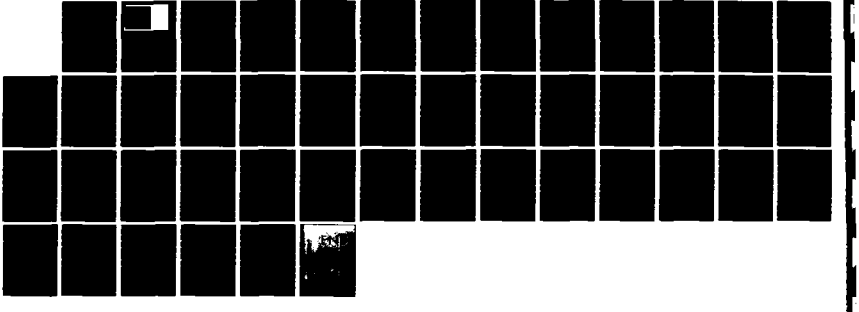
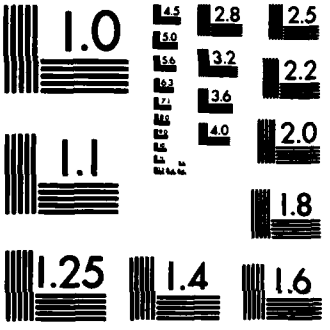
AD-A132 838 MEASURES OF NONLINEARITY FOR SEGMENTED REGRESSION MODELS(U) WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER M L GOLDBERG AUG 83 MRC-TSR-2549 1/1

UNCLASSIFIED DAAG29-80-C-0041 F/G 12/1 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

MRC Technical Summary Report #2549

MEASURES OF NONLINEARITY FOR SEGMENTED REGRESSION MODELS

Miriam L. Goldberg

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53705

August 1983

(Received January 11, 1983)

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

National Science Foundation
Washington, DC 20550

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

MEASURES OF NONLINEARITY FOR SEGMENTED REGRESSION MODELS

Miriam L. Goldberg

Technical Summary Report #2549
August 1983

ABSTRACT

A simple physical model of residential energy consumption provides the framework for an exploration of segmented regression models fit by least squares. The energy model is a generalization of a linear, single change-point model such as that considered by Hinkley (1971).

Some simple geometric measures of nonlinearity and nondifferentiability are proposed. These measures are related to the construction of approximate confidence regions for the parameters of a general segmented model. In addition, the relation shown between these measures and those proposed by Bates and Watts (1980) may be useful in analyzing continuously differentiable models.

AMS (MOS) Subject Classification: 62J02

Key Words: Nonlinearity, Regression, Confidence Intervals, Least Squares, Nondifferentiable Models.

Work Unit Number 4 - Statistics and Probability

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041. This material is based upon work supported by the National Science Foundation under Grant No. MCS-7927062, Mod. 2, and by joint funding from the Ford Foundation's State Environmental Management Program and the New Jersey Department of Energy.

SIGNIFICANCE AND EXPLANATION

Simple procedures are presented for assessing the severity of nonlinearity in a regression model involving a function which is nondifferentiable with respect to the unknown parameters. The nonlinearity measures proposed indicate the validity of standard approximations which may be used to determine the accuracy of parameter estimates. The proposed measures are related to existing measures of nonlinearity, but can be applied to a broader class of models, and in some cases may be easier to calculate.

The methods developed are motivated and illustrated by a simple model of residential energy consumption. This model has been the basis for measurements of energy conservation in several studies.

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the author of this report.

Table of Contents

MEASURES OF NONLINEARITY FOR SEGMENTED REGRESSION MODELS

Miriam L. Goldberg

1. Introduction.

The object of this paper is to develop the geometry of nondifferentiable least squares problems, and within that framework to indicate same simple procedures for assessing the effects of nonlinearity. Our exploration of piecewise differentiable regression models is based on a simple model which arises in the context of residential energy analyses. This model is a generalization of a linear change-point model.

We begin by describing the motivating model. After reviewing some basic elements of the geometry of nonlinear regressions, we then examine the behavior of the residual sum of squares function for the energy model, and relate this function to approximate confidence intervals for the unknown parameters. Finally, we consider some measures of nonlinearity, which indicate the validity of these approximations, and which are appropriate for nondifferentiable models. In addition, these measures may be useful for certain types of continuously differentiable models.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041. This material is based upon work supported by the National Science Foundation under Grant No. MCS-7927062, Mod. 2, and by joint funding from the Ford Foundation's State Environmental Management Program and the New Jersey Department of Energy.

2. The Energy Model

A simple model of residential energy consumption assumes daily consumption is constant at the baseload level α as long as the average outdoor temperature T is above a reference temperature τ, and increases in proportion to $\tau - T$ for $T < \tau$. With Y_m representing average daily fuel consumption for the N_m days of month m, our model is expressed formally as

$$Y_m = \alpha + \beta H_m(\tau) + \varepsilon_m \quad , \tag{2.1}$$

where

$$H_m(\tau) = \frac{1}{N_m} \sum_{j=1}^{N_m} (\tau - T_{mj}) \, I(T_{mj} < \tau) \quad , \tag{2.2}$$

I is the indicator function and ε is a random disturbance. The variable $H_m(\tau)$ represents the average daily base-τ degree-days for the month. The temperature τ is interpreted as the maximum outdoor temperature at which the furnace is required to heat the house, and β as the house's effective heat loss rate.

Models of this type have been the basis for analyses of energy consumption patterns in a large number of gas heated houses, and in a smaller number of oil- and electrically-heated houses.* The consumption data Y_m are derived from a customer's fuel bills. The daily temperature data T_{mj}, in integer degrees Farenheit, are obtained from a nearby U.S. Weather Bureau station (National Oceanic and Atmospheric Administration, monthly).

Equation (2.1) has also been applied to aggregate data,** with Y_m representing fuel consumption per household for month m. For utility- or state-wide aggregates, a different definition of the degree-day variable

---

* See, e.g., Fels et al (1981), Dutt, Lavine et al (1982), and Socolow (1978).

** See Fels and Goldberg (1982) and Goldberg and Fels (1982).

$H_m(\tau)$ is used, to account for the lag introduced by meters' being read on different days throughout the month:

$$H_m(\tau) = \frac{\sum_{j=1}^{N_m} (N_m+1-j)(\tau-T_{mj})I(T_{mj}<\tau) + \sum_{j=1}^{N_{m-1}} j(\tau-T_{m-1,j})I(T_{m-1,j}<\tau)}{\sum_{j=1}^{N_m} (N_m+1-j) + \sum_{j=1}^{N_{m-1}} j} . \quad (2.3)$$

For both single-house and aggregate analyses, the major use of the model defined by Equation (2.1) is in determining the normalized annual consumption Γ. The index Γ is given by

$$\Gamma = 365\ (\alpha+\beta H_o(\tau)) \quad (2.4)$$

where $H_o(\tau)$ is the long-term (several-year) average of daily degree-days base τ.

If consumption data were available on a daily, rather than monthly, basis, so that $N_m \equiv 1$, Equation (2.1) would represent a simple change-point regression with slope zero over one region. Such a model has been analyzed in detail by Hinkley (1971). In addition to the summation in Equation (2.2) or (2.3), a second important difference between the energy model considered here and Hinkley's change-point model is in the restriction placed on the temperature data T_{mj}, as discussed below in Section 5.

For the energy model, we will consider estimation of the reference temperature τ, baseload α, and heating rate β by the method of least squares. Placing the model in a more general context, we treat it as a special case of a piecewise differentiable model. We will explore the behavior of such models in the framework of general nonlinear models. Naturally, many existing results for simple change-point models relate closely to this problem. We will continually return to the energy model defined by Equation (2.1) for illustration.

Our emphasis is on methods for assessing the validity of approximate confidence intervals for the model parameters. Two approximation methods are considered. One is based on the asymptotic normality of the least squares estimates, and implicitly on a linearization of the model function. The other is based on the asymptotic chi-squared distribution of the likelihood ratio, and uses regions bounded by contours of constant Residual Sum of Squares (RSS). In developing methods for assessing the adequacy of these approximations, we will rely on the geometry of nonlinear least squares.

3. The Geometry of Nonlinear Least Squares

The geometrical approach has been developed extensively for continuously differentiable models, and will be applied here to the general piecewise differentiable model. The general nonlinear model with unknown p-dimensional parameter θ can be written in matrix form as

$$Y = \eta(\theta) + \varepsilon \tag{3.1}$$

$$E(\varepsilon) = 0 \tag{3.2}$$

$$E(\varepsilon'\varepsilon) = \sigma^2 I \quad . \tag{3.3}$$

Here, Y, η, and ε are n-dimensional vectors, such that η_m, the mth component of η, depends on observations x_m as well as on θ. We further assume that the random disturbance ε has a Gaussian distribution. Note that ε enters the model linearly; the nonlinearity is only in the model function η.

For the energy model given by Equation (2.1), $\theta = [\alpha,\beta,\tau,]'$ (with the apostrophe denoting the transpose) and

$$\eta(\theta) = \alpha 1 + \beta H(\tau), \tag{3.4}$$

where H is the n-dimensional vector with components H_m given by Equation (2.2) or Equation (2.3) and 1 signifies an n-dimensional vector of ones. The observations x_m are vectors of daily temperatures T_{mj}.

For the general model, as θ ranges over the parameter space Θ, the function $\eta(\theta)$ sweeps out a p-dimensional surface or "solution locus" L in the n-dimensional sample space:

$$L = \{\eta(\theta) : \theta \in \Theta\}.$$

We define the derivative vector $\dot{\eta}$ and the Hessian matrix $\ddot{\eta}_m$ as

$$\dot{\eta}(\theta) = \frac{\partial \eta}{\partial \theta'} = \left[\frac{\partial \eta_m}{\partial \theta_i}\right]_{nxp},$$

$$\ddot{\eta}_m(\theta) = \frac{\partial^2 \eta_m}{\partial \theta \partial \theta'} = \left[\frac{\partial^2 \eta_m}{\partial \theta_i \partial \theta_j}\right]_{pxp}.$$

(Throughout this paper an expression in square brackets indicates a matrix with components given by the subscripted expression, such that $m=1,2,\ldots,n$; $i,j=1,2,\ldots,p$.)

The least squares estimate $\hat{\theta}$ is the solution to the normal equation

$$\dot{\eta}'(\theta)\,(Y - \eta(\theta)) = 0.$$

That is, the residual vector $Y-\eta(\hat{\theta})$ is normal to the tangent plane at $\eta(\hat{\theta})$, the tangent plane being the linear span of the column vectors which make up the matrix $\dot{\eta}(\theta)$. In this sense, $\eta(\hat{\theta})$ is the projection $\hat{Y}_L$ of Y onto the solution locus L. The estimate θ is determined by pulling back the projection $\eta(\hat{\theta}) = \hat{Y}_L$ to the parameter space Θ.

To quantify the severity of departures from linearity, Bates and Watts (1980) propose measuring the nonlinearity of a model in terms of the curvature of the solution locus. For any direction v in the parameter space, the tangent t_v and acceleration vector a_v at θ are defined by

$$t_v = \dot{\eta}(\theta)v \tag{3.5}$$

$$a_v = [v'\ddot{\eta}_m(\theta)v]_{n\times 1}. \tag{3.6}$$

The curvature K_v in the v direction is then defined as

$$K_v = \frac{|a_v|}{|t_v|^2} . \qquad (3.7)$$

The relative curvature γ_v is obtained by multiplying K_v by the standard radius $\sqrt{ps^2}$, where s^2 is an estimate of the error variance σ^2. In the present work, s^2 is always obtained from the residual sum of squares RSS from the regression, as $s^2 = RSS(\hat{\theta})/(n-p)$. Decomposing the acceleration vector into components $a_v^{\parallel}$ and $a_v^{\perp}$, respectively parallel and normal to the tangent plane, yields analogous definitions for tangential and normal curvatures $K_v^{\parallel}$ and $K_v^{\perp}$, and relative curvatures $\gamma_v^{\parallel}$ and $\gamma_v^{\perp}$.

Noting that the tangential acceleration component is caused by the parameterization chosen, while the normal component is independent of this choice, Bates and Watts refer to the parallel $K_v^{\parallel}$ as the "parameter-effects" curvature, and to the normal $K_v^{\perp}$ as the "intrinsic" curvature. That is, the acceleration component normal to the solution locus L describes the bending of the p-dimensional surface L in n-dimensional Euclidean space. The acceleration component parallel to the tangent plane simply reflects the meandering within the solution locus of the "lifted line"

$$\eta_v = \{\eta(\theta+rv) : r \in R\}.$$

The parameter-effects curvature can in principle be reduced or removed by an appropriate reparameterization (Bates and Watts, 1981). By contrast, we may consider a model to be intrinsically nonlinear with respect to the parameter θ_j if the normal curvature (or acceleration) in the direction of θ_j is nonzero. This is equivalent to requiring that the vector $\partial^2\eta(\theta)/\partial\theta_j^2$ (composed of the jth diagonal elements of the matrices $\ddot{\eta}_m(\theta)$) does not lie in the plane spanned by the columns of $\dot{\eta}(\theta)$.

4. Nonlinearity in the Energy Model.

Many of the problems introduced by nondifferentiability can be understood in terms of the general geometrical framework just described. For the energy model defined by Equation (2.1), the model function η is given by Equation (3.4). Thus,

$$\begin{aligned} \dot{\eta}(\alpha,\beta,\tau) &= \left[\frac{\partial \eta}{\partial \alpha} \frac{\partial \eta}{\partial \beta} \frac{\partial \eta}{\partial \tau} \right] \\ &= [1 \mid H(\tau) \mid \beta F(\tau)]. \end{aligned} \tag{4.1}$$

The degree-day derivative F is obtained by dropping the terms of the form $(\tau - T_{mj})$ from Equation (2.2) or (2.3). For the single house, we have

$$F_m(\tau) = \frac{1}{N_m} \sum_{j=1}^{N_m} I(T_{mj} \leq \tau).$$

That is, F is (arbitrarily) defined to be right-continuous at discontinuity points T_{mj}, which occur only at integer Farenheit degrees. The step-function F_m is thus the empirical distribution function of outdoor temperatures T_{mj} for month m, and H_m is the convolution of temperature with F_m.

The Hessian $\ddot{\eta}_m$ for Equation (2.1) is given by

$$\ddot{\eta}_m(\alpha,\beta,\tau) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & F_m(\tau) \\ 0 & F_m(\tau) & \beta \partial F_m/\partial \tau \end{bmatrix} . \tag{4.2}$$

The only nonzero diagonal element is $\partial^2 \eta / \partial \tau^2 = \beta \partial F / \partial \tau$, which is a delta function with spikes at discontinuity points of F (i.e., at integers). Thus, the energy model is intrinsically linear with respect to α and β, but intrinsically nonlinear with respect to τ. Between any two successive integers, however, the model is also intrinsically linear in τ, since $\beta \partial F / \partial \tau = 0$. Hence, the solution locus L (and in this sense the model) is piecewise planar. However, the model function η is nonlinear in β as well

as in τ, since $\partial^2\eta/\partial\beta\partial\tau$ is nonzero. Thus, in addition to whatever intrinsic nonlinearity results from the discontinuities in $F(\tau)$, we expect to find effects of nonlinearity in the parameterization.

5. Impact of Nonlinearities on Approximate Confidence Regions

Approximate confidence regions for a q-dimensional linear combination $C\theta$ based on the asymptotic normality of $\hat{\theta}$ are given by the set of θ satisfying

$$(\theta-\hat{\theta})'C'\{C'(\dot{\eta}'\dot{\eta})^{-1}C\}^{-1}C(\theta-\hat{\theta}) \leq s^2 F^{\pi}_{q,n-p} \quad . \tag{5.1}$$

In Equation (5.1), $\dot{\eta}$ is the derivative evaluated at $\hat{\theta}$, π denotes a probability, and $F^{\pi}_{q,n-p}$ the $1-\pi$ quantile of the F-distribution with q and $n-p$ degrees of freedom. If the model function η is linear, the region defined by Equation (5.1) has exact confidence level $1-\pi$ even in finite samples. The small-sample validity of such confidence regions is affected by both parameter-effect and intrinsic nonlinearity.

By contrast, the sum-of-squares based regions are unaffected by parameter effects. For continuously differentiable q-dimensional functions g, such a region is the set of $g(\theta)$ such that

$$\frac{RSS(\theta) - RSS(\hat{\theta})}{RSS(\hat{\theta})} \frac{n-p}{q} \leq F^{\pi}_{q,n-p} \tag{5.2}$$

or

$$RSS(\theta) \leq RSS(\hat{\theta}) \left(1+ \frac{q}{n-p} F^{\pi}_{q,n-p}\right) . \tag{5.3}$$

Note that confidence intervals for a single component θ_j are obtained from Equation (5.3) by taking $q = 1$ and $g(\theta) = \theta_j$.

The region defined by Equation (5.3) is the inverse image in the parameter space θ of the intersection of the solution locus L with a sphere centered at Y whose squared radius is $RSS(\hat{\theta})(1 + \frac{q}{n-p} F^{\pi}_{q,n-p})$. If the solution locus is flat, the region determined by Equation (5.3) has exact confidence level $1-\pi$ in finite samples. Thus, the small-sample validity of the sum-of-squares approximate confidence region depends on how sharply the solution locus departs from the approximating tangent plane over the region of interest.

For continuously differentiable models the use of confidence regions based on Equation (5.1) or (5.2) is well-established. The asymptotic validity of confidence regions defined by Equation (5.1) was proved by Fisher (1925), and the validity of regions determined by Equation (5.2) by Wilks (1938).

For small samples, Beale (1960) proposed an inflation factor μ for the right-hand side of Equation (5.2) which yields a conservative confidence region for the case $q = p$. Beale's factor μ is given by

$$\mu = 1 + \frac{n\,(p+2)}{(n-p)\,p} N^{\perp} . \qquad (5.4)$$

Bates and Watts (1980) showed Beale's nonlinearity measure $N^{\perp}$ to be equal to one quarter the mean square relative intrinsic curvature $\gamma^{\perp}$, and showed also that the factor μ was very close to one for a wide variety of data sets.

For cases where the parameter-effects curvature is slight, Hamilton, Watts and Bates (1982) showed how to approximate the sum-of-squares region given by Equation (5.2) using an elliptical region similar to that given by Equation (5.1), but with a correction for the intrinsic nonlinearity. Bates and Watts (1981) suggested ways of choosing parameter transformations to reduce the parameter-effects curvature, rendering the Gaussian approximation

regions defined by Equation (5.1) more accurate. Bates and Watts (1980) indicated that the sum-of-squares regions may be considered reliable if the intrinsic curvature $K^{\perp}$ is small compared to $1/\sqrt{ps^2 F^{*}_{p,n-p}}$, the normal-based regions if the parameter-effects curvature $K^{\parallel}$ is also small compared to this quantity.

In using the approximate confidence regions defined by Equation (5.1) or (5.2) for a piecewise differentiable model, we have two main concerns of a theoretical nature. The first is to establish the asymptotic validity of these confidence regions for our non-regular case. The second is to find ways of assessing the severity of both parameter-effects and intrinsic nonlinearity for nondifferentiable models with finite samples. We will deal briefly with the first concern before proceeding to our main purpose, the development of nonlinearity measures for segmented models.

Hinkley (1969) proved the asymptotic normality of least squares estimates for the simple change-point linear regression. His methods are not quite applicable to the model defined by Equation (2.1), because the observations (temperature data) for this model are taken only at certain fixed points (integers), while Hinkley's proof assumes the observations may come arbitrarily close to the change point. For the general piecewise differentiable model with discontinuities in the derivative at fixed points, the present author (Goldberg, 1982) has shown that the least squares estimates are asymptotically normal, except when the true parameter value is at a point of discontinuity; in that case, the normal approximation yields asymptotically conservative confidence intervals.

The asymptotic normality justifies the use of sum-of-squares contours to define likelihood regions. For higher confidence levels or for more strongly skewed RSS functions, the likelihood approach should be more accurate, in the

sense of giving regions with coverage probability closer to the nominal level. Hinkley (1969) found empirically that for his model the likelihood-based regions were indeed more accurate than the normal-approximation regions.

6. Visual Indicators of Nonlinearity

We turn now to the question of how to assess the severity of nonlinearity in a particular small sample. We begin by considering some useful display techniques.

One way to study the effects of nonlinearity in the energy model is to examine the residual sum of squares RSS as a function of the "nonlinear" parameter τ. If our model function η were linear in τ, then RSS(τ) would be a quadratic function. Instead, we expect to see a more irregular function, with kinks at discontinuity points of $\frac{\partial \eta}{\partial \tau}$, that is, at each integer value of τ.

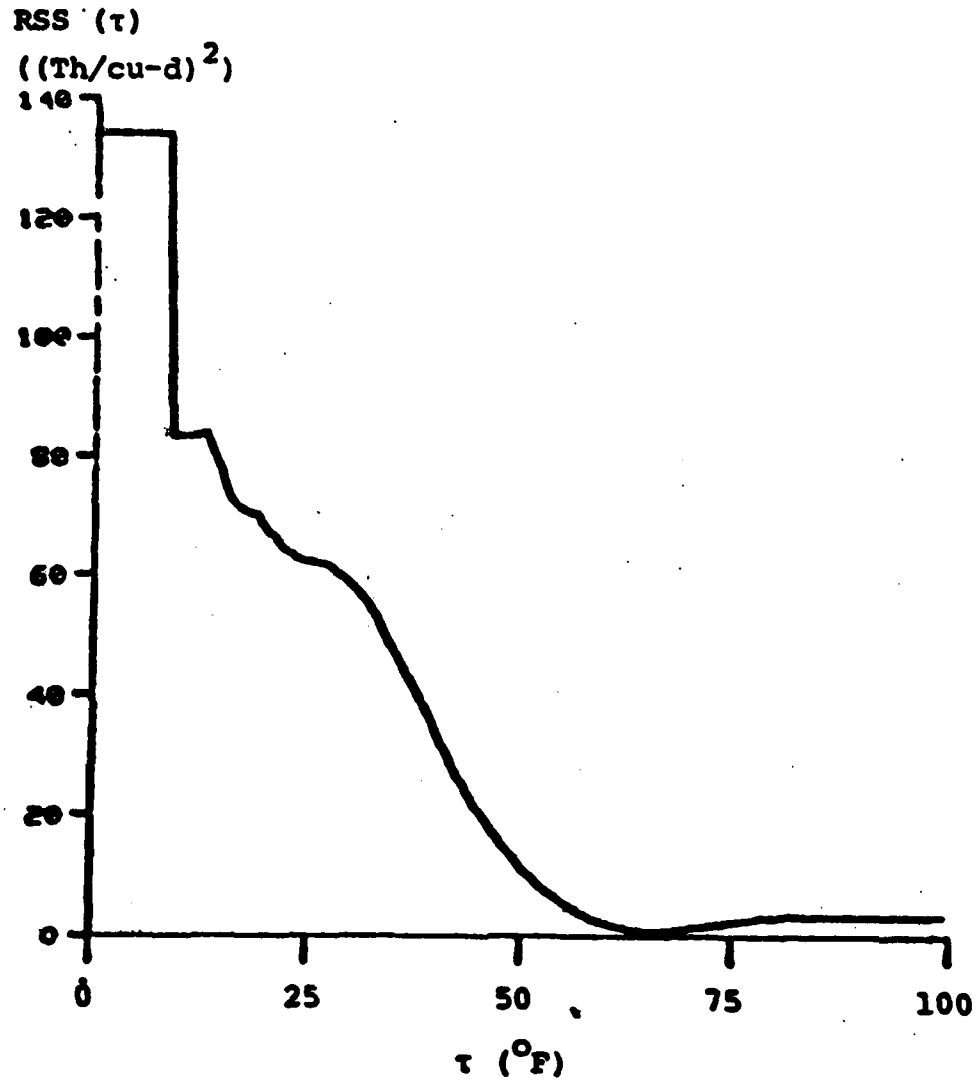
Figure 1 shows a plot of RSS versus the change point τ for a typical data set fit to Equation (2.1).* Above the maximum and below the minimum observed temperature T_{mj}, the function is flat, indicating that the reference temperature τ is not identifiable if it falls outside the range of the temperature data. In the region of low τ, where these data are very sparse, we do see the somewhat jagged behavior anticipated. For similar change point models, Hudson (1966) and Hinkley (1971) have also shown RSS curves of this general shape, but with a more pronounced scalloped appearance.

Overall, and especially in the neighborhood of the minimum (i.e., in the neighborhood of the least squares estimate $\hat{\tau}$) the RSS function looks fairly

---

* The data are for the the New Jersey Residential Gas Heating Sector, August 1969-July 1970.

Figure 1: Residual Sum of Squares RSS(τ)
Versus Reference Temperature τ.



Based on a fit of Equation (2.1) to data for the state aggregate, August 1969 - July 1970. The abbreviations are Th for therms, cu for customer, and d for day.

smooth, offering some hope that procedures which have been developed for continuously differentiable models may still be useful in the present application.

In particular, in addition to the inference procedures which are the focus of the present work, fitting procedures for smooth models can be extended to the model defined by Equation (2.1). The fitting procedure used in this study, discussed by Dutt, Fels et al (1982) and in more detail by the present author (Goldberg, 1982), is based on Newton's method. This procedure represents a modification of a method described by Hinkley (1969) for simple change-point models, and in most cases is more efficient.

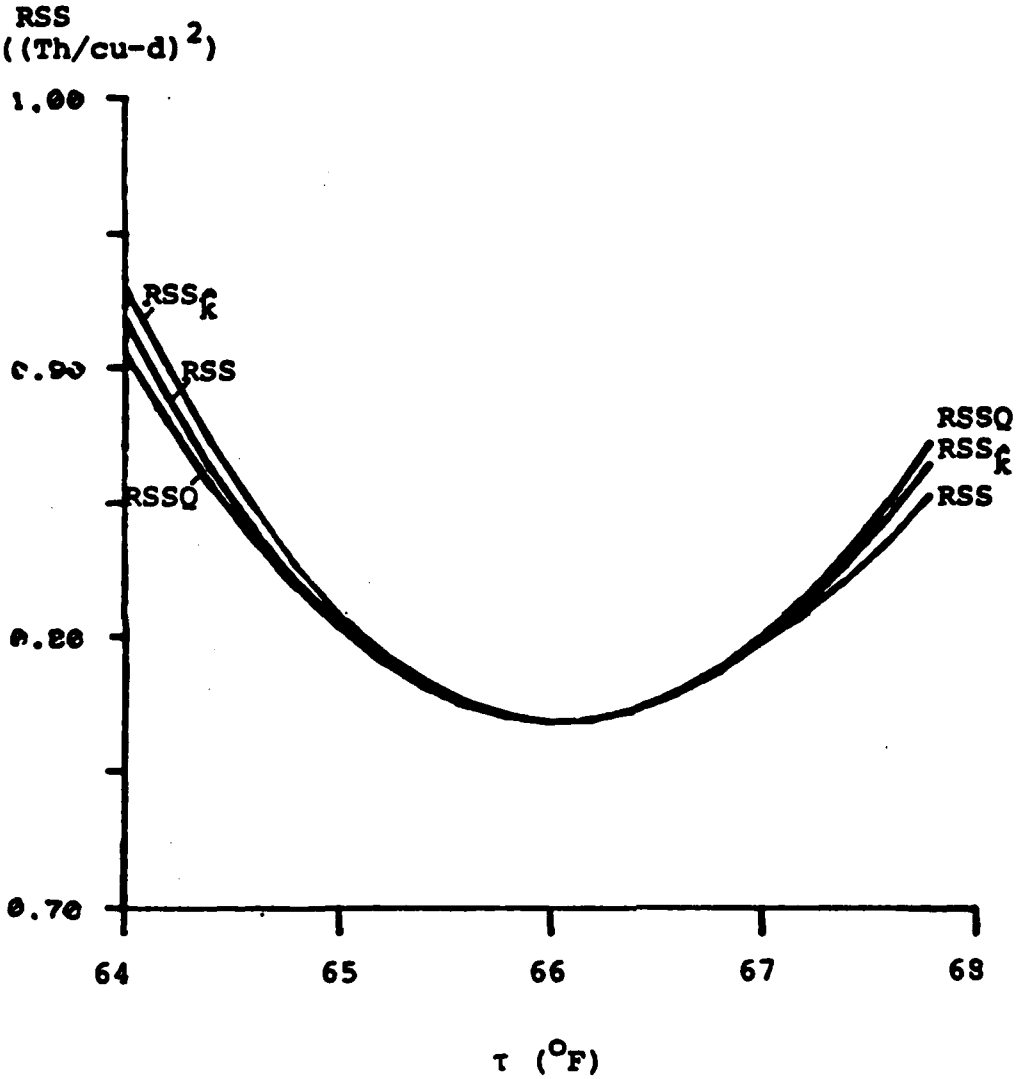
Figure 2 shows three residual-sum-of-squares curves. The first is the original curve $RSS(\tau)$. The second $RSS_{\hat{k}}$ is the residual sum of squares for an approximate model function

$$\tilde{\eta}_{\hat{k}}(\alpha,\beta,\tau) = \alpha 1 + \beta\{H(\hat{k}) + (\tau-\hat{k})\, F(\hat{k})\}.$$

The function $\tilde{\eta}_k$ extends to the whole real line the planar function which defines $\eta(\alpha,\beta,\tau)$ in the integer interval $[\hat{k},\hat{k}+1]$ containing $\hat{\tau}$. For any integer k, the approximation RSS_k coincides with the original function RSS for values of τ in $[k,k+1]$. The curve $RSS_{\hat{k}}$ shown coincides with RSS in the interval containing the minimum. The third curve shown is the quadratic approximation RSSQ based on a linearization of the model function η.

The discrepancy between the original $RSS(\tau)$ and the extension $RSS_{\hat{k}}(\tau)$ stems from the departure of the solution locus from the plane spanned by 1, $H(\hat{k})$, and $F(\hat{k})$. Thus, the divergence of RSS from $RSS_{\hat{k}}$ is an indicator of intrinsic nonlinearity. The discrepancy between $RSS_{\hat{k}}$ based on the planar extension and RSSQ based on the linear approximation to the model function reflects parameter-effects nonlinearity. Both types of nonlinearity appear from Figure 2 to be slight.

Figure 2: Residual Sum of Squares Function RSS, Extension $RSS_{\hat{k}}$, and Quadratic Approximation RSSQ



See caption to Figure 1.

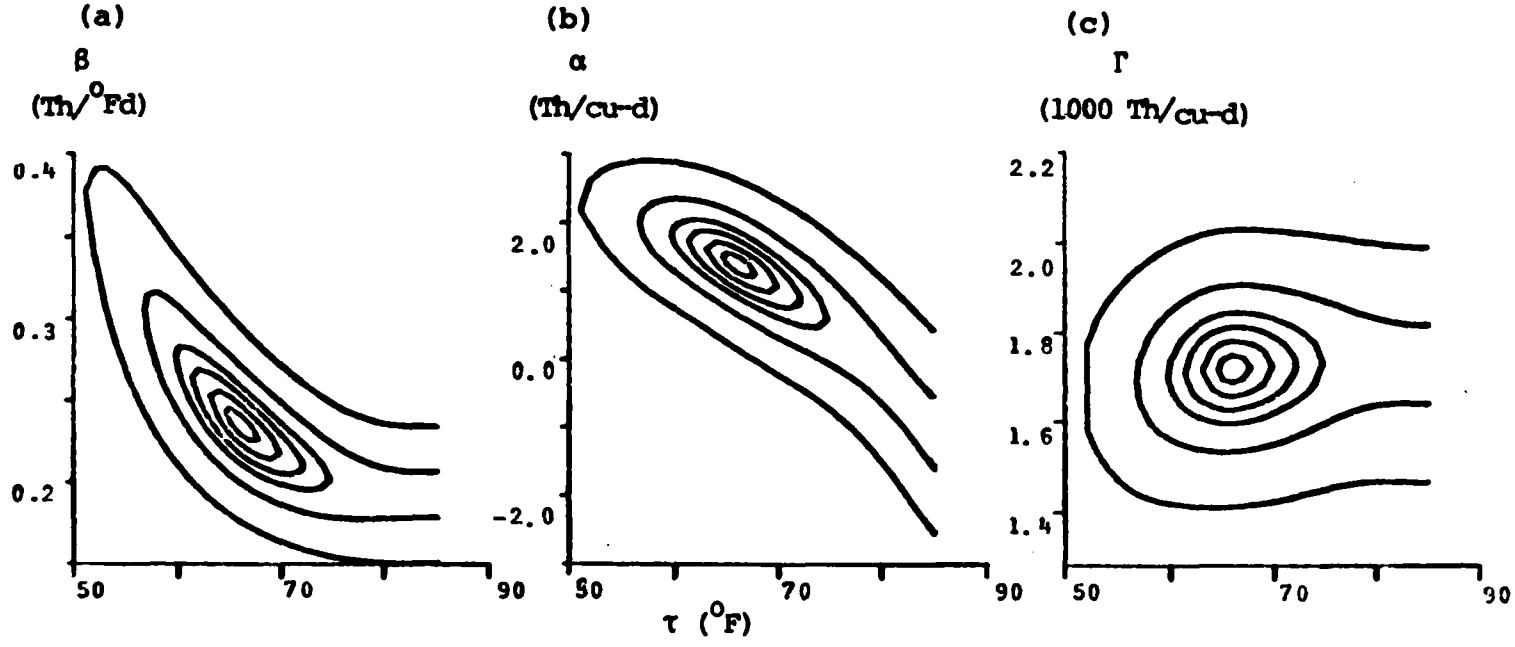
Another way to see nonlinearity is to examine two-dimensional projections of the sum-of-squares regions defined by Equation (5.3). If the boundary of such a projection is not elliptical, this is evidence of strong nonlinearity. Figure 3 shows several such regions for parameters of the energy model, for varying values of $F^{\pi}_{1,9}$, for the data set displayed in the first two figures. For small values of $F^{\pi}_{1,9}$, corresponding to confidence levels of 0.99 or less ($\pi > 0.01$) the regions shown in Figure 3 all look fairly elliptical. Only at rather high confidence levels, which are of little practical interest, do the contours become appreciably distorted from the elliptical ideal.

By itself, unfortunately, the shape of the sum-of-squares regions gives only limited information about the nature of the nonlinearity. If these regions are not elliptical, the Gaussian approximation (Equation (5.1)) is clearly inadequate to give confidence regions. At the same time, interpreting the sum-of-squares regions themselves as confidence regions (of the indicated confidence level) may or may not be valid. The reason for this ambiguity is that the distortion from the ellipse may reflect the shape of the solution locus L itself, (indicating strong intrinsic nonlinearity), or might simply result from a nonlinear mapping between L and the parameter space θ (parameter-effects nonlinearity).

Conversely, certain types of intrinsic and parameter-effects nonlinearities will still yield elliptical contours. Thus, the breakdown of either approximation (5.1) or (5.3) may not be manifest in simple examination of regions such as those drawn in Figure 3.

Somewhat more informative is a comparison of the (projected) sum-of-squares regions defined by Equation (5.3) with the elliptical regions defined by Equation (5.1), for various values of π. Here again, though, the implications of the visual comparison are ambiguous. A particular effect may

Figure 3: Sum-of-Squares Contours for Several Confidence Levels, for All Parameters of the Energy Model



From the innermost contours outward, the regions shown are determined by Equation (5.3), with $F^{\pi}_{1,9}$ = 1, 4, 9, 36, and 100. See caption to Figure 1.

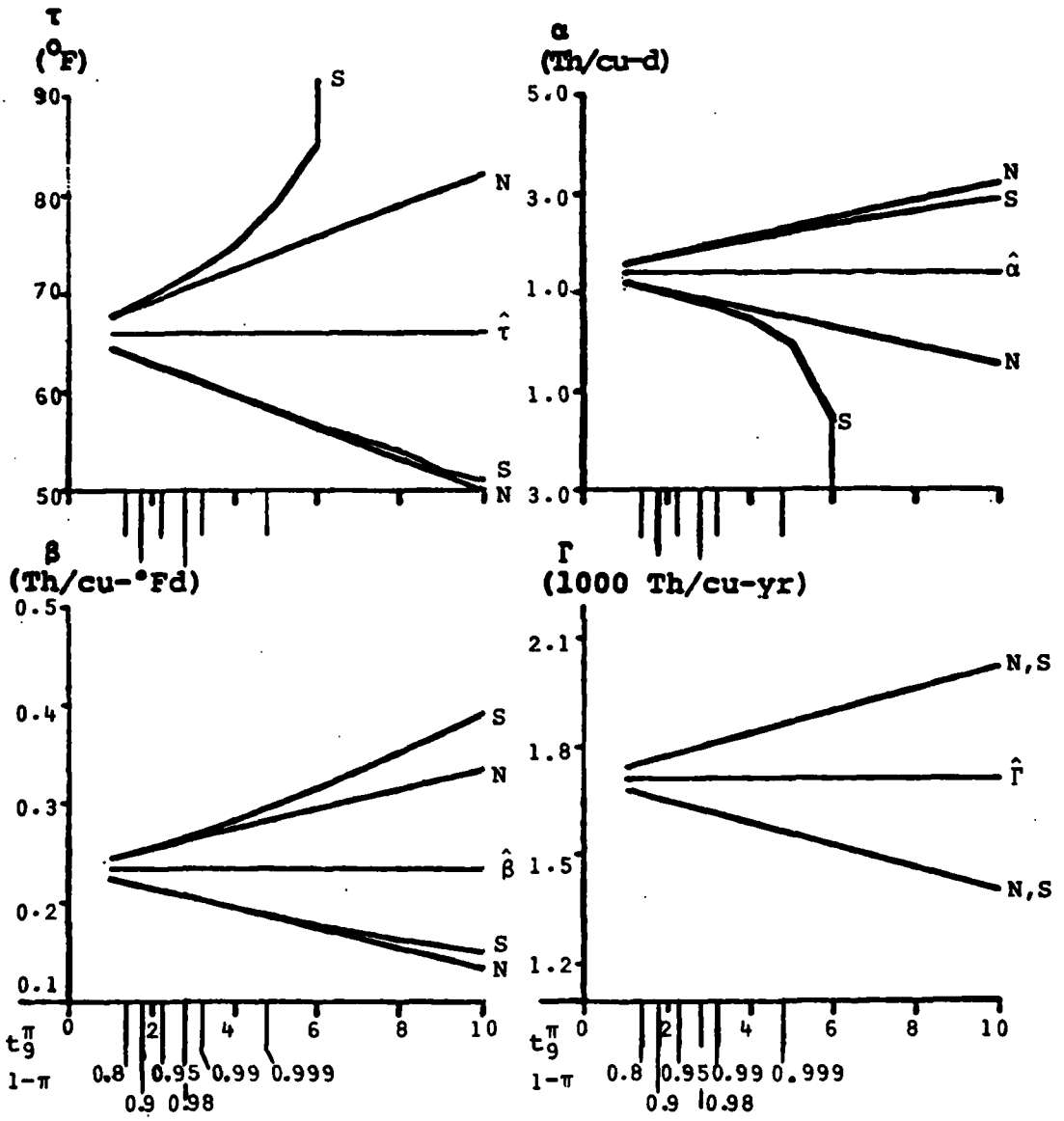
result either from the falling off of the solution locus from the tangent plane (intrinsic nonlinearity) or from the distortion of coordinate lines within the plane (parameter-effect nonlinearity).

What such a comparison does reveal is how close the normal-theory regions come to the sum-of-squares regions, which in general are more reliable. The two sets of confidence regions may be compared more conveniently by plotting for each parameter component the one-dimensional projections of the two regions onto the coordinate axis, as a function of $\sqrt{F^{\pi}_{1,n-p}}$. A set of such plots for the parameters of the energy model is shown in Figure 4, for our example data set.

Consistent with the indication from the previous figure, Figure 4 shows that for confidence levels of practical interest, say $1-\pi < 0.999$, the Gaussian-based confidence intervals (indicated by 'N' for Normal) are in good agreement with those based on the sum-of-squares methods for all three parameters α, β, and τ of the basic model. For the important index Γ, the two sets of confidence intervals are in virtually perfect agreement even for extremely high confidence levels. Thus, provided the sum-of-squares method gives accurate confidence intervals for this data set, the Gaussian approximation also appears to be trustworthy.

The visual indicators just described are unsatisfying in two major respects. First, they are only qualitative, giving no firm basis for determining whether the Gaussian or sum-of-squares regions are justified as confidence regions. Secondly, they require evaluation of sum-of-squares contours. In many cases, a justification for the normal approximation is sought precisely because evaluation of sum-of-squares contours is difficult or costly.

Figure 4: Confidence Intervals for Parameters of the Energy Model by the Gaussian Approximation (N) and the Sum-of-Squares Method (S) for Various Confidence Levels



The factor $t_9^\pi = \sqrt{F_{1,9}^\pi}$. See caption to Figure 1.

We address these difficulties in the remainder of this paper. First, in Section 7 we introduce two simple quantitative measures of intrinsic nonlinearity which are particularly suitable for nondifferentiable models. We then relate these measures to "effective curvatures" for segmented models, in Section 8, and apply the effective curvatures to the energy model in Section 9. Finally, in Section 10, we suggest an alternate approach, which yields effective parameter-effects as well as intrinsic curvatures.

7. Quantifying Nonlinearity

As noted previously, the small-sample validity of the approximate confidence intervals determined by Equation (5.3) depends on how nearly planar the solution locus L is over the region of interest. For the continuously differentiable model, the intrinsic curvature K_v^i measures the departure from the tangent plane in terms of the rate of change, normal to that plane, of a tangent vector t_v. For both segmented and smoother models, this departure can be measured in other ways, two of which are considered in this section.

A direct measure of the departure from the plane is the distance Δ between the tangent plane at $\eta(\hat{\theta})$ and a point $\eta(\theta)$ at the edge of the region of interest - that is, for θ lying on a sum-of-squares contour as defined by Equation (5.3). The gap is easily evaluated at a point $\eta(\theta)$ as the square root of the residual sum of squares from a regression of the secant $\eta(\theta) - \eta(\hat{\theta})$ on the derivative matrix $\dot{\eta}(\hat{\theta})$, which defines the tangent plane at $\eta(\hat{\theta})$. The solution locus may be considered "nearly planar", and the sum-of-squares region an adequate approximate confidence interval, if the gap Δ is small compared to the radius of the sphere defining the region. This radius, as given by Equation (5.3), is $\sqrt{(1+f)RSS(\hat{\theta})}$, where

$f = qF^{\alpha}_{q,n-p}/(n-p)$. Alternatively, following Bates and Watts (1980), we may simply compare the gap Δ with $\sqrt{fRSS(\hat{\theta})}$, the radius of the sphere's intersection with the tangent plane at $\eta(\hat{\theta})$.

For each of 75 aggregate data sets fit to Equation (2.1), the maximum gap Δ_{max} was evaluated on the "one-standard-error" contour defined by $f = 1/(n-p)$ (i.e., $q = 1$, $F^{\alpha}_{q,n-p} = 1$), for which $\sqrt{f\ RSS(\hat{\theta})} = s$. The procedure used to compute Δ_{max} is described below in Section 7. The ratio of Δ_{max} to $s = \sqrt{RSS(\hat{\theta})/(n-3)}$ ranged from 0.02 to 0.25, with a median of 0.08. Thus, along this contour, the greatest departure of the solution locus from the approximating tangent plane was typically less than 10% of the distance from a point on the contour to $\eta(\hat{\theta})$, and at worst was 25% of this distance. On this basis, the planar approxima- tion appears to be reasonable for most data sets arising for the energy model.

The gap Δ can be used as a measure of intrinsic nonlinearity for either a segmented or a smooth model. Note also that the gap indicates the total intrinsic nonlinearity in a particular direction, whether caused by a continuous or a discontinuous change in the derivative η.

A second measure, which reflects the effect of nondifferentiability alone, is the angle ϕ between the two limiting tangent planes at a point of discontinuity of $\dot{\eta}$. In the case of a piecewise planar model such as that given by Equation (2.1), ϕ is simply the angle between planar segments. For the general model with discontinuous derivative in τ, we denote by U_- and U_+, respectively, the left and right-hand limits of $\frac{\partial \eta}{\partial \tau}$ at a point of discontinuity, and by V the matrix of derivatives of η with respect to the other parameters at that point. Then the angle at the discontinuity is given by

$$\cos(\phi) = \frac{(U_- \perp V)'(U_+ \perp V)}{|U_- \perp V|\ |U_+ \perp V|}$$

where the notation $_{\perp}V$ denotes the component orthogonal to V.

For the same 75 aggregate data sets fit by Equation (2.1), the angle ϕ evaluated at the integers $\hat{k}$ and $\hat{k}+1$ bracketing the estimate $\hat{\tau}$ ranged from 0.02 to 0.13 radians, with a median of 0.06 radians. These small angles again indicate that the intrinsic nonlinearity is slight for this model. However, the impact of the bend in the solution locus depends not just on the magnitude of a single bend, but also on how many bends there are in a region of interest.

Certainly other direct measures of nonlinearity could be considered for segmented models. The appeal of the two proposed here will emerge as we proceed.

8. Effective Curvature Measures for Segmented Models

The measures described in the previous section allow us to associate numbers with nonlinearity, but still leave us with the question of what the numbers mean. How small must the gap Δ or angle ϕ be for the intrinsic nonlinearity to be considered negligible? As noted above, the angles ϕ, and the spacing between points of discontinuity $\dot{\eta}$ together indicate the severity of intrinsic nondifferentiablity. The present author (Goldberg, 1982) has related the angles and spacing to the shape of the observed likelihood function, and to the performance of fitting procedures. For purposes of inference, however, we are concerned with the total intrinsic nonlinearity. Hence, we focus now on the gap Δ, which incorporates both instantaneous and continuous changes in the derivative $\dot{\eta}$.

By considering the relation of the gap Δ to the intrinsic curvature, as defined by Equation (3.7), for smooth models, we will obtain an expression for the effective curvature of segmented models. Effective curvatures make it

possible to think of such models in the same terms as the more familiar smooth models.

In the smooth case, we can approximate the geodesic curve from $\eta(\theta)$ to $\eta(\hat{\theta})$ by a parabola centered at $\eta(\hat{\theta})$, as illustrated in Figure 5. A parabola defined by $y_2 = cy_1^2$ has curvature at $y_1 = 0$ given by

$$\frac{|a|}{|t|^2} = \frac{|[0,2c]'|}{|[1,0]'|^2}$$
$$= 2c$$
$$= 2\frac{y_2}{y_1^2} .$$

The curvature at the center can therefore be determined from any point (y_1,y_2) of the parabola.

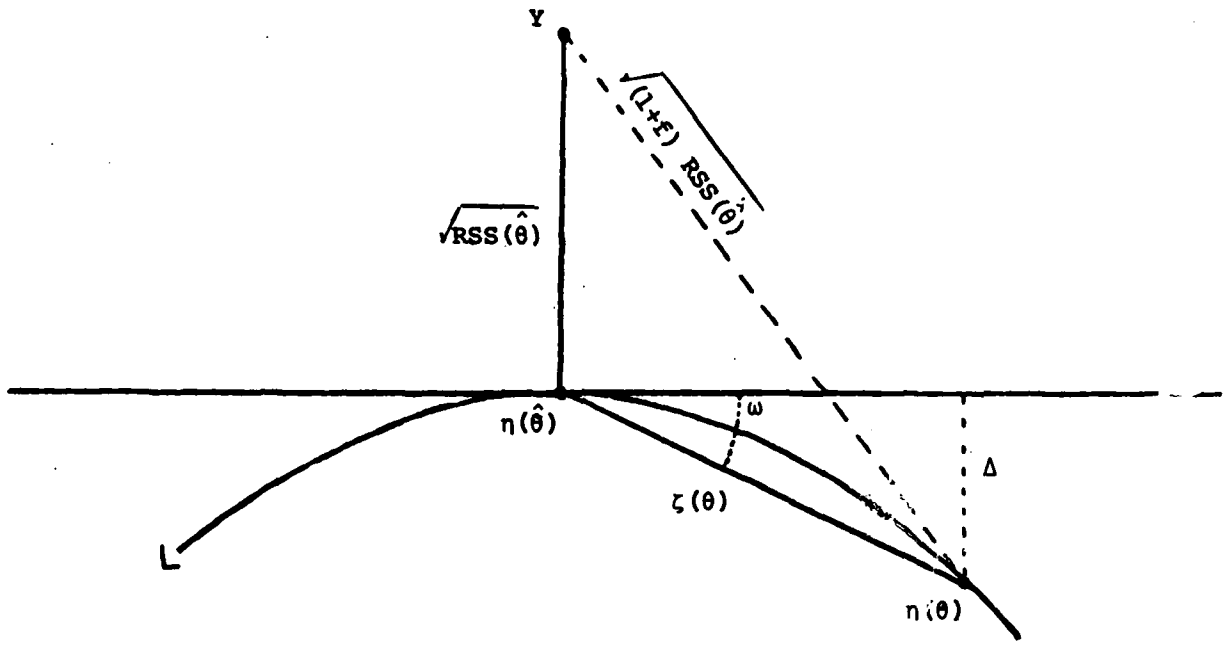
For the parabola which ideally represents a cross-section of the solution locus, y_1 and y_2 correspond respectively to the tangential and normal components of the secant $\eta(\theta) - \eta(\hat{\theta})$. Thus, denoting by P the projection matrix onto $\dot{\eta}(\hat{\theta})$, and by $\zeta(\theta)$ the secant $\eta(\theta) - \eta(\hat{\theta})$, we have

$$K^{\perp} = 2\frac{|(I-P)\zeta(\theta)|}{|P\zeta(\theta)|^2}$$
$$= \frac{2\Delta}{|P\zeta(\theta)|^2} . \qquad (8.1)$$

Hence
$$\gamma^{\perp} = \frac{2\sqrt{ps^2}\,\Delta}{|P\zeta(\theta)|^2} . \qquad (8.2)$$

To complete the connection between intrinsic curvature $K^{\perp}$ and the direct measure of intrinsic nonlinearity Δ, it is necessary to specify the direction v associated with $K^{\perp}$ as given by Equation (3.7). This direction is simply the coordinate with respect to $\dot{\eta}(\hat{\theta})$ of the projection onto the tangent plane of the secant $\zeta(\theta)$.

Figure 5: Idealized Cross Section of the Solution Locus L as a Parabola Centered at $\eta(\hat{\theta})$



Thus, for a given point $\eta(\theta)$, the squared gap Δ^2 at $\eta(\theta)$ is the residual sum of squares from a regression of $\zeta(\theta)$ on $\dot{\eta}(\hat{\theta})$, while the direction v is given by the coefficients of this regression. Further, the multiple correlation for the regression is the cosine of the angle ω between the secant $\zeta(\theta)$ and the tangent plane defined by $\dot{\eta}(\hat{\theta})$. For a piecewise planar model, such as the energy model, the angle ϕ defined above can be related to this secant angle ω. Specifically, whenever $\eta(\theta)$ and $\eta(\hat{\theta})$ lie on adjacent planar segments, the angle ϕ represents an upper bound on the secant angle ω in the direction v.

It is important to note that the direction v, which indicates the line in the tangent plane pointing toward $\eta(\theta)$, will not in general coincide with $\theta - \hat{\theta}$. The reason is that the sample-space image of the parameter-space segment $\overline{\theta\hat{\theta}}$ is in general a curve, not a straight line. Thus, the vector $t_{\theta-\hat{\theta}} = \dot{\eta}(\theta-\hat{\theta})$, which is the tangent at $\hat{\theta}$ to the curved image of $\overline{\theta\hat{\theta}}$, does not point toward $\eta(\theta)$.

A more precise relationship between v and $\theta - \hat{\theta}$ is determined by expanding $\eta(\theta)$ about $\hat{\theta}$. We have

$$\zeta(\theta) = \dot{\eta}(\theta-\hat{\theta}) + (\tfrac{1}{2})[(\theta-\hat{\theta})'\ddot{\eta}_m(\theta-\hat{\theta})] \quad ,$$

so that

$$v = \theta-\hat{\theta} + (\tfrac{1}{2})(\dot{\eta}'\eta)^{-1}\dot{\eta}'a^{\perp}_{\theta-\hat{\theta}} \quad , \tag{8.3}$$

with $a^{\perp}_{\theta-\hat{\theta}}$ the tangential component of the acceleration as defined by Equation (3.6). The difference between v and $\theta-\hat{\theta}$ is thus closely related to the parameter-effects curvatures. Recall that the gap Δ and the secant angle ω themselves measure intrinsic nonlinearity only.

For a smooth model, Equation (8.1) or (8.2) can be regarded as an approximation to the actual curvature. For a segmented model, we will take these equations as the definitions of effective curvatures K and γ. In the

latter case, the effective curvatures so obtained will depend strongly on the size of the regions of interest and on the proximity of $\hat{\theta}$ and θ to a discontinuity point of $\dot{\eta}$. One may question the value of a curvature measure which is so sensitive to the points chosen for its evaluation. In fact, however, such a dependence is entirely appropriate for nondifferentiable models.

Nondifferentiability means that a single number indicating a local rate of change (i.e., a derivative or curvature) does not adequately represent more global behavior. Describing nonlinearity in terms of curvature amounts to approximating the solution locus L by a spherical or parabolic surface, which coincides with L at the point of the fit. For a smooth model, the same approximation is valid over a wide range, essentially until the second-order expansion of the model function η breaks down. For the segmented model, on the other hand, a different smooth approximation is relevant depending on the width of the region of interest. For inferences in a close neighborhood of a discontinuity point θ_o, it is wise to consider a surface of small radius of curvature, which approximates L well in that neighborhood. For inferences over a wider region, a sphere of larger radius, which might be relatively far from L in the immediate vicinity of θ_o, would be more appropriate.

To apply Beale's formula (Equation (5.4)), the root mean square intrinsic relative curvature $\gamma^{\perp}_{rms}$ is required, while to use the methods of Bates and Watts (1981), Hamilton, Watts and Bates (1982), or Box (1971) requires the entire acceleration array $[\ddot{\eta}_m]$. Thus, the approximation given by Equation (8.2), which provides estimates of the relative intrinsic curvature in a particular direction, still leaves much work to be done if the procedures which make the concept of curvature so appealing in general applications are to be used.

In many cases, however, nonlinearities are slight, so that the corrections offered by these procedures are negligible. In particular, the experience of Bates and Watts (1980) indicates that the relative intrinsic nonlinearity of most models is quite small. Thus, a quick method of establishing that the maximum relative intrinsic curvature is sufficiently small could frequently obviate the need for more complicated computations. This is the approach taken below in applying effective curvature measures to the energy model.

9. Effective Curvature of the Energy Model

Above, we have seen several indications that the nonlinearities in the energy model are slight: the small discrepancies among the RSS functions in Figure 2, the close correspondence between the Gaussian and sum-of-squares confidence intervals in Figure 4, the mild tangent angles ϕ, and the small ratios Δ_{max}/s of the gap to the radius of a sum-of-squares region. Hence, to determine that the Gaussian approximation is adequate, it should be sufficient in most cases to verify that the maximum curvatures are small. In this section, we consider only the intrinsic curvature $\gamma^{\perp}$.

Appendix A describes how the maximum effective intrinsic curvature $\gamma^{\perp}_{max}$ can be found for the energy model, on a sum-of-squares contour defined by Equation (5.3). Using the approximation given by Equation (A.1) for the case $F^{W}_{q,n-p} = 1$, Equation (8.2) yields

$$\gamma^{\perp}_{max} \approx \frac{2\sqrt{p}\,\Delta_{max}}{s} \quad . \tag{9.1}$$

The results in Section 7 on the ratio Δ_{max}/s of the maximum gap to the radius of a one-standard-error sum-of-squares region can now be translated into statements about effective curvatures for the energy model. For the 75

data sets, the maximum (effective) relative intrinsic curvature $\gamma^{\perp}_{max}$ ranges from 0.07 to 0.85, with a median of 0.29.

When the second-derivative array $[\ddot{\eta}]$ has only one non-zero vector, on the diagonal, it is possible to show that the root mean square curvature γ_{rms} and the maximum curvature γ_{max} are related by

$$\gamma_{rms} = \sqrt{\frac{3}{p(p+2)}}\,\gamma_{max} \quad . \tag{9.2}$$

For the energy model, with η given by (3.4), the normal component of $[\ddot{\eta}]$ has a single non-zero vector, the orthogonal component of $\partial^2\eta/\partial\tau^2 = \beta\partial F/\partial\tau$. It is therefore possible to evaluate Beale's inflation factor μ, given by (5.4), knowing the effective intrinsic curvature $\gamma^{\perp}_{max}$ only in the direction of maximum curvature.

For the worst case then ($\gamma^{\perp}_{max} = 0.85$), Equation (9.2) yields an inflation factor $\mu = 1.08$, while for the median value ($\gamma^{\perp}_{max} = 0.29$) we get $\mu = 1.01$. Thus, if Beale's formula holds approximately in the non-differentiable case, with the effective relative intrinsic curvature defined by (8.2), then the correction required to make the sum-of-squares regions (5.3) conservative is minimal in most cases. In fact, the factor μ is greater than 1.03 for only two of the 75 cases studied.

10. Smoothing the Model Function

The effective curvatures defined for segmented models by Equations (8.1) and (8.2) are based implicitly on an approximation to the solution locus L by some smooth surface. Another approach is to approximate the model function η explicitly over the region of interest by a smooth function $\tilde{\eta}$, then consider the curvature of $\tilde{\eta}$. Obviously, this procedure provides both parameter-effect and intrinsic curvatures.

The approximation method must be left as an ad hoc procedure to be chosen for the particular model studied, and in general will involve considerably more computation than the measures suggested above. On the other hand, the curvature of a close smooth approximation is arguably the best definition of curvature for a segmented model. Furthermore, if the same approximation $\tilde{\eta}$ applies over a wide range of values of θ, then the curvature array may need to be evaluated only once for all confidence levels of interest.

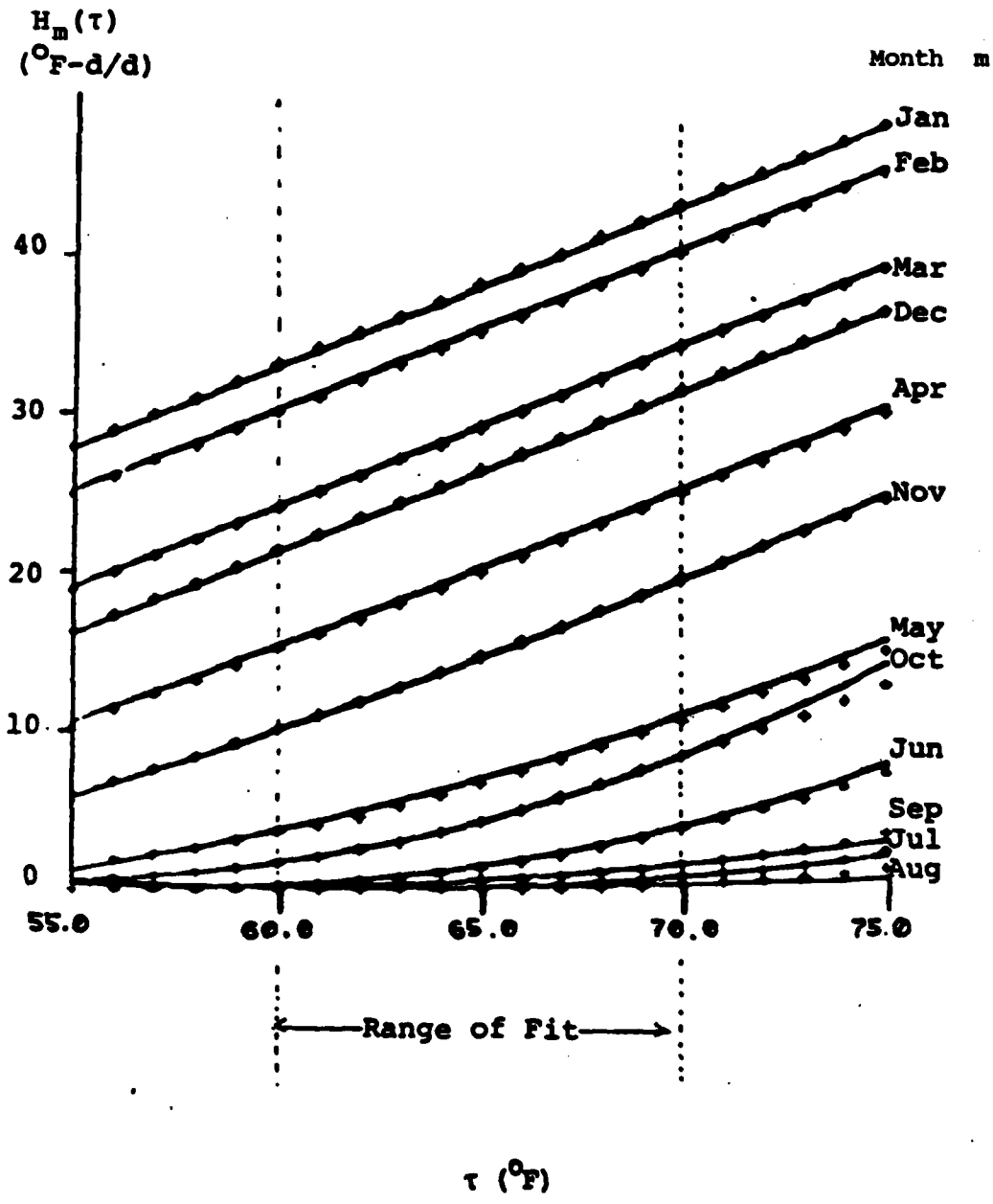
In the case of the energy model, a very close approximation to the model function was obtained for each data set by smoothing the nondifferentiable degree-day variable $H(\tau)$. For each month m, a smooth function $\tilde{H}_m(\tau)$ was obtained by fitting a quadratic function

$$H_m(\tau_i) = H_m(\hat{\tau}) + b_m(\tau_i - \hat{\tau}) + c_m(\tau_i - \hat{\tau})^2 + e_m \quad . \qquad (10.1)$$

The coefficients b_m and c_m in Equation (10.1) were found by the method of least squares, using values $\tau_i = \hat{\tau} + k$, $k = -5,-4,\ldots,4,5$.

Figure 6 shows the actual and smoothed degree-days for each month m, from August 1969 to July 1970, obtained by this procedure with $\hat{\tau}$ set equal to 65°F. The figure shows that the approximation $\tilde{H}_m$ is quite close to the actual H_m, not only over the range of the fit, but also considerably beyond. Table 1 shows the results of the regressions for the twelve months. The R^2 values are quite high in all cases. In addition, the coefficient b_m is generally very close to the derivative $F_m(\hat{\tau})$, so that the derivative $\dot{\tilde{H}}_m(\hat{\tau})$ is also close to the derivative of the original function. In all data sets studied, the least squares estimates $\tilde{\theta}$ found by using the approximation $\tilde{\eta}(\alpha,\beta,\tau) = \alpha 1 + \beta\tilde{H}(\tau)$ were also quite close to the original estimates $\hat{\theta}$ corresponding to the true model; the differences $\tilde{\theta}_j - \hat{\theta}_j$ were found to be on the order of 10% of the standard errors of $\hat{\theta}_j$.

Figure 6: Heating Degree-Days $\underset{\sim}{H}$ and Smooth Approximation $\underset{\sim}{H}$ Versus Reference Temperature τ.



Actual degree-days $\underset{\sim}{H}_m$ are indicated by '+', the approximations $\underset{\sim}{H}_m$ by the continuous curves. The fits are for aggregate degree-days, defined by Equation (2.3), August 1969 - July 1970.

Table 1: Coefficients from Fitting Equation (10.1).

Month (m)	H(65)	F(65)	b_m	c_m	R^2_m
August	0.00	0.000	0.016	0.0040	0.89479
September	0.46	0.148	0.138	0.0111	0.99863
October	4.18	0.760	0.677	0.0311	0.99944
November	14.47	0.950	0.941	0.0083	0.99998
December	26.15	1.000	0.999	0.0002	1.00000
January	37.83	1.000	1.000	0.0000	1.00000
February	35.20	1.000	1.000	0.0000	1.00000
March	29.11	1.000	1.000	0.0000	1.00000
April	20.14	0.990	0.982	0.0031	1.00000
May	6.91	0.727	0.721	0.0154	0.99970
June	1.39	0.386	0.371	0.0274	0.99984
July	0.06	0.052	0.069	0.0130	0.99116

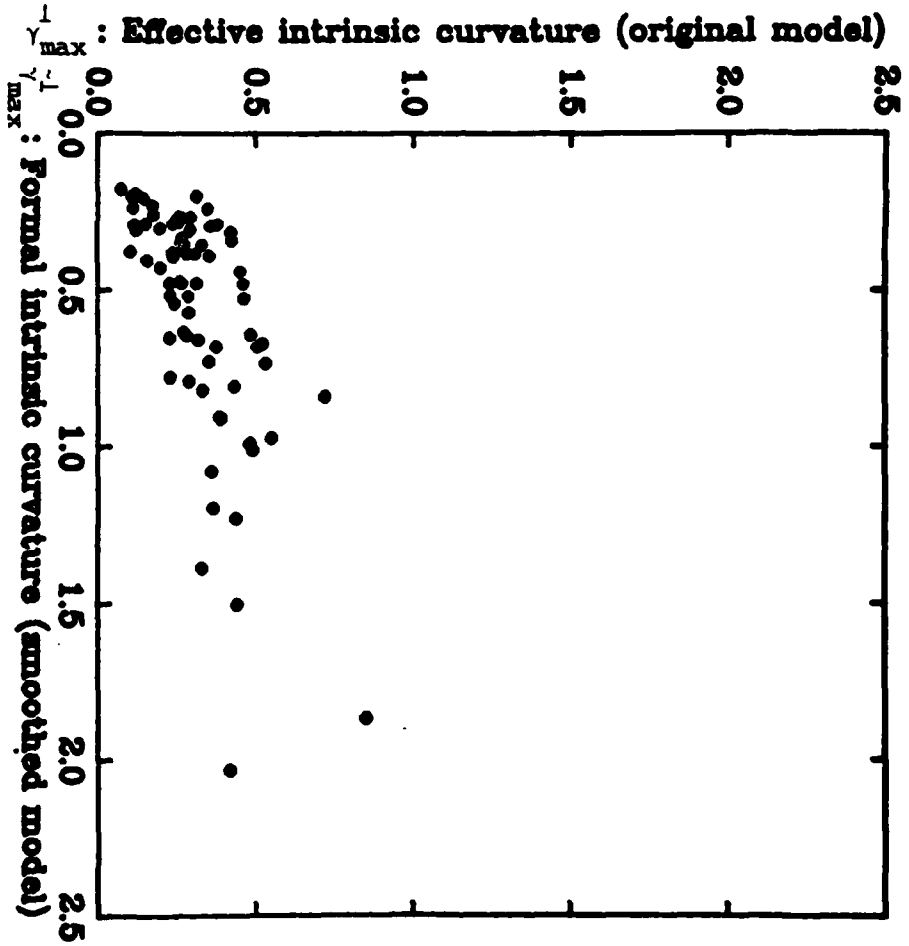
See caption to Figure 6.

Figure 7 shows a plot of the maximum effective relative intrinsic curvatures $\gamma^{\perp}_{max}$, computed from the gap Δ using Equation (9.1), versus the relative intrinsic curvatures $\tilde{\gamma}^{\perp}_{max}$ for the smoothed model function $\tilde{\eta}$. The figure shows a weak, positive relation between the two curvature measures over the 75 data sets. However, the formal curvature $\tilde{\gamma}^{\perp}_{max}$ for the smoothed model is almost always larger than the effective curvature $\gamma^{\perp}_{max}$ derived from the gap Δ_{max}. Evidently, then, for θ at a distance of one standard error from $\hat{\theta}$, (where the gap Δ was evaluated) the original function η tends to be closer to the tangent plane at $\eta(\hat{\theta})$ than is the approximation $\tilde{\eta}$.

The disparity between the two measures $\gamma^{\perp}_{max}$ and $\tilde{\gamma}^{\perp}_{max}$ does seem to depend on the standard error of $\hat{\tau}$, which determines how many bends in the solution locus occur between $\hat{\theta}$ and the point θ where the gap Δ_{max} was evaluated. In general, the larger disparities are associated with larger standard errors (around $3^{o}F$), while for the data sets for which $\gamma^{\perp}_{max}$ and $\tilde{\gamma}^{\perp}_{max}$ are roughly equal the standard error of $\hat{\tau}$ is relatively small (less than or equal to $1^{o}F$).

The curvatures $\tilde{\gamma}^{\perp}_{max}$ computed for the smooth model $\tilde{\eta}$ not only tend to be larger than the effective curvatures $\gamma^{\perp}_{max}$ based on the gap, but are also more spread out. The gap-based effective curvature $\gamma^{\perp}_{max}$ is derived from a single point $\eta(\theta)$, where the value of τ corresponding to that point is anywhere from one to three (or in one case seven) degrees from $\hat{\tau}$, and each degree represents a point of discontinuity of $\dot{\eta}$. It is therefore somewhat surprising that smoothing η over ten integer values of τ yields a measure $\tilde{\gamma}^{\perp}_{max}$ which is more erratically behaved than that based on the gap. Whatever the reason for this behavior, the relatively large values of $\tilde{\gamma}^{\perp}_{max}$ found for a few data sets serve as a warning that at some confidence levels the impact of intrinsic nonlinearity may be greater than is indicated by the gap evaluated on a one-standard-error sum-of-squares contour.

Figure 7: Effective Relative Intrinsic Curvature $\gamma^{\perp}$ Versus Relative Intrinsic Curvature $\tilde{\gamma}^{\perp}$ of the Smoothed Energy Model



Based on fits of 75 aggregate data sets to Equation (2.1).

As discussed above, it may not make sense to try to describe the shape of a piecewise differentiable surface in terms of the curvature of a single quadratic approximation. Even though the approximation $\tilde{\eta}$ is quite close to the original function η, the correspondence between the two functions must vary with distance, as well as direction, from $\eta(\hat{\theta})$. Whether the approximation $\tilde{\eta}$, and its curvatures, can be considered adequate to describe the behavior of η depends, of course, on the degree of precision required. The ratio $\tilde{\gamma}^{\perp}_{max}/\gamma^{\perp}_{max}$ of the two measures is less than two for most of the data sets, and is greater than four for only two.

The approximation $\tilde{\eta}$ also offers a measure of parameter-effects curvature $\tilde{\gamma}^{\parallel}$. Unfortunately, there is not simple way to determine the direction v in which $\tilde{\gamma}^{\parallel}_{v}$ is maximized. However, as explained in Appendix A, a good indication of the strength of parameter effects is given by $\tilde{\gamma}^{\parallel}_{v}$ for $v = [-\beta\bar{F},0,1]'$, corresponding to $t_v = \beta F \perp 1$.

For the 75 data sets studied here, $\tilde{\gamma}^{\parallel}_{v}$ for this direction ranged from 0.1 to 1.4, with a median of 0.2. The small median value indicates that parameter-effects nonlinearities are slight in most cases. For the data set used as an example throughout this paper, $\tilde{\gamma}^{\parallel}_{v} = 0.46$, which is the 85^{th} percentile of the 75 observed values. Thus, for most data sets, the parameter-effects nonlinearities appear to be smaller than was seen for the example data set. As a result, the Gaussian approximation may typically be expected to perform as well or better than is indicated in Figure 4 over that range of confidence levels.

11. Conclusion

We have presented several methods for examining nonlinearity in awkward models. Although the primary focus has been a segmented model, the visual indicators and the curvature measures proposed may also be used for

continuously differentiable models. The visual indicators can reveal a great deal about the behavior of the model function, but computation of the required quantities may be quite cumbersome. By constrast, the quantitative measures proposed may be easier to compute than formal curvatures based on second derivatives. In addition, in cases where the second-order approximation does not hold over an entire region of interest, an effective curvature based on points at the edge as well as the interior of that region may be more meaningful than the formal curvature evaluated at the point of the fit.

For the energy model which motivated this study, all the measures explored indicate that the nonlinearities are generally small for data sets like those examined in this work. The intrinsic nonlinearity, as measured by the tangent angle ϕ and by approximate curvatures, is small enough that the sum-of-squares method gives good approximate confidence regions. A combination of direct comparisons of Gaussian and sum-of-squares regions (for a particular data set) and examination of parameter-effect curvatures $\tilde{\gamma}^{[}$ (for a large number of sets) leads to the conclusion that the more easily computed Gaussian approximation should be acceptable in most cases.

Further study is needed to assess the performance of the effective curvature measures proposed here, for a variety of segmented and smooth models. In this context, both the validity of the approximations and the degree to which these methods actually facilitate computations are important. Also useful would be efficient means of finding the maximum intrinsic or parameter-effects curvature, on the basis of which more detailed computations might be forgone. A paper currently in preparation describes procedures for obtaining mean square effective curvatures, both intrinsic and parameter-effect, based on methods developed here, with an emphasis on applications to smooth models.

Acknowledgement

The author thanks Peter Bloomfield and Margaret Fels, who supervised the dissertation on which this work is based, for their help and encouragement.

Appendix A. Finding Maximum Curvatures for the Energy Model

For the energy model defined by (2.1), the maximum intrinsic curvature at $\hat{\theta} = [\hat{\alpha},\hat{\beta},\hat{\tau}]'$ is in the direction of $F(\hat{\tau}) \perp [1,H(\hat{\tau})]$, the component of $F(\hat{\tau})$ orthogonal to the vectors 1 and $H(\hat{\tau})$. The formal curvature $\tilde{\gamma}^{\perp}_{max}$ was therefore evaluated in the direction of $F(\tilde{\tau}) \perp [1,H(\tilde{\tau})]$, using the estimate $\tilde{\tau}$ and the second derivative $\ddot{\tilde{\eta}}$ for the smoothed model function $\tilde{\eta}$. Finding the effective curvature $\gamma^{\perp}_{max}$ in the corresponding direction for the original model is more complicated; as noted in the text, evaluating the gap Δ at $\eta(\theta+v)$ does not in general yield the effective curvature $K^{\perp}_{v}$ in the desired direction v. Rather than searching for points on the sum-of-squares contour in the indicated direction, a more ad hoc procedure was used to find Δ_{max}.

To find the maximum gap Δ around the sum-of-squares contour, it is necessary to maximize the residual sum-of-squares from the regression of $\eta(\theta) - \eta(\hat{\theta})$ on $\dot{\eta}(\hat{\theta})$. For the energy model, this is a regression of $(\alpha-\hat{\alpha})1 + \beta H(\tau)-\hat{\beta}H(\hat{\tau})$ on $[1,H(\hat{\tau}),\hat{\beta}F(\hat{\tau})]$. The terms $(\alpha-\hat{\alpha})1$ and $\hat{\beta}H(\hat{\tau})$ leave no residual, while β is a scalar. Hence, maximizing the residual from a regression of $H(\tau)$ on $[1,H(\hat{\tau}), \beta F(\hat{\tau})]$, then multiplying by β^{+}, the maximum of β along the contour, yields an overestimate of the maximum gap Δ_{max}. The maximal divergence of $H(\tau)$ from the $H(\hat{\tau}) - \beta F(\hat{\tau})$ plane occurs at values of τ farthest from $\hat{\tau}$. Thus, the maximal gap was found by finding the extreme values τ^{-} and τ^{+} along the contour, obtaining the residuals from regressions of $H(\tau^{-})$ and $H(\tau^{+})$ on $[1,H(\hat{\tau}), \beta F(\hat{\tau})]$, then multiplying the larger magnitude residual by β^{+}.

Having obtained (overestimates of) the maximum gap, we still need the tangential secant component $|P\zeta(\theta)|$ to derive effective curvatures from Equation (8.1) or (8.2). According to the linear approximation,

$|P\zeta(\theta)| = \sqrt{f\ RSS(\hat{\theta})}$ for θ on a contour defined by (5.3), with $f = F^{\alpha}_{q,n-p}/(n-p)$. Making this approximation, Equation (8.2) becomes

$$\gamma^{\perp} \doteq 2\sqrt{\frac{p}{F^{\alpha}_{q,n-p}}}\, s^2 \Delta \quad . \tag{A.1}$$

The maximum parameter-effects curvature is also of interest. For the parameter effects, only the formal curvature $\tilde{\gamma}^{\parallel}$ of the smoothed model is available. As noted in the text, there is no simple way to determine the direction v in which $\tilde{\gamma}^{\parallel}_{v}$ is maximized. We do know that the tangent vector t_v must be orthogonal to $\partial\eta/\partial\alpha = 1$, since the model function has zero curvature in the α direction. In addition, it is clear from Equation (4.2) that the τ-component of the maximizing v (hence the βF-component of t_v) must be non-zero. A reasonable measure of the strength of parameter effects is therefore offered by $\tilde{\gamma}^{\parallel}_{v}$ for $v = [-\beta\bar{F},0,1]'$, corresponding to $t_v = \beta F \perp 1$.

REFERENCES

1. Bates, D. M. and Watts, D. G. (1980). "Relative Curvature Measures of Nonlinearity." J.R.S.S. B 42(1):1-25.

2. Bates, D. M. and Watts, D. G. (1981). "Parameter Transformations for Improved Approximate Confidence Regions in Nonlinear Least Squares". Annals of Statistics 9:1152-1167.

3. Beale, E. M. L. (1960). "Confidence Regions in Nonlinear Estimation" (with discussion). J.R.S.S. B 22:1-88.

4. Box, M. J. (1971). "Bias in Nonlinear Estimation" (with discussion). J.R.S.S. B 32:171-201.

5. Dutt, G. S., Fels, M. F., Goldberg, M. L., and Stram, D. (1982). "The Scorekeeping Model for Residential Energy Consumption: Procedures and Problems". Princeton University Center for Energy and Environmental Studies. (Presented at ACEEE Summer Study "What Works? Documenting the Results of Energy Conservation in Buildings". Santa Cruz, CA, August 21-28, 1982.)

6. Dutt, G. S., Lavine, M., Levi, B., and Socolow, R. (1982). "The Modular Retrofit Experiment: Exploring the House Doctor Concept." Princeton University Center for Energy and Environmental Studies Report No. 130.

7. Fels, M. F. and Goldberg, M. L. (1982). "Measuring Household Fuel Consumption on the Standard Living Cycle." Energy 7 :489-504.

8. Fels, M. F., Goldberg, M. L., Lavine, M. L., Socolow, R. H., and Abrams, P. S. (1981). "Exploratory Analysis of Oil-Heated Homes." Princeton University Center for Energy and Environmental Studies (report in draft).

9. Fisher, R. A. (1925). "Theory of Statistical Estimation." Proc. Camb. Phil. Soc. 22:700-732.

10. Goldberg, M. L. (1980). "Residential Energy Consumption: Use of Aggregate Data to Measure Physical Properties of Housing Stock." Princeton University Center for Energy and Environmental Studies Report No. 105. (As Masters Thesis for Department of Mechanical and Aerospace Engineering, No. 1476-T).

11. Goldberg, M. L (1982). A Geometrical Approach to Nondifferentiable Regression Models as Related to Methods for Assessing Residential Energy Conservation. Princeton University Center for Energy and Environmental Studies Report No. 142. (As Ph.D. Thesis for Department of Statistics.)

12. Goldberg, M. L. and Fels, M. F. (1982). "Post-embargo Conservation: A New Jersey Case Study." Princeton University Center for Energy and Environmental Studies Report No. 143.

13. Hamilton, D. C., Watts, D. G., and Bates, D. M. (1982). "Accounting for Intrinsic Nonlinearity in Nonlinear Regression Parameter Inference Regions." Annals of Statistics 10:386-393.

14. Hinkley, D. V. (1969). "Inference about the Intersection in Two-Phase Regression." Biometrika 56 (3):495-504.

15. Hinkley, D. V. (1971). "Inference in Two-Phase Regression." J.A.S.A. 66:736-743.

16. Hudson, D. J. (1960). "Fitting Segmented Curves Whose Join Points Have to be Estimated." J.A.S.A. 61:1097-1129.

17. National Oceanic and Atmospheric Administration, U.S. Dept. of Commerce (Monthly Publication). "Local Climatalogical Data for Newark, New Jersey."

18. Socolow, R. H. (ed.), (1978). Saving Energy in the Home. Ballinger Pub. Co., Cambridge.

19. Wilks, S. S. (1938). "The Large-sample Distribution of the Likelihood Ratio for Testing Hypotheses." Annals of Mathematical Statistics 9:501-513.

MLG/jvs

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 2549	2. GOVT ACCESSION NO.: AD-A132 838
3. RECIPIENT'S CATALOG NUMBER:	
4. TITLE (and Subtitle): Measures of Nonlinearity for Segmented Regression Models	5. TYPE OF REPORT & PERIOD COVERED: Summary Report - no specific reporting period
	6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): Miriam L. Goldberg	8. CONTRACT OR GRANT NUMBER(s): MCS-7927062, Mod. 2 DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: Work Unit Number 4 - Statistics & Probability
11. CONTROLLING OFFICE NAME AND ADDRESS: See Item 18 below	12. REPORT DATE: August 1983
	13. NUMBER OF PAGES: 40
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

National Science Foundation
Washington, DC 20550

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Nonlinearity, Regression, Confidence Intervals, Least Squares, Nondifferentiable Models

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A simple physical model of residential energy consumption provides the framework for an exploration of segmented regression models fit by least squares. The energy model is a generalization of a linear, single change-point model such as that considered by Hinkley (1971).

Some simple geometric measures of nonlinearity and nondifferentiability are proposed. These measures are related to the construction of approximate confidence regions for the parameters of a general segmented model. In addition, the relation shown between these measures and those proposed by Bates and Watts (1980) may be useful in analyzing continuously differentiable models.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

END

FILMED

10-83

DTIC