Title

Linearization Confidence Intervals in Nonlinear Regression: Diagnostics and Transformations

Authors
Jian-Shen Chen
Robert Jennrich

Publication Date
2011-10-24

Diagnostics for Linearization Confidence Intervals in Nonlinear Regression

Jian-Shen CHEN and Robert I. JENNRICH*

We investigate linear approximation (LA) confidence intervals for functions $g(\boldsymbol{\theta})$ of the parameters $\boldsymbol{\theta}$ in a nonlinear regression model. These intervals are almost universally used and generally perform well, but at times they have poor coverage probabilities. A diagnostic plot and index are developed to detect these failures. We show how these diagnostics may be used to estimate coverage probabilities and these are used to calibrate the diagnostics. The performance of the coverage probability estimates in a variety of nonlinear regression problems is investigated via simulation; for these problems, they work quite well. Conditions are identified under which the estimates are exact. Finally, we discuss the use of the profile t plot and asymmetry and bias indices as diagnostics for LA intervals and show how to calibrate them in terms of coverage probabilities.

KEY WORDS: Coverage probability approximation; Estimator differential; Generalized linear function; Gradient direction plot; Invariance; Nonlinearity index.

1. INTRODUCTION

Linear approximation (LA) confidence intervals are very popular in statistical applications. They appear most often as an estimate and standard error obtained by linear approximation methods. Their use is in fact many times more extensive than all alternative methods combined. These alternatives include intervals obtained by transformations of LA intervals, likelihood ratio intervals, and intervals of the simulation, jackknife, and bootstrap types. Motivated by their ubiquity, we propose diagnostics designed to detect one of their serious failures—the failure of their coverage probabilities to achieve nominal levels. Unlike most diagnostics, which are calibrated by rules of thumb, or not at all, we have calibrated ours in terms of coverage probability estimates that have a clear quantitative interpretation.

Linearity is the key issue. To focus our work on detecting departures from linearity as opposed, say, to departures from normality, we have formulated our development in terms of normal theory nonlinear regression. Consider the nonlinear regression model

$$\mathbf{y} = \mathbf{f}(\boldsymbol{\theta}) + \mathbf{e}, \qquad \mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I}).$$

Here $\mathbf{y}$ is an n vector of observed responses, and $\mathbf{f}(\boldsymbol{\theta})$ is a known vector-valued function of an unknown parameter vector $\boldsymbol{\theta}$ that ranges over a p-dimensional parameter space Θ. We assume that $\mathbf{f}$ is one-to-one and has partial derivatives to order 3. (Actually, Lipschitz continuous second-order derivatives are sufficient.)

We are interested in confidence intervals for real-valued functions $g(\boldsymbol{\theta})$ and, more specifically, in LA confidence intervals. As the name implies, LA intervals are based on linear approximations. Here we carefully identify these approximations, discuss some new diagnostics for departures from linearity, and study the effect of these departures on coverage probabilities.

In Section 2 we define the LA interval and identify two conditions, weaker than the linearity of $\mathbf{f}$ and g, that make it exact. In Section 3 we introduce a diagnostic plot and index to detect failures in what we believe to be the more important of these two conditions, and discuss methods for constructing this plot. We show how to use the plot to estimate actual coverage probabilities in Section 4. The accuracy of these estimates for a variety of nonlinear regression models is evaluated by simulation. The coverage probability estimates are used as a natural calibration for the diagnostic plot. In Section 5 we extend the use of the profile t plot of Bates and Watts (1988) as a diagnostic for LA intervals and show how it can be calibrated in terms of coverage probabilities. In Section 6 we give a new interpretation for the bias index of Box (1971) and show how to use it, and bias indices in general, as diagnostics for LA intervals and how to calibrate them. In Section 7 we give a new interpretation of the asymmetry measure of Lowry and Morton (1983) and show how it can be calibrated and used as a diagnostic for LA intervals. Finally, in the Appendix we give a variety of derivations deferred from the previous sections.

2. LINEAR APPROXIMATION INTERVALS AND ASSUMPTIONS

The LA confidence interval for $g(\boldsymbol{\theta})$ with nominal level $1 - \alpha$ is

$$g(\boldsymbol{\theta}) = g(\hat{\boldsymbol{\theta}}) \pm t_0\hat{\sigma}\alpha(\hat{\boldsymbol{\theta}}), \tag{1}$$

where $\hat{\boldsymbol{\theta}}$ is the least squares estimator of $\boldsymbol{\theta}$, t_0 is the $\alpha/2$ upper quantile of Student's t distribution with $n - p$ degrees of freedom, $\hat{\sigma}^2 = \|\mathbf{y} - \mathbf{f}(\hat{\boldsymbol{\theta}})\|^2/(n - p)$, and

$$\alpha^2(\boldsymbol{\theta}) = \frac{dg}{d\boldsymbol{\theta}}\left(\frac{d\mathbf{f}}{d\boldsymbol{\theta}}^T\frac{d\mathbf{f}}{d\boldsymbol{\theta}}\right)^{-1}\frac{dg^T}{d\boldsymbol{\theta}}, \tag{2}$$

where $d\mathbf{f}/d\boldsymbol{\theta}$ and $dg/d\boldsymbol{\theta}$ are the Jacobians of $\mathbf{f}$ and g at $\boldsymbol{\theta}$ and the superscript T denotes transpose. We assume that $\alpha(\boldsymbol{\theta}) > 0$. Here $\hat{\boldsymbol{\theta}}$ and hence $\hat{\sigma}^2$ are functions defined on $\Re^n$, so, for example, $\hat{\boldsymbol{\theta}}(\mathbf{y})$ is the least squares estimate of $\boldsymbol{\theta}$ corresponding to a response vector $\mathbf{y}$. Although they will be explicitly defined in each case, the "$\hat{\ }$" notation will be used throughout to indicate functions defined on $\Re^n$.

Let $\hat{g} = g(\hat{\boldsymbol{\theta}})$ and $\hat{\alpha} = \alpha(\hat{\boldsymbol{\theta}})$. It is easily verified that $\hat{g}$, $\hat{\alpha}$, and $\hat{\sigma}$ are invariant under reparameterization of $\boldsymbol{\theta}$, and hence

* Jian-Shen Chen is Associate Professor, Department of Insurance Operations, Chaoyang Institute of Technology, Taichung, Taiwan, R.O.C. Robert I. Jennrich is Professor, Division of Statistics, Department of Mathematics, University of California, Los Angeles, CA 90024. The work of both authors was supported by National Science Foundation Grant MCS-8301587; Jennrich's work was also supported by NSF Grant DMS92-13264.

Journal of the American Statistical Association
September 1995, Vol. 90, No. 431, Theory and Methods

the LA interval is also invariant. We will frequently exploit this invariance.

We say the LA interval is exact if its coverage probability is $1 - \alpha$. We are interested in diagnosing departures from assumptions that lead to exact intervals. The first step is to identify appropriate assumptions. The simplest assumptions are that $\mathbf{f}$ and g are linear. These are easily checked, and when they hold we are in the context of linear regression where it is known that (1) is exact. There are reasons why we are not interested in detecting departures from these assumptions, however. First, the LA interval can be exact when $\mathbf{f}$ and g are very far from linear (e.g., $\mathbf{f}(\theta) = \theta^3\mathbf{x}$, $g(\theta) = \theta^3$), so that departures from linearity do not by themselves indicate trouble. Second, the LA intervals are invariant under reparameterization, but the linearity of $\mathbf{f}$ and g is not. Finally, because of these problems, it is difficult to calibrate departures from linearity of $\mathbf{f}$ and g in terms of their effect on the LA interval.

To develop alternative assumptions, let $\mathcal{M}$ be the range of $\mathbf{f}$. The set $\mathcal{M}$ is sometimes called an expectation surface. We might replace the assumptions that $\mathbf{f}$ and g are linear with the assumptions that $\mathcal{M}$ and $\hat{g}$ are linear—or, more precisely, that $\mathcal{M}$ is a linear manifold in $\Re^n$ and $\hat{g}$ is a linear function on $\Re^n$. These assumptions are invariant under reparameterization and, as we will see, lead to exact LA intervals. A slightly more basic assumption than the linearity of $\hat{g}$ is that

$$\hat{z} = \frac{\hat{g} - g(\theta)}{\hat{\alpha}} \tag{3}$$

is linear. This assumption, together with the linearity of $\mathcal{M}$, also leads to exact LA intervals.

In our technical development we make frequent use of the differential of a vector-valued function at a point. For example, the differential of $\mathbf{f}$ at θ is the linear function, $d\mathbf{f}_\theta$, defined by

$$d\mathbf{f}_\theta(d\theta) = \frac{d\mathbf{f}}{d\theta}\, d\theta$$

for all $d\theta$ in $\Re^p$. Thus $d\mathbf{f}_\theta$ is a linear transformation from $\Re^p$ into $\Re^n$. By the adjoint $d\mathbf{f}_\theta^*$ of $d\mathbf{f}_\theta$, we mean the adjoint of this linear transformation with respect to the standard inner products on $\Re^p$ and $\Re^n$, which means that $d\mathbf{f}_\theta^*$ is the linear transformation from $\Re^n$ into $\Re^p$ defined by

$$d\mathbf{f}_\theta^*(d\mathbf{y}) = \frac{d\mathbf{f}}{d\theta}^T d\mathbf{y}$$

for all $d\mathbf{y}$ in $\Re^n$.

We also use the second differential. The second differential of $\mathbf{f}$ at θ is $dd\mathbf{f}_\theta$, defined by

$$dd\mathbf{f}_\theta(d\theta_1, d\theta_2) = \sum_{r=1}^{p}\sum_{s=1}^{p} d\theta_{r1}\frac{\partial^2 \mathbf{f}}{\partial\theta_r\partial\theta_s}\, d\theta_{s2}$$

for all $d\theta_1 = (d\theta_{r1})$ and $d\theta_2 = (d\theta_{s2})$ in $\Re^p$. If f is a scalar-valued function on $\Re^p$, then ddf_θ is a bilinear form on $\Re^p$. By $\mathrm{tr}(ddf_\theta)$, we mean the trace of the matrix $(\partial^2 f/\partial\theta_r\partial\theta_s)$ that defines this bilinear form. (A discussion of differentials and differentiation using them may be found in most advanced calculus texts. We use Buck 1956 and Loomis and Sternberg 1968.)

Let $\mu = \mathbf{f}(\theta)$ and $\hat{\mu} = \mathbf{f}(\hat{\theta})$. We show in the Appendix (Thm. A.1) that there is a simple relation between $\hat{\alpha}$ and the differential of $\hat{g}$ at $\hat{\mu}$:

$$\hat{\alpha} = \| d\hat{g}_{\hat{\mu}} \|. \tag{4}$$

From this and (3), it follows that the linearity of $\hat{g}$ implies the linearity of $\hat{z}$.

We next show that the linearity of $\mathcal{M}$ and $\hat{z}$ are sufficient for exactness of the LA interval. Note that $\hat{g}(\mu) = g(\hat{\theta}(\mathbf{f}(\theta))) = g(\theta)$. Thus $\hat{z}(\mu) = 0$. If $\hat{z}$ is linear, then its first-order Taylor expansion about μ is exact and has the form

$$\hat{z}(\mathbf{y}) = d\hat{z}_\mu(\mathbf{y} - \mu), \tag{5}$$

where $d\hat{z}_\mu$ denotes the differential of $\hat{z}$ at μ. It follows from (3) and (4) that $d\hat{z}_\mu = d\hat{g}_\mu/\| d\hat{g}_\mu\|$, and hence $d\hat{z}_\mu$ has length 1. Using (5), $\hat{z} \sim N(0, \sigma^2)$. Clearly, if $\mathcal{M}$ is linear of dimension p, then $(n-p)\hat{\sigma}^2/\sigma^2 \sim \chi^2(n-p)$. From (3), $\hat{z} = \hat{z}(\hat{\mu})$ and hence is independent of $\hat{\sigma}^2$. Thus $\hat{z}/\hat{\sigma}$ has a Student's t distribution with $n-p$ degrees of freedom. Using (1), we have proved the following.

Theorem 2.1. If $\hat{z}$ is linear, and $\mathcal{M}$ is linear of dimension p, then the LA interval is exact.

We consider the linearity of $\hat{z}$ and $\mathcal{M}$ to be basic assumptions. We do not claim that these are necessary conditions for exactness of the LA interval, but it is not easy, for us at least, to give a counter example. Because the linearity of $\mathbf{f}$ and g imply linearity of $\mathcal{M}$ and $\hat{z}$ and the converse is not true, linearity assumptions on $\mathcal{M}$ and $\hat{z}$ are weaker than those on $\mathbf{f}$ and g.

Bates and Watts (1988) have proposed an index, the RMS relative intrinsic curvature, to measure the nonlinearity in $\mathcal{M}$. They have computed this index for a fairly extensive collection of problems (1988, p. 257) and have come to the very pleasant conclusion that for most of these problems, $\mathcal{M}$ is fairly linear. Motivated by this, and by the desire to focus our investigation, we look for departures from linearity in $\hat{z}$ rather than in $\mathcal{M}$.

3. LINEARITY DIAGNOSTICS FOR $\hat{z}$

Recall that $\hat{z}$ is a real-valued function on $\Re^n$. If we had $n+1$-dimensional eyes, we would plot this function and look at it. To make a plot we can see, let $\mathbf{d}$ be the direction of the gradient of $\hat{z}$ at μ and for any real x, let

$$\tilde{z}(x) = \hat{z}(\mu + x\mathbf{d}). \tag{6}$$

This is a real-valued function on the real line that can be plotted and examined. We call plots like this *gradient direction plots.* Clearly, if $\tilde{z}$ is not linear, then $\hat{z}$ is not linear. The converse is far from true, but we can learn interesting things from looking at $\tilde{z}$ and related plots, as we show.

We now describe a method for constructing $\tilde{z}$ plots. Theorem A.1 in the Appendix gives a computing formula for the gradient of $\hat{g}$ at μ:

$$\nabla\hat{g}(\mu) = \frac{d\mathbf{f}}{d\theta}\left(\frac{d\mathbf{f}}{d\theta}^T\frac{d\mathbf{f}}{d\theta}\right)^{-1}\frac{dg^T}{d\theta}. \tag{7}$$

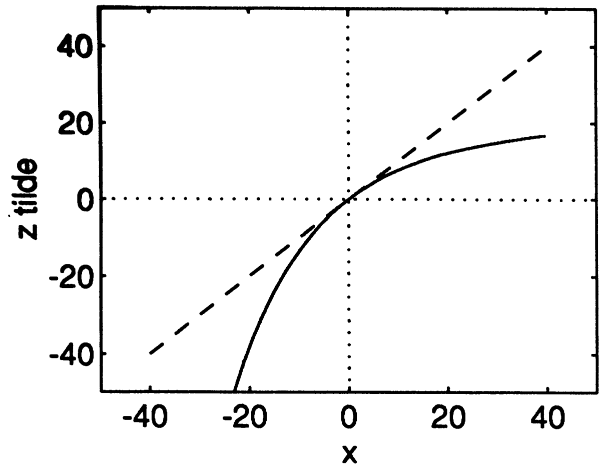

Figure 1. A Typical $\tilde{z}$ Plot. A 45-degree line has been inserted for reference. It connects the points $\pm(2\sigma, 2\sigma)$. Here $g(\theta) = \theta_1$ from the CGM model defined in the Appendix.

Because, as noted earlier, $d\hat{z}_\mu$ is a multiple of $d\hat{g}_\mu$, $\nabla\hat{z}(\mu)$ is a multiple of $\nabla\hat{g}(\mu)$, and hence $\mathbf{d}$ may be obtained by scaling $\nabla\hat{g}(\mu)$. To compute $\tilde{z}$ at a point, construct the pseudodata $\mathbf{y}^* = \mu + x\mathbf{d}$ and find the corresponding least squares estimate $\theta^* = \hat{\theta}(\mathbf{y}^*)$. Then $\tilde{z}(x) = (g(\theta^*) - g(\theta))/\alpha(\theta^*)$. This requires a separate least squares fit for each value of x, but no special software.

Figure 1 shows a $\tilde{z}$ plot made in the manner just described. Its value at the origin is zero, and its slope is 1. These facts follow from (6), (3), and (4). Because, as in Figure 1, plots of $\tilde{z}$ tend to be very smooth, good plots often can be made with as few as four points using cubic spline interpolation, a flexible ruler, or a steady hand. If $\hat{z}$ is linear, then the plot will be a 45-degree line through the origin.

We view the plot of $\tilde{z}$ as a nonlinearity diagnostic for $\hat{z}$ and view $\ddot{\tilde{z}}(0)$, the second derivative of $\tilde{z}$ at 0, as a nonlinearity index. To compute $\ddot{\tilde{z}}(0)$, note that when $\tilde{z}$ is quadratic, $\tilde{z}(x) = x + \frac{1}{2}\ddot{\tilde{z}}(0)x^2$. Computing $\tilde{z}(x)$ for one small nonzero x and solving for $\ddot{\tilde{z}}(0)$ gives a good approximation for $\ddot{\tilde{z}}(0)$. The value $\ddot{\tilde{z}}(0)$ can be computed exactly in terms of the first and second derivatives of $\mathbf{f}$ and g at θ using Theorem A.2 in the Appendix. We call

$$\gamma = \sigma\ddot{\tilde{z}}(0) \tag{8}$$

a relative nonlinearity index, because it is the second derivative at the origin of a plot of $\tilde{z}/\sigma$ on x/σ. Table 1 gives values of γ computed using Theorem A.2 for 40 parametric functions $g(\theta)$ associated with 14 models defined in the Appendix (Table A.1).

4. COVERAGE PROBABILITY

We give an estimate of the coverage probability,

$$CP = P(\hat{g} - t_0\hat{\sigma}\hat{\alpha} < g(\theta) < \hat{g} + t_0\hat{\sigma}\hat{\alpha}), \tag{9}$$

of the LA interval and use it to calibrate the diagnostics $\tilde{z}$ and γ.

We say that $\hat{z}$ is a generalized linear function if

$$\hat{z}(\mathbf{y}) = h((\mathbf{l}, \mathbf{y})) \tag{10}$$

for some function h and n vector $\mathbf{l}$. The notation $(\mathbf{l}, \mathbf{y})$ denotes the inner product of $\mathbf{l}$ and $\mathbf{y}$. If $\hat{z}$ is a generalized linear function, then, using (10), $\nabla\hat{z}(\mu) = \dot{h}((\mathbf{l}, \mu))\mathbf{l}$ where $\dot{h}$ denotes the derivative of h. Hence $\mathbf{d}$ is a multiple of $\mathbf{l}$. Let $\mathbf{r}$ be the residual in $\mathbf{y} - \mu$ after projection on $\mathbf{d}$. Then $\mathbf{y} = \mu + (\mathbf{d}, \mathbf{y} - \mu)\mathbf{d} + \mathbf{r}$ and $\mathbf{r}$ is perpendicular to $\mathbf{l}$. Thus $\tilde{z}((\mathbf{d}, \mathbf{y} - \mu)) = \hat{z}(\mu + (\mathbf{d}, \mathbf{y} - \mu)\mathbf{d}) = \hat{z}(\mathbf{y} - \mathbf{r}) = h((\mathbf{l}, \mathbf{y} - \mathbf{r})) = h((\mathbf{l}, \mathbf{y})) = \hat{z}(\mathbf{y})$. Hence if $\hat{z}$ is a generalized linear function, then $\hat{z}(\mathbf{y}) = \tilde{z}((\mathbf{d}, \mathbf{y} - \mu))$. A similar statement holds for $\hat{g}$.

In the Appendix we show the following.

Theorem 4.1. If $\hat{z}$ is a generalized linear function and $\mathcal{M}$ is linear of dimension p, then

$$CP = P\left(\left|\frac{\tilde{z}(\sigma z)}{\sigma}\right| < t_0\sqrt{u}\right), \tag{11}$$

where $z \sim N(0, 1)$, $u \sim \chi^2(n - p)/(n - p)$, and they are independent.

In general, of course, $\hat{z}$ is not a generalized linear function and $\mathcal{M}$ is not linear. In this case we use the right side of (11)

Table 1. Values of γ, CP, and $\widehat{CP}.Q$ for $\alpha = .05$ and 40 Parametric Functions $g(\theta)$ in 14 Models

$g(\theta)$	θ_1			θ_2			θ_3			θ_4		
Models	γ	CP	$\widehat{CP}.Q$	γ	CP	$\widehat{CP}.Q$	γ	CP	$\widehat{CP}.Q$	γ	CP	$\widehat{CP}.Q$
MMM	−.032	95.0	94.99	−.107	94.7	94.95						
BOD	−.346	94.0	94.38	−.351	94.2	94.38						
BOX	−.121	93.7	94.88	−.075	95.0	94.93						
EAM	−.169	95.5	94.80	−.038	95.9	94.90						
CGM	−1.102	86.3	88.32	−.056	95.2	95.00						
ARM	−1.750	85.1	85.56	1.797	84.1	85.21	.017	94.1	94.97			
LOG	−.103	94.4	94.92	−.143	95.2	94.87	.040	95.8	94.96			
GOM	−.262	93.5	94.58	−.126	94.9	94.89	.214	95.5	94.69			
CTM	−.002	94.2	95.00	−.004	94.3	95.00	−.002	94.3	95.00			
MRM	−.024	94.4	94.96	−.006	94.5	95.00	−.004	94.7	95.00			
FQM	−.000	95.2	95.00	−.206	94.1	94.75	.000	94.1	95.00			
MMF	−.406	95.2	93.72	.127	95.7	94.94	−1.559	83.2	85.72	−.060	95.6	94.99
FMM	−.406	95.2	93.72	.127	95.7	94.94	−.085	95.7	94.97	−.060	95.6	94.99
WBT	−.279	95.5	94.53	−.237	96.6	94.66	−.773	89.8	91.55	−.030	95.4	94.98

NOTE: Simulation error is approximately 1%.

as an estimate $\widehat{CP}$ of CP. Thus

$$\widehat{CP} = P\left(\left|\frac{\tilde{z}(\sigma z)}{\sigma}\right| < t_0\sqrt{u}\right), \qquad (12)$$

where z and u are as defined in Theorem 4.1.

The estimate $\widehat{CP}$ can be computed by a one-dimensional integral:

$$\widehat{CP} = \int_{-\infty}^{\infty} S\left[(n-p)\left(\frac{\tilde{z}(\sigma z)}{\sigma t_0}\right)^2\right] \phi(z)\, dz,$$

where ϕ is the standard normal density and $S = 1 - F$, where F is the cumulative distribution function for χ^2 $(n - p)$.

Clearly, $\widehat{CP}$ provides a natural calibration for the $\tilde{z}$ plot. There are, however, simpler and more visual calibrations that work quite well. For these, we consider approximations to $\widehat{CP}$.

The easiest and perhaps most useful approximation is obtained by setting $n = \infty$. Let $\widehat{CP}.I$ denote this approximation. Then $\widehat{CP}.I = P(|\tilde{z}(x)| < z_0\sigma)$, where $x \sim N(0, \sigma^2)$ and z_0 is the $\alpha/2$ upper quantile of the standard normal distribution. Using this, one can read $\widehat{CP}.I$ from the $\tilde{z}$ plot. No numerical integration is required. For example, if $\alpha = .05$ and $\sigma = 19.31$, as it is for the CGM model in Table A.1, then $z_0\sigma = 37.9$, so $\widehat{CP}.I = P(|\tilde{z}(x)| < 37.9)$. Inverting this using the plot of $\tilde{z}$ in Figure 1 gives $\widehat{CP}.I \doteq P(x > -20) = .851$. Our simulation estimate for the actual coverage probability CP for this example was .863. This is typical, not unusual, agreement.

If one does not wish to construct the $\tilde{z}$ plot, he or she can approximate it quadratically as described earlier. Let $\widehat{CP}.Q$ be the value of $\widehat{CP}$ using this approximation to $\tilde{z}$. It is easy to see, using (12), that $\widehat{CP}.Q$ depends only on $|\gamma|$ and $n - p$. Figure 2 gives a plot of the failure rate $1 - \widehat{CP}.Q$ on γ for $\alpha = .05$ and $n - p = 5$, 10, 20, and ∞. To use this, one need only compute one value of $\tilde{z}$ to estimate γ and then read the coverage probability estimate $\widehat{CP}.Q$ from Figure 2. The figure also provides a coverage probability calibration for γ. Clearly, from Figure 2 little is lost if n is set to infinity when $n - p \geq 20$. Our current mode of operation is to approximate $\widehat{CP}$ by $\widehat{CP}.I$ if we have the complete $\tilde{z}$ plot, and by $\widehat{CP}.Q$ if we have only the index value γ.

The real comparison of interest, of course, is between CP and its estimates. Table 1 gives $\alpha = .05$ values of CP and $\widehat{CP}.Q$ for 40 parametric functions $g(\boldsymbol{\theta})$ associated with 14 models defined in Table A.1 of the Appendix. Each CP was computed by simulation, using 1,000 fits. In every case but one, the interval for θ_2 in the WBT model, $\widehat{CP}.Q$, is within 2.19 simulation standard errors of CP. The computations were done using MATLAB on a Sun SPARCstation 2 computer. The normal random generator was the MATLAB generator. It is hard to evaluate the numerical precision of all of our calculations, but in a number of cases where we were able to compare confidence intervals that we computed with those computed by others, they were essentially identical, agreeing to at least five significant figures.

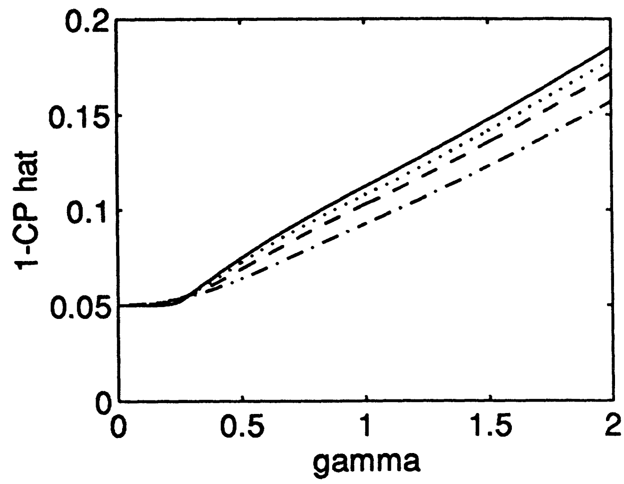

Figure 2. Plots of $1 - \widehat{CP}.Q$ on γ for $\alpha = .05$ and $n - p = 5$ (– · – · –), 10 (– – –), 20 (· · ·), and $n - p = \infty$ (——).

All of the parameter functions $g(\boldsymbol{\theta})$ in Table 1 are coordinate functions of the form $g(\boldsymbol{\theta}) = \theta_j$ for some j. These are functions of interest in the examples, and their use greatly simplifies the presentation of the results. But one need not view the results in Table 1 as coordinate function results. Simply reparameterize all of the models. The results in the body of Table 1 are invariant and hence will not change. But the coordinate functions $g(\boldsymbol{\theta}) = \theta_j$ will in general transform into noncoordinate functions. Actually, we exploited this in carrying out our simulations by reparameterizing when our fitting algorithm ran into trouble. For example, the ARM model generated a number of $\mathbf{y}$ vectors that were numerically difficult to fit. Rather than looking at these individually and attempting to nurse a fit in each case, we reparameterized to reduce the numerical difficulty for our Gauss–Newton algorithm for problems generated by the ARM model. The parameterization used was $f_i(\boldsymbol{\theta}) = \theta_1 + \theta_2(\exp(\theta_3 x_i) - 1)/\theta_3$. The functions to be estimated become $g_1(\boldsymbol{\theta}) = \theta_1 - \theta_2/\theta_3$, $g_2(\boldsymbol{\theta}) = \theta_2/\theta_3$, and $g_3(\boldsymbol{\theta}) = \theta_3$. These correspond to the coordinate functions under the original parameterization.

Constructing the diagnostics discussed requires a value for $\boldsymbol{\theta}$. In a specific nonlinear regression analysis, the value of $\boldsymbol{\theta}$ of greatest interest is that used to define the population sampled. Because this value is usually unknown, one might replace it by $\hat{\boldsymbol{\theta}}$ and view the diagnostics obtained as approximations to those that would have been obtained using the population $\boldsymbol{\theta}$.

5. USING THE PROFILE *t* PLOT

The profile t plot (Bates and Watts 1988) is used as a device to construct likelihood ratio confidence intervals. It has also been suggested that the plot can be used as a diagnostic for LA intervals. We expand on this suggestion here, by showing how the profile t plot is related to the gradient direction plot for $\hat{g}$.

To define the profile t plot, let $\hat{\boldsymbol{\mu}}_c = \mathbf{f}(\hat{\boldsymbol{\theta}}_c)$, where $\hat{\boldsymbol{\theta}}_c$ is the least squares estimate of $\boldsymbol{\theta}$ given $g(\boldsymbol{\theta}) = c$. The profile t plot is a plot of the function

$$t^*(c) = \operatorname{sign}(c - \hat{g})\left(\frac{\|\mathbf{y} - \hat{\boldsymbol{\mu}}_c\|^2 - \|\mathbf{y} - \hat{\boldsymbol{\mu}}\|^2}{\hat{\sigma}^2}\right)^{1/2}. \quad (13)$$

Bates and Watts noted the following.

Theorem 5.1. If $\mathbf{f}$ and g are linear, then t^* is linear.

This suggests using the plot of t^* as a diagnostic for LA intervals. If t^* is not linear, then $\mathbf{f}$ or g must be nonlinear. To our knowledge there has been no further investigation of the use of profile t plot as a diagnostic for LA intervals. Because, as noted earlier, LA intervals can be exact when $\mathbf{f}$ and g are nonlinear, it would be helpful to have a stronger version of Theorem 5.1.

To relate the profile t plot to the gradient direction plot for $\hat{g}$, recall that $\mathbf{d}$ is the direction of the gradient of $\hat{g}$ at $\boldsymbol{\mu}$. Thus the gradient direction plot for $\hat{g}$ is a plot of

$$\tilde{g}(x) = \hat{g}(\boldsymbol{\mu} + x\mathbf{d}), \quad (14)$$

which can be constructed from pseudodata exactly as $\tilde{z}$ was constructed.

We say $\hat{g}$ is *linearizable* if there is a strictly monotone transformation h on the real line such that $h(\hat{g})$ is linear. Clearly, a linearizable function is a generalized linear function. Moreover, one can show that $\tilde{g}$ must be invertible if $\hat{g}$ is linearizable.

In the Appendix we show the following theorem.

Theorem 5.2. If $\mathcal{M}$ is linear and $\hat{g}$ is linearizable, then $\tilde{g}^{-1} = \hat{x} + \hat{\sigma}t^*$, where $\hat{x} = (\mathbf{d}, \hat{\boldsymbol{\mu}} - \boldsymbol{\mu})$.

The next theorem follows at once.

Theorem 5.3. If $\mathcal{M}$ is linear and $\hat{g}$ is linear, then t^* is linear.

This provides stronger motivation than Theorem 5.1 for using t^* as a diagnostic for LA intervals. It is difficult to produce examples for which the LA interval is exact when $\mathcal{M}$ or $\hat{g}$ are not linear.

Finally, one could also use the profile t plot to estimate LA interval coverage probabilities. We show the following in the Appendix.

Theorem 5.4. If $\mathcal{M}$ is linear and $\hat{g}$ is linearizable, then $\tilde{z} = (\tilde{g} - g(\boldsymbol{\theta}))/\dot{\tilde{g}}$.

Thus starting with t^*, one can produce $\tilde{g}$ using Theorem 5.2. Then, produce $\tilde{z}$ using Theorem 5.4 and use this as in Section 4 to estimate coverage probabilities. This is a bit roundabout so we have not investigated it, but we would expect its performance to be similar to that of using $\tilde{z}$ directly.

6. BOX'S BIAS APPROXIMATION

Box (1971) derived an approximation for the bias in $\hat{\boldsymbol{\theta}}$ that was extended by Ratkowsky (1983) to an approximation for the bias in $\hat{g}$. A much simpler argument that leads directly to these approximations goes as follows. Approximate $\hat{g}$ by a second-order Taylor expansion at $\boldsymbol{\mu}$, giving

$$\hat{g} \overset{\circ}{=} g(\boldsymbol{\theta}) + d\hat{g}_{\mu}(\mathbf{e}) + \frac{1}{2}\, dd\hat{g}_{\mu}(\mathbf{e}, \mathbf{e})$$

where $dd\hat{g}_{\mu}$ denotes the second differential of $\hat{g}$ at $\boldsymbol{\mu}$. An approximation for the bias in $\hat{g}$ is then given by

$$E(\hat{g} - g(\boldsymbol{\theta})) \overset{\circ}{=} \frac{\sigma^2}{2}\operatorname{tr}(dd\hat{g}_{\mu}). \quad (15)$$

This is the Box–Ratkowsky (BR) approximation, except that those authors expressed it in a fairly complex form involving the first and second partial derivatives of $\mathbf{f}$ and g. Their expressions may be viewed as computing formulas for $\operatorname{tr}(dd\hat{g}_{\mu})$.

We show how to use our methods to calibrate the BR bias estimate. If $\hat{g}$ is a generalized linear function, then $\hat{g}(\mathbf{y}) = \tilde{g}((\mathbf{d}, \mathbf{y} - \boldsymbol{\mu}))$ and $\operatorname{tr}(dd\hat{g}_{\mu}) = \ddot{\tilde{g}}(0)$. We show in the Appendix (Thm. A.4) that $\ddot{\tilde{z}}(0) = -\ddot{\tilde{g}}(0)/\alpha(\boldsymbol{\theta})$. Using (8),

$$\gamma = -\frac{2\text{Box}}{\sigma\alpha(\boldsymbol{\theta})},$$

where Box is the BR bias estimate given by (15). One can use this and Figure 2 to calibrate the BR index in terms of estimated coverage probabilities.

Instead of using coverage probabilities to calibrate the BR index, one might do this the other way around and use the index to estimate coverage probabilities. But it is easier to base these directly on $\ddot{\tilde{z}}(0)$, which is much easier to compute than the BR index.

7. THE ASYMMETRY MEASURE OF LOWRY AND MORTON

Lowry and Morton (1983) introduced an asymmetry measure for the distribution of each component $\hat{\theta}_j$ of $\hat{\boldsymbol{\theta}}$ that may be viewed as a diagnostic for the LA interval for θ_j. This was discussed further by Morton (1987). To define this measure, let $\hat{\theta}_j^+ = \hat{\theta}_j(\boldsymbol{\mu} + \mathbf{e})$, $\hat{\theta}_j^- = \hat{\theta}_j(\boldsymbol{\mu} - \mathbf{e})$, and $\hat{\varphi}_j = \frac{1}{2}(\hat{\theta}_j^+ + \hat{\theta}_j^-) - \theta_j$. The Lowry and Morton asymmetry measure is $\lambda_j = \operatorname{var}(\hat{\varphi}_j)/\operatorname{var}(\hat{\theta}_j)$. Because this generally cannot be computed, they also given an asymptotic approximation,

$$\lambda_j^{(a)} = \frac{\operatorname{avar}(\hat{\varphi}_j)}{\operatorname{avar}(\hat{\theta}_j)},$$

where $\operatorname{avar}(\hat{\varphi}_j)$ and $\operatorname{avar}(\hat{\theta}_j)$ are asymptotic approximations to $\operatorname{var}(\hat{\varphi}_j)$ and $\operatorname{var}(\hat{\theta}_j)$ defined by fairly complex formulas involving the first four partial derivatives of each f_i. Using the methods developed here, it can be shown that $\operatorname{avar}(\hat{\varphi}_j) = \sigma^4\| dd\hat{\theta}_{j\mu}\|_F^2/2$ and $\operatorname{avar}(\hat{\theta}_j) = \sigma^2\| d\hat{\theta}_{j\mu}\|^2$, where $\|\cdot\|_F$ denotes the Frobenius norm. This provides a new interpretation for $\lambda_j^{(a)}$. It measures the local nonlinearity of $\hat{\theta}_j$ at $\boldsymbol{\mu}$.

It is easy to extend Lowry and Morton's measure for $\hat{\theta}_j$ to arbitrary functions $g(\hat{\boldsymbol{\theta}})$. Let λ_g denote this extended measure and let $\lambda_g^{(a)}$ denote its asymptotic approximation. Then

$$\lambda_g^{(a)} = \frac{\sigma^2}{2}\frac{\| dd\hat{g}_{\mu}\|_F^2}{\| d\hat{g}_{\mu}\|^2}.$$

It can be shown that when $\hat{g}$ is a generalized linear function,

$$\lambda_g^{(a)} = \frac{\gamma^2}{2}. \quad (16)$$

Thus the results of Section 4 can be used directly to calibrate $\lambda_g^{(a)}$ in terms of coverage probabilities.

Given (16), which index $\lambda_g^{(a)}$ or γ should we prefer? As a diagnostic for LA intervals we prefer γ, because it is considerably easier to compute and because it is simply related to the $\tilde{z}$ plot.

8. COMMENTS

It is possible that the methods used here may be as important as the results derived. We believe that we have made a number of choices that worked well. The decision to look at confidence intervals rather than higher-dimensional regions was motivated by the fact that they are far more frequently used and are simpler. There are, however, many interesting things that one can say about them. The decision to consider $g(\boldsymbol{\theta})$ rather than a component θ_j of $\boldsymbol{\theta}$ seems not only more general, but also more natural and conceptually simpler.

Perhaps the most important observation is that the linear approximation interval is invariant under reparameterization. This motivated our development in terms of invariant structures such as $\hat{g}$, $\hat{z}$, and $\mathcal{M}$ rather than in terms of $\mathbf{f}$ and g, which are not invariant. This in turn has led to substantial simplifications and insights. Consider the expression

$$\frac{dg}{d\theta}\left(\frac{d\mathbf{f}^T}{d\theta}\frac{d\mathbf{f}}{d\theta}\right)^{-1}\frac{dg^T}{d\theta}$$

that appears in the definition of the LA interval. Using (7), this is a formula for the square of the norm of the differential of $\hat{g}$ at $\boldsymbol{\mu}$. In symbols,

$$\|d\hat{g}_{\mu}\|^2 = \frac{dg}{d\theta}\left(\frac{d\mathbf{f}^T}{d\theta}\frac{d\mathbf{f}}{d\theta}\right)^{-1}\frac{dg^T}{d\theta}.$$

From a theoretical viewpoint, the expression on the right is unnecessarily complicated, because it involves a specific parameterization for $\mathbf{f}$ and g. It cannot be avoided entirely, because in a specific application $\hat{g}$ is defined by a specific $\mathbf{f}$ and g. However, it can, and we claim should, be viewed as a computing formula for the theoretical expression on the left.

In general it is a good idea to separate computing formulas from conceptual formulas. We have given conceptually much simpler formulations for the bias index of Box and Ratkowsky and the asymmetry index of Lowry and Morton by expressing them in terms of $\hat{g}$ as simple multiples of $\mathrm{tr}(dd\hat{g}_{\mu})$ and $\|dd\hat{g}_{\mu}\|_F^2$. The formulas given by Box and Ratkowsky and by Lowry and Morton are useful computing formulas but in our opinion are too complicated to provide useful theoretical insight.

Finally, we have made extensive use of differentials rather than attempting to replace them by gradients and Jacobians. Our development has involved a fair number of rather delicate differentiations that would have been awkward to carry out without the simple and precise differential notation.

APPENDIX: DEFERRED THEOREMS AND PROOFS

Theorem A.1. In the notation of Section 2,

$$\nabla\hat{g}(\boldsymbol{\mu}) = \frac{d\mathbf{f}}{d\theta}\left(\frac{d\mathbf{f}^T}{d\theta}\frac{d\mathbf{f}}{d\theta}\right)^{-1}\frac{dg^T}{d\theta} \tag{A.1}$$

and

$$\alpha(\boldsymbol{\theta}) = \|d\hat{g}_{\mu}\|. \tag{A.2}$$

Proof. By definition, $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}(\mathbf{y})$ minimizes $\|\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})\|^2$. Assuming that $\hat{\boldsymbol{\theta}}(\mathbf{y})$ is an interior point of the domain of $\mathbf{f}$,

$$(d\mathbf{f}_{\hat{\theta}(\mathbf{y})}(d\boldsymbol{\theta}), \mathbf{y} - \mathbf{f}(\hat{\boldsymbol{\theta}}(\mathbf{y}))) = 0, \tag{A.3}$$

for all $\mathbf{y} \in \Re^n$ and $d\boldsymbol{\theta} \in \Re^p$. Computing the differential of both sides of (A.3) with respect to $\mathbf{y}$ and replacing $\mathbf{y}$ by $\boldsymbol{\mu}$ gives

$$(d\mathbf{f}_{\theta}(d\boldsymbol{\theta}), d\mathbf{y} - d\mathbf{f}_{\theta}(d\hat{\boldsymbol{\theta}}_{\mu}(d\mathbf{y}))) = 0 \tag{A.4}$$

for all $d\mathbf{y} \in \Re^n$ and $d\boldsymbol{\theta} \in \Re^p$. Solving (A.4) gives $d\hat{\boldsymbol{\theta}}_{\mu} = (d\mathbf{f}_{\theta}^* d\mathbf{f}_{\theta})^{-1} d\mathbf{f}_{\theta}^*$, where $d\mathbf{f}_{\theta}^*$ denotes the adjoint of $d\mathbf{f}_{\theta}$. Because $\hat{g} = g(\hat{\boldsymbol{\theta}})$,

$$d\hat{g}_{\mu} = dg_{\theta} d\hat{\boldsymbol{\theta}}_{\mu} = dg_{\theta}(d\mathbf{f}_{\theta}^* d\mathbf{f}_{\theta})^{-1} d\mathbf{f}_{\theta}^*.$$

When written in terms of Jacobians, this is (A.1). Using (A.1) and (2), $\|d\hat{g}_{\mu}\| = \|\nabla\hat{g}(\boldsymbol{\mu})\| = \alpha(\boldsymbol{\theta})$, which proves (A.2).

The following theorem was proved by Chen (1991, p. 93).

Theorem A.2. If

$$\mathbf{u} = \left(\frac{d\mathbf{f}^T}{d\theta}\frac{d\mathbf{f}}{d\theta}\right)^{-1}\frac{dg^T}{d\theta} \quad \text{and} \quad \mathbf{A} = \sum_{i=1}^{n} d_i \frac{d^2 f_i}{d\theta d\theta},$$

where d_i and f_i are the ith components of $\mathbf{d}$ and $\mathbf{f}$, and if

$$\mathbf{B} = (\alpha(\boldsymbol{\theta}))^{-2}\mathbf{A} - (\alpha(\boldsymbol{\theta}))^{-3}\frac{d^2 g}{d\theta d\theta},$$

then $\ddot{\tilde{z}}(0) = \mathbf{u}^T\mathbf{B}\mathbf{u}$.

Proof of Theorem 4.1.

From (9) and (3), $CP = P(|\hat{z}| < t_0\hat{\sigma}) = P(|\tilde{z}((\mathbf{d}, \mathbf{y} - \boldsymbol{\mu}))| < t_0\hat{\sigma})$. Recall that $\mathbf{d}$ is the direction of the gradient of $\hat{z}$ at $\boldsymbol{\mu}$. Using (3), $\nabla\hat{z}(\boldsymbol{\mu}) = \hat{\alpha}^{-1}(\boldsymbol{\mu})\nabla\hat{g}(\boldsymbol{\mu})$, so $\mathbf{d}$ is a multiple of $\nabla\hat{g}(\boldsymbol{\mu})$. By (7), $\mathbf{d}$ is in the column space of $d\mathbf{f}/d\boldsymbol{\theta}$, and hence $\mathbf{d}$ is tangent to $\mathcal{M}$ at $\boldsymbol{\mu}$. Because $\mathcal{M}$ is linear, $\mathbf{y} - \hat{\boldsymbol{\mu}}$ is perpendicular to $\mathcal{M}$. Thus $(\mathbf{d}, \mathbf{y} - \hat{\boldsymbol{\mu}}) = 0$ and $(\mathbf{d}, \mathbf{y} - \boldsymbol{\mu}) = (\mathbf{d}, \hat{\boldsymbol{\mu}} - \boldsymbol{\mu})$. Also because $\mathcal{M}$ is linear, $\hat{\boldsymbol{\mu}}$ and $\mathbf{y} - \hat{\boldsymbol{\mu}}$ are independent. Thus $(\mathbf{d}, \mathbf{y} - \boldsymbol{\mu})$ and $\mathbf{y} - \hat{\boldsymbol{\mu}}$ are independent, as are $(\mathbf{d}, \mathbf{y} - \boldsymbol{\mu})$ and $\hat{\sigma}^2$. Again because $\mathcal{M}$ is linear, $(n-p)\hat{\sigma}^2/\sigma^2 \sim \chi^2(n-p)$. The theorem follows from the fact that $(\mathbf{d}, \mathbf{y} - \boldsymbol{\mu}) \sim N(0, \sigma^2)$.

Proof of Theorem 5.2.

Let $\hat{\boldsymbol{\mu}}_c(\mathbf{y})$ be the projection of $\mathbf{y}$ onto $\mathcal{M}_c = \{\mathbf{f}(\boldsymbol{\theta}) : g(\boldsymbol{\theta}) = c\}$. Because $\hat{g}(\mathbf{f}(\boldsymbol{\theta})) = g(\hat{\boldsymbol{\theta}}(\mathbf{f}(\boldsymbol{\theta}))) = g(\boldsymbol{\theta})$,

$$\mathcal{M}_c = \{\boldsymbol{\mu}' \in \mathcal{M} : \hat{g}(\boldsymbol{\mu}') = c\}. \tag{A.5}$$

Because $\hat{g}$ is a generalized linear function, $\hat{g}(\mathbf{y}) = \tilde{g}((\mathbf{d}, \mathbf{y} - \boldsymbol{\mu}))$, and because $\tilde{g}$ is invertible, $\mathcal{M}_c = \{\boldsymbol{\mu}' \in \mathcal{M} : (\mathbf{d}, \boldsymbol{\mu}' - \boldsymbol{\mu}) = \tilde{g}^{-1}(c)\}$. Because this is a linear manifold,

$$\hat{\boldsymbol{\mu}}_c - \hat{\boldsymbol{\mu}} = x\mathbf{d} \tag{A.6}$$

for some x. Using (13) and (A.6),

$$t^*(c) = \mathrm{sign}(c - \hat{g})\frac{|x|}{\hat{\sigma}}. \tag{A.7}$$

Using (A.6) and the definition of $\hat{x}$,

$$(\mathbf{d}, \hat{\boldsymbol{\mu}}_c - \boldsymbol{\mu}) = x + \hat{x}. \tag{A.8}$$

Using (A.5) and the fact that $\hat{g}$ is a generalized linear function, $c = \hat{g}(\hat{\boldsymbol{\mu}}_c) = \tilde{g}((\mathbf{d}, \hat{\boldsymbol{\mu}}_c - \boldsymbol{\mu}))$. Using (A.8),

$$c = \tilde{g}(x + \hat{x}). \tag{A.9}$$

Thus

Table A.1. Models Used in the Simulation

Models[a]	θ	σ	$f(\theta)$
MMM	$(212.7, .06412)^T$	10.93	$\theta_1 x/(\theta_2 + x)$
	$\mathbf{x} = (.02, .02, .06, .06, .11, .11, .22, .22, .56, .56, 1.10, 1.10)^T$		
BOD	$(19.14, .5311)^T$	2.549	$\theta_1(1 - \exp(-\theta_2 x))$
	$\mathbf{x} = (1, 2, 3, 4, 5, 7)^T$		
BOX	$(205.3, .4306)^T$	7.100	$\theta_1(1 - \exp(-\theta_2 x))$
	$\mathbf{x} = (1, 2, 3, 4, 5, 6, 7)^T$		
EAM	$(.4801, 1.603)^T$	1.560	$\theta_1 x^{\theta_2}$
	$\mathbf{x} = (4, 10, 17, 22, 25)^T$		
CGM	$(.009228, 1.826)^T$	19.31	$\theta_1 x^{\theta_2}$
	$\mathbf{x} = (50, 60, 60, 70, 70, 80, 80, 90, 90, 90, 95, 100, 100, 100, 105, 105, 110, 110, 110, 115, 115, 115, 120, 120, 120, 125, 130, 130, 135, 135, 140, 140, 145, 150, 150, 155, 155, 160, 160, 160, 165, 170, 180)^T$		
ARM	$(188.6, -193.2, -.006247)^T$	3.167	$\theta_1 + \theta_2 e^{\theta_3 x}$
LOG	$(72.46, 13.71, -.06736)^T$	1.159	$\theta_1/(1 + \theta_2 e^{\theta_3 x})$
GOM	$(82.83, 1.224, .03707)^T$	1.906	$\theta_1 e^{-e^{(\theta_2 - \theta_3 x)}}$
MMF	$(80.96, 8.895, 49,577, 2.828)^T$	1.647	$(\theta_2\theta_3 + \theta_1 x^{\theta_4})/(\theta_3 + x^{\theta_4})$
FMM	$(80.96, 8.895, 10.81, 2.828)^T$	1.647	$(\theta_2 e^{\theta_3} + \theta_1 x^{\theta_4})/(e^{\theta_3} + x^{\theta_4})$
WBT	$(69.96, 61.68, .0001001, 2.378)^T$	1.294	$\theta_1 - \theta_2 e^{-\theta_3 x^{\theta_4}}$
	$\mathbf{x}^b = (9, 14, 21, 28, 42, 57, 63, 70, 79)^T$		
CTM	$(5.145, 6,149, 344.1)^T$	1.72E−4	$-\theta_1 + \theta_2/(\theta_3 + x)$
MRM	$(.00561, 6,181, 345.2)^T$	2.608	$\theta_1 e^{\theta_2/(\theta_3 + x)}$
	$\mathbf{x}^b = (50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 105, 110, 115, 120, 125)^T$		
FQM	$(.4889, 3.047, -.1008)^T$	.266	$\theta_1 - 2\theta_2\theta_3 x + \theta_3 x^2$
	$\mathbf{x} = (-4, -3, -2, -1, 0, 1, 2, 3, 4, 5)^T$		

[a] MMM, BOD: Bates and Watts (1988, p. 329); BOX, EAM, CGM: Draper and Smith (1981, pp. 522, 521, 519); ARM, LOG, GOM, MMF, FMM, WBT, CTM: Ratkowsky (1983, pp. 88, 71, 69, 88, 88, 88, 120); MRM: Meyer and Roth (1972, p. 232); FQM: Fieller (1954, p. 180).

[b] The foregoing models use the same **x**.

$$x = \tilde{g}^{-1}(c) - \hat{x}. \tag{A.10}$$

We need to find the relation between the sign of x and that of $c - \hat{g}$. Because $\tilde{g}$ has an inverse and $\dot{\tilde{g}}(0) = d\hat{g}_\mu(\mathbf{d}) = \| d\hat{g}_\mu \| > 0$, $\tilde{g}$ is strictly increasing. Because $\hat{x} = (\mathbf{d}, \hat{\mu} - \mu)$, $\tilde{g}(\hat{x}) = \tilde{g}((\mathbf{d}, \hat{\mu} - \mu)) = \hat{g}(\hat{\mu}) = \hat{g}$. Using (A.9), $c - \hat{g} = \tilde{g}(x + \hat{x}) - \tilde{g}(\hat{x})$, and hence $c - \hat{g}$ has the same sign as x. Finally, from (A.7) and (A.10), $\hat{\sigma}t^* = x = \tilde{g}^{-1} - \hat{x}$, which completes the proof.

Let

$$\tilde{\alpha}(x) = \hat{\alpha}(\mu + x\mathbf{d}). \tag{A.11}$$

Theorem A.3. $\tilde{\alpha}(0) = \dot{\tilde{g}}(0)$ and $\dot{\tilde{\alpha}}(0) = \ddot{\tilde{g}}(0)$.

Proof. Because $\mathbf{d}$ is the direction of $\nabla\hat{g}(\mu)$, $\tilde{\alpha}(0) = \| d\hat{g}_\mu \| = d\hat{g}_\mu(\mathbf{d}) = \dot{\tilde{g}}(0)$. From (7), $\nabla\hat{g}(\mu)$ is tangent to $\mathcal{M}$ at μ. Let γ be a curve on $\mathcal{M}$ such that $\gamma(0) = \mu$ and $\dot{\gamma}(0) = \mathbf{d}$. Because $\hat{\mu}(\gamma) = \gamma$,

$$\hat{\alpha}^2(\gamma) = (d\hat{g}_\gamma, d\hat{g}_\gamma) = d\hat{g}_\gamma(\nabla\hat{g}(\gamma)).$$

Differentiating,

$$\hat{\alpha}(\gamma)d\hat{\alpha}_\gamma(\dot{\gamma}) = dd\hat{g}_\gamma(\nabla\hat{g}(\gamma), \dot{\gamma}).$$

Evaluating at zero,

$$\hat{\alpha}(\mu)d\hat{\alpha}_\mu(\mathbf{d}) = dd\hat{g}_\mu(\mathbf{d}, \nabla\hat{g}(\mu)).$$

Because $\nabla\hat{g}(\mu) = \|\nabla\hat{g}(\mu)\|\mathbf{d} = \hat{\alpha}(\mu)\mathbf{d}$, $d\hat{\alpha}_\mu(\mathbf{d}) = dd\hat{g}_\mu(\mathbf{d}, \mathbf{d})$, and finally, using (A.11) and (14),

$$\dot{\tilde{\alpha}}(0) = d\hat{\alpha}_\mu(\mathbf{d}) = dd\hat{g}_\mu(\mathbf{d}, \mathbf{d}) = \ddot{\tilde{g}}(0),$$

which proves the second assertion of the theorem.

Theorem A.4. $\ddot{\tilde{z}}(0) = -\ddot{\tilde{g}}(0)/\alpha(\theta)$.

Proof. Using (3), (6), (14), and (A.11), $\tilde{z} = (\tilde{g} - g(\theta))/\tilde{\alpha}$. Differentiating twice and using Theorem A.3 gives $\ddot{\tilde{z}}(0) = -\ddot{\tilde{g}}(0)/\tilde{\alpha}(0)$. Using (A.11) again, $\tilde{\alpha}(0) = \hat{\alpha}(\mu) = \alpha(\theta)$, which completes the proof.

Proof of Theorem 5.4

Using (3), (6), and (14), $\tilde{z}(x) = (\tilde{g}(x) - g(\theta))/\hat{\alpha}(\mu + x\mathbf{d})$. Using (4) and the linearity of $\mathcal{M}$, $\hat{\alpha}(\mu + x\mathbf{d}) = \| d\hat{g}_{\mu+x\mathbf{d}} \|$. Thus it is sufficient to show that $\| d\hat{g}_{\mu+x\mathbf{d}} \| = \dot{\tilde{g}}(x)$. Because $\hat{g}$ is a generalized linear function, $\hat{g}(\mathbf{y}) = \tilde{g}((\mathbf{d}, \mathbf{y} - \mu))$. Thus $d\hat{g}_{\mu+x\mathbf{d}}(d\mathbf{y}) = \dot{\tilde{g}}(x)(\mathbf{d}, d\mathbf{y})$, and hence $\| d\hat{g}_{\mu+x\mathbf{d}} \| = |\dot{\tilde{g}}(x)|$. Because $\dot{\tilde{g}}(0) = d\hat{g}_\mu(\mathbf{d}) > 0$ and $\tilde{g}$ is invertible, $\dot{\tilde{g}}(x) \geq 0$ for all x. Thus $\| d\hat{g}_{\mu+x\mathbf{d}} \| = \dot{\tilde{g}}(x)$, which completes the proof.

[Received August 1992. Revised July 1994.]

REFERENCES

Bates, D. M., and Watts, D. G. (1988), *Nonlinear Regression Analysis and Its Applications,* New York: John Wiley.

Box, M. J. (1971), "Bias in Nonlinear Estimation" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 32, 171–201.

Buck, R. C. (1956), *Advanced Calculus,* New York: McGraw-Hill.

Chen, J. S. (1991), "Confidence Intervals for Parametric Functions in Nonlinear Regression," Ph.D. dissertation, University of California, Los Angeles, Dept. of Mathematics.

Draper, N. R., and Smith, H. (1981), *Applied Regression Analysis* (2nd ed.), New York: John Wiley.

Fieller, E. C. (1954), "Some Problems on Interval Estimation," *Journal of the Royal Statistical Society,* Ser. B, 16, 175–185.

Loomis, L. H., and Sternberg, S. (1968), *Advanced Calculus,* Reading, MA: Addison-Wesley.

Lowry, R. K., and Morton, R. (1983), "An Asymmetry Measure for Estimators in Nonlinear Regression Models," in *Proceedings of the 44th Session of the International Statistical Institute, Madrid, Conbtribued Papers 1983,* Vol. 1, pp. 351–354.

Meyer, R. R., and Roth, P. M. (1972), "Modified Damped Least Squares: An Algorithm for Least Squares Nonlinear Estimation," *Journal of Institute of Mathematics and Its Applications,* 9, 218–233.

Morton, R. (1987), "Asymmetry of Estimators in Nonlinear Regression," *Biometrika,* 74, 679–685.

Ratkowsky, D. A. (1983), *Nonlinear Regression Modeling,* New York: Marcel Dekker.