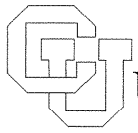
Computational Experience with Confidence Regions and Confidence Intervals for Nonlinear Least Squares*

Janet R. Donaldson
Robert B. Schnabel

CU-CS-302-85



University of Colorado at Boulder
DEPARTMENT OF COMPUTER SCIENCE

*Research supported by ARO contract DAAG 29-84-K-0140

COMPUTATIONAL EXPERIENCE WITH CONFIDENCE REGIONS AND CONFIDENCE INTERVALS FOR NONLINEAR LEAST SQUARES

Janet R. Donaldson *
Robert B. Schnabel **

CU-CS-302-84 May, 1985

* Center for Applied Mathematics, National Bureau of Standards, Boulder, Colorado 80303 and Department of Computer Science, University of Colorado, Boulder, Colorado, 80309

** Department of Computer Science, University of Colorado, Boulder, Colorado, 80309 and Center for Applied Mathematics, National Bureau of Standards, Boulder, Colorado 80303
Research supported by ARO contract DAAG 29-84-K-0140

Abstract

We present the results of a Monte Carlo study of the most commonly discussed methods for constructing approximate confidence regions and confidence intervals for parameters estimated by nonlinear least squares. The methods we examine are the three variants of the linearization method, the likelihood method, and the lack-of-fit method. The linearization method is the most frequently implemented method. It is computationally inexpensive and produces easily understandable results. The likelihood and lack-of-fit methods both are much more expensive and more difficult to report. Based on our results, we conclude that among the three variants of the linearization method, the variant based solely on the Jacobian appears preferable because it is simpler, less expensive, more numerically stable, and at least as accurate as the other two variants which utilize the full Hessian. In our tests, however, all three variants of the linearization method often produce gross underestimates of confidence regions and sometimes produce significant underestimates of confidence intervals. Both the likelihood and lack-of-fit methods, on the other hand, perform very reliably. For the datasets analyzed, the Bates and Watts curvature measures reliably predict when the linearization method confidence regions will be poor, and for the most part are consistent with our results for the likelihood method.

1. Introduction

This paper presents the results of an empirical study comparing several methods for constructing confidence regions and confidence intervals about parameters estimated by nonlinear least squares. The methods compared are the lack-of-fit method, the likelihood method, and three variants of the linearization method.

The need for confidence regions and intervals commonly arises in data fitting applications, where a response variable y_i observed with unknown error $\dot{e}_i$ is fit to m fixed predictor variables $\mathbf{x}_i$ using a function $f(\mathbf{x}_i;\boldsymbol{\theta})$ which can be either linear or nonlinear in the p parameters $\boldsymbol{\theta}$. The function $f(\mathbf{x}_i;\boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}$ if it can be written

$$f(\mathbf{x}_i;\boldsymbol{\theta}) = \mathbf{x}_i\boldsymbol{\theta} = \sum_{j=1}^{p} x_{i,j}\theta_j, \; i=1,\ldots,n.$$

Otherwise, it is nonlinear. The methods analyzed in this study are identical when $f(\mathbf{x}_i;\boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}$; otherwise they are not.

When the error $\dot{e}_i$ is additive, the response variable can be modeled by

$$y_i = f(\mathbf{x}_i;\dot{\boldsymbol{\theta}}) + \dot{e}_i, \; i=1,\ldots,n,$$

where $\dot{\boldsymbol{\theta}}$ denotes the true but unknown value of the parameters. The least squares estimator of $\dot{\boldsymbol{\theta}}$ is the parameter value, denoted $\hat{\boldsymbol{\theta}}$, which minimizes the sum of the squares of the residuals, where the residuals, $r_i(\boldsymbol{\theta})$, are estimates of the random error, $\dot{e}_i$,

$$r_i(\boldsymbol{\theta}) = y_i - f(\mathbf{x}_i;\boldsymbol{\theta}).$$

Thus,

$$\hat{\boldsymbol{\theta}} = \arg\min S(\boldsymbol{\theta})$$

where $S(\boldsymbol{\theta})$ is the residual sum of squares,

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n} r_i(\boldsymbol{\theta})^2 = \mathbf{R}(\boldsymbol{\theta})^T\mathbf{R}(\boldsymbol{\theta})$$

with $\mathbf{R}(\boldsymbol{\theta})$ denoting a column vector with i^{th} component $r_i(\boldsymbol{\theta})$, and $\mathbf{R}(\boldsymbol{\theta})^T$ denoting the transpose of $\mathbf{R}(\boldsymbol{\theta})$.

In our study, we assume that the model is correct and that the errors are normal, independent, identically distributed random variables with zero mean and variance $\dot{\sigma}^2$, i.e., distributed as $N(\mathbf{0},\dot{\sigma}^2\,\mathbf{I})$. Then, the least squares estimator $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator of the parameters $\dot{\boldsymbol{\theta}}$ of the p-variate normal density function,

$$L(\mathbf{Y}) = (2\pi\dot{\sigma}^2)^{-n/2} \; e^{(-\dot{e}^T\dot{e}/2\dot{\sigma}^2)}$$

where $\mathbf{Y}$ is a column vector with i^{th} component y_i, and $\dot{e}$ is a column vector with i^{th} component $\dot{e}_i$.

Nearly normally distributed errors are, in fact, encountered quite frequently in practice. This is because measurement errors are often the sum of a number of random errors from unknown sources, and, by the central limit theorem, the sum of these errors is approximately normally distributed whatever the distribution of the individual errors that make up the sum.

In practice, the estimated values of the parameters $\hat{\boldsymbol{\theta}}$ will not equal the true values $\dot{\boldsymbol{\theta}}$ because of the random errors, $\dot{e}_i$, in the data. Since $\hat{\boldsymbol{\theta}}$ is a random variable, however, it may be possible to indicate with some specific probability $1-\alpha$ in what region about $\hat{\boldsymbol{\theta}}$ we might reasonably expect $\dot{\boldsymbol{\theta}}$ to be. Such regions are known as $100\cdot(1-\alpha)\%$ confidence regions. A joint confidence region about all of the parameters is defined using a function

$$CR_\alpha : \mathbf{Y} \rightarrow \text{a region in } R^p$$

which satisfies

$$Pr[\; \dot{\boldsymbol{\theta}} \in CR_\alpha(\mathbf{Y}) \;] = 1-\alpha.$$

Similarly, a confidence interval about an individual parameter $\hat{\theta}_j$ is defined using a function

$$CI_{j,\alpha} : \mathbf{Y} \rightarrow \text{an interval in } R$$

which satisfies

$$Pr[\; \dot{\theta}_j \in CI_{j,\alpha}(\mathbf{Y}) \;] = 1-\alpha.$$

The above definitions state that, before the data are sampled, the probability that the confidence regions and confidence intervals to be constructed will contain the true parameter values is $1-\alpha$. Thus, if we repeatedly draw samples and construct confidence regions and intervals about the least squares estimates for each sample, then in the long run $100\cdot(1-\alpha)\%$ of these confidence regions and intervals should contain the true values. Methods that, for all functions $f(\mathbf{x}_i;\boldsymbol{\theta})$ and confidence levels $1-\alpha$, are statistically guaranteed asymptotically to contain the true value $100\cdot(1-\alpha)\%$ of the time are called exact; all other methods are called approximate.

Various methods have been proposed for calculating confidence regions and intervals for parameter estimation by nonlinear least squares. These include several variants of the linearization method, as well as the likelihood and lack-of-fit methods. [See e.g. Bard (1974), Gallant (1976), Draper and Smith (1981).] We review all these methods briefly in Section 2. They all are equivalent, and exact, for linear models. For nonlinear models, only the lack-of-fit

method for computing confidence regions is exact; the other methods for computing confidence regions and all the methods for computing confidence intervals are approximate. The linearization regions and intervals appear to be the most approximate for nonlinear models, but they also are far less expensive to compute than the likelihood or lack-of-fit regions and intervals, and are the predominant methods implemented in production software. Some nonlinear least squares packages, including NL2SOL [Dennis, Gay, and Welsch (1981)], include three variants of the linearization method, which differ only in that the variance-covariance matrix of the estimated parameters is approximated in three different ways, namely

$$\hat{\mathbf{V}}_a = s^2 \left(\mathbf{J}(\hat{\boldsymbol{\theta}})^T \mathbf{J}(\hat{\boldsymbol{\theta}})\right)^{-1},$$

$$\hat{\mathbf{V}}_b = s^2 \, \mathbf{H}(\hat{\boldsymbol{\theta}})^{-1},$$

or

$$\hat{\mathbf{V}}_c = s^2 \, \mathbf{H}(\hat{\boldsymbol{\theta}})^{-1} \left(\mathbf{J}(\hat{\boldsymbol{\theta}})^T \mathbf{J}(\hat{\boldsymbol{\theta}})\right) \mathbf{H}(\hat{\boldsymbol{\theta}})^{-1},$$

where $s^2 = S(\hat{\boldsymbol{\theta}})/(n-p)$ is the estimated residual variance; $\mathbf{J}(\hat{\boldsymbol{\theta}})$ is the Jacobian of $f(\mathbf{x}_i;\boldsymbol{\theta})$, $i=1,\ldots,n$, at $\hat{\boldsymbol{\theta}}$; and $\mathbf{H}(\hat{\boldsymbol{\theta}})$ is the Hessian of $S(\boldsymbol{\theta})$ at $\hat{\boldsymbol{\theta}}$.

Sections 3-6 of this paper describe and analyze a Monte Carlo study that compares all of these methods for computing confidence regions and intervals on 20 nonlinear models. The study is used to empirically observe how often the true parameter values are contained in the confidence regions and confidence intervals constructed using a given method. The actual percent of the nominally $100\cdot(1-\alpha)\%$ confidence regions and intervals which are found to contain the true value is known as the observed coverage. The observed coverage will generally depend on the method used to construct the confidence regions and confidence intervals, on the nominal confidence level, $1-\alpha$, on the degree of nonlinearity of the function, $f(\mathbf{x}_i;\boldsymbol{\theta})$, and to a small extent, on the number of replications in the simulation. If the experiment used to generate the data is repeated a large number of times under the same conditions, and if CR_α and $CI_{j,\alpha}$ are exact and the model is correct, then the observed coverage will approach the nominal coverage. When CR_α and $CI_{j,\alpha}$ are only approximate, the observed coverage will not necessarily approach the nominal coverage, although one would hope that the difference between the observed and nominal coverage for a reasonable approximate method would be small for most functions.

No similar study of this magnitude appears to have been reported previously. The properties of confidence regions and confidence intervals computed using the linearization,

likelihood, and lack-of-fit methods have been analyzed by several authors, including Jennrich (1959), Beale (1960), Guttman and Meeter (1965), Gallant (1976), Duncan (1978), and Bates and Watts (1980). While the literature includes numerous warnings regarding the possible inaccuracy of the approximate methods, it contains little empirical data to illustrate the size of the discrepancies between observed and nominal coverage that might be expected. In those studies which do contain empirical data on confidence regions and intervals, the largest reported differences between the observed and nominal coverage is only 9% for a 95% confidence region computed using the linearization method, and is even smaller for the likelihood method [Gallant (1976)]. In many practical applications, potential differences of 9% might not be cause for concern. Evidence of much larger differences, however, would indicate the need for improved methods. Our results provide such evidence.

Our Monte Carlo study has several purposes. First, we wish to determine whether the observed coverage of the linearization method is significantly affected by how the variance-covariance matrix is computed. Second, we wish to determine whether the approximate confidence regions and confidence intervals constructed using the linearization and likelihood methods, and the approximate confidence intervals constructed using the lack-of-fit method have observed coverage significantly different from nominal. In particular, we want to know whether the frequently used linearization method is significantly better or worse than the more expensive likelihood and lack-of-fit methods. Section 3 describes how we designed our study to answer these questions. The results are presented and discussed in Section 4. We have also investigated how effective the diagnostics of Bates and Watts (1980) are in predicting when the confidence regions produced by the linearization and likelihood methods should be reliable; this part of the study is the subject of Section 5.

Our study is oriented toward nonlinear least squares software developers who need assurance that the methods they implement are reasonable for a wide variety of problems. We make only the customary assumptions that the model is correct and that the errors are normally distributed. We do not assume that we can change the representation of the parameters, e.g., by reparameterizing θ as $\log(\theta)$, in order to reduce the difference between the observed and nominal coverage, because reparameterization is not a technique that can be routinely implemented by software developers who have no control over the functions analyzed. Readers interested in using reparameterization to improve their results are refered to Ratkowsky (1983).

The conclusions we draw from this study are presented in Section 6. The first conclusion is that among the variants of the linearization method, the one using $\hat{\mathbf{V}}_a$ is the best choice because it is the cheapest, and is always at least as reliable as the other two variants and

sometimes more reliable. The second conclusion is that even the best linearization method can be very poor; confidence regions with observed coverage as low as 12.4% for a nominal 95% region, and confidence intervals with observed coverage as low as 75.0% for a nominal 95% interval are reported. In contrast, for each of the datasets tested, the confidence regions and confidence intervals constructed using the likelihood method and lack-of-fit methods are quite close to nominal. Finally, our study indicates that the diagnostics of Bates and Watts (1980) appear quite successful at predicting when linearization confidence regions will be poor. Our recommendations as to how nonlinear least squares software should calculate confidence regions and intervals, in light of these conclusions, also are given in Section 6.

2. Background

This section briefly discusses methods for constructing confidence regions and confidence intervals. First, we give a very quick survey of confidence regions and confidence intervals for linear least squares. Next, we describe the two different ways function nonlinearity can affect the solution locus. Then, we review the linearization, likelihood, and lack-of-fit methods for constructing confidence regions and confidence intervals when the model is nonlinear. For a more complete discussion, see Bard (1974), Gallant (1976), Draper and Smith (1981), or Donaldson (1985).

Linear least squares

When $f(\mathbf{x}_i;\boldsymbol{\theta})$ is linear in the parameters $\boldsymbol{\theta}$, then $f(\mathbf{x}_i;\boldsymbol{\theta}) = \mathbf{x}_i\,\boldsymbol{\theta}$. Consequently, the Jacobian of $\mathbf{F}(\boldsymbol{\theta})$ is $\mathbf{X}$, an n by p matrix with i^{th} row $\mathbf{x}_i$. If we assume that $\mathbf{X}$ is of full rank, then $\mathbf{X}^T\mathbf{X}$ is nonsingular, and the linear least squares estimators can be expressed in closed form by

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{Y}\ .$$

When $\dot{e} \sim N(\mathbf{0},\dot{\sigma}^2\,\mathbf{I})$, a $100\cdot(1-\alpha)\%$ confidence region about $\hat{\boldsymbol{\theta}}$ contains those values $\tilde{\boldsymbol{\theta}}$ for which

$$S(\tilde{\boldsymbol{\theta}}) - S(\hat{\boldsymbol{\theta}}) \leq s^2\ p\ F_{p,n-p,1-\alpha}. \tag{2.1}$$

Equation (2.1) is equivalent to

$$(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\,\mathbf{X}^T\mathbf{X}\,(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) \leq s^2\ p\ F_{p,n-p,1-\alpha} \tag{2.2}$$

for all linear models, which shows that the shape of the confidence regions about $\hat{\boldsymbol{\theta}}$ is ellipsoidal for all linear models.

A $100\cdot(1-\alpha)\%$ confidence interval about $\hat{\theta}_j$ contains those values $\tilde{\theta}_j$ for which

$$|\tilde{\theta}_j - \hat{\theta}_j| \leq s \sqrt{(\mathbf{X}^T\mathbf{X})_{jj}^{-1}}\, t_{n-p,1-\alpha/2} \tag{2.3}$$

where $(\mathbf{X}^T\mathbf{X})_{jj}^{-1}$ is the $(j,j)^{th}$ element of the inverse of $\mathbf{X}^T\mathbf{X}$. The limits of this confidence interval can be shown to be those values θ_j which

$$\text{maximize } (\theta_j - \hat{\theta}_j)^2 \text{ subject to} \tag{2.4}$$

$$S(\boldsymbol{\theta}) - S(\hat{\boldsymbol{\theta}}) = s^2 \left(t_{n-p,1-\alpha/2}\right)^2 = s^2 F_{1,n-p,1-\alpha} \ .$$

Nonlinearity and the Solution Locus

The solution locus, or estimation space, of $f(\mathbf{x}_i;\boldsymbol{\theta})$, $i=1,\ldots,n$, consists of all points with coordinates expressible as

$$\left(f(\mathbf{x}_1;\boldsymbol{\theta}), f(\mathbf{x}_2;\boldsymbol{\theta}), \ldots, f(\mathbf{x}_n;\boldsymbol{\theta})\right)$$

where the $\mathbf{x}_i$, $i=1,\ldots,n$, are the fixed values of the predictor variables, and $\boldsymbol{\theta}$ is allowed to vary over all possible values of the p unknown parameters. The solution locus is planar if there exists a reparameterization of $f(\mathbf{x}_i;\boldsymbol{\theta})$ that makes the function linear in the p parameters. Otherwise, the solution locus is curved.

A coordinate grid on the solution locus can be formed by tracing the paths obtained when each parameter is individually allowed to vary while all other parameters are held fixed. The coordinate grid is curvilinear whenever the function $f(\mathbf{x}_i;\boldsymbol{\theta})$ is nonlinear in one or more of its parameters. It is linear only when the function itself is linear.

Curvature of the solution locus is called "intrinsic" curvature [Beale (1960); Bates and Watts (1980)]. Curvature of the coordinate grid is called "parameter-effects" or simply "parameter" curvature [Bates and Watts (1980)]. Intrinsic curvature is not affected by reparameterization. Parameter-effects curvature is. Linear functions have zero parameter-effects curvature and zero intrinsic curvature. Nonlinear functions always have nonzero parameter-effects curvature, and can have either zero or nonzero intrinsic curvature, i.e., a planar or curved solution locus, respectively.

Nonlinear Least Squares

When the function is nonlinear, the least squares estimators of the parameters cannot in general be expressed in closed form, and must instead be computed by iterative techniques. Construction of exact confidence regions and confidence intervals also is much more difficult, and so approximate methods are frequently used. The leading methods, linearization, likelihood, and lack-of-fit, are described briefly below.

Linearization methods. Linearization methods for constructing confidence regions and confidence intervals assume that the nonlinear function can be adequately approximated by an affine, or linear, approximation to the function at the solution. That is, this method assumes that the solution locus is planar, and that the coordinate grid is linear throughout the area to be covered by the confidence regions and confidence intervals. Under this assumption, linear least squares theory tells us that the confidence region about $\hat{\boldsymbol{\theta}}$ consists of those values $\tilde{\boldsymbol{\theta}}$ for which

$$(\tilde{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}})^T\hat{\mathbf{V}}^{-1}(\tilde{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}) \leq p\, F_{p,n-p,1-\alpha}$$

while a confidence interval about $\hat{\theta}_j$, $j=1,\ldots,p$, consists of those values $\tilde{\theta}_j$ for which

$$|\tilde{\theta}_j-\hat{\theta}_j| \leq \hat{\mathbf{V}}_{jj}^{1/2}\, t_{n-p,1-\alpha/2},$$

where $\hat{\mathbf{V}}$ is the estimated variance-covariance matrix of the parameters, and $\hat{\mathbf{V}}_{jj}$ is the $(j,j)^{th}$ element of $\hat{\mathbf{V}}$.

Three approximations to $\hat{\mathbf{V}}$ are frequently used. These are

$$\hat{\mathbf{V}}_a = s^2\left(\mathbf{J}(\hat{\boldsymbol{\theta}})^T\mathbf{J}(\hat{\boldsymbol{\theta}})\right)^{-1}, \tag{A}$$

$$\hat{\mathbf{V}}_b = s^2\,\mathbf{H}(\hat{\boldsymbol{\theta}})^{-1}, \tag{B}$$

and

$$\hat{\mathbf{V}}_c = s^2\,\mathbf{H}(\hat{\boldsymbol{\theta}})^{-1}\left(\mathbf{J}(\hat{\boldsymbol{\theta}})^T\mathbf{J}(\hat{\boldsymbol{\theta}})\right)\mathbf{H}(\hat{\boldsymbol{\theta}})^{-1}, \tag{C}$$

where $\mathbf{J}(\hat{\boldsymbol{\theta}})$ is the Jacobian of $\mathbf{F}(\boldsymbol{\theta})$ at $\hat{\boldsymbol{\theta}}$; $\mathbf{H}(\hat{\boldsymbol{\theta}})$ is the Hessian of $S(\boldsymbol{\theta})$ at $\hat{\boldsymbol{\theta}}$; and s^2 is the residual variance, $s^2 = S(\hat{\boldsymbol{\theta}})/n-p$.

Approximation (A) is the most common approximation to $\hat{\mathbf{V}}$. It is computed by approximating $\mathbf{F}(\boldsymbol{\theta})$ by the affine approximation around $\hat{\boldsymbol{\theta}}$,

$$\mathbf{F}(\boldsymbol{\theta}) \approx \mathbf{F}(\hat{\boldsymbol{\theta}})+\mathbf{J}(\hat{\boldsymbol{\theta}})\,(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}),$$

where $\mathbf{F}(\boldsymbol{\theta})$ is a column vector with i^{th} component $f(\mathbf{x}_i;\boldsymbol{\theta})$, and then directly applying the

linear least squares theory.

Approximation (B) can be obtained using maximum likelihood theory. For large samples, maximum likelihood estimators are asymptotically distributed as the p-variate normal with variances and covariances given by $\dot{\mathbf{V}}$ where

$$\dot{\mathbf{V}}^{-1} = -\mathrm{E}\left(\frac{\partial^2 \log L(\mathbf{Y})}{\partial\theta_j \partial\theta_k}\right)$$

It is straightforeward to show that $\dot{\mathbf{V}}^{-1}$ approaches $\hat{\mathbf{V}}_b{}^{-1}$ as $n \to \infty$.

Approximation (C) can be obtained from sensitivity analysis. If the observations $\mathbf{Y}$ are changed to $\mathbf{Y} + \dot{e}$, then, to within $O(\dot{e})$ terms, $\hat{\boldsymbol{\theta}}$ will be changed to

$$\dot{\boldsymbol{\theta}}(\dot{e}) = \hat{\boldsymbol{\theta}} - \mathbf{H}(\hat{\boldsymbol{\theta}})^{-1}\, \mathbf{J}(\hat{\boldsymbol{\theta}})^T \dot{e}.$$

Solving

$$\hat{\mathbf{V}} = cov\left(\dot{\boldsymbol{\theta}}(\dot{e})\right) = \mathrm{E}\left[\left(\dot{\boldsymbol{\theta}}(\dot{e}) - \hat{\boldsymbol{\theta}}\right)\left(\dot{\boldsymbol{\theta}}(\dot{e}) - \hat{\boldsymbol{\theta}}\right)^T\right]$$

yields $\hat{\mathbf{V}} = \hat{\mathbf{V}}_c$.

When certain regularity conditions are met [Jennrich (1959)], each of these approximations to $\hat{\mathbf{V}}$ asymptotically will approach the true variance-covariance matrix of the model. Note also that these approximations differ only when

$$\sum_{i=1}^{n} r_i(\boldsymbol{\theta})\, \frac{\partial^2 f(\mathbf{x}_i;\boldsymbol{\theta})}{\partial\theta_j \partial\theta_k}$$

is nonzero. In particular, for linear functions, each of these representations of $\hat{\mathbf{V}}$ is equal to

$$s^2 \left(\mathbf{J}(\hat{\boldsymbol{\theta}})^T \mathbf{J}(\hat{\boldsymbol{\theta}})\right)^{-1} = s^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1} .$$

For nonlinear functions, $\hat{\mathbf{V}}_b$ is said to utilize observed information and $\hat{\mathbf{V}}_a$ is said to utilize expected information.

Linearization methods have the advantage that their resulting confidence regions and intervals are simple and inexpensive to construct, and that they produce bounded, convex confidence regions. In addition, the information needed to construct confidence regions and intervals using this method can be parsimoniously summarized by the p by p matrix $\hat{\mathbf{V}}$, and is well understood by users familiar with linear least squares. Because the linearization methods assume that both the intrinsic curvature and the parameter-effects curvature of $f(\mathbf{x}_i;\boldsymbol{\theta})$ are

zero, however, we expect that the linearization methods could sometimes produce observed coverages very far from the expected nominal coverage. The results of our Monte Carlo study show this to be true.

Likelihood method. The likelihood method is another approximate method for producing confidence regions and confidence intervals. The likelihood method confidence region about $\hat{\boldsymbol{\theta}}$ consists of those values $\tilde{\boldsymbol{\theta}}$ for which

$$S(\tilde{\boldsymbol{\theta}}) - S(\hat{\boldsymbol{\theta}}) \leq s^2 \, p \, F_{p,n-p,1-\alpha} \; .$$

This is analogous to equation (2.1) for confidence regions for the parameters of a linear function, although when $f(\mathbf{x}_i;\boldsymbol{\theta})$ is nonlinear in the parameters the resulting confidence region is no longer ellipsoidal. The likelihood method confidence interval about $\hat{\theta}_j$ is the interval bounded by the points which

$$\text{maximize } (\theta_j - \hat{\theta}_j)^2 \text{ subject to}$$

$$S(\boldsymbol{\theta}) - S(\hat{\boldsymbol{\theta}}) \leq s^2 \, F_{1,n-p,1-\alpha} .$$

This confidence interval is the projection onto the appropriate parameter axis of the above region, and is analogous to equation (2.4) for confidence intervals in the case of linear least squares.

When the solution locus is planar, the confidence regions (but not the confidence intervals) constructed using the likelihood method are equivalent to the lack-of-fit confidence regions, and therefore are exact. In addition, likelihood method confidence regions and intervals have the desirable property that they are constructed from contours of constant likelihood, and that the regions and intervals are not affected by reparameterization of the function $f(\mathbf{x}_i;\boldsymbol{\theta})$. Thus we might expect the likelihood method to produce confidence regions and confidence intervals with observed coverage closer to nominal than those produced using the linearization methods. However, the likelihood method has several practical disadvantages. Both the confidence regions and confidence intervals produced using the likelihood method can be disjoint and unbounded because the contours of a nonlinear function can be disjoint and unbounded. The method also is very expensive to use, and, when the data arrays are large, it can be awkward to publish the information necessary to reconstruct the confidence region because this information is not succinctly summarized as it is in the case of the linearization method.

Lack-of-fit method. The lack-of-fit method can be used to produce exact joint confidence regions for all p of the parameters, and to produce approximate confidence

intervals and confidence regions for subsets of the parameters. The lack-of-fit method is based on the fact that the quadratic forms

$$Q_1(\dot{\boldsymbol{\theta}}) = \frac{\mathbf{R}(\dot{\boldsymbol{\theta}})^T\mathbf{P}(\dot{\boldsymbol{\theta}})\mathbf{R}(\dot{\boldsymbol{\theta}})}{\dot{\sigma}^2}$$

and

$$Q_2(\dot{\boldsymbol{\theta}}) = \frac{\mathbf{R}(\dot{\boldsymbol{\theta}})^T\big(\mathbf{I}-\mathbf{P}(\dot{\boldsymbol{\theta}})\big)\mathbf{R}(\dot{\boldsymbol{\theta}})}{\dot{\sigma}^2}$$

where

$$\mathbf{P}(\dot{\boldsymbol{\theta}}) = \mathbf{J}(\dot{\boldsymbol{\theta}})\Big(\mathbf{J}(\dot{\boldsymbol{\theta}})^T\mathbf{J}(\dot{\boldsymbol{\theta}})\Big)^{-1}\mathbf{J}(\dot{\boldsymbol{\theta}})^T$$

are independent chi-square random variables with p and $n-p$ degrees of freedom respectively. Therefore,

$$\frac{Q_1(\dot{\boldsymbol{\theta}})/p}{Q_2(\dot{\boldsymbol{\theta}})/(n-p)}$$

is distributed as $F_{p,n-p,1-\alpha}$, so an exact $100\cdot(1-\alpha)\%$ confidence region consists of all values $\tilde{\boldsymbol{\theta}}$ such that

$$\frac{\mathbf{R}(\tilde{\boldsymbol{\theta}})^T\mathbf{P}(\tilde{\boldsymbol{\theta}})\mathbf{R}(\tilde{\boldsymbol{\theta}})}{\mathbf{R}(\tilde{\boldsymbol{\theta}})^T\big(\mathbf{I}-\mathbf{P}(\tilde{\boldsymbol{\theta}})\big)\mathbf{R}(\tilde{\boldsymbol{\theta}})} \leq \frac{p}{n-p}F_{p,n-p,1-\alpha}$$

Note that the lack-of-fit method does not require that the least squares solution be found prior to constructing the confidence region.

Similarly, a lack-of-fit method confidence interval for the j^{th} parameter consists of those values $\tilde{\theta}_j$ for which there exists values of θ_k, $k=1,\ldots,j-1,j+1,\ldots,p$, such that for these p parameter values, $\tilde{\boldsymbol{\theta}}$,

$$\frac{S^L(\hat{\boldsymbol{\theta}}_{\mathbf{J}_{k\neq j}(\tilde{\boldsymbol{\theta}})}) - S^L(\hat{\boldsymbol{\theta}}_{\mathbf{J}(\tilde{\boldsymbol{\theta}})})}{S^L(\hat{\boldsymbol{\theta}}_{\mathbf{J}(\tilde{\boldsymbol{\theta}})})/(n-p)} \leq F_{1,n-p,1-\alpha}$$

where $S^L(\hat{\boldsymbol{\theta}}_{\mathbf{J}_{k\neq j}(\tilde{\boldsymbol{\theta}})})$ is the residual sum of squares obtained when $\mathbf{R}(\tilde{\boldsymbol{\theta}})$ is *linearly* fit to all the columns of $\mathbf{J}(\tilde{\boldsymbol{\theta}})$ excluding the j^{th}, and $S^L(\hat{\boldsymbol{\theta}}_{\mathbf{J}(\tilde{\boldsymbol{\theta}})})$ is the residual sum of squares obtained when $\mathbf{R}(\tilde{\boldsymbol{\theta}})$ is *linearly* fit to $\mathbf{J}(\tilde{\boldsymbol{\theta}})$. This interval is exact if $f(\mathbf{x}_i;\boldsymbol{\theta})$ is linear in

θ_k, $k=1,...,j-1,j+1,...,p$; otherwise it is approximate. [See Halperin (1983).]

The lack-of-fit method is even more expensive to use than the likelihood method, and, as is the case for the likelihood method, the information needed to construct the confidence regions cannot be succinctly summarized for publication. Also, the confidence regions and confidence intervals constructed using the lack-of-fit method are guaranteed to contain every minimum, maximum, and/or saddle point of the likelihood surface. This makes the lack-of-fit method structurally undesirable.

3. The Monte Carlo Study

This section briefly describes how our Monte Carlo study was constructed. Full details are provided by Donaldson (1985).

The Monte Carlo method uses the computer to simulate the results of repeating an experiment many times in order to obtain a large sample from which the statistical properties of a system can be examined. For each simulation, we first generated the errors and response variables. The errors, $\dot{e}$, were produced using the Marsaglia and Tsang pseudo-normal random number algorithm (1984) as implemented by James Blue and David Kahanar of the National Bureau of Standards Scientific Computing Division. The response variable, $\mathbf{Y}$, was then constructed so that its i^{th} component is

$$y_i = f(\mathbf{x}_i;\dot{\boldsymbol{\theta}}) + \dot{e}_i \ .$$

Then the least squares estimate, $\hat{\boldsymbol{\theta}}$, was calculated using NL2SOL, an unconstrained quasi-Newton code for nonlinear least squares [Dennis, Gay, and Welsch (1981)]. Starting values for NL2SOL were set to the true values of the parameters, $\dot{\boldsymbol{\theta}}$, and the stopping criteria for the convergence tests based on the relative change in the parameters and in the sum of squares both were set to 10^{-5}.

Finally, for each confidence region or interval method and each derivative configuration being analyzed, we recorded whether the true values of the parameters were contained within the confidence regions and confidence intervals for this realization of the data. Determining whether the true parameter values lay within the confidence regions and confidence intervals about the least squares estimates fortunately did not require that we construct the full confidence regions and confidence intervals for each confidence level and method. Instead, we simply calculated the smallest confidence level, $1-\omega$, such that a $100\cdot(1-\omega)\%$ confidence region or confidence interval constructed using the method being analyzed will contain the

true parameter values. When $\omega > \alpha$, the true value did not lie in the $100 \cdot (1-\alpha)\%$ confidence region or confidence interval; when $\omega \leq \alpha$, it did. The values $1-\omega$ were obtained using the hypothesis tests corresponding to the formulas for confidence regions and intervals given in Section 2, and the appropriate cumulative distribution functions; the procedures are described in detail in Donaldson (1985). The cumulative distribution functions were obtained from the STARPAC subprogram library [Donaldson and Tryon (1983)].

The observed coverage, γ_α, for the particular nominal confidence level, method and system under analysis is the percentage of the total number of realizations of the data, N, for which $\omega \leq \alpha$. When N is large, the standard deviation of γ_α can be approximated using the normal approximation to the binomial distribution. In this study we used $N = 500$, so the maximum standard deviation of the observed coverage at any coverage level is approximately 2.2%.

Note that substituting a new realization of the data for one which could not be completely analyzed because either (a) the nonlinear least squares algorithm did not converge, or (b) the test statistics could not be computed for every method being analyzed, is a form of censoring which will bias the observed coverages obtained. In our analysis, we adjusted the value of $\dot{\sigma}$ for each dataset so that every realization could be completely analyzed, and therefore the results reported in this paper are not derived from censored data.

We computed the observed coverage for four nominal confidence levels, 0.50, 0.75, 0.95, and 0.99. In this paper we only include our data for the level 0.95, although we comment briefly in Section 4 on our results at the other levels. Data for the full study are given in Donaldson (1985).

The references for the datasets used in our Monte Carlo study are given in Appendix A and described in detail in Donaldson (1985). With only two exceptions, the functions and data which comprise our datasets have been taken from Ratkowsky (1983), Himmelblau (1970), Guttman and Meeter (1965), and Duncan (1978). The standard deviation of the errors of some of the datasets has been adjusted in order to allow us to successfully analyze each realization of the data for each dataset. The two datasets not taken from the published literature are identified as 8ACA and 9AAG. Dataset 8ACA was created especially for this study by generalizing function 3 to a larger number of parameters. Dataset 9AAG involves a microwave absorption line function taken from a consulting session at the National Bureau of Standards in Boulder, Colorado.

The number of parameters in the 20 datasets analyzed range from 2 to 8 and the ratio of the number of parameters to the number of observations range from 2/42 to 3/5. While these

datasets are often troublesome, they are mostly real world problems that have not been made artificially difficult.

Each dataset was analyzed twice to allow us to examine the effect of increasing the standard deviation of the errors. In the first analysis, $\dot{e} \sim N(\mathbf{0},\dot{\sigma}^2\,\mathbf{I})$; in the second analysis, $\dot{e} \sim N\big(\mathbf{0},(\eta\,\dot{\sigma})^2\,\mathbf{I}\big)$, where η is approximately the largest number ≤ 10 for which every realization of the data could be successfully analyzed. The methods analyzed in the second analysis were the same as in the first except that variants B and C of the linearization method were excluded from the second analysis because, when $\eta > 1.0$, we were frequently unable to compute the required test statistics using these two variants.

Computation of the linearization method and the lack-of-fit method requires that certain derivatives be available. The Jacobian of $\mathbf{F}(\boldsymbol{\theta})$ is used by both the linearization and lack-of-fit methods. Variants B and C of the linearization method use the Hessian of $S(\boldsymbol{\theta})$ as well. In practice, analytic derivatives often are not available. Therefore, in our study each method was implemented and analyzed using three different derivative configurations. These configurations are (1) the Jacobian and Hessian both approximated by finite-differences, (2) the Jacobian computed analytically and the Hessian computed by finite-differences, and (3) both the Jacobian and the Hessian computed analytically. For derivative configurations (1) and (2), the variance-covariance matrix needed by the linearization method was returned directly from NL2SOL [Dennis, Gay and Welsch (1981)]. For configuration (3), it was constructed outside of NL2SOL. For details on the formulas used to compute the finite-difference derivative approximations, see Donaldson (1985).

We ran our Monte Carlo study in single precision on a 60 bit word length computer. All subroutines extracted from other sources were used without modification except for NL2SOL, which was changed for this study in two important ways. First we disabled the two tests within NL2SOL used to detect near singularity. Second, we used the STARPAC [Donaldson and Tryon (1983)] front end to NL2SOL. With this front end, the finite difference approximation to the Jacobian is computed with the optimal derivative step sizes selected using the algorithm developed by Schnabel (1981), thus maximizing the number of correct digits in each element of the finite difference Jacobian.

4. Results and Observations

This section presents the results of our Monte Carlo study of the lack-of-fit method, the likelihood method, and the three variants of the linearization method. The section is divided into a discussion of confidence regions and confidence intervals. For each, we also make a number of observations about the results. The conclusions we draw from our analysis are discussed in the next chapter.

The material in this chapter includes a number of figures. These are printed at the end of the paper.

Confidence Regions

Results. The results for nominally 95% confidence regions constructed using each of the methods analyzed in this study with $\dot{e} \sim N(\mathbf{0}, \dot{\sigma}^2 \mathbf{I})$ are graphically displayed in Figure 1. For each dataset, the observed coverage is plotted against the method and derivative configuration used to obtain it.

The three derivative configurations are labeled DC1, DC2, and DC3 in these and the following figures and tables, as well as in Appendix B, Here DC1 denotes use of finite difference approximations for both the Jacobian and the Hessian, DC2 denotes use of analytic Jacobian and finite difference Hessian, and DC3 denotes use of analytic Jacobian and Hessian. Since the computations required to calculate the lack-of-fit method results and the likelihood method results using derivative configurations DC2 and DC3 are exactly the same, these results are displayed together.

Figure 2 shows the analogous results for $\dot{e} \sim N\left(\mathbf{0}, (\eta\ \dot{\sigma})^2\ \mathbf{I}\right)$. As noted in Section 3, variants B and C of the linearization method are excluded from the analysis displayed in Figure 2 because computational difficulties were encountered for these variants when the variance of the errors was increased.

A conservative 95% confidence interval about the nominal confidence level is indicated on each plot by a pair of horizontal lines which represent the values $100 \cdot (1-\alpha) \pm 4.4$, where 4.4 is two times the maximum standard deviation of the observed coverage at any coverage level. This confidence interval provides a quick means of determining whether any of the observed coverages for each method are significantly different from the nominal confidence level at the 5% level. When the method used to construct the confidence regions and confidence intervals is exact, we expect that the observed coverage for 95% of all possible datasets will lie within this confidence interval.

Observations. Figures 1 and 2 show that the lack-of-fit and likelihood method confidence regions are quite reliable, and that the results are not affected by use of finite difference derivatives. In all our tests, they produced observed coverages which seldom vary from nominal by an amount that is significant at the 5% level. In fact, for these datasets, there is only one instance (dataset 3AAA, $\dot{e} \sim N(\mathbf{0},(\eta\ \dot{\sigma})^2\ \mathbf{I})$) where the difference between the nominal and observed coverages produced using these two methods is greater than 5%, and in this instance, the observed coverage is greater than nominal, not less.

The three variants of the linearization method, on the other hand, frequently produced far less reliable confidence regions, although, as discussed below, the results still do not appear to be affected by the use of finite-difference derivatives. The difference between the nominal and observed coverages obtained using the linearization methods often are considerably more than 20%, which is a difference that many if not most users would find unacceptable.

By comparing Figure 1 to Figure 2, it is apparent that increasing the variance of the errors does, in fact, increase the differences between observed and nominal coverage for all methods. Our tests at confidence levels 0.50, 0.75, and 0.99, which are not reported in detail here, also showed that the spread between the observed and nominal coverage obtained using the linearization method increases as the nominal confidence level is increased.

The large differences for some datasets between the observed coverage of confidence regions constructed using the likelihood method and those obtained using the linearization method may be explained by the difference in the shape of the two regions. The likelihood method confidence region corresponds to the boundary and interior of a contour of the sum of squares surface, i.e., a contour of constant likelihood, whereas the linearization method confidence regions are always ellipsoidal. We plotted these contours for various datasets, and the difference sometimes were very large. Examples for datasets 3AAA and 14AAG are given in Donaldson (1985).

Figure 1 also indicates that the observed coverage obtained using variants A, B, and C of the linearization method are nearly identical. The results of two-sided paired-sample t-tests indicate that there is no statistically significant differences at the 5% level between the observed coverages obtained using any of the variants of the linearization method with any of the derivative configurations. The same results were obtained for our tests at the 0.50, 0.75, and 0.99 confidence levels.

Confidence Intervals

Results. Figures 3 and 4 provide information for confidence intervals which is analogous to that shown in figures 1 and 2 for confidence regions. The observed coverages plotted are the *smallest* of the p confidence interval coverages obtained for each dataset. Figure 3 displays the observed confidence interval results for nominally 95% confidence levels, when $\dot{e} \sim N(\mathbf{0}, \dot{\sigma}^2\, \mathbf{I})$; figure 4 shows the results when $\dot{e} \sim N\left(\mathbf{0}, (\eta\, \dot{\sigma})^2\, \mathbf{I}\right)$, excluding linearization method variants B and C as was done for the linearization method confidence regions.

Observations. Figure 3 shows that for confidence intervals, the best results are obtained using the lack-of-fit and likelihood methods, and the worst results are obtained using the linearization method, as was the case for confidence regions. The lack-of-fit and likelihood methods produce confidence intervals which seldom vary from nominal by an amount that is significant at the 5% level, and never are less than nominal by more than 5.0%. Again, use of finite difference Jacobians does not appear to affect the results for these two methods.

The three variants of the linearization method, on the other hand, frequently produce far less reliable confidence intervals than the lack-of-fit and likelihood methods. Disturbing differences between observed and nominal coverages occur when each of the variants of the linearization method is used to construct confidence intervals. The observed coverage for a nominally 95% confidence interval is as low as 75.0%, 44.0%, and 10.8% for variants A, B, and C, respectively. For most of the datasets tested in our study, however, the span between observed and nominal coverage produced by the three variants of the linearization method is considerably less for confidence intervals than for linearization method confidence regions constructed about the parameters of the same dataset. This is especially true when derivative configurations DC2 and DC3 are used.

One reason why linearization method confidence intervals have better coverage than linearization method confidence regions is that, when the parameter estimates are correlated with each other, a number of points may be included in the linearization method confidence intervals but not in the confidence regions. Note, however, that if a confidence interval was computed for the linear combination of the parameters given by the eigenvector corresponding to the minor axis of the linearization method confidence region ellipsoid, then the linearization method confidence interval observed coverage should approximately equal that of the linearization method confidence region. In our Monte Carlo study, we actually computed the linearization method confidence interval observed coverage for this linear combination of the parameters. In every case, the observed coverage we obtained for the confidence interval

about this linear combination was approximately equal to that of the linearization method confidence region observed coverage.

The use of finite differences to approximate both the Jacobian and the Hessian appears to significantly degrade the confidence interval results for linearization variants B and C. Figure 3 shows that, while there is no striking difference in the results obtained using the three variants of the linearization method with derivative configurations DC2 and DC3, variants B and C degrade significantly more than variant A when using DC1, i.e., finite difference Jacobian and Hessian. A two-sided paired-sample t-test was used to determine whether, for a given derivative configuration, the observed coverages obtained using the different linearization method variants are statistically different at the 5% significance level. The results indicate that when derivative configuration DC2 and DC3 are used, the differences in the results obtained using variants A, B, and C are seldom statistically significant at the 5% level, but that when the Jacobian and Hessian are approximated using finite differences (derivative configuration DC1) then the differences in results are often significant.

Comparing Figures 3 and 4 shows that as the variance of the errors is increased, the differences between observed and nominal coverage also are increased, as was the case for the confidence region results. However, this increase is not as pronounced for confidence intervals as for confidence regions. The results at confidence levels 0.50, 0.75, 0.95, and 0.99 also showed that as the nominal confidence level approaches 100%, the spread between observed and nominal coverages obtained using the linearization method is increased.

5. Diagnostic tools

The preceding section demonstrates a pressing need for diagnostics to warn users when the commonly used linearization method confidence region will not have adequate coverage. In addition, it would be useful to have a warning to indicate when the approximate likelihood method may be inadequate. Bates and Watts (1980) have proposed measures of nonlinearity that provide such diagnostics.

According to Bates and Watts, when their relative measure of parameter effects curvature is small compared to the critical value $(F_{p,n-p,0.05})^{-1/2}$, then the linear coordinate grid assumption is valid over the region of interest, and therefore the linearization method confidence region should be adequate. Similarly, when their relative measure of intrinsic curvature is small compared to the same critical value, then the assumption that the solution

locus is planar is valid over the region of interest, and therefore the likelihood method confidence region should be adequate.

In Figure 5 we plot the 20 confidence region observed coverages obtained using linearization method variant A with analytic derivatives (derivative configuration DC3) and $\dot{e} \sim N(\mathbf{0},(\eta\ \dot{\sigma})^2\ \mathbf{I})$ against the Bates and Watts relative measure of parameter effects curvature. Likewise, in figure 6 we plot the corresponding 20 likelihood method confidence region observed coverages against the Bates and Watts relative measure of intrinsic curvature. The relative curvature measures were computed at the true parameter values using the true variance of the errors. In these plots, we have scaled the measures of parameter effects curvature and intrinsic curvature by dividing the measure by the appropriate critical value. Thus, in both of these plots, a scaled curvature measure less than 1 indicates the relative measure was less than the critical value, while a value greater than 1 indicates the curvature exceeded the critical value.

It is clear from figure 5 that the Bates and Watts parameter effects curvature measure is strongly correlated with the observed coverage obtained using the linearization method. In fact, for our data, as the parameter effects curvature increases, the observed coverage for the linearization method confidence regions decreases nearly monotonically and linearly as the logarithm of the scaled parameter effects curvature. Furthermore, in all datasets where the parameter effects curvature is less than the critical value, the observed confidence region is very close to nominal, while in all cases where the parameter effects curvature is greater than ten times the critical value, the observed coverage is unsatisfactorily low. Datasets with parameter effects curvature between one and ten times the critical value had observed confidence region coverage between 83.2% and 91.6%. From these results, it appears that the Bates and Watts parameter effects curvature is a reliable, if perhaps stringent, indicator of when the linearization method will produce reliable confidence regions.

Figure 6 shows that all but one of the 20 datasets tested in this study have intrinsic curvature which is less than the critical value, which means that each of these datasets is nearly planar. For nearly planar datasets we expected good observed coverage from the likelihood method, and, as figure 6 shows, that is what we got. Since none of our datasets have high intrinsic curvature, however, we do not know how the likelihood method will perform when the solution locus is not nearly planar. We cannot assume that the accurate results obtained in our study using the likelihood method will necessarily carry over to datasets with large intrinsic curvature.

Cook, Tsai and Wei (1984) provide an example which has scaled parameter effects curvature of 934.5 and scaled intrinsic curvature of 8.4. Both the parameter effects curvature and intrinsic curvature of this dataset exceed any curvature measure we observed in the 20 datasets in our study. For this dataset, we computed observed confidence region coverages of 19.0% and 95.0% using the linearization method and likelihood methods, respectively. While the linearization method confidence region observed coverage is very far from nominal as we would expect based on the parameter effects curvature of this model, the likelihood method confidence region observed coverage is not. We cannot conclude anything from this one observation. It is clear, however, that additional analysis of datasets with high intrinsic curvature would be useful to further assess the effect of a non-planar solution locus on the likelihood method.

6. Conclusions

Based on our computational study, we can draw conclusions about : i) the comparison between the three variants of the linearization method; ii) the reliability of linearization methods for calculating confidence regions and confidence intervals; and iii) the reliability of the likelihood and lack-of-fit methods for calculating confidence regions and confidence intervals.

When using the linearization method to construct confidence regions and intervals, our Monte Carlo study has shown no clearcut difference in the observed coverage of one variant as compared to another. In our tests, the only statistically significant difference among the results produced by the three linearization variants was in constructing confidence intervals with finite difference Jacobians and Hessians; here variant A was superior to variants B and C. We found no empirical evidence that one should prefer variants B or C, even though they may be appealing from a theoretical point of view. Therefore we conclude that variant A of the linearization method, which is computed using

$$\hat{\mathbf{V}}_a = s^2 \left(\mathbf{J}(\hat{\boldsymbol{\theta}})^T \mathbf{J}(\hat{\boldsymbol{\theta}}) \right)^{-1}, \tag{6.1}$$

is the best variant to use for constructing both confidence regions and confidence intervals, because it is simpler, less expensive, and more numerically stable to compute than variants B or C, which use

$$\hat{\mathbf{V}}_b = s^2 \, \mathbf{H}(\hat{\boldsymbol{\theta}})^{-1} \tag{6.2}$$

and

$$\hat{\mathbf{V}}_c = s^2\, \mathbf{H}(\hat{\boldsymbol{\theta}})^{-1} \left(\mathbf{J}(\hat{\boldsymbol{\theta}})^T \mathbf{J}(\hat{\boldsymbol{\theta}}) \right) \mathbf{H}(\hat{\boldsymbol{\theta}})^{-1} , \tag{6.3}$$

respectively. Variant A is simpler and less expensive because it only requires the Jacobian of the model function at the solution and not the additional second order terms that are also required to form the Hessian. It is more stable because it can be formed by inverting the upper triangular factor R of the QR factorization of the Jacobian rather than by calculating the inverse of the Hessian; the former calculation can be expected to lose roughly half as many digits as the latter in finite precision arithmetic.

The linearization method is not always an adequate method for approximating confidence regions and confidence intervals for the parameters of a nonlinear model, however. The results presented in the preceding section show just how poor the linearization method can be in some cases. Although there are many examples where the linearization method's observed coverage differs from nominal by only a very small amount, there are also many cases where the observed coverage is far lower than the nominal. In our tests, the best linearization method variant, A, produced observed coverages as low as 12.4% for nominal 95% confidence regions and 75.0% for nominal 95% confidence intervals.

Users will continue to use the linearization method, however, because it is readily available in software packages and provides a concise representation of the information needed to construct confidence regions and intervals. The erratic results obtained in our study when using the linearization method lead us to conclude that users of nonlinear least squares software must be helped to cautiously assess the results they obtain using the linearization method. The results of the preceding section show that the diagnostic tools proposed by Bates and Watts (1980) are very successful in indicating cases where the linearization method confidence regions are likely to be unreliable. In these cases, more reliable methods, such as the likelihood or lack-of-fit methods, are required to produce accurate confidence regions or intervals.

Our study shows that the lack-of-fit and likelihood methods both produce observed coverages acceptably close to nominal in every test case. Although the difficulties and expense associated with using these two methods to compute confidence regions make it unlikely that they will ever routinely replace the commonly used linearization method for this purpose, they appear to be a reliable alternative that should be considered when diagnostics show that linearization confidence regions are unreliable. It is not as difficult and expensive to construct confidence intervals using the lack-of-fit or likelihood methods, and we believe that producers of nonlinear least squares software should consider this possibility. (Constructing these

intervals requires the solution of a series of nonlinearly constrained optimization problems; it may be necessary to construct special purpose software to solve these problems as efficiently as possible.) Performing hypothesis tests using the likelihood or lack-of-fit methods is computationally simple for both confidence regions and intervals, so we recommend that one of these two methods be employed for hypothesis tests whenever possible.

Users may prefer the likelihood method to the lack-of-fit method even though it is approximate and the lack-of-fit method is exact, because the likelihood method has more desirable structural characteristics than the lack-of-fit method. Our study provides no empirical evidence that the results produced by the likelihood method are inferior to those produced by the lack-of-fit method. This does not guarantee that similar results will be obtained on other datasets, however. In particular, the results of the diagnostic test proposed by Bates and Watts showed that all our datasets have low intrinsic curvature, which is precisely the situation when likelihood methods are expected to be very reliable. The additional dataset we analyzed with high intrinsic curvature also produced likelihood method confidence region observed coverage close to nominal. Additional analysis is required to determine whether the likelihood method is reliable for datasets with high intrinsic curvature, and to determine whether the Bates and Watts measure of intrinsic curvature is a useful tool for indicating when the likelihood method confidence regions are likely to be unreliable.

In addition to diagnostics, it appears that there is a need for new methods for estimating confidence regions that are both reliable and easy to report. We are especially interested in investigating two methods that would result in conservative elliptical confidence regions. The first method is to find the minimal magnification of the (95%) linearization confidence region that encloses the (95%) likelihood or lack-of-fit confidence region. This would require the solution of a constrained optimization problem with one nonlinear equality constraint. The second method is to find the smallest volume ellipse that encloses the desired likelihood or lack-of-fit confidence region. This would require the solution of a semi-infinite programming poblem, i.e. an optimization problem with an infinite set of constraints.

7. Summary

We have presented the results of a Monte Carlo study comparing the linearization, likelihood and lack-of-fit methods for constructing confidence regions and confidence intervals. Our results indicate that the linearization method should be constructed using the simplest approximation to the variance-covariance matrix, (6.1), as it is simpler, less expensive, more

numerically stable, and at least as accurate as the other two linearization variants, which are constructed using (6.2) and (6.3). We have also given considerable evidence that confidence regions, and to some extent confidence intervals, constructed using the linearization method can be essentially meaningless.

Our study shows that the likelihood and lack-of-fit methods, on the other hand, produced consistently good results for the datasets tested. However, because the likelihood method is approximate it is not clear that the good results we obtained with it will necessarily be characteristic of all datasets. Also, because of the undesirable structural characteristics of the lack-of-fit method, it is unlikely to be used routinely, although in cases where accuracy is of extreme importance, it may be a useful tool to have.

Because of the uncertainty associated with the linearization and likelihood methods, we also have briefly examined how the Bates and Watts curvature measures relate to the confidence region observed coverages we obtained in this study. Our results show that the Bates and Watts parameter effects curvature appears to provide excellent indication of when the linearization method may produce less than satisfactory results. Our results are not as conclusive, however, about the relation between intrinsic curvature and likelihood method coverage since the solution locus for all of our datasets were nearly planar.

References

Bard, Y. (1974), *Nonlinear Parameter Estimation,* Academic Press, New York.

Bates, D. M., Watts, D. G. (1980), "Relative curvature measures of nonlinearity," *Journal of the Royal Statistical Society (B),* 42(1), pp. 1-25.

Beale, E. M. L. (1960), "Confidence regions in non-linear estimation," *Journal of the Royal Statistical Society (B),* 22(1), pp. 41-76.

Cook, R. D., Tsai, C. L., Wei, B. C. (1984), "Bias in nonlinear regression," Mathematics Research Center Technical Summary Report No. 2645, University of Wisconsin, Madison, Wisconsin.

Dennis, J. E. Jr., Gay, D. M., Welsch, R. E. (1981), "Algorithm 573: NL2SOL - An Adaptive Nonlinear Least Squares Algorithm [E4]," *ACM Transactions on Mathematical Software,* 7(3), pp. 369-383.

Donaldson, J. R. (1985), "Computational experience with confidence regions and confidence intervals for nonlinear least squares," MS Thesis, Department of Computer Science, University of Colorado, Boulder, Colorado.

Donaldson, J. R., Tryon, P. V. (1983), "Nonlinear least squares regression using STARPAC," National Bureau of Standards Technical Note 1068-2, U.S. Government Printing Office, Washington, D. C.

Draper, N. R., Smith, H. (1981), *Applied Regression Analysis (Second Edition),* John Wiley and Sons, New York.

Duncan, G. T. (1978), "An empirical study of jackknife-constructed confidence regions in nonlinear regression," *Technometrics,* 20(2), pp. 123-129.

Gallant, R. A. (1976), "Confidence regions for the parameters of a nonlinear regression model," Institute of Statistics Mimeograph Series No. 1077, North Carolina State University, Raleigh, North Carolina.

Guttman, I., Meeter, D. A. (1965), "On Beale's measures of non-linearity," *Technometrics,* 7(4), pp. 623-637.

Himmelblau, D. M. (1970), *Process Analysis by Statistical Methods,* John Wiley and Sons, Inc., New York.

Jennrich, R. I. (1968), "Asymptotic properties of non-linear least squares estimators," *The Annals of Mathematical Statistics,* 40(2), pp. 633-643.

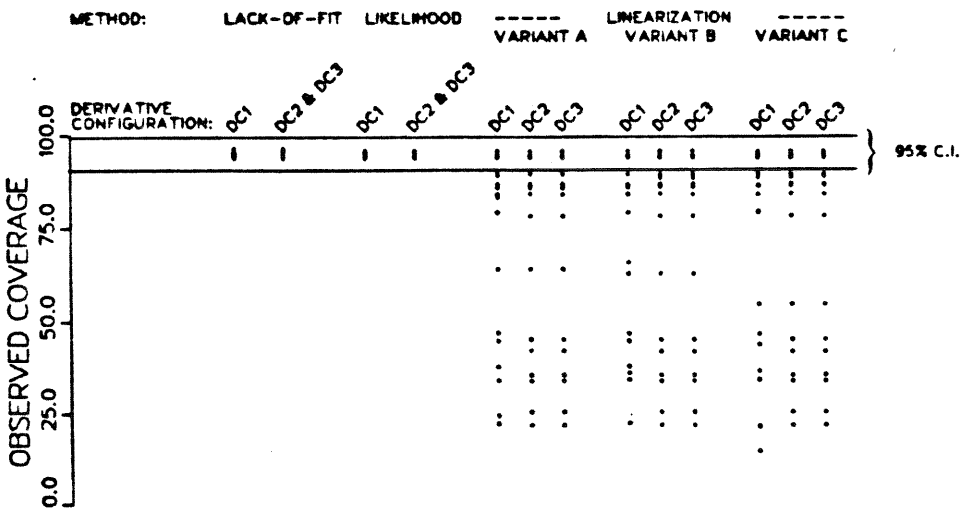
Marsaglia, G., Tsang, W. W. (1984), "A fast, easily implemented method for sampling from decreasing or symmetric unimodal density functions," *SIAM Journal on Scientific and Statistical Computing,* 5(2), pp. 349-359.

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

Schnabel, R. B. (1982), "Finite difference derivatives - theory and practice," Internal Report.

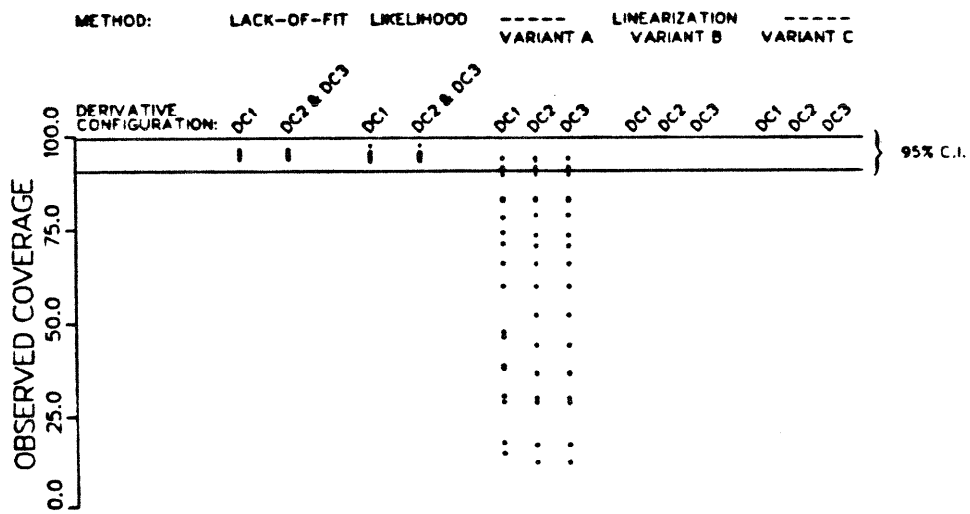
Appendix

Dataset	Id.	p/n	Reference
1	2AAA	2/12	Guttman and Meeter (1965) ; model η_2 , page 628
2	3AAA	2/12	Guttman and Meeter (1965) ; model η_3 , page 628
3	4AAA	2/24	Duncan (1978) ; model III , page 127
4	5AAF	4/18	Himmelblau (1970) ; model 6.2-3 , page 183
5	6AAA	3/13	Himmelblau (1970) ; model 6.2-4 , page 188
6	8ACA	4/24	None
7	9AAG	8/25	Inghold Hertel; Microwave Absorption Line Function (personal communication)
8	11AAB	4/9	Ratkowsky (1983) ; model 4.4 , page 62
9	12AAB	4/9	Ratkowsky (1983) ; model 4.14 , page 77
10	14ACG	3/10	Ratkowsky (1983) ; model 3.5 , page 51 and 58
11	14ABG	3/21	Ratkowsky (1983) ; model 3.5 , page 51 and 58
12	14AAG	3/42	Ratkowsky (1983) ; model 3.5 , page 51 and 58
13	15AAA	3/16	Ratkowsky (1983) ; model 6.11 , page 120 and 58
14	16AAF	5/27	Ratkowsky (1983) ; model 6.12 , page 122, 123 and 125
15	17AAA	2/42	Ratkowsky (1983) ; model 3.4 , page 50 and 58
16	18AAA	3/9	Ratkowsky (1983) ; model 4.1 , page 61 and 88
17	19AAA	3/9	Ratkowsky (1983) ; model 4.2 , page 61 and 88
18	20AAG	4/9	Ratkowsky (1983) ; model 4.3 , page 62 and 88
19	21AAA	4/9	Ratkowsky (1983) ; model 4.5 , page 63 and 88
20	22AAB	3/5	Ratkowsky (1983) ; model 5.1 , page 93 and 102



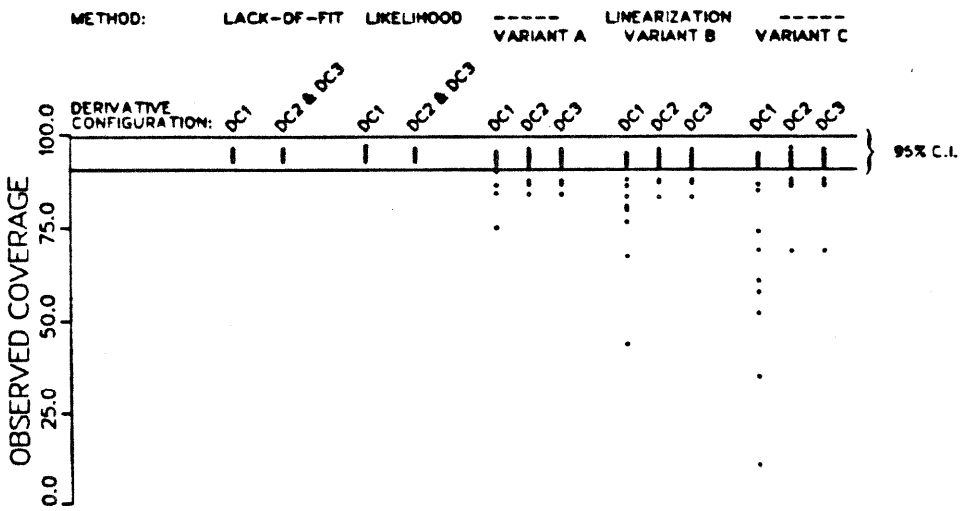
Observed coverage for nominally 95% confidence regions
with $\dot{e} \sim N(\mathbf{0}, \dot{\sigma}^2 \mathbf{I})$
versus
method by derivative configuration

Figure 1



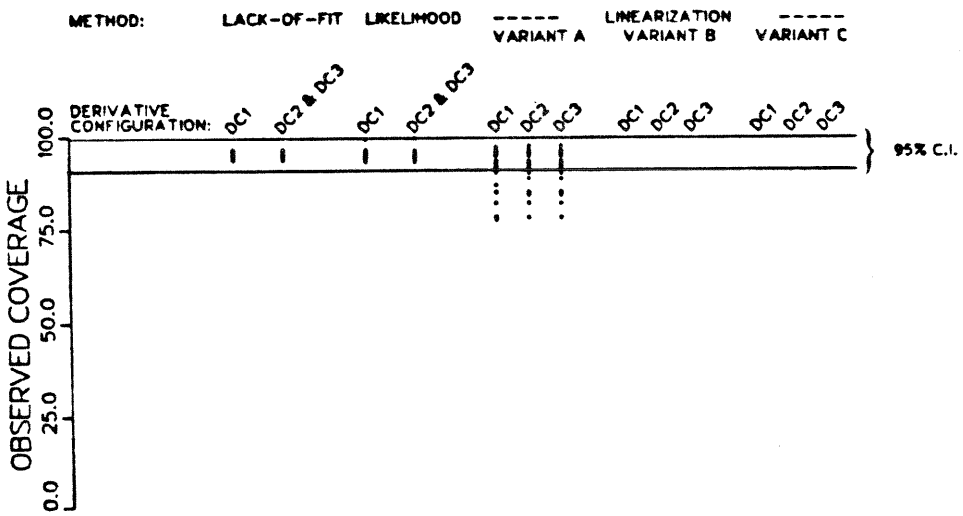
Observed coverage for nominally 95% confidence regions
with $\dot{e} \sim N\left(\mathbf{0}, (\eta\, \dot{\sigma})^2 \mathbf{I}\right)$
versus
method by derivative configuration

Figure 2



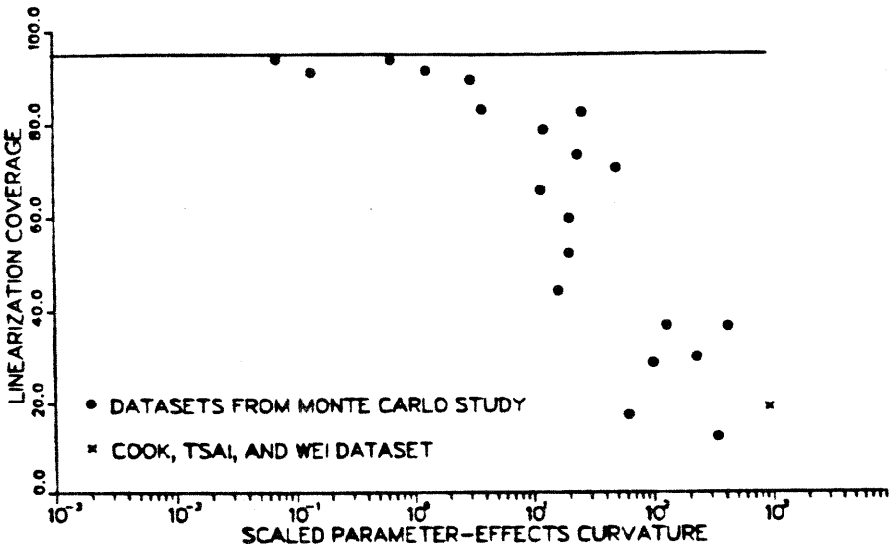
Observed coverage for nominally 95% confidence intervals
with $\dot{e} \sim N(\mathbf{0}, \dot{\sigma}^2 \mathbf{I})$
versus
method by derivative configuration

Figure 3



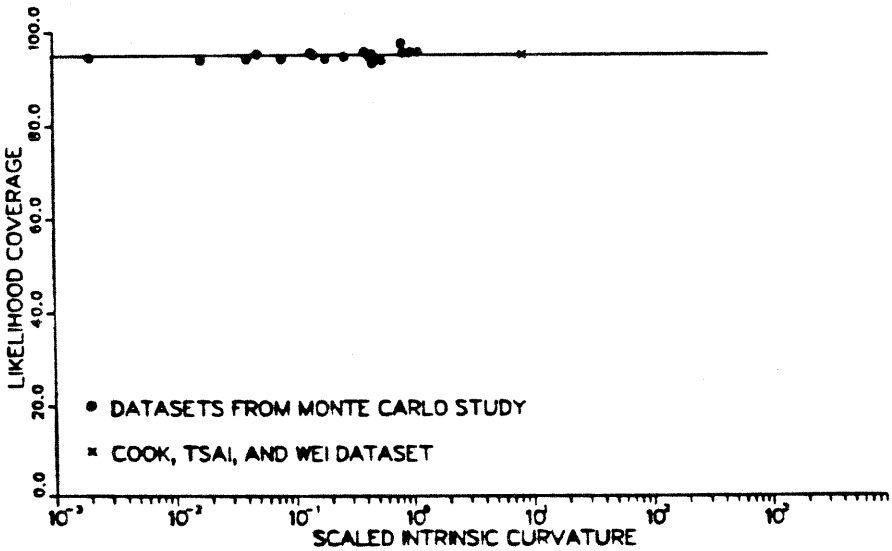
Observed coverage for nominally 95% confidence intervals
with $\dot{e} \sim N\left(\mathbf{0}, (\eta\, \dot{\sigma})^2 \mathbf{I}\right)$
versus
method by derivative configuration

Figure 4



Linearization method confidence region observed coverage
versus
parameter effects curvature scaled by $(F_{p,n-p,0.05})^{-1/2}$

Figure 5



Likelihood method confidence region observed coverage
versus
intrinsic curvature scaled by $(F_{p,n-p,0.05})^{-1/2}$

Figure 6