^T Z(XZ)^T(y - Xb^\circ) + c^T ZH(I - H)^{-1}(XZ)^T(y - Xb^\circ)$$

(13b) $$= c^T b^\circ + c^T Z(XZ)^T(y - Xb^\circ) + c^T Z(I - H)^{-1}H(XZ)^T(y - Xb^\circ).$$

This suggests separately bounding

(14) $$\square(\{(ZZ^T c)^T X^T(y - Xb^\circ): X \in [\underline{X}, \overline{X}],\ y \in [\underline{y}, \overline{y}]\})$$

and

(15) $$\square(\{c^T Z(I - H)^{-1}H(XZ)^T(y - Xb^\circ): X \in \mathbf{X},\ y \in \mathbf{y},\ H = I - (ZX)^T(ZX)\}).$$

To bound (14), we incur no loss of accuracy if we separately bound the contribution to (14) of each row of $[\mathbf{X}, \mathbf{y}]$, then sum these bounds. To bound (15), we can add parentheses to (13) in half a dozen different ways, solve some interval linear equations, then do some interval arithmetic; but in all cases, it is convenient to start by computing an enclosure of

$$\square(\{Z^T X^T(y - Xb^\circ): X \in [\underline{X}, \overline{X}],\ y \in [\underline{y}, \overline{y}]\}).$$

Thus we are led to bounding sets of the form

(16) $$\square(\{(a^T x)(\gamma - b^T x): \underline{x} \le x \le \overline{x},\ \underline{\gamma} \le \gamma \le \overline{\gamma}\})$$

for a, b, $\underline{x}$, $\overline{x} \in \mathbb{R}^p$ and $\underline{\gamma}$, $\overline{\gamma} \in \mathbb{R}$. Since

$$(17a) \qquad (a^T x)(\gamma - b^T x) = (a^T x^\circ)(\gamma^\circ - b^T x^\circ) + (\gamma - \gamma^\circ)(a^T x^\circ) + [(\gamma - b^T x^\circ)a - (a^T x)b]^T (x - x^\circ)$$

$$(17b) \qquad = (a^T x^\circ)(\gamma^\circ - b^T x^\circ) + (\gamma - \gamma^\circ)(a^T x^\circ) + [(\gamma - b^T x)a - (a^T x^\circ)b]^T (x - x^\circ),$$

we can bound (16) accurately to first order by using interval arithmetic to evaluate one of the right-hand sides of (17) for $\underline{x} \le x \le \overline{x}$, $\underline{\gamma} \le \gamma \le \overline{\gamma}$. The computing reported below uses (17a).

I first presented a *QR* approach to solving interval least squares problems at the SIAM Conference on Applied Linear Algebra in 1982. Unfortunately, I now distrust the scheme I mentioned there for overcoming the simultaneity problem (subtracting off a linear estimate of the excess width caused by simultaneity): the bounds it delivers are not rigorous. Recently Neumaier [28] has also proposed a *QR* approach to interval least squares problems, but he only considers bounding B^*, and his bounds suffer from simultaneity.

Nonlinearity Indices

Below I report the values of three nonlinearity indices on some sample problems. The first, denoted by ρ_N, is (9) with M given by (11b) and (8b) and with $\mathbf{A} \supset \{X^T X : X \in \mathbf{X}\}$, i.e.,

$$(18a) \qquad \rho_N := \sup \| \tilde{\mathbb{I}}(M\mathbf{X} - I) \|,$$

where

$$(18b) \qquad M = (X^{\circ T} X^\circ)^{-1}.$$

The second, ρ_Q, is one of the two I prefer as a nonlinearity diagnostic. It is the analog of (9) for the interval linear equations suggested by (12c):

$$(19) \qquad \rho_Q := \sup \| \tilde{\mathbb{I}}(\{(XZ)^T (XZ^\circ) - I : X \in \mathbf{X}\}) \|.$$

The third nonlinearity index, ρ_S, my other choice for a nonlinearity diagnostic, comes from singular value theory and is motivated by Stewart (see §3 and the appendix of [30]). The idea is to use singular-value analysis to compute an index that will

behave comparably to ρ_Q, with values ≥ 1 warning of situations where the perturbations permitted X could make it rank deficient, i.e., could make its columns linearly dependent. To compute ρ_S, we partition X as in (4), i.e., $X = [K, W]$ with $K \in \mathbb{R}^{n \times p_K}$ and $W \in \mathbb{R}^{n \times p_W}$ $(0 \leq p_K \leq p,\ p_K + p_W = p)$, and we scale the columns of W (i.e., linearly change variables) so that the maximum Euclidean norm of the error in each column is 1. Then

$$\rho_S := \frac{\sqrt{p_W}}{\text{svmin}(\Pi_K W)}, \tag{20}$$

where Π_K projects orthogonally onto the orthogonal complement of the column space of K (with $\Pi_K \equiv I$ if $p_K = 0$), and svmin$(\cdot)$ denotes the least singular value of $(\cdot)$. That is, we "regress out" of W any constant columns K, compute the least singular value of what remains, and scale it appropriately (dividing it into $\sqrt{p_W}$). The numerator of (20) is both the Euclidean and the Frobenius norm of the largest perturbation to which W could be subject, given the scaling of W; svmin$(\Pi_K W)$ is both the smallest Euclidean and smallest Frobenius norm of any perturbation to W that reduces its rank; hence their ratio, ρ_S, is the nonlinearity index we seek. Note that if we factor X as in (4), then

$$\text{svmin}(\Pi_K W) = \text{svmin}(R_{WW}).$$

Test Data

Table 1 summarizes my test data, showing the problem dimensions, sources, and nominal error radii: $\frac{1}{2}(\bar{X}_1 - \underline{X}_1)$, $\cdots$, $\frac{1}{2}(\bar{X}_p - \underline{X}_p)$, $\frac{1}{2}(\bar{y} - \underline{y})$. Except as explained below and summarized in Table 2, I used half of each variable's least significant reported digit as the nominal error radius. Some of the exogenous variables X_i were computed from other data. For example, as on page 598 of [13], the Anderson problem uses the model

$$y = \beta_1 + X_2\beta_2 + X_3\beta_3 + \beta_4 X_2^2;$$

that is, $X_{i,4} = X_{i,2}^2$. Since $\underline{X}_{i,2} > 0$ for all i in this problem, it is appropriate to use $\bar{X}_{i,2}^2 - \underline{X}_{i,2}^2$ as the error radius for $X_{i,4}$, as indicated in Table 2. (This is the only instance in my test data where there is an obvious correlation between the errors in any pair of variables.) The data-dependent error radii in problem Spray arise because $X_{i,j}$, $2 \leq j \leq 4$ and y_i are logarithms of observed data; those in Turnover come from the statement in [18] that "Both the independent variables are subject to significant

measurement error, and although the size of this is not accurately known, it is believed to be within ±15 per cent.''

Name	n	p	Source	Pages	Error radii
Anderson	27	4	[13]	406, 598	0, .0005, .05, *, .05
Asphalt	31	7	[9]	109, 95–100	0, .005, .005, .005, 0, .05, .0005, .00005
Fig1–4	6	2	(3)		.25, 0, .25
Fuel	48	5	[32]	33	0, 0, .0005, .5, .5, .5
Longley	16	7	[30]	§4	0, .5, .5, .5, .5, .5, 0, .5
Shell	50	2	[23]	198	0, 7, 86.6
Spray	35	4	[13]	405, 598	0, *, *, *, *
Turnover	30	3	[18]	191	0, *, *, .05
Water	17	5	[13]	353	0, .05, .5, 0, 0, .5

* = computed error radii — see Table 2.
Table 1: data sizes, sources, nominal error radii

Problem	Variable	Radius
Anderson	X_4	$(X_2 + .0005)^2 - X_2^2$
Spray	X_2	$\log(1 + .00005/\exp(X_2))$
Spray	X_3	$\log(1 + 50/\exp(X_3))$
Spray	X_4	$\log(1 + .005/\exp(X_4))$
Spray	y	$\log(1 + .05/\exp(y))$
Turnover	X_2	$0.15 \cdot X_2$
Turnover	X_3	$0.15 \cdot X_3$

Table 2: formulae for data-dependent error radii

All the problems have an intercept term, i.e., a column of ones. Aside from Fig1–4 (i.e., (3), where I had in mind the formula $y = ax + b$), I have taken X_1 to be the column of ones, which is obviously not subject to error. Partly because of the intercept term — and because I start with 1 when numbering the columns of X — the column numbers I mention often differ by one from those in the original sources.

It seems reasonable to regard a few other variables as not subject to error. In problem Fuel, X_2 is a tax rate, which is known and not measured; in problem Water, X_4 and X_5 are integers: the number of plant operating days in the month and the number of people on the monthly plant payroll; in problem Longley, X_7 is a column of years. (My decision to regard the latter as constant is based on [11]. See also [2], and [3].)

Table 1 cites Stewart [30] rather than Longley's original paper [22] because I use Stewart's scaling of X_2 (which makes X a matrix of integers, for which one hopes there is no roundoff error on input). [However, following [22], I used 554,894 rather than 554,984 as the value for $X_{16,3}$.]

Test Results

My computational testing used the standard unit vectors (columns of I) for the linear form c, so I computed bounds on the individual components of $B^* = \mathbf{X}^\dagger \mathbf{y}$, i.e., I computed $[\underline{b}, \overline{b}] \supset B^*$. Table 3 shows the values of the nonlinearity indices (18–20) when the error radii have the nominal values shown in Table 1; below I discuss the effects of scaling these radii. The choices of c play a role in fifth column, which displays a measure of the sharpness of $[\underline{b}, \overline{b}]$, i.e., an approximate maximum excess width percentage over all components of b. Specifically, the fifth column shows

$$100 \cdot \max\{\frac{(\overline{b}_i - \underline{b}_i) - (\overline{b}_i^{\#} - \underline{b}_i^{\#})}{\overline{b}_i^{\#} - \underline{b}_i^{\#}} : 1 \le i \le p\}, \tag{21}$$

where $[\underline{b}^{\#}, \overline{b}^{\#}] \approx \mathbb{I}(\mathbf{X}^\dagger \mathbf{y})$ is computed as follows. First, based on the signs of (5a,b), i.e., $\frac{\partial b_k}{\partial X_{i,j}}$ and $\frac{\partial b_k}{\partial y_i}$, both evaluated at (X°, y°), we choose $X^{(\pm,k)}$ and $y^{(\pm,k)}$ with $X_{i,j}^{(\pm,k)} \in \{\underline{X}_{i,j}, \overline{X}_{i,j}\}$ and $y_i^{(\pm,k)} \in \{\underline{y}_i, \overline{y}_i\}$ to approximately maximize $(X^{(+,k)\dagger} y^{(+,k)})_k$ and minimize $(X^{(-,k)\dagger} y^{(-,k)})_k$. Then

$$\underline{b}_i^{\#} := \min\{(X^{(\pm,k)\dagger} y^{(\pm,k)})_i : 1 \le k \le p\}$$

and

$$\overline{b}_i^{\#} := \max\{(X^{(\pm,k)\dagger} y^{(\pm,k)})_i : 1 \le k \le p\}.$$

The values of (5a,b) at $(X^{(\pm,k)}, y^{(\pm,k)})$ suggest that the values I report of (21) overestimate the true excess width percentage (i.e., (21) with $[\underline{b}^{\#}, \overline{b}^{\#}] := \mathbb{I}(\mathbf{X}^\dagger \mathbf{y})$) by at most a few percent.

Problem	ρ_N	ρ_Q	ρ_S	max % excess width
Anderson	7.501	0.371	0.169	105.7
Asphalt	9.739	0.457	0.191	171.9
Fig1–4	0.322	0.158	0.076	38.2
Fuel	0.353	0.029	0.017	6.7
Longley	381.900	0.198	0.125	31.3
Shell	5.427	0.700	0.262	231.8
Spray	2.408	0.147	0.082	31.6
Turnover	2.570	1.106	0.551	—
Water	0.096	0.013	0.006	2.5

Table 3: values at nominal error radii

Since $\rho_Q > 1$ for problem Turnover, the QR approach described above can deliver

no bounds on Turnover's B^*. But on four of the seven other problems, it delivers bounds that overestimate the components of B^* by less than 40%. Whether one regards such bounds as sufficiently tight surely depends on context, but it is probably safe to say that they are not wild overestimates.

It is interesting to see how the values in Table 3 behave when one scales the error radii. For $\rho_Q < 1$, the ratios ρ_N / ρ_Q and ρ_S / ρ_Q are approximately constant. Figure 5 shows some typical plots of ρ_S / ρ_Q versus ρ_Q. Plots of ρ_N / ρ_Q versus ρ_Q are qualitatively similar, but the scale on the y axis is sometimes much larger: for some problems, $\rho_N / \rho_Q \gg 1$ because of simultaneity. Thus one can compute good bounds (i.e., bounds with a small excess-width ratio (21)) for fewer problems by the normal equations (11) than by the QR approach (12).

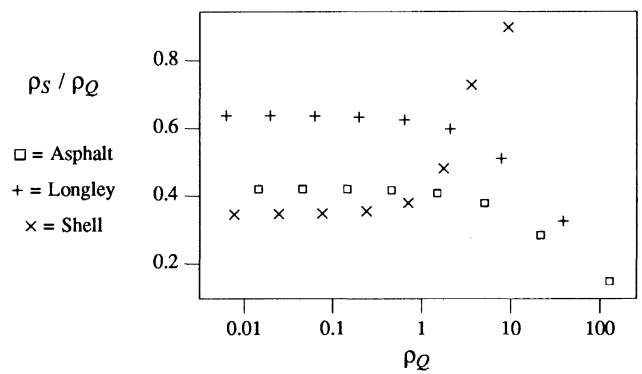

Figure 5: ρ_S / ρ_Q versus ρ_Q for Asphalt, Longley, and Shell

My calculations of ρ_Q use (17a) (with $\gamma = 0$) to bound the components of $\{(XZ)^T(XZ): X \in \mathbf{X}\}$. Although this avoids simultaneity to first order in these component bounds, there remains a simultaneity problem when the bounds are summed to compute ρ_Q. Thus it should come as no surprise that usually $\rho_S < \rho_Q$.

As mentioned earlier, the choice of D in (10) strongly affects ρ_Q. The stopping tests in the iteration that determines D are of the form "accept D^k as D if $\rho(D^k) > \tau_{rel}\rho(D^{k-1})$ or $\rho(D^k) < \tau_{abs}$." On some problems I got a reduction of 20% or more in ρ_Q by tightening τ_{rel} from the value 0.9 recommended in [14] to 0.999; I also reduced τ_{abs} from 0.1 to 0.001. This made plots like Figure 5 much smoother, but sometimes made the iteration that determines D take several times as many iterations. In one instance of a 20% reduction, the iteration count went from 5 to 19. All

this makes ρ_S more appealing than ρ_Q as a nonlinearity diagnostic.

For $\rho_Q < 1/10$, the excess-width ratios divided by ρ_Q, i.e.,

$$(22) \qquad \frac{(\bar{b}_i - \underline{b}_i) - (\bar{b}_i^{\#} - \underline{b}_i^{\#})}{(\bar{b}_i^{\#} - \underline{b}_i^{\#}) \cdot \rho_Q}$$

are approximately constant. Figures 6 and 7 show typical plots of these values against ρ_Q; the dotted lines indicate the nominal values of ρ_Q corresponding to Table 1. The small excess-width ratios for $\rho_Q < 1/10$ accord with my arguments above that (13–15) overcomes (to first order and for small ρ_Q) simultaneity problems in bounding C^*, i.e., (7).

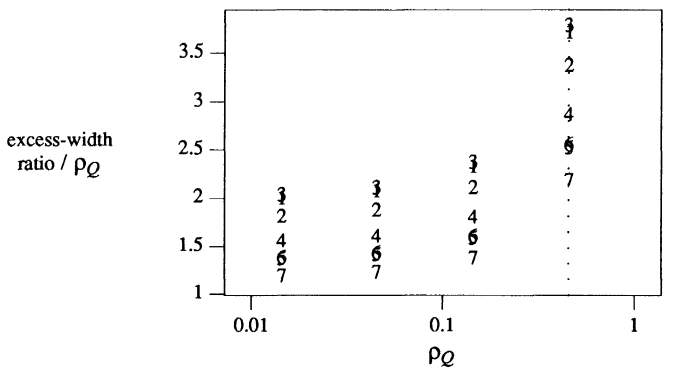

Figure 6: (22) versus ρ_Q for Asphalt

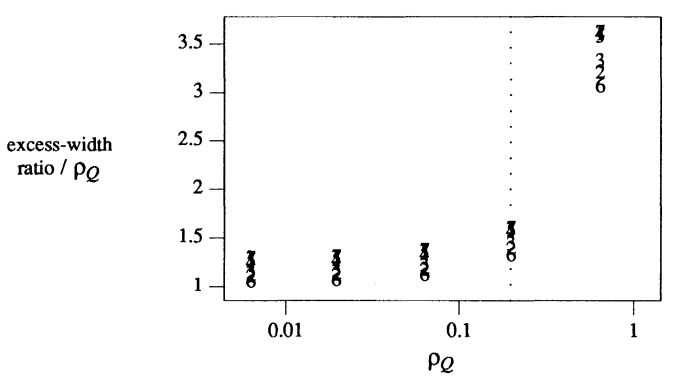

Figure 7: (22) versus ρ_Q for Longley

Bias is a common issue in papers on errors in variables. That is, one would like to know whether the expected value of the least-squares estimate, $E(X^\dagger y)$, differs significantly from the true parameter vector β (or perhaps whether $| c^T E(X^\dagger y) - c^T \beta |$ is large enough to worry about). Hodges and Moore [18] argue that if we observe X, where $E((X - X^\circ)^T (X - X^\circ))$ is the diagonal matrix S, $E(X - X^\circ) = \mathbf{0}$, $E(e) = \mathbf{0}$, and $X - X^\circ$ is independent of the error e in the least-squares model $y = X^\circ \beta + e$, then

$$E(X^\dagger y) \approx \beta - (n - p - 1)(X^{\circ T} X^\circ)^{-1} S \beta; \tag{23}$$

they suggest approximating the right-hand side of (23) to roughly bound the possible bias. The nonlinearity indices ρ_Q and ρ_S can also warn of possible bias, because the last term in (13) may contribute significantly to the bias when these indices are largish, say on the order of a tenth or more. To see how well correlated ρ_Q and ρ_S are to bias, I did some Monte Carlo calculations, with perturbations to X chosen uniformly in the intervals defined, as in Figure 5, by scaled versions of the error radii in Table 1. (There were 10000 trials for each error scaling. I used a random number generator provided by Eric Grosse and based on exercise 23 in §3.2.2 of [21]. The plots look about the same when I substitute the research UNIX® *frand* generator or the UNI generator from the PORT library.) Figures 8–15 plot the relative bias, i.e.,

$$(E(X^\dagger y) - X^{\circ\dagger} y)_i \,/\, (X^{\circ\dagger} y)_i, \tag{24}$$

determined in my Monte Carlo calculations against ρ_S for the problems in Table 1. The dotted lines indicate the nominal values of ρ_S corresponding to the error bounds in Table 1.

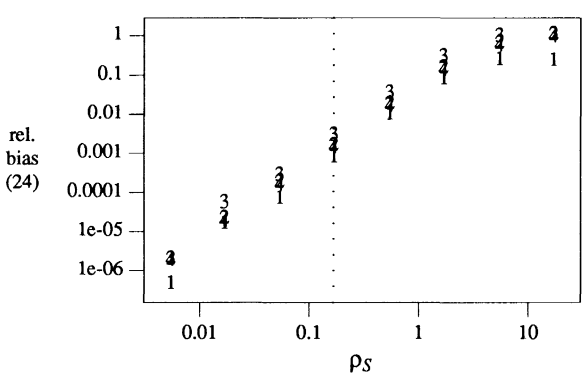

Figure 8: bias for Anderson

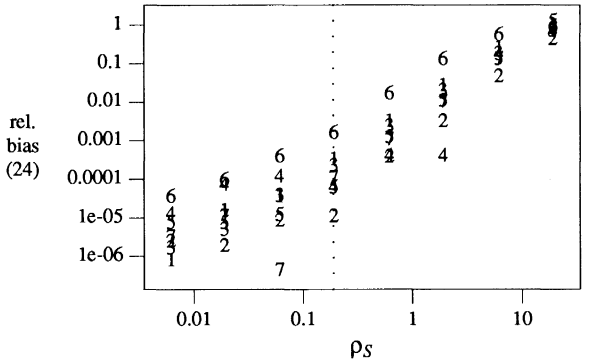

Figure 9: bias for Asphalt

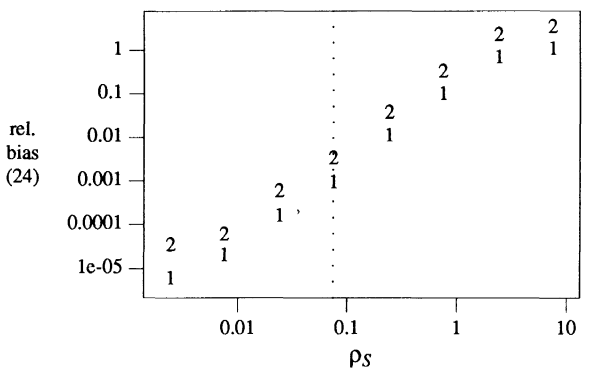

Figure 10: bias for Fig1−4

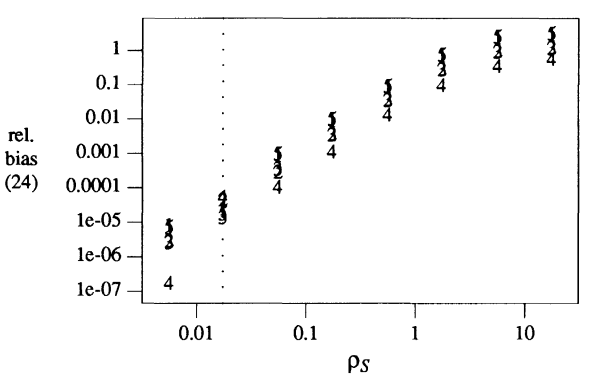

Figure 11: bias for Fuel

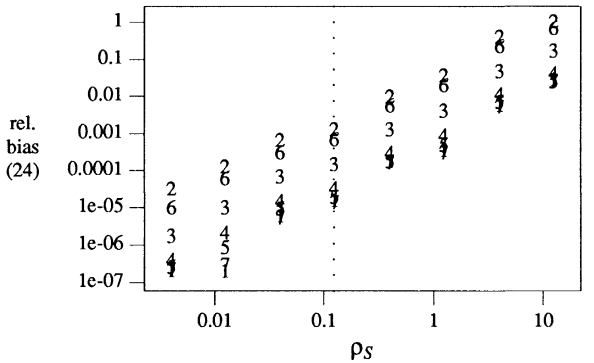

Figure 12: bias for Longley

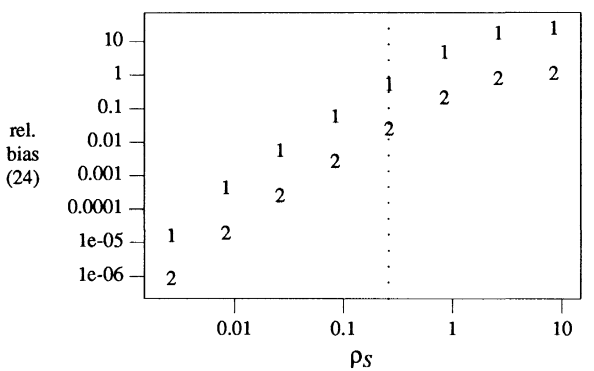

Figure 13: bias for Shell

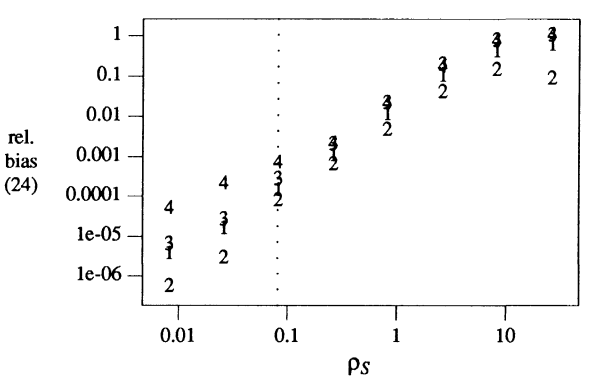

Figure 14: bias for Spray

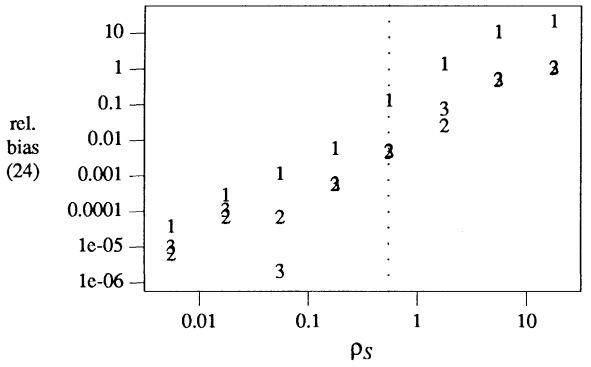

Figure 15: bias for Turnover

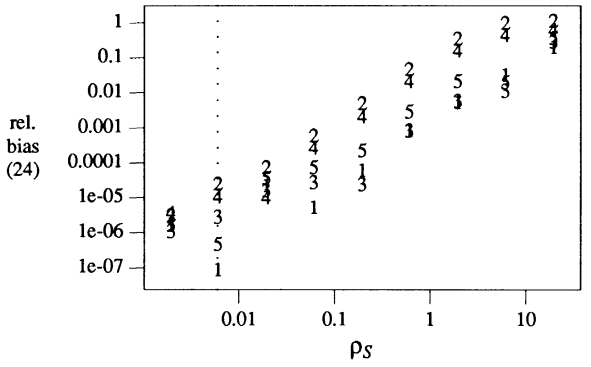

Figure 16: bias for Water

Although (23) often gives roughly similar bias estimates to those from my Monte Carlo calculations, I distrust (23). Simple examples suggest that the last two expectations on p. 195 of [18] are incorrect.

Diagnosing Collinearity

When working any mathematical model, one should worry a bit about how appropriate the model is. A linear least-squares model may be inappropriate if the exogenous variables are too close to being linearly dependent. Trouble caused by near linear dependence is generally called "collinearity" in the statistical literature. As discussed above, collinearity can lead to biased estimates and forecasts. It may also lead to incorrect choices among models. There seems to be no agreement on a precise

definition of (approximate) collinearity [4, p. 85], but it is sometimes taken to be synonymous with ill-conditioning [5, p. 93], i.e., with a small (relative) change in X or in y making a big change in $X^{\dagger}y$. To assess collinearity, Belsley, Kuh, and Welsch [4] recommend scaling the columns of X to have unit Euclidean length, then computing the spectral condition number (ratio of largest to smallest singular values) of X. Let us call this κ_{BKW}. Belsley argues persuasively for evaluating κ_{BKW} on the original X matrix [6]. So far as assessing numerical ill conditioning is concerned, i.e., inaccuracy caused by roundoff errors in the computed approximation of $X^{\dagger}y$, I fully agree with him. But for assessing the effects observational errors may have on parameter estimates and forecasts, I think one should consider only the perturbations that these errors could introduce. Specifically, when there are dummy variables or other variables not subject to observational error (e.g., the column of ones in problems that have an intercept term), it is inappropriate to consider perturbations in them. Moreover, I think it important to consider the size of the perturbations that the errors could introduce. This argues for examining ρ_S as a nonlinearity diagnostic (source of bias), first-order or interval bounds on $c^{\mathrm{T}}\mathbf{X}^{\dagger}\mathbf{y}$ to assess whether the perturbations could be large enough to worry about, and κ_{BKW} as a diagnostic only for numerical ill-conditioning — the tolerance for which should be considerably larger than the values 15 to 30 recommended in [4]. When ρ_S is small, one can easily use (5) to compute approximate standard deviations for $c^{\mathrm{T}}\mathbf{X}^{\dagger}\mathbf{y}$ (after assuming a distribution for $X \in \mathbf{X}$ and $y \in \mathbf{y}$).

When all exogenous variables save the intercept are subject to error, one obtains Π_W in (20) by scaling the exogenous variables so the maximum Euclidean norm of the error possible in each column is one, *then* centering these variables (subtracting their means). This is to be contrasted with the procedure that disquiets Belsley [6] — *first* centering, then examining the condition of the centered exogenous variables.

Disallowing impossible perturbations can substantially change one's assessment of nonlinearity and perhaps of collinearity. The infamous Longley data provide an example. If, as above, we disallow perturbations in the years (X_7), then some linear combinations of the parameters are reasonably well determined and, as Figure 12 indicates, there is only modest bias. In support of this statement about "reasonably well determined" parameters, Table 4 shows the nominal parameter estimate $b^{\circ} = X^{\circ\dagger}y$, the error radii, i.e., bounds on $|\, b_i - b^{\circ}_i \,|$, delivered by the *QR* approach, and the error radii as percentages of b°_i for Longley with the nominal errors in Table 1.

i	b_i°	radius	% rad
1	−3.48226e+06	5.432e+04	1.6
2	1.50619	1.671	110.9
3	−0.0358192	0.004688	13.1
4	−2.02023	0.06117	3.0
5	−1.03323	0.019	1.8
6	−0.0511041	0.03132	61.3
7	1829.15	26.32	1.4

Table 4: bounds for Longley with nominal error radii

On the other hand, if, as in [2], we allow perturbations of up to .5 in the column of years, then $\rho_S = 7.32$ and $\rho_Q = 7.137$, so the *QR* approach cannot deliver any bounds. Moreover, as in [2], simulations like those behind Figures 8−15 then reveal substantial bias at the nominal error radii, as shown below in Figure 16.

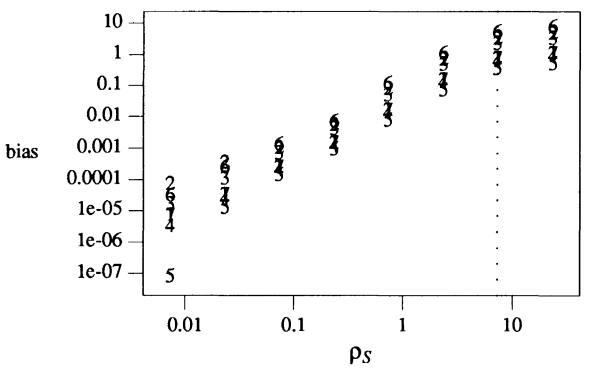

Figure 16: Longley bias when years are subject to error

Concluding Remarks

Interval least-squares estimates can furnish rigorous and reasonably tight bounds on the the effect of errors in the exogenous variables on forecasts and, as special cases, on parameter estimates. In cases where ρ_S (see (20)) is small, first-order assessments (5) of the effects of exogenous variable errors are probably acceptably accurate for most purposes. Stewart's [30] sensitivity coefficients, which are based on (5), may also be helpful in this case. The main thing to get out of this paper is that ρ_S (or ρ_Q, (20), if one wants to compute interval linear least-squares estimates) may be worth computing as a nonlinearity diagnostic — a warning of potential bias.

References

[1] G. Alefeld and J. Herzberger, *Introduction to Interval Computations,* Academic Press, 1983.

[2] A. E. Beaton, D. B. Rubin, and J. L. Barone, "The Acceptability of Regression Solutions: Another Look at Computational Accuracy," *J. Amer. Statist. Assoc.* **71** (1976), pp. 158–168.

[3] A. E. Beaton, D. B. Rubin, and J. L. Barone, "Comment," *J. Amer. Statist. Assoc.* **72** (1977), pp. 600–601.

[4] D. A. Belsley, E. Kuh, and R. E. Welsch, *Regression Diagnostics,* Wiley, 1980. Subtitle: Identifying Influential Data and Sources of Collinearity.

[5] D. A. Belsley, "Reply," *American Statistician* **38** (1984), pp. 90–93.

[6] D. A. Belsley, "Demeaning Conditioning Diagnostics Through Centering," *American Statistician* **38** (1984), pp. 73–77.

[7] P. T. Boggs, R. H. Byrd, and R. B. Schnabel, "A Stable and Efficient Algorithm for Nonlinear Orthogonal Distance Regression," *SIAM J. Sci. Statist. Comput.* **8** (1987), pp. 1052–1078.

[8] W. G. Cochran, "Errors of Measurement in Statistics," *Technometrics* **10** (1968), pp. 637–665.

[9] C. Daniel and F. S. Wood, *Fitting Equations to Data,* Second Edition; John Wiley & Sons, 1980. Subtitle: Computer Analysis of Multifactor Data.

[10] R. B. Davies and B. Hutton, "The Effect of Errors in the Independent Variables in Linear Regression," *Biometrika* **62** (1975), pp. 383–391.

[11] W. T. Dent and D. C. Cavander, "Rejoinder," *J. Amer. Statist. Assoc.* **72** (1977), pp. 601–602.

[12] W. T. Dent and D. C. Cavander, "More on Computational Accuracy in Regression," *J. Amer. Statist. Assoc.* **72** (1977), pp. 598–600.

[13] N. Draper and H. Smith, *Applied Regression Analysis, Second Edition,* John Wiley and Sons, 1981.

[14] D. M. Gay, "Solving Interval Linear Equations," *SIAM J. Numer. Anal.* **19** (1981), pp. 858–870.

[15] D. M. Gay, "Computing Perturbation Bounds for Nonlinear Algebraic Equations," *SIAM J. Numer. Anal.* **20** (1983), pp. 638–651.

[16] G. H. Golub and C. F. Van Loan, "An Analysis of the Total Least Squares Problem," *SIAM J. Numer. Anal.* **17** (1980), pp. 883–893.

[17] G. H. Golub and C. F. Van Loan, *Matrix Computations,* Johns Hopkins University Press, Baltimore, MD, 1983.

[18] S. D. Hodges and P. G. Moore, "Data Uncertainties and Least Squares Regression," *J. Roy. Statist. Soc. Ser. C* **21** (1972), pp. 185–195.

[19] S. Van Huffel and J. Vandewalle, "Subset Selection Using the Total Least-Squares Approach in Collinearity Problems with Errors in the Variables," *Lin. Alg. Applic.* **88/89** (1987), pp. 695–714.

[20] S. Inman, "The Probability of a Given Error Being Exceeded in Approximate Computation," *Math. Gazette* **34** (1950), pp. 99–113.

[21] D. E. Knuth, *Seminumerical Algorithms,* second edition; Addison-Wesley, 1981. Volume 2 of *The Art of Computer Programming.*

[22] J. W. Longley, "An Appraisal of Least Squares Programs for the Electronic Computer from the Point of View of the User," *J. Amer. Statist. Soc.* **62** (1967), pp. 819–841.

[23] A. Madansky, "The Fitting of Straight Lines when Both Variables are Subject to Error," *J. Amer. Statist. Assoc.* **54** (1959), pp. 173–205.

[24] R. E. Moore, *Methods and Applications of Interval Analysis,* SIAM, Philadelphia, 1979.

[25] A. Neumaier, "New Techniques for the Analysis of Linear Interval Equations," *Lin. Algebra Applic.* **58** (1984), pp. 273–325.

[26] A. Neumaier, "Further Results on Linear Interval Equations," *Lin. Algebra Applic.* **87** (1987), pp. 155–179.

[27] A. Neumaier, "Overestimation in Linear Interval Equations," *SIAM J. Numer. Anal.* **24** (1987), pp. 207–214.

[28] A. Neumaier, "Solving Linear Least Squares Problems by Interval Arithmetic," *Freiburger Intervall-Berichte* **87/6** (1987), pp. 37–42.

[29] G. W. Stewart, *Introduction to Matrix Computations,* Academic Press, 1973.

[30] G. W. Stewart, "Assessing the Effects of Variable Error in Linear Regression," Tech. Rep. 818 (Oct. 1979), Dept. of Computer Science, Univ. of Maryland, College Park, MD 20742.

[31] B. F. Swindel and D. R. Bower, "Rounding Errors in the Independent Variables in a General Linear Model," *Technometrics* **14** (1972), pp. 215–218.

[32] S. Weisberg, *Applied Linear Regression,* John Wiley & Sons, 1980.

[33] J. Ziegler, "Detecting Sensitivity to Variable Errors in Linear Regression," Tech. Rep. 13 (Oct. 1980), Center for Computational Research in Economics and Management Science, Massachusetts Institute of Technology, Cambridge, MA 02139.