SIAM J. SCI. STAT. COMPUT.
Vol. 8, No. 6, November 1987

014

A STABLE AND EFFICIENT ALGORITHM FOR NONLINEAR ORTHOGONAL DISTANCE REGRESSION*

PAUL T. BOGGS†, RICHARD H. BYRD‡ AND ROBERT B. SCHNABEL§

Abstract. One of the most widely used methodologies in scientific and engineering research is the fitting of equations to data by least squares. In cases where significant observation errors exist in the independent variables as well as the dependent variables, however, the ordinary least squares (OLS) approach, where all errors are attributed to the dependent variable, is often inappropriate. An alternate approach, suggested by several researchers, involves minimizing the sum of squared orthogonal distances between each data point and the curve described by the model equation. We refer to this as orthogonal distance regression (ODR). This paper describes a method for solving the orthogonal distance regression problem that is a direct analog of the trust region Levenberg-Marquardt algorithm. The number of unknowns involved is the number of model parameters plus the number of data points, often a very large number. By exploiting sparsity, however, our algorithm has a computational effort per step which is of the same order as required for the Levenberg-Marquardt method for ordinary least squares. We prove our algorithm to be globally and locally convergent, and perform computational tests that illustrate some differences between ODR and OLS.

Key words. least squares, trust-region methods, Levenberg–Marquardt methods, errors in variables, total least squares

AMS(MOS) subject classifications. Primary 65K99; secondary 62J99, 65D10

1. Introduction.

The problem of fitting a model to data with errors in the observations has a rich history and a considerable literature. The problem where there are also errors in the independent variables at which these observations are made, however, has only relatively recently been given attention. In this paper, we consider a general form of this extended problem and provide an efficient and stable algorithm for its solution. Several names for this extended problem have been suggested; we prefer orthogonal distance regression (ODR).

Errors in independent variables virtually always occur, but are often ignored in order that classical or ordinary (linear or nonlinear) least squares (OLS) techniques can be applied (see, e.g., [LawH74], [Ste73], [Mor77], [DenGW81]). Also, if these errors are small with respect to those in the observed variables, then ignoring them does not usually seriously degrade the accuracy of the estimates. In some fields, however, measurement techniques are sufficiently accurate that errors in the independent variables are not insignificant compared to those in the observations. Examples at the National Bureau of Standards (NBS) include the calibration of electronic devices, flow-meters and calorimeters. Another class of examples comes from curve and surface fitting problems.

We first develop a formal statement of the ODR estimation problem and briefly discuss its application to statistical estimation and to curve fitting. The main contributions of the paper are the derivation and convergence analysis of a highly efficient algorithm for solving ODR problems (§§2 and 3). In §4, the results of some compu-

* Received by the editors December 24, 1985; accepted for publication (in revised form) March 20, 1987.

† Scientific Computing Division, National Bureau of Standards, Gaithersburg, Maryland 20899.

‡ Computer Science Department, University of Colorado, Boulder, Colorado 80309. The research of this author was supported in part by National Science Foundation grant DCR-8403483.

§ Computer Science Department, University of Colorado, Boulder, Colorado 80309 and Scientific Computing Division, National Bureau of Standards, Boulder, Colorado 80303. The research of this author was supported in part by Army Research Office Contract DAAG 29-84-K-0140.

tations are shown which illustrate the performance of the algorithm and allow some comparisons with ordinary least squares.

Observations in applied science are often thought of as satisfying a mathematical model of the form

$$y^a = f(x, \beta^a) \tag{1.1}$$

where y^a, (a for "actual" or "true") is taken to be the dependent variable that is to be measured as a function of the independent variable x; and $\beta^a \in R^p$ is the true parameter vector to be estimated based on the measured values, or data pairs, $(x_i, y_i), i = 1, \cdots, n$. The function f is not assumed to be linear, but is assumed to be smooth. Typically the number of data points, n, is far greater than the number of parameters, p.

In the classical case, only the observations y_i are assumed to contain errors. If these errors, designated for convenience by $-\epsilon_i^a$, are additive and the mathematical model is exact then

$$y_i = f(x_i, \beta^a) - \epsilon_i^a, \qquad i = 1, \cdots, n. \tag{1.2}$$

If in addition the errors are normally distributed with mean 0 and variance $\sigma^2 I$, then the maximum likelihood estimate, $\hat{\beta}$, is the solution to the least squares problem

$$\min_{\beta} \sum_{i=1}^{n} [f(x_i, \beta) - y_i]^2 . \tag{1.3}$$

If f is a linear function of β then this is a classical linear least squares problem; otherwise it is a classical nonlinear least squares problem. Even when the above assumptions on the model or the errors are not satisfied, problem (1.3) is the most frequently used method for parameter estimation.

In the more general situation, the measurements of the independent variables x_i are also assumed to contain errors. If we assume that y_i has unknown additive error ϵ_i^a and that x_i has unknown additive error δ_i^a, then (1.2) becomes

$$y_i = f(x_i + \delta_i^a; \beta^a) - \epsilon_i^a. \tag{1.4}$$

A reasonable way to select the parameters in this case is to choose the β that causes the sum of the squares of the orthogonal distances from the data points (x_i, y_i) to the curve $f(x, \beta)$ to be minimized. (See Fig. 1.) If r_i is the orthogonal distance from (x_i, y_i) to the curve, then

$$r_i^2 = \epsilon_i^2 + \delta_i^2, \qquad i = 1, \cdots, n$$

where ϵ_i and δ_i solve

$$\begin{aligned} &\min_{\epsilon_i, \delta_i} (\epsilon_i^2 + \delta_i^2) \\ &\text{subject to } f(x_i + \delta_i; \beta) - \epsilon_i = y_i. \end{aligned} \tag{1.5}$$

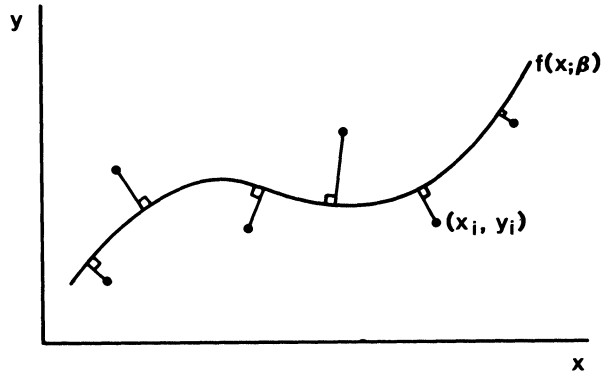

Figure 1

The constraint in (1.5) ensures that the distance r_i connects the point (x_i, y_i) to the curve. The minimization ensures that r_i is the radius of the smallest circle centered at (x_i, y_i) which is tangent to the curve $f(x_i; \beta)$. (See Fig. 2.) Therefore, the parameter estimate $\hat{\beta}$ and the error estimates $\hat{\epsilon}$ and $\hat{\delta}$ that cause the sum of the

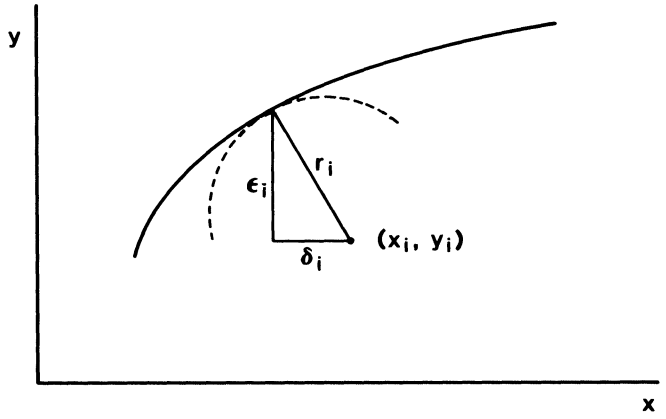

Figure 2

squares of the orthogonal distances from the data points to the curve to be minimized are found by solving

$$\min_{\beta,\epsilon,\delta} \sum_{i=1}^{n} r_i^2 = \min_{\beta,\epsilon,\delta} \sum_{i=1}^{n} (\epsilon_i^2 + \delta_i^2) \tag{1.6}$$
$$\text{subject to } y_i = f(x_i + \delta_i; \beta) - \epsilon_i, \qquad i = 1, \cdots, n.$$

Since the constraints in (1.6) are simple linear constraints in ϵ_i, we solve for ϵ_i and eliminate both the ϵ_i and all of the constraints thereby obtaining

$$\min_{\beta,\delta} \sum_{i=1}^{n} \left[(f(x_i + \delta_i; \beta) - y_i)^2 + \delta_i^2 \right] \tag{1.7}$$

which is now an unconstrained minimization problem.

Two slight extensions to this form constitute the ultimate problem to be considered. The first allows the possibility that $x_i \in R^m$ rather than R^1. Therefore, $\delta_i \in R^m$ and instead of δ_i^2 in (1.7) we have $\delta_i^T \delta_i = \sum_{j=1}^m \delta_{ij}^2$. (The superscript T denotes transpose.) The second extension merely admits a general weighting scheme on the problem. The form we have chosen results in the general nonlinear ODR problem

$$\text{(ODR)} \qquad \min_{\beta,\delta} \sum_{i=1}^{n} w_i^2 \left[(f(x_i + \delta_i; \beta) - y_i)^2 + \delta_i^T D_i^2 \delta_i \right]$$

where $w_i > 0, i = 1, \cdots, n$ and $D_i = \text{diag}\{d_{ij} > 0,\ j = 1, \cdots, m\}$, $i = 1, \cdots, n$, i.e., D_i is a diagonal matrix of order m. It follows that the vectors $y, w \in R^n$ and $x, \delta \in R^{nm}$ and that $\delta_i^T D_i^2 \delta = \sum_{j=1}^m \delta_{ij}^2 d_{ij}^2$.

While we have not assumed that f is linear, it is important to note that (ODR) is a nonlinear optimization problem even if f is the simple linear function

$$y = \beta_1 x + \beta_2$$

since we then have that

$$y_i = \beta_1 (x_i + \delta_i) + \beta_2 - \epsilon_i.$$

Clearly the product of β_1 and δ_i is an unavoidable nonlinearity.

ODR problems have been considered by statisticians, usually under the rubric errors in variables. Most of this effort, however, has been devoted to linear models, i.e., when f is linear in β. (See e.g.,[Mor71], [KenSO83], [Bar74,p.67] and [Ful86].) As in the classical nonlinear least squares case, little theory on the statistical properties of the solution appears to exist. It is known that if both ϵ and δ are normally distributed with mean zero and variances $\sigma_\epsilon^2 I$ and $\sigma_\delta^2 I$, respectively, then the solution of (ODR) with $w_i = 1$ and $D_i = (\sigma_\epsilon / \sigma_\delta) I$, $i = 1, \cdots, n$ is a maximum likelihood estimate of the parameters. Unfortunately, as in the nonlinear classical case, no generally valid, computationally efficient, inferential statistical tests are known.

Independent of statistical considerations, ODR has potentially significant applications in curve and surface fitting. Consider, for example, the problem of finding the parabola which best fits the given set of points (see Fig. 3). (We have seen this problem arise from a dental application.) Here it is clear that ordinary least squares will unduly weight the top data points, while fitting in the horizontal direction would unduly weight the bottom data points. An orthogonal measure of distance alleviates

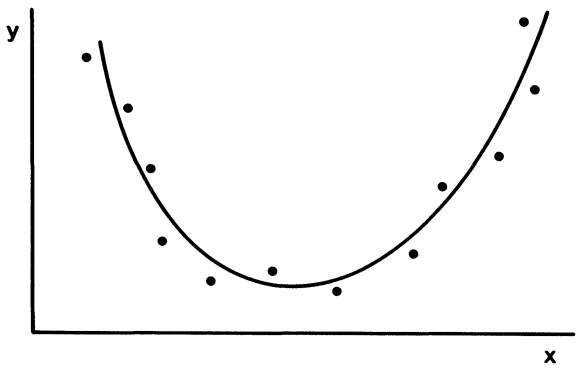

Figure 3

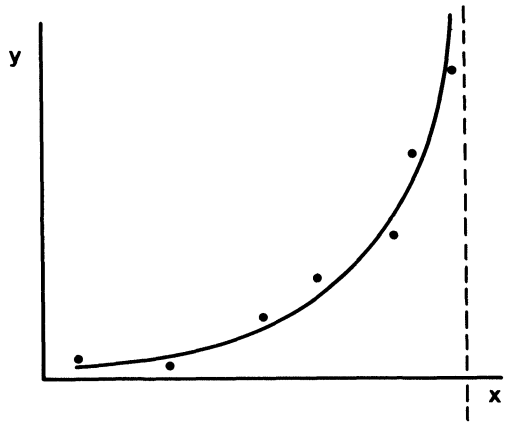

Figure 4

these problems and provides a reasonable fit. A related case is the problem of fitting near an asymptote as illustrated in Fig. 4. Orthogonal distances here prevent the undue influence of points close to the asymptote. This problem is discussed further in §4.

The literature contains several algorithms for solving (ODR) and related problems. For example, Golub and Van Loan [GolV83] give a singular value decomposition procedure for the problem when f is linear. They refer to this problem as total least squares. Britt and Luecke [BriL73] consider the nonlinear case as well as the nonlinear implicit case and present an algorithm. Powell and Macdonald [PowM72] also consider (ODR) and propose an algorithm which ignores the mixed partial derivatives and approximates the other derivatives by (central) finite differences. They report some good results on small problems. More recently, Schwetlick and Tiller [SchT85] propose an algorithm similar to the one here for the nonlinear problem. Our algorithm, however, does not make use of the singular value decomposition and it does incorporate a full trust region strategy. Finally, Watson [Wat85] also proposes a structured approach for the Gauss-Newton step, but does not propose a trust region strategy.

2. The Algorithm. In order to solve the minimization problem (ODR),

$$\min_{\beta,\delta} \sum_{i=1}^{n} w_i^2 \left[(f(x_i + \delta_i; \beta) - y_i)^2 + \delta_i^T D_i^2 \delta_i \right] \tag{2.1}$$

we first express it in a more convenient form and simplify the notation. Next, we give an overview of the iteration which is based on the trust region Levenberg-Marquardt strategy popularized by Moré [Mor77]. (See also [Heb73], [DenS83].) We then show how to modify this technique to obtain an algorithm which requires the same order of work per iteration as these algorithms applied to the same problem without allowing changes to x_i. That is, if the $\delta_i's$ are held fixed at zero, ODR reduces to OLS and trust region methods require $O(np^2)$ operations per iteration. Our algorithm, by exploiting the structure of (ODR), still requires only $O(np^2) + O(nm)$ operations per iteration to solve the problem.

While we have designed and implemented the algorithm to handle the full generality of (2.1), the notation is considerably simplified by assuming $x_i \in R^1$. We

temporarily make this assumption and rewrite (2.1) into the form of an OLS problem by the following device. Let

$$g_i(\beta,\delta) = \begin{cases} w_i\left(f(x_i+\delta_i;\beta)-y_i\right), & i=1,\cdots,n, \\ w_{i-n}d_{i-n}\delta_{i-n}, & i=n+1,\cdots,2n. \end{cases} \tag{2.2}$$

Also let $G: R^{p+n} \to R^{2n}$ have component functions $g_i(\eta)$ where $\eta = \binom{\beta}{\delta}$. Now (ODR) becomes

$$\min_{\eta} \|G(\eta)\|^2 = \min_{\beta,\delta} \sum_{i=1}^{2n} \left(g_i(\beta,\delta)\right)^2 \tag{2.3}$$

which is an OLS problem with $(p+n)$ parameters and $2n$ equations. (In all cases in this paper, $\|\cdot\|$ denotes the ℓ_2 vector or matrix norm.) Direct application of trust region methods to (2.3) would require $O(2n(n+p)^2)$ operations per iteration which rapidly becomes prohibitive if n is large. (Recall that n is usually far greater than p in practice.)

The basic idea of a trust region strategy is to choose as the step that vector which minimizes a linear approximation to G over a region in which the linearization is a "reasonable" approximation to G. Specifically, if $G'(\eta^c) \in R^{2n\times(n+p)}$ is the Jacobian matrix of G evaluated at the current iterate, η^c, then the step z is chosen by solving

$$\begin{aligned} &\min_{z} \|G(\eta^c)+G'(\eta^c)z\|^2 \\ &\text{subject to} \|Zz\| \le \tau \end{aligned} \tag{2.4}$$

where Z is a nonsingular (usually diagonal) scaling matrix and τ is the trust region radius. It is easy to show that the solution to (2.4) is given by the $z(\alpha)$ satisfying

$$\left(G'(\eta^c)^T G'(\eta^c) + \alpha Z^T Z\right) z(\alpha) = -G'(\eta^c)^T G(\eta^c) \tag{2.5}$$

where $\alpha > 0$ is the Lagrange multiplier for the inequality constraint. Note that if $\|Zz(0)\| \le \tau$, $\alpha = 0$ and the constraint is inactive. Otherwise $\alpha > 0$ and the constraint is active. Equation (2.5) is the famous Levenberg-Marquardt formula, but this derivation has given rise to more stable and robust implementations. (See, e.g., [Mor77] and [DenS83]). Clearly (2.5) can be regarded as the"normal equations" for the extended least squares problem,

$$\begin{bmatrix} G'(\eta^c) \\ \alpha^{1/2}Z \end{bmatrix} z =_2 - \begin{bmatrix} G \\ 0 \end{bmatrix} \tag{2.6}$$

where "$=_2$" means "equal in the least squares sense."

Our implementation is based on the careful exploitation of the structure of the extended Jacobian matrix in (2.6). From (2.2) we have that

$$G'(\eta^c) = \begin{pmatrix} J & V \\ 0 & D \end{pmatrix}$$

where

$$J \in R^{n\times p}: \quad J_{ij} = \frac{\partial g_i(\beta,\delta)}{\partial \beta_j} = \frac{w_i \partial f(x_i+\delta_i;\beta)}{\partial \beta_j},$$
$$i = 1,\cdots,n, \qquad j = 1,\cdots,p;$$
$$V \in R^{n\times n}: \quad V_{ij} = \frac{\partial g_i(\beta,\delta)}{\partial \delta_j} = \frac{w_i \partial f(x_i+\delta_i;\beta)}{\partial \delta_j},$$
$$i = 1,\cdots,n, \qquad j = 1,\cdots,n;$$
$$D \in R^{n\times n}: \quad D = \operatorname{diag}\{w_i d_i, i = 1,\cdots,n\}.$$

Here, we have omitted the arguments of J and V for the sake of clarity. Observe that since g_i only depends on δ_i, $i = 1,\cdots,n$,

$$V = \operatorname{diag}\left\{\frac{\partial g_i(\beta,\delta)}{\partial \delta_i}, i = 1,\cdots,n\right\}.$$

Commensurate with this partitioning of $G'(\eta^c)$, η^c is naturally partitioned into components $(\beta^c,\delta^c)^T$ and the step z into a step in β, say s, and a step in δ, say t. Furthermore, we allow for s to be scaled by a nonsingular diagonal scaling matrix S and t by a nonsingular diagonal matrix T. Thus (2.6) becomes

$$\begin{bmatrix} J & V \\ 0 & D \\ \alpha^{1/2}S & 0 \\ 0 & \alpha^{1/2}T \end{bmatrix} \begin{bmatrix} s \\ t \end{bmatrix} =_2 - \begin{bmatrix} G_1 \\ G_2 \\ 0 \\ 0 \end{bmatrix} \tag{2.7}$$

where G_1 is the first n components of G and G_2 is the last n components.

Now, if $x_i \in R^m$, then (2.7) will have the same form except that $V \in R^{n\times nm}$; $T, D \in R^{nm\times nm}$ are still diagonal; and V, instead of being diagonal, has the "staircase" structure which is illustrated for $n = 4$ and $m = 3$ as follows:

$$V = \begin{bmatrix} xxx & & & \\ & xxx & & \\ & & xxx & \\ & & & xxx \end{bmatrix} = \begin{bmatrix} V_1 & & & \\ & V_2 & & \\ & & V_3 & \\ & & & V_4 \end{bmatrix}$$

where $V_i \in R^{1\times m}$. The rest of the development now allows $x_i \in R^m$.

To derive an efficient procedure for solving (2.7), we first form the normal equations associated with (2.7):

$$\begin{pmatrix} J^TJ + \alpha S^2 & J^TV \\ V^TJ & V^TV + D^2 + \alpha T^2 \end{pmatrix} \begin{pmatrix} s \\ t \end{pmatrix} = - \begin{pmatrix} J^TG_1 \\ V^TG_1 + DG_2 \end{pmatrix}. \tag{2.8}$$

Let $P = V^TV + D^2 + \alpha T^2$. Solving the bottom equations of (2.8) for t in terms of s yields

$$t = -P^{-1}(V^TG_1 + DG_2 + V^TJs) \tag{2.9}$$

and then substituting (2.9) into the top equations of (2.8) gives

$$[J^T(I - VP^{-1}V^T)J + \alpha S^2]\, s = -J^T G_1 + J^T V P^{-1}(V^T G_1 + DG_2). \tag{2.10}$$

Thus s is the solution to the least squares problem

$$\begin{pmatrix} (I - VP^{-1}V^T)^{1/2} J \\ \alpha^{1/2} S \end{pmatrix} s \\ =_2 \begin{pmatrix} (I - VP^{-1}V^T)^{-1/2} \left[-G_1 + VP^{-1}(V^T G_1 + DG_2)\right] \\ 0 \end{pmatrix}. \tag{2.11}$$

The following propositions provide the necessary tools for the accurate and efficient solution of (2.9) and (2.11).

PROPOSITION 2.1. *Let* $E = D^2 + \alpha T^2$. *Then,* $VP^{-1}V^T = diag\{\omega_i/(1+\omega_i), i = 1, \cdots n\}$ *where*

$$\omega_i = \sum_{j=1}^{m} \frac{V_{ij}^2}{E_{(i-1)m+j}}, \qquad i = 1, \cdots, n.$$

Proof. We have that E is a nonsingular diagonal matrix and that $V^T V$ is block diagonal where the blocks are $m \times m$ rank 1 matrices. Thus P^{-1} can be calculated by the Sherman-Morrison-Woodbury formula (see, e.g., [DenS83, p. 188] or [OrtR70, p. 50]) as

$$\begin{aligned} P^{-1} &= (V^T V + E)^{-1} \\ &= E^{-1} - E^{-1} V^T (I + VE^{-1}V^T)^{-1} VE^{-1}. \end{aligned} \tag{2.12}$$

By direct calculation, $VE^{-1}V^T = \text{diag}\,\{\omega_i,\ i = 1, \cdots, n\}$ with ω_i as defined above. The calculation of $VP^{-1}V^T$ follows directly from (2.12).

Diagonal matrices involving the ω_i frequently enter into the formulas which follow. We simplify the notation by defining

$$M = \text{diag}\left\{ \left[\frac{1}{1+\omega_i} \right]^{1/2},\ i = 1, \cdots, n \right\}.$$

PROPOSITION 2.2. $(I - VP^{-1}V^T)^{1/2} = M$ *where* M *is as above.*

Proof. The result follows immediately from Proposition 2.1.

PROPOSITION 2.3. *With* M *as above*

$$VP^{-1} = M^2 VE^{-1}$$

and

$$P^{-1}V^T = E^{-1}V^T M^2.$$

Proof. The results follow directly from (2.12) and Propositions 2.1 and 2.2.

PROPOSITION 2.4. *The right-hand side of* (2.11) *is given by*

$$(I - VP^{-1}V^T)^{-1/2} \left[-G_1 + VP^{-1}(V^T G_1 + DG_2)\right] = -M(G_1 - VE^{-1}DG_2).$$

Proof. We have that

$$\begin{aligned} &-G_1 + VP^{-1}(V^TG_1 + DG_2) \\ &\quad = -(I - VP^{-1}V^T)G_1 + VP^{-1}DG_2 \\ &\quad = -M^2G_1 + VP^{-1}DG_2. \end{aligned}$$

Now the result follows immediately from Propositions 2.2 and 2.3.

With the above formulas, we can solve (2.11) for s by first calculating E and ω in $O(nm)$ operations and then forming

$$\bar{J} = MJ$$

and

$$y = -M(G_1 - VE^{-1}DG_2).$$

Now rewrite the least squares problem (2.11) as

$$\begin{pmatrix} \bar{J} \\ \alpha^{1/2}S \end{pmatrix} s =_2 \begin{pmatrix} y \\ 0 \end{pmatrix} \tag{2.13}$$

which requires $O(np + nm)$ operations. The solution of (2.13) then involves a QR decomposition of $\bar{J}$ (accomplished by Householder transformations with column pivoting) and then a sequence of plane rotations to eliminate $\alpha^{1/2}S$. The cost for this phase is dominated by the $O(np^2)$ operations for the QR decomposition of $\bar{J}$.

Using Proposition 2.3 and (2.12), it is easily verified that

$$t = -E^{-1}\left[V^TM^2(G_1 + Js - VE^{-1}DG_2) + DG_2\right]$$

which is dominated in cost by the $O(np)$ operations needed to form Js and several $O(nm)$ terms. Thus the leading cost of calculating a step for ODR is the same $O(np^2)$ operations needed to do the factorization of an $n \times p$ matrix as in OLS. The only additional costs are a small number of calculations costing $O(nm)$ or $O(np)$ operations.

It may occur to the reader that an efficient QR factorization of the matrix in (2.7) might yield a procedure with the same order of work. By re-ordering the upper 2×2 blocks, one can, indeed, do the factorization of this part in $O(np^2)$ operations. The subsequent elimination of the αS and αT blocks, however, would require $O((nm+p)^2)$ operations for each α. It is for this reason, as well as others, that Schwetlick and Tiller [SchT85] do only a "partial" trust region strategy, i.e., their trust region only applies to the step in the β variables. In some badly scaled problems, however, (e.g., Example 3 in §4) the ability to scale and constrain the step in δ is essential to solve the problem.

The above formulas for s and t are used for each α value in (2.5). Thus in order to complete the specification of the algorithm, we need to provide the procedure for computing the trust region parameter α to satisfy (2.4) and to discuss a few miscellaneous details.

Moré [Mor77], following Hebden [Heb73] (see also [DenS83]), suggested a procedure for computing α in (2.5) so that

$$\|Zz(\alpha)\| \approx \tau$$

when $\|Zz(0)\| > \tau$. This procedure is based on approximating the function

$$\phi(\alpha) = \|Zz(\alpha)\| - \tau$$

by a rational function $\theta(\alpha) = \gamma/(\mu - \alpha)$ where the constants γ and μ are chosen by making $\theta(\alpha^c) = \phi(\alpha^c)$ and $\theta'(\alpha^c) = \phi'(\alpha^c)$. This results in the iteration

$$\alpha^+ = \alpha^c - \frac{\phi(\alpha^c)}{\phi'(\alpha^c)} \times \frac{\phi(\alpha^c) - \tau}{\tau}$$

which is a modified Newton step for the equation $\phi(\alpha) = 0$. In our case, the derivative of $\phi(\alpha)$ is not as simple to compute as in OLS and thus we opt to compute γ and μ by making $\theta(\alpha^c) = \phi(\alpha^c)$ and $\theta(\alpha^-) = \phi(\alpha^-)$. ($\alpha^-$ is the previous estimate.) Then

$$\alpha^+ = \alpha^c - \frac{\phi(\alpha^c)(\alpha^c - \alpha^-)}{\phi(\alpha^c) - \phi(\alpha^-)} \times \frac{\phi(\alpha^c) + \tau}{\tau}$$

which is clearly a modified secant step for the equation $\phi(\alpha) = 0$. Moré found it necessary to safeguard his procedure by computing and updating upper and lower bounds on α. Similarly, we maintain such bounds, but with formulas appropriate to the secant-like method. These will be provided in a subsequent paper.

The trust region bound τ is updated according to well-tested ideas which are in several existing codes. (See e.g., [Mor77], [Gay84], [SchKW86].) A valuable feature in our code has been the "internal doubling" step. For a given η^c and τ, suppose z_τ is generated such that τ restricts z and the reduction in $\|G\|$ predicted by the linear approximation agrees with the actual reduction in $\|G\|$ to a high precision. Normally, one would accept z_τ , set $\eta^+ := \eta^c + z_\tau$ and double τ for the next step. The internal doubling procedure is to remember $\eta^r := \eta^c + z_\tau$, double τ and generate a new z_τ from η^c. Note that this procedure only costs an evaluation of G and, if successful, may save several evaluations of J . In practice, it has been successful often enough to warrant leaving it in. Its main advantage is that it permits rapid and cheap increases in τ based on an overly conservative initial value of τ or when the iterates are moving away from a highly nonlinear region in parameter space.

Since many users will want to compare the results of OLS with ODR, an option to do OLS has been implemented. Enabling this option merely initializes the δ vector to zero and sets V to zero whenever it is computed. It is easily verified that, in this case, (2.11) reduces to the OLS Levenberg-Marquardt step and (2.9) yields $t = 0$ leaving $\delta = 0$. Using this procedure to do OLS, therefore, is equivalent to a standard OLS algorithm with a moderate extra algebraic overhead.

3. Local and global convergence analysis. The global convergence properties of trust region-Levenberg-Marquardt methods applied to the general nonlinear least squares problem (2.3) are well known (see e.g., [Pow75], [Mor77], [MorS81], [SchSB85]). As long as the sequence of Jacobian matrices, $\{G'(\eta_k)\}$, is uniformly bounded, then

$$\lim_{k \to \infty} G'(\eta_k)^T G(\eta_k) = 0,$$

so that any cluster point satisfies the first order necessary conditions for a local minimizer. These results apply to our algorithm and nothing more needs to be said regarding global convergence.

The local convergence behavior of general trust region-Levenberg-Marquardt methods for nonlinear least squares is discussed by Byrd and Schnabel [ByrS87] who show that, if there is a cluster point η_* where $G'(\eta_*)$ is nonsingular, then the iterates converge at least linearly to η_* independent of the size of $G(\eta_*)$. This theory also applies to our algorithms. If, in addition, the residual $G(\eta_*)$ is sufficiently small, Byrd and Schnabel show that asymptotically the trust region constraint becomes inactive, and that the Levenberg-Marquardt algorithm reduces to the Gauss-Newton iteration

$$\eta_{k+1} = \eta_k - \left[G'(\eta_k)^T G'(\eta_k)\right]^{-1} G'(\eta_k)^T G(\eta_k)$$

and is linearly convergent to η_*. The linear convergence analysis of the Gauss-Newton method is well known (see e.g., [OrtR70], [DenS83]). The constant of linear convergence depends upon the smallest singular value of $G'(\eta_*)$, the residual $G(\eta_*)$, and the nonlinearity of $G(\eta)$ near η_*.

The small residual analysis is particularly relevent to ODR because most applications of ODR will have small residuals. This is especially true when ODR is used to consider errors in independent variables in parameter estimation, because errors in the independent variables are most likely to be considered when the model and the dependent variable measurements are accurate, which implies that the residuals will be small.

It turns out that the application of the local Gauss-Newton analysis to ODR is nontrivial, although the expected results can be proven. This is the main contribution of this section. To simplify the algebra here, we consider a version of the ODR problem (2.1) with the simplified weighting scheme $w_i = 1$ and $d_i = \sigma$ for all i, i.e.,

$$\min_{\beta,\,\delta} \sum_{i=1}^{n} \left[(f(x_i + \delta_i; \beta) - y_i)^2 + \sigma^2 \delta_i^T \delta_i \right] \tag{3.1}$$

where $\sigma > 0$. This weighting still allows the metric of distance from the curve $f(x;\beta)$ to the data points (x_i, y_i) to vary from vertical (as $\sigma \to \infty$) to orthogonal ($\sigma = 1$) to horizontal (as $\sigma \to 0$). (We explain this statement more carefully later in this section.) This is all the generality in the weighting that is usually used in practice, and precisely what we use in most of our computational results in §4.

In fact, as we illustrate in §4, in practical ODR applications, the user may wish to solve (3.1) for various values of σ. A second contribution of this section is to produce a bound on the constant of linear convergence of the Gauss-Newton method applied to (3.1) that is independent of the value of σ.

To further simplify notation, we rewrite (3.1) as

$$\min_{\eta} R(\eta)^T R(\eta) + \sigma^2 \delta^T \delta \tag{3.2a}$$

or equivalently,

$$\min_{\eta} G(\eta)^T G(\eta) \tag{3.2b}$$

where $\delta = (\delta_1^T, \delta_2^T, \cdots, \delta_n^T)^T$, $\eta = (\beta^T, \delta^T)^T$, $R(\eta)_i = f(x_i + \delta_i; \beta) - y_i$, $i = 1, \cdots, n$, and $G(\eta) = (R(\eta)^T, \sigma\delta^T)^T$. Our analysis will not depend upon the special form of

$R(\eta)$ in any way. Recall that

$$G'(\eta) = \begin{pmatrix} J(\eta) & V(\eta) \\ 0 & \sigma I \end{pmatrix}$$

where $J(\eta)$ and $V(\eta)$ are, as in §2, the derivatives of $R(\eta)$ with respect to β and δ respectively.

The difficulty in applying standard Gauss-Newton analysis to (3.2) is that $G(\eta)$ and $G'(\eta)$ are functions of σ. In Theorem 3.5 we show that the convergence can be analyzed in terms of the properties of $J(\eta)$, $V(\eta)$, $R(\eta_*)$, and δ_* only, i.e., independent of σ except for its role in determining η_*. Lemmas 3.3 and 3.4 are used in the proof of Theorem 3.5 to bound the linear and quadratic terms in the convergence results, respectively, independent of σ. Lemmas 3.1 and 3.2 are technical results used in the proof of Lemma 3.3. Theorem 3.5 can be applied to the more generally weighted problem (2.1) by making an appropriate change of variables.

LEMMA 3.1. *Let $A \in R^{p\times p}$, $C \in R^{q\times q}$, $B \in R^{p\times q}$, and let*

$$H = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}$$

be symmetric and positive definite. Then $\|H\| \le \|A\| + \|C\|$.

Proof. Define $a = \|A\|, c = \|C\|$. For any $v \in R^p$ and $w \in R^q$, consider

$$y = \begin{bmatrix} v\|w\|\sqrt{c} \\ -w\|v\|\sqrt{a} \end{bmatrix}.$$

Then, since H is positive definite,

$$0 \le y^T H y = v^T A v \|w\|^2 c - 2v^T B w \|w\| \|v\| \sqrt{ac} + w^T C w \|v\|^2 a$$

and therefore

$$\begin{aligned} 2v^T B w &\le \frac{v^T A v \|w\|^2 c + w^T C w \|v\|^2 a}{\|w\|\|v\|\sqrt{ac}} \\ &\le \frac{\|v\|^2 a \|w\|^2 c + \|w\|^2 c \|v\|^2 a}{\|w\|\|v\|\sqrt{ac}} \\ &= 2\|v\|\|w\|\sqrt{ac} \\ &\le \|v\|^2 c + \|w\|^2 a. \end{aligned}$$

Thus

$$\begin{aligned} \|H\| &= \max_{v\in R^p,\ w\in R^q} \frac{v^T A v + 2v^T B w + w^T C w}{v^T v + w^T w} \\ &\le \max_{v\in R^p,\ w\in R^q} \frac{\|v\|^2 a + (\|v\|^2 c + \|w\|^2 a) + \|w\|^2 c}{v^T v + w^T w} \\ &= a + c. \end{aligned}$$

LEMMA 3.2. *Let $A \in R^{n\times p}$ have full column rank and $B \in R^{n\times n}$ be positive definite. Then*

$$\|(A^T B^{-1} A)^{-1}\| \le \|(A^T A)^{-1}\| \|B\|.$$

Proof.

$$\|(A^TB^{-1}A)^{-1}\| = \max_{v\in R^p} \frac{v^Tv}{v^TA^TB^{-1}Av}$$
$$\leq \max_v \frac{v^Tv\|B\|}{v^TA^TAv} \leq \|(A^TA)^{-1}\|\,\|B\|.$$

LEMMA 3.3. *Let $J \in R^{n\times p}$ have full column rank, $V \in R^{n\times q}$, I the $q \times q$ identity, and σ a positive scalar. Let*

$$M(\sigma) = \begin{pmatrix} J & V \\ 0 & \sigma I \end{pmatrix}, \quad \textit{and} \quad N(\sigma) = M(\sigma)^TM(\sigma).$$

Then $N(\sigma)$ is nonsingular and

$$\|N(\sigma)^{-1}\| \leq \|(J^TJ)^{-1}\| + \sigma^{-2}\left(1 + \|(J^TJ)^{-1}\|\,\|V\|^2\right). \tag{3.3}$$

Proof. Since J has full column rank, $M(\sigma)$ has full column rank. Thus $N(\sigma)$ is positive definite, and it is straightforward to verify that

$$N(\sigma)^{-1} = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix},$$

where

$$A = \left[\sigma^2J^T\left(V^TV + \sigma^2I\right)^{-1}J\right]^{-1},$$
$$C = \left[\sigma^2I + V^T\left(I - J(J^TJ)^{-1}J^T\right)V\right]^{-1},$$
$$B = -(J^TJ)^{-1}J^TVC.$$

From Lemma 3.2,

$$\begin{aligned}\|A\| &\leq -\sigma^{-2}\|(J^TJ)^{-1}\|\,\|V^TV + \sigma^2I\| \\ &= \sigma^{-2}\|(J^TJ)^{-1}\|(\|V^TV\| + \sigma^2) \\ &= \sigma^{-2}\|(J^TJ)^{-1}\|\,\|V\|^2 + \|(J^TJ)^{-1}\|.\end{aligned} \tag{3.4}$$

Since $C^{-1} - \sigma^2I = V^T\left[I - J(J^TJ)^{-1}J^T\right]V$ is positive semi-definite, the smallest eigenvalue of C^{-1} is at least σ^2 which shows

$$\|C\| \leq \sigma^{-2}. \tag{3.5}$$

Thus, since $N(\sigma)^{-1}$ is positive definite, applying Lemma 3.1 and using (3.4) and (3.5) gives (3.3).

LEMMA 3.4. *Let the assumptions of Lemma 3.3 hold, and let $u \in R^n$. Let z be the solution to*

$$\min_{z\in R^{p+q}} \left\|\begin{pmatrix} J & V \\ 0 & \sigma I \end{pmatrix} z - \begin{pmatrix} u \\ 0 \end{pmatrix}\right\|. \tag{3.6}$$

Let $J^+ = (J^TJ)^{-1}J^T$, let $Z \in R^{n\times(n-p)}$ satisfy $Z^TZ = I$ and $Z^TJ = 0$, let $\tilde{V} = Z^TV$, and let $\tilde{V}^+$ denote the pseudoinverse of $\tilde{V}$. Then

$$\|z\| \leq \left[\|\tilde{V}^+\| + \|J^+\|\left(1 + \|V\|\,\|\tilde{V}^+\|\right)\right]\|u\|.$$

Proof. Define $y \in R^{p+q}$ by

$$z = \begin{pmatrix} I & -J^+V \\ 0 & I \end{pmatrix} y$$

and let

$$z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}, \qquad y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

where $z_1, y_1 \in R^p$, and $z_2, y_2 \in R^q$. Then (3.6) is equivalent to

$$\min_{y \in R^{p+q}} \left\| \begin{pmatrix} J & ZZ^TV \\ 0 & \sigma I \end{pmatrix} y - \begin{pmatrix} u \\ 0 \end{pmatrix} \right\| \tag{3.7}$$

since $(I - JJ^+) = ZZ^T$. From the normal equations and using $J^TZ = 0$, the solution to (3.7) is

$$\begin{aligned} y_1 &= J^+u, \\ y_2 &= (V^TZZ^TV + \sigma^2 I)^{-1}V^TZZ^Tu \\ &= (\tilde{V}^T\tilde{V} + \sigma^2 I)^{-1}\tilde{V}^TZ^Tu. \end{aligned}$$

It is well known, for instance from the theory of Levenberg-Marquardt methods, that

$$\|y_2\| \le \left\|\tilde{V}^+Z^Tu\right\|,$$

and since $\left\|Z^Tu\right\| \le \|u\|$,

$$\|y_2\| \le \left\|V^+\right\| \|u\|, \qquad \|y_1\| \le \left\|J^+\right\| \|u\|.$$

Finally since $z_2 = y_2$ and $z_1 = y_1 - J^+Vy_2$,

$$\|z_2\| \le \left\|\tilde{V}^+\right\| \|u\|$$

and

$$\begin{aligned} \|z_1\| &\le \|y_1\| + \left\|J^+\right\| \|V\| \|y_2\| \\ &\le \left\|J^+\right\| \left(1 + \|V\| \left\|\tilde{V}^+\right\|\right) \|u\|. \end{aligned}$$

Using $\|z\| \le \|z_1\| + \|z_2\|$ gives the desired result.

Throughout the statement and proof of Theorem 3.5, we will often omit the argument η; i.e., we will denote $G(\eta_*)$ and $G(\eta_0)$ by G_* and G_0, respectively, and likewise for other symbols in place of G. Also for J having full column rank, J^+ will denote $(J^TJ)^{-1}J^T$, and for $\tilde{V}$ having full row rank, $\tilde{V}^+$ will denote $\tilde{V}^T(\tilde{V}\tilde{V}^T)^{-1}$. Note that $\|J^+\|^2 = \left\|(J^TJ)^{-1}\right\|$.

THEOREM 3.5. *Let $R(\eta) : R^t \to R^n$ be continuously differentiable in an open convex set $D \subset R^t$. Let $\eta^T = (\beta^T, \delta^T)$, $\beta \in R^p$, $\delta \in R^q$, let σ be a positive scalar, and let*

$$G(\eta) = \begin{pmatrix} R(\eta) \\ \sigma\delta \end{pmatrix}.$$

Assume there exists $\eta_* \in D$ such that $G'(\eta_*)^T G(\eta_*) = 0$, and that there exists $\gamma \geq 0$ for which

$$\|R'(\eta) - R'(\eta^*)\| \leq \gamma \|\eta - \eta_*\| \tag{3.8}$$

for all $\eta \in D$. Assume that $J(\eta_*)$ and $R'(\eta_*)$ have full column and row rank, respectively. Let $Z_* \in R^{n\times(n-p)}$ satisfy $Z_*^T Z_* = I$, $Z_*^T J_* = 0$, and define $\tilde{V}_* = Z_*^T V_*$. Define

$$c_1 = \gamma \left[\left\| (J_*^T J_*)^{-1} \right\| \|R_*\| + \left(1 + \left\| (J_*^T J_*)^{-1} \right\| \|V_*\|^2\right) \left\| \tilde{V}_*^+ \right\| \|\delta_*\| \right],$$
$$c_2 = (\gamma/2) \left[\left\| \tilde{V}_*^+ \right\| + \left\| J_*^+ \right\| \left(1 + \|V_*\| \left\| \tilde{V}_*^+ \right\|\right)\right].$$

If $c_1 < 1$, then for any $c \in (1, 1/c_1)$, there exists $\epsilon > 0$ such that for all η_0 for which $\|\eta_0 - \eta_*\| \leq \epsilon$, the sequence generated by the Gauss-Newton method

$$\eta_{k+1} = \eta_k - \left(G'(\eta_k)^T G'(\eta_k)\right)^{-1} G'(\eta_k)^T G(\eta_k)$$

is well defined, converges to η_* and obeys

$$\|\eta_{k+1} - \eta_*\| \leq c\,(c_1 + c_2 \|\eta_k - \eta_*\|) \|\eta_k - \eta_*\|. \tag{3.9}$$

Proof. The crucial remaining part of the proof is to note that since the optimality condition $G_*' G_* = 0$ gives $V_*^T R_* + \sigma^2 \delta_* = 0$, and since $R_* = Z_* Z_*^T R_*$ because $J_*^T R_* = 0$, we have

$$V_*^T Z_* Z_*^T R_* = \tilde{V}_*^T Z_*^T R_* = -\sigma^2 \delta_*.$$

Since R_*' has full row rank, $\tilde{V}_*$ has full row rank so that

$$Z_*^T R_* = \sigma_*^2 (\tilde{V}_* \tilde{V}_*^T)^{-1} \tilde{V}_*^T \delta_* = -\sigma_*^2 (\tilde{V}_*^+)^T \delta_*$$

and

$$\|R_*\| = \left\| Z_* Z_*^T R_* \right\| \leq \left\| Z_*^T R_* \right\| \leq \sigma_*^2 \left\| \tilde{V}_*^+ \right\| \|\delta_*\|.$$

Thus from (3.3) of Lemma 3.2,

$$\begin{aligned} &\left\| \left((G_*')^T G_*'\right)^{-1} \right\| \|R_*\| \\ &\quad \leq \left\| \left(J_*^T J_*\right)^{-1} \right\| \|R_*\| + \sigma_*^{-2} \left(1 + \left\| \left(J_*^T J_*\right)^{-1} \right\| \|V_*\|^2\right) \|R_*\| \\ &\quad \leq \left\| \left(J_*^T J_*\right)^{-1} \right\| \|R_*\| + \left(1 + \left\| \left(J_*^T J_*\right)^{-1} \right\| \|V_*\|^2\right) \left\| \tilde{V}_*^+ \right\| \|\delta_*\| \\ &\quad = c_1/\gamma. \end{aligned} \tag{3.10}$$

The remainder of the proof is by induction. Let c be a fixed constant in $(1, 1/c_1)$. Then there exists $\epsilon_1 > 0$ such that for all η for which $\|\eta - \eta_*\| \leq \epsilon_1$, $J(\eta)$and$V(\eta)$ have full column and row rank, respectively, and

$$\left\| \left(G'(\eta)^T G'(\eta)\right)^{-1} \right\| \leq c \left\| \left((G_*')^T G_*'\right)^{-1} \right\|, \tag{3.11}$$

and

$$\begin{aligned}(3.12)\qquad &\left\|\tilde{V}(\eta)^+\right\| + \|J(\eta)^+\| \left(1 + \|V(\eta)\| \left\|\tilde{V}(\eta)^+\right\|\right) \\ &\quad \le c\left[\left\|\tilde{V}_*^+\right\| + \|J_*^+\| \left(1 + \|V_*\| \left\|\tilde{V}_*^+\right\|\right)\right].\end{aligned}$$

Let $\epsilon = \min\{\epsilon_1, (1 - cc_1)/(2cc_2)\}$. Then at the first iteration $\left((G_0')^T G_0'\right)$ is nonsingular, and

$$\begin{aligned}(3.13)\qquad \eta_1 - \eta_* &= \eta_0 - \eta_* - \left((G_0')^T G'_0\right)^{-1} (G_0')^T G_0 \\ &= a + b\end{aligned}$$

where

$$\begin{aligned}a &= -\left((G_0')^T G_0'\right)^{-1} (G_0')^T G_*, \\ b &= \left((G_0')^T G_0'\right)^{-1} (G_0')^T (G_* - G_0 - G_0'(\eta_* - \eta_0)).\end{aligned}$$

Since $(G_*')^T G_* = 0$,

$$(3.14)\qquad (G_0')^T G_* = (G_0' - G_*')^T G_* = (R_0' - R_*')^T R_*$$

due to the linearity of the final q components of $G(\eta)$. Using (3.14) plus (3.8), (3.10), and (3.11), we have

$$\begin{aligned}(3.15)\qquad \|a\| &\le \left\|\left((G_0')^T G'_0\right)^{-1}\right\| \|R_0' - R_*'\| \|R_*\| \\ &\le c\gamma \left\|\left((G_*')^T G_*'\right)^{-1}\right\| \|R_*\| \|\eta_0 - \eta_*\| \\ &\le cc_1 \|\eta_0 - \eta_*\|.\end{aligned}$$

Also note that b is the solution to

$$\min \|G_0' b - [G_* - G_0 - G_0'(\eta_* - \eta_0)]\|,$$

and that

$$(3.16)\qquad G_* - G_0 - G_0'(\eta_* - \eta_0) = \begin{pmatrix} R_* - R_0 - R_0'(\eta_* - \eta_0) \\ 0 \end{pmatrix}.$$

From (3.8) and standard results

$$(3.17)\qquad \|R_* - R_0 - R_0'(\eta_* - \eta_0)\| \le (\gamma/2) \|\eta_0 - \eta_*\|^2.$$

Thus using (3.16), (3.17), Lemma 3.4, and (3.12),

$$\begin{aligned}(3.18)\qquad \|b\| &\le \left[\left\|\tilde{V}_0^+\right\| + \|J_0^+\| \left(1 + \|V_0\| \left\|\tilde{V}_0^+\right\|\right)\right] (\gamma/2) \|\eta_0 - \eta_*\|^2 \\ &\le cc_2 \|\eta_0 - \eta_*\|^2.\end{aligned}$$

Substituting (3.15) and (3.18) into (3.13) and recalling $\|\eta_0 - \eta_*\| \le (1 - cc_1)/2cc_2$,

$$\begin{aligned}\|\eta_1 - \eta_*\| &\le c(c_1 + c_2 \|\eta_0 - \eta_*\|) \|\eta_0 - \eta_*\| \\ &\le [(1 + cc_1)/2] \|\eta_0 - \eta_*\|\end{aligned}$$

which proves (3.9) in the case $k = 0$ and shows that $\|\eta_1 - \eta_*\| < \|\eta_0 - \eta_*\|$. The proof of the induction step is identical.

Byrd and Schnabel [ByrS87] show that the trust region-Levenberg-Marquardt algorithm described in §2 will reduce asymptotically to Gauss-Newton if the constant of linear convergence, c_1, in Theorem 3.5 is sufficiently small. In particular, our computer code will accept the trial step if the ratio of actual to "predicted" reduction,

$$\frac{\|G(\eta_c + z_T)\| - \|G(\eta_c)\|}{\|G(\eta_c) + G'(\eta_c)z_T\| - \|G(\eta_c)\|}$$

is at least 0.001. In this case, as long as $c_1 < .9999$, the trust region in the Levenberg-Marquardt algorithm is inactive asymptotically, so that the algorithm becomes Gauss-Newton.

Now we consider the behavior of the ODR problem (3.2) as the parameter σ is varied. For this purpose, let us denote the global minimizer to (3.2) by $\eta_*(\sigma)$. Then by standard analyses of barrier function methods, (see e.g., [FiaM68] or [Lue73]) we know that the limit of $\eta_*(\sigma)$ as $\sigma \to \infty$ is the solution to

$$\min_{\eta} \|R(\eta)\|^2 \quad \text{subject to } \delta = 0,$$

i.e., the standard OLS problem

$$\min_{\beta} \|R(\beta, 0)\|^2. \tag{3.19}$$

Similarly, the limit of $\eta_*(\sigma)$ as $\sigma \to 0$ is the solution to the implicit least squares (ILS) problem

$$\min_{\eta} \|\delta\|^2 \quad \text{subject to } R(\eta) = 0. \tag{3.20}$$

In the data fitting context where $R(\eta)_i = f(x_i + \delta_i; \beta) - y_i$, (3.19) is the standard problem where the independent variables x_i are assumed exact so that the metric of distance is in the y (vertical) direction only. In constast (3.20) is the case where the dependent variables y_i are assumed exact and the independent variables x_i inexact, so that the metric is entirely in the x (horizontal) direction.

The standard analysis of barrier function methods also shows that $\|R(\eta_*(\sigma))\|$ is a monotonically increasing function of σ, and that $\|\delta_*(\sigma)\|$ is a monotonically decreasing function of σ. This means that for all $\sigma \in (0, \infty)$, the values of $\|R(\eta_*(\sigma))\|$ and $\|\delta_*(\sigma)\|$ are bounded above by the optimal objective function values for problems (3.19) and (3.20), respectively. In data fitting terms, for any σ, the norm of the optimal vertical residuals in ODR is bounded above by the norm of the optimal residuals in OLS, and the norm of the optimal horizontal residuals in ODR is bounded above by the norm of the optimal residual for the ILS problem. The computational results of Section 4 demonstrate these relationships.

Combining the above facts with Theorem 3.5 shows that, if the optimal objective function values for problems (3.19) and (3.20) are sufficiently small, and if $J(\eta_*(\sigma))$ and $V(\eta_*(\sigma))$ are sufficiently well-conditioned for all $\sigma \in (0, \infty)$, then the Gauss-Newton algorithm applied to (3.2) is linearly convergent for any $\sigma \in (0, \infty)$.

COROLLARY 3.6. *Let* η, β, δ, $R(\eta)$, $G(\eta)$, $J(\eta)$, *and* $V(\eta)$ *be defined as in Theorem* 3.5 *For any* $\sigma \in (0, \infty)$, *let* $\eta_*(\sigma) = (\beta_*(\sigma)^T, \delta_*(\sigma)^T)^T$ *denote the global solution to*

$$\min_{\beta, \delta} \|R(\beta, \delta)\|^2 + \sigma^2 \|\delta\|^2 . \tag{3.21}$$

Also let β_{OLS} *denote the global solution to the ordinary least squares problem*

$$\min_{\beta} \|R(\beta, 0)\|^2$$

and let $(\beta_{\mathrm{ILS}}, \delta_{\mathrm{ILS}})$ *denote the global solutions to the implicit least squares problem*

$$\min_{\beta, \delta} \|\delta\|^2 \text{ subject to } R(\beta, \delta) = 0.$$

Let $R_{\mathrm{OLS}} = R(\beta_{\mathrm{OLS}})$. *Assume that there exist* $\hat{\epsilon} > 0$, $\hat{\gamma} \geq 0$ *such that for each* $\sigma \in (0, \infty)$,

$$\|R'(\eta) - R'(\eta_*(\sigma))\| \leq \hat{\gamma} \|\eta - \eta_*(\sigma)\|$$

for all η *for which* $\|\eta - \eta_*(\sigma)\| \leq \hat{\epsilon}$. *Assume also that for all* $\sigma \in (0, \infty)$, $J(\eta_*(\sigma))$ *and* $R'(\eta_*(\sigma))$ *have full column and row rank, respectively, and let* $\hat{J}$, $\hat{J}^+$, $\hat{V}$, *and* $\hat{V}^+$ *be uniform bounds on* $\|J(\eta_*(\sigma))\|$, $\|J(\eta_*(\sigma))^+\|$, $\|V(\eta_*(\sigma))\|$, *and* $\left\|\tilde{V}(\eta_*(\sigma))^+\right\|$, *respectively, over all* $\sigma \in (0, \infty)$ *where* $\tilde{V}$ *is defined analogously to the definition in Theorem* 3.5. *Define*

$$\hat{c}_1 = \hat{\gamma}\left[\left(\hat{J}^+\right)^2 R_{\mathrm{OLS}} + \left(1 + \left(\hat{J}^+\right)^2 \hat{V}^2\right) \hat{V} \delta_{\mathrm{ILS}}\right],$$

$$\hat{c}_2 = (\hat{\gamma}/2)\left[\hat{V}^+ + \hat{J}^+\left(1 + \hat{V}\hat{V}^+\right)\right].$$

If $\hat{c}_1 < 1$, *then for any* $c \in (1, 1/\hat{c}_1)$, *there exists* $\hat{\epsilon} > 0$ *such that for any* $\sigma \in (0, \infty)$, *the sequence* $\{\eta_k\}$ *generated by the Gauss-Newton method applied to* (3.2) *starting from any* η_0 *for which* $\|\eta_0 - \eta_*(\sigma)\| \leq \hat{\epsilon}$ *is well-defined, converges to* $\eta_*(\sigma)$, *and obeys*

$$\|\eta_{k+1} - \eta_*(\sigma)\| \leq c\left[\hat{c}_1 + \hat{c}_2 \|\eta_k - \eta_*(\sigma)\|\right] \|\eta_k - \eta_*(\sigma)\| .$$

4. Computational testing. In this section we report the results of preliminary computational testing. These tests, consisting of two contrived problems and one real problem, were selected in order to illustrate the effectiveness of the implementation and to demonstrate the performance of the basic algorithm. They also point out the need for good starting values in δ as well as in β.

The tests allow a contrast between OLS and ODR which is best brought out in terms of the parameter σ and the function $\beta(\sigma)$ from §3. (Recall that $\beta(\infty)$ corresponds to the OLS solution.) Since, in practice, the correct value of σ may not be known exactly, it is of interest to compute $\beta(\sigma)$ for various values of σ. In each example below we have computed several points on the trajectories $\beta(\sigma)$. Note that the solutions for subsequent values of σ are always obtained using the previous solution as the initial condition. Thus only the initial condition for the first problem in each sequence is important. In the first two examples, more than one trajectory is located.

The algorithm was coded in Fortran 77 and run in single precision on the CDC Cyber 855 at the National Bureau of Standards (NBS). The graphics were done on

the Evans and Sutherland PS-300, also located at NBS. For all of the examples, we used $.84 \times 10^{-7}$ for the f-convergence test and $.37 \times 10^{-9}$ for the x-convergence test. These are $\epsilon_{\text{mach}}^{2/3}$ and $\epsilon_{\text{mach}}^{1/2}$, respectively, where ϵ_{mach} is machine presision. The first two examples were run without scaling, i.e., $S = I$ and $T = I$; the third example was scaled as described below.

Example 1. Consider

$$y = \frac{1}{x - 1}$$

and define $x_i = .01 + (i - 1) * .05, i = 1, \cdots, 40$. Next let

$$y_i = \frac{1}{x_i - 1}, \qquad i = 1, \cdots, 40.$$

Now we perturb the data points as follows:

$$x_i := x_i + rx, \qquad y_i := y_i + ry$$

where the rx are uniformly distributed on $(-.05, .05)$ and the ry are uniformly distributed on $(-.25, .25)$. The model for the data was taken to be

$$y = \frac{\beta_1}{x - \beta_2}$$

and the ODR program was run with several values of σ.

The results are reported in Tables 1 and 2, and Figs. 5, 6 and 7. Table 1 was generated by taking $\beta^0 = (1, 1)^T$ and $\delta^0 = 0$. The second trajectory, reported in

TABLE 1

σ	$\beta_1(\sigma)$	$\beta_2(\sigma)$	Evaluations of G	Evaluations of G'	Final Value of $\|G_1(\eta(\sigma))\|$	Final Value of $\|G_2(\eta(\sigma))\|$
∞	0.6866	0.9909	52	15	21.773	0.0
1000	0.8248	0.9936	10	8	18.881	6.980
500	0.9486	0.9953	12	10	15.524	8.775
300	0.9881	0.9927	1	11	10.409	10.611
100	0.9246	0.9971	8	6	3.203	6.111
25	0.9847	1.001	6	4	1.280	2.003
5	1.015	1.003	6	4	0.711	0.637
2	1.021	1.005	6	4	0.454	0.446
1	1.022	1.005	6	4	0.228	0.354

TABLE 2

σ	$\beta_1(\sigma)$	$\beta_2(\sigma)$	Evaluations of G	Evaluations of G'	Final Value of $\|G_1(\eta(\sigma))\|$	Final Value of $\|G_2(\eta(\sigma))\|$
∞	-0.3172	1.099	20	9	104.708	0.0
1000	-0.3355	1.095	21	13	104.233	6.947
700	-0.3840	1.092	25	19	103.661	10.768
500	0.9486	0.9953	30	16	15.524	8.775

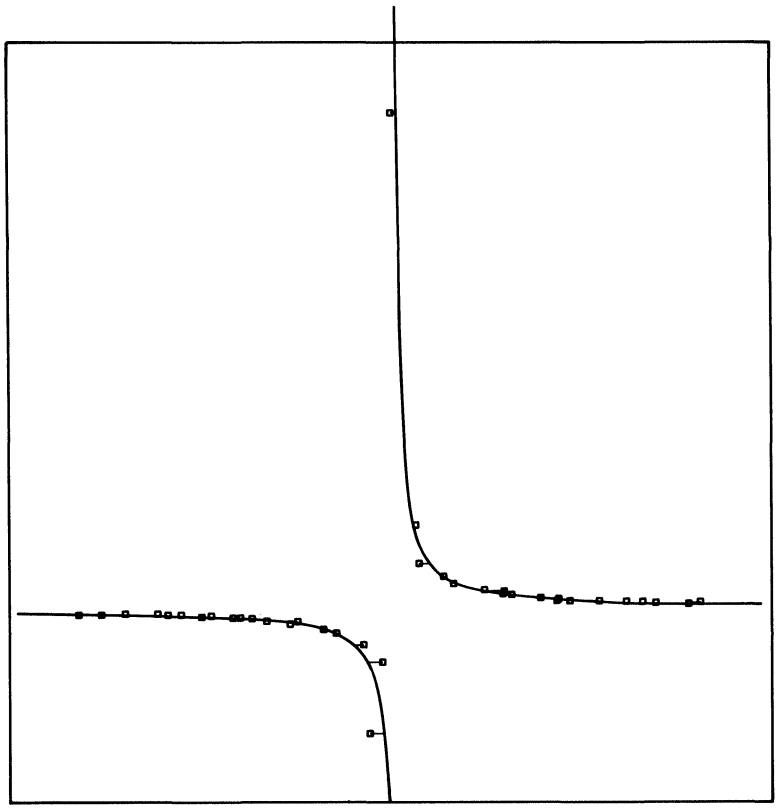

Figure 5
$\sigma = 5$

Table 2, was generated by taking $\beta^0 = (1, 1.1)^T$ and $\delta^0 = 0$. In addition to the values of $\beta(\sigma)$, Tables 1 and 2 contain the number of evaluations of the extended residual function G (cf (2.3)) and its Jacobian, and the optimal values of $\|G_1(\eta(\sigma))\|$ and $\|G_2(\eta(\sigma))\|$ for each value of σ. (Note that since $\|G_2(\eta(\sigma))\| = \sigma \|\delta(\sigma)\|$, dividing the entries in this column by the corresponding value of σ shows the monotonic behavior of $\|\delta(\sigma)\|$ as predicted in §3.) The graphs are as follows: Fig. 5 corresponds to the $\sigma = 2$ fit; Fig. 6 corresponds to the OLS fit from Table 1; and Fig. 7 to the OLS fit from Table 2. Note that on all of these plots, the y-axis has been scaled by a factor of approximately 100 in order to get all of the points on the plot. Because of this scaling, the error pegs (connecting the data points to their predicted values) appear to be much more horizontal than they really are.

Obviously, Tables 1 and 2 exhibit a nonuniqueness of the solutions. It appears that there are two local solutions for the OLS problem corresponding to the asymptote to $+\infty$ being on the left or right half of the curve, and that either the trajectories emanating from these solutions come together near $\sigma = 600$ or the trajectory represented in Table 2 fails to be continuous near $\sigma = 600$. A possible means of investigating this phenomenon is to write the differential equation describing the trajectory $\beta(\sigma)$ and to study possible bifurcation points. This is not pursued here.

Observe that β_2 determines the position of the asymptote and thus the data locate this parameter very well. Note, however, that the data point nearest the asymptote, corresponding to $(x_i^a, y_i^a) = (1.01, 100)$, completely dominates the fitting process for

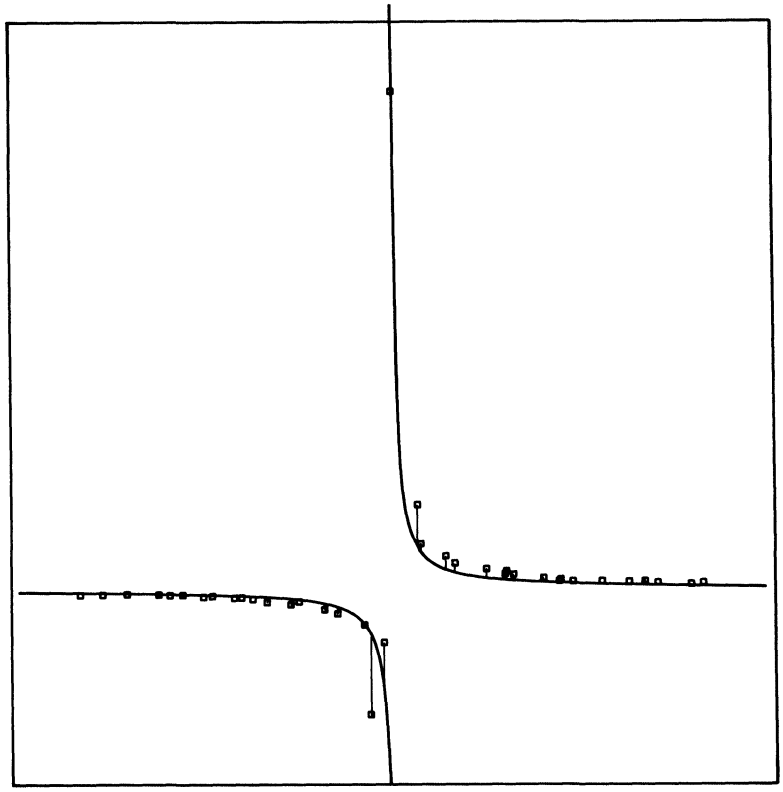

Figure 6
OLS from Table 1

large values of σ, including OLS ($\sigma = \infty$). For smaller values of σ, the ODR fit is not nearly so influenced by this data point and, for a broad range of σ, does a very good job of fitting the data. This last point is important, namely that the parameter values do not vary much as a function of σ, which means that the true value of σ may not need to be known with much accuracy. The stability of $\beta(\sigma)$ has been noticed on all of our examples and on problems not reported here. This is not, of course, a proof that this phenomenon holds more generally, but see Lakshminarayanan and Gunst [LakG84] who investigate this question in linear models. Finally, although the fit "looks" better for ODR, we stress that only one set of data was generated so that no statements on the statistical effectiveness of this procedure can be made. This aspect is currently under investigation.

An examination of the results reveals that, with proper initial values, none of the problems was difficult to solve and that subsequent problems were in fact quite easy. We do point out that the problems tended to be a little more difficult for large values of σ.

While it seems intuitive to use the initial value $\delta^0 = 0$, this is not always adequate. A glance at Fig. 5 reveals that the point nearest the asymptote, corresponding to (x_{21}, y_{21}), should correspond to the right half of the curve if the sign of β_1 is positive. Using $\sigma = 1$, $\beta^0 = (1,1)^T$, and $\delta^0 = 0$ implies the need to cross a discontinuity in δ_{21} in order to find the solution reported here. Initializing δ_{21} to the horizontal distance to

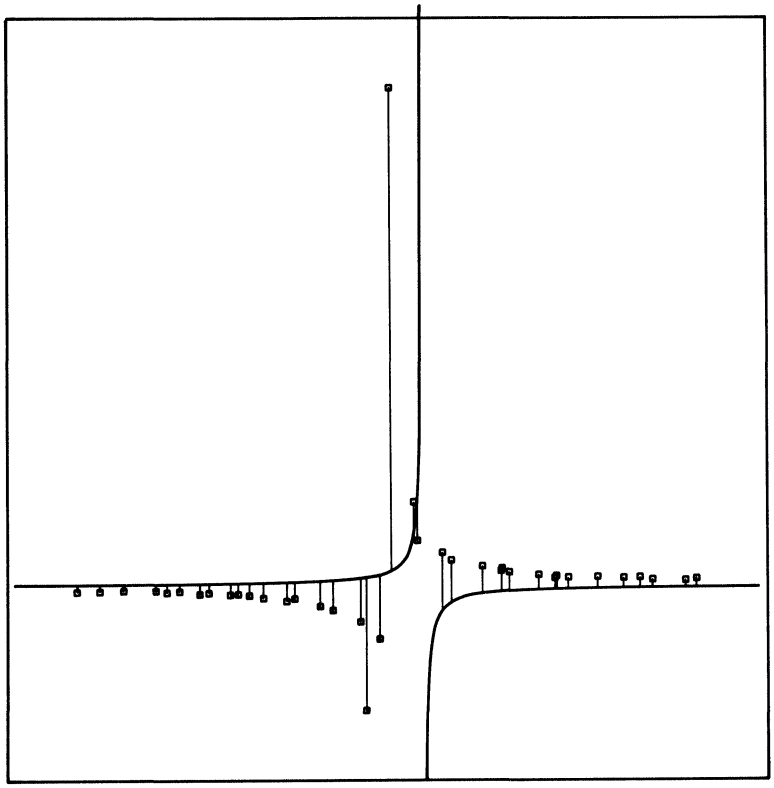

Figure 7
OLS from Table 2

the curve (easily accomplished in this case by algebraically solving for the appropriate values) alleviates this problem.

A final comment on the initial value of δ^0 is in order. Even without a discontinuity, a poor choice of δ^0 can cause the algorithm to converge very slowly. We are indebted to one of the referees for the following simple example. Let $y = e^{\beta x}$. Choose $\beta = 2$ and generate data and errors as for Example 1. Then choose $\beta^0 = 1$ and initialize δ^0 to the horizontal distances. Set $\sigma^2 = .001$ and solve the problem. The algorithm will converge very slowly since $G(\delta, \beta)$ must always remain smaller than $\sigma^2 \sum(\delta_i^0)^2$. With the starting value of $\delta^0 = 0$, however, convergence is very rapid.

Example 2. This example is a two-dimensional version of Example 1. Here we take $x \in R^2$ and

$$y = \frac{-1}{x_1 + x_2 - 1}.$$

This function has a line of singularities along $x_1 + x_2 = 1$. We take the data to be on the rectangular grid of width .1 in the x_1 direction and width .2 in the x_2 direction. The first point is $(.01, .01)^T$ and there are 10 points in the x_1 direction and 5 points in the x_2 direction. y is the evaluated at these points and the data are then perturbed

TABLE 3

σ	$\beta_1(\sigma)$	$\beta_2(\sigma)$	$\beta_3(\sigma)$	Evaluations of G	Evaluations of G'	Final Value of $\|G_1(\eta(\sigma))\|\|G_2(\eta(\sigma))\|$	
1	0.8988	0.9482	1.014	16	9	0.183	0.669
2	0.9222	0.9478	1.018	8	6	0.427	1.236
4	0.9344	0.9505	1.027	8	6	0.988	2.160
10	0.9048	0.9510	1.047	9	7	2.378	4.289
40	0.7147	0.9568	1.043	10	8	6.410	12.618
100	0.3645	0.9343	0.9893	17	14	19.933	14.422
500	0.09138	0.8830	0.9675	24	15	30.423	19.464
∞	0.1191	0.8882	0.9338	10	6	77.443	0.0

according to the following:

$$(x_1)_i := (x_1)_i + rx,$$
$$(x_2)_i := (x_2)_i + rx,$$
$$y_i := y_i + ry$$

where rx are normally distributed with mean 0 and standard deviation .1 and the ry are distributed normally with mean 0 and standard deviation .2.

The form of the model is

$$y = \frac{-\beta_1}{\beta_2 x_1 + \beta_3 x_2 - 1}.$$

The results are given in Table 3 which is organized as Tables 1 and 2. The initial starting value for the function parameters is $\beta_0 = (1, 1, 1)^T$, and δ^0 is taken to be 0 except for values near the asymptote which were initialized to the horizontal distance as described above. Again we observe that the values of $\beta(\sigma)$ do not vary quickly and that as σ increases the fits depend more and more on the points near the asymptote. Note that the location of the asymptote is well-determined by the data and that only β_1 changes much as σ increases. Here, as in Example 1, the insistence on near vertical measures of the error forces β_1 to assume smaller values which has the effect of flattening the function near the asymptote. This, of course, tends to minimize the vertical component of the error.

The nonuniqueness observed in Example 1 was again observed here. The details are not reported, but we found a second OLS solution which led to a trajectory of solutions that finally joined the above trajectory at $\sigma = 2$.

Example 3. The data here are actual measurements from a calibration run on an electronic device which was intended to give a flat response over a wide range of frequencies. In the (x, y)-data, the x-values are in units of frequency squared and the y-data are the gain. The x-values are scaled to the interval $(0, 1)$ with several measurements made in each decade from 10^{-8} to 1. More measurements were taken at the higher frequencies since most of the important information is obtained there. The data are plotted in Fig. 8 with a log scale on the x-axis in order to see the situation better. The y-scale on these graphs has been magnified to accentuate the differences. The peak at the right side of the data is at 1.001 while the low point at the extreme right end is at .9473. The flat part at the left side is all near .9882.

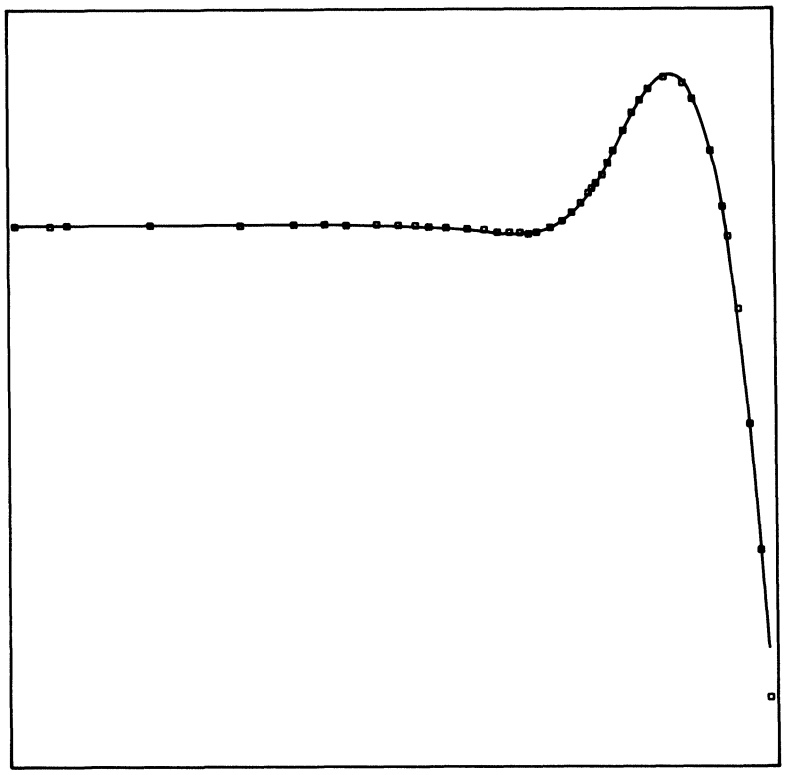

Figure 8
$\sigma = 0.01$

The model for this data was obtained from theoretical considerations and has the form

$$y_i = \sum_{j=1}^{4} \frac{\alpha_j}{x_i + \gamma_j} + \mu, \qquad i = 1, \cdots, 44$$

where the parameters to be determined are

$$\beta = (\alpha_1, \cdots, \alpha_4, \mu, \gamma_1, \cdots, \gamma_4)^T.$$

Initial estimates of the pole locations—the negative γ-values—were obtained quite accurately from other analyses. These γ-values are approximately

$$\gamma_1 = 1.38 \times 10^{-3}, \qquad \gamma_3 = 6.71 \times 10^{1},$$
$$\gamma_2 = 5.96 \times 10^{-2}, \qquad \gamma_4 = 1.07 \times 10^{9}.$$

Starting values for the α-values and μ were obtained by fixing the γ-values at their initial values and then solving for these linear terms using OLS. A feature of the ODR software allowed this step to be accomplished trivially. δ^0 was initialized to 0.

Since all of the poles are negative and all of the data have positive x-values, there is no problem with being close to the asymptotes. The range of the x-values, however,

TABLE 4

σ	Evaluations of G	Evaluations of G'	Final Value of $\|G_1(\eta(\sigma))\|$	Final Value of $\|G_2(\eta(\sigma))\|$
0.01	19	13	0.000505	0.000594
0.10	11	6	0.00170	0.000572
1.00	6	4	0.00190	0.000068
∞	4	2	0.00190	0.0

implies the need to scale the trust region. We used for the diagonal scaling matrices S and T the following:

$$S_i = \frac{1}{|(\beta^0)_i|}, \qquad T_i = \frac{1}{|x_i|}.$$

It turns out that the measurements are proportionately more accurate at the lower frequencies and we therefore took the d-weights to be the same as the t-weights.

While the data were measured quite accurately, there were simply no data at a sufficiently high frequency to warrent including the term corresponding to $j = 4$, and to allow γ_3 to be estimated. This situation was evidenced by the fact that the Jacobian J had five almost identical columns.

With these parameters removed, the resulting problem was easily solved. The correct σ-value is estimated to be 0.01 since the gain measurements in this data set were 100 times more accurate than the frequency measurements. Other values of σ were subsequently used for comparison. The results are in Table 4. Figure 8 depicts the ODR fit for $\sigma = .01$. Virtually no difference appears between the ODR and OLS fits at the lower frequencies, but some differences occur at the higher frequencies. In the enlargements (Figs. 9 and 10) one can easily see that the contribution of the

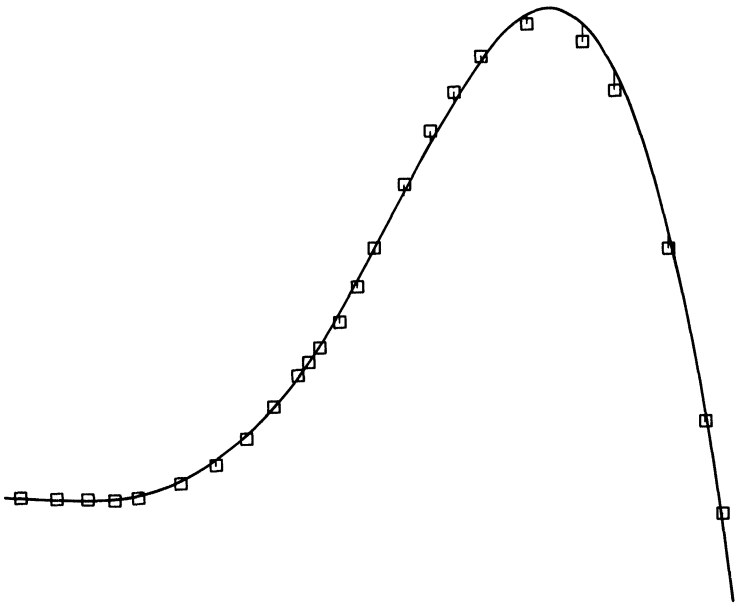

Figure 9
OLS

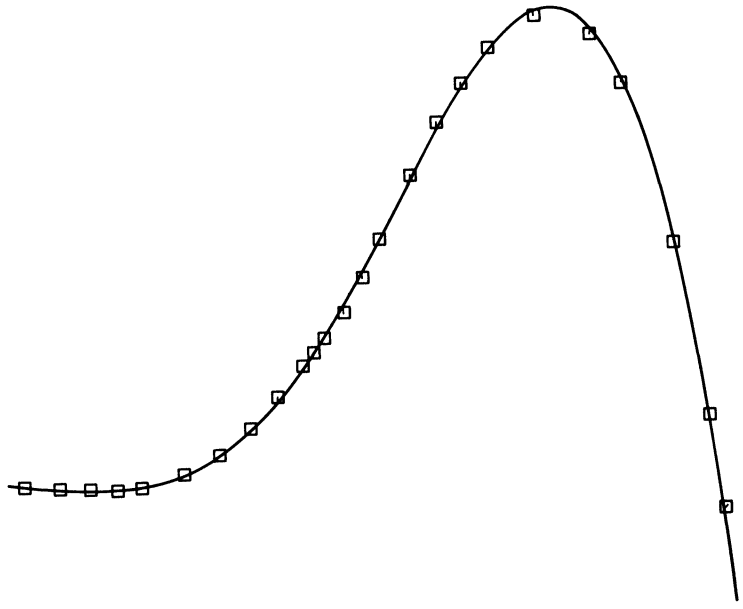

Figure 10
$\sigma = 0.01$

error in the x-values causes ODR to get a significantly better fit than OLS. While the β-values are not reported here, there were, again, very slow changes in $\beta(\sigma)$.

In this section we have shown that our algorithm is effective on highly nonlinear problems, but that these problems themselves often have multiple solutions and other difficulties which imply that potential solutions need to be studied carefully. In subsequent papers, we will provide a more complete description of our implementation and further results on its performance.

Acknowledgments. The authors wish to thank C. Spiegelman and R. Carroll (University of North Carolina) for many useful discussions of the statistical aspects of ODR; J. Donaldson for her help in enhancing the software and conducting the numerical tests; E. Bromberg and T. Griffin for their help in the graphics aspects of these experiments; and T. Griffin for his help in the writing of the data handling and driver routines. All of the above except Carroll are of the Center for Applied Mathematics, National Bureau of Standards.

REFERENCES

[Bar74] Y. BARD, *Nonlinear Parameter Estimation,* Academic Press, New York, 1974.

[BriL73] H. I. BRITT AND R. H. LUECKE, *The estimation of parameters in nonlinear, implicit models,* Technometrics, 15, (1973), pp. 233-247.

[ByrS87] R. H. BYRD AND R. B. SCHNABEL, *A unified local and global convergence analysis of Levenberg-Marquardt methods,* (in preparation).

[DenGW81] J. E. DENNIS, JR., D. M. GAY AND R. E. WELSCH, *An adaptive nonlinear least-squares algorithm,* TOMS, 7 (1981), pp. 348-368.

[DenS83] J. E. DENNIS, JR. AND R. B. SCHNABEL, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations,* Prentice-Hall, Englewood Cliffs, NJ, 1983.

[FiaM68] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques,* John Wiley, New York, 1968.

[Ful86] W. A. FULLER, *Measurement Error Models*, John Wiley, New York, 1986.

[Gay84] D. M. GAY, *Algorithm* 611. *Subroutines for unconstrained minimization using a model/trust-region approach*, TOMS, 9 (1983), pp. 503-524.

[GolV83] G. H. GOLUB AND C. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[Heb73] M. D. HEBDEN, *An algorithm for minimization using exact second derivatives*, Rept. TP515m, AERE, Harwell, England, 1973.

[KenSO83] M. G. KENDALL, A. STEWART AND J. K. ORD, *The Advanced Theory of Statistics*, Fourth Edition, MacMillan, New York, 1983.

[LakG84] M. LAKSHMINARAYANAN AND R. F. GUNST, *Estimation in linear structural relationships: sensitivity to the choice of the ratio of error variances*, Biometrika, 71 (1984), pp. 569-573.

[LawH74] C. L. LAWSON AND R. J. HANSON, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, NJ, 1974.

[Lue73] D. G. LUENBERGER, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, 1973.

[Mor71] T. A. P. MORAN, *Estimating structural and functional relationships*, J. Mult. Anal., 1 (1971), pp. 232-255.

[Mor77] J. J. MORÉ, *The Levenberg-Marquardt algorithm: implementation and theory*, in Numerical Analysis, G. A. Watson, ed., Lecture Notes in Mathematics 630, Springer-Verlag, Berlin, 1977, pp. 105-116.

[MorS81] J. J. MORÉ AND D. C. SORENSON, *Computing a trust region step*, this Journal, 4 (1981), pp. 553-572.

[OrtR70] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[PowM72] D. R. POWELL AND J. R. MACDONALD, *A rapidly convergent iterative method for the solution of the generalised nonlinear least squares problem*, Computer J., 15 (1972), pp. 148-155.

[Pow75] M. J. D. POWELL, *Converging properties of a class of minimization algorithms*, in Nonlinear Programming 2, O. Mangasarian, R. Meyer and S. Robinson, eds., Academic Press, New York, 1975, pp. 1-27.

[SchKW86] R. B. SCHNABEL, J. E. KOONTZ AND B. E. WEISS, *A modular system of algorithms for unconstrained minimization*, TOMS, (1986).

[SchSB85] G. A. SCHULTZ, R. B. SCHNABEL AND R. H. BYRD, *A family of trust-region algorithms for unconstrained minimization with strong global convergence properties*, SIAM J. Numer. Anal., 22 (1985), pp. 47-67.

[SchT85] H. SCHWETLICK AND V. TILLER, *Numerical methods for estimating parameters in nonlinear models with errors in the variables*, Technometrics, 27 (1985), pp. 17-24.

[Ste73] G. W. STEWART, III, *Introduction to Matrix Computations*, Academic Press, New York, 1973.

[Wat85] G. A. WATSON, *The solution of generalized least squares problems*, International Ser. Numer. Math., 75 (1985), pp. 388-400.