Local Likelihood Estimation

ROBERT TIBSHIRANI and TREVOR HASTIE*

A *scatterplot smoother* is applied to data of the form $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ and uses *local fitting* to estimate the dependence of Y on X. A simple example is the *running lines smoother*, which fits a least squares line to the y values falling in a window around each x value. The value of the estimated function at x is given by the value of the least squares line at x. A smoother generalizes the least squares line, which assumes that the dependence of Y on X is linear.

In this article, we extend the idea of local fitting to likelihood-based regression models. One such application is to the class of *generalized linear models* (Nelder and Wedderburn 1972). We enlarge this class by replacing the covariate form $\beta_0 + x\beta_1$ with an unspecified smooth function $s(x)$. This function is estimated from the data by a technique we call *local likelihood estimation*. The method consists of maximum likelihood estimation for β_0 and β_1, applied in a window around each x value. Multiple covariates are incorporated through an iterative algorithm.

We also apply the local likelihood technique to the proportional hazards model of Cox (1972), a model for censored data. The proportional hazards assumption $\lambda(t \mid x) = \lambda_0(t)\exp(x\beta)$ is replaced by $\lambda(t \mid x) = \lambda_0(t)\exp(s(x))$, and the function $s(x)$ is estimated from the data by the local likelihood method.

In some real data examples, the local likelihood technique proves to be effective in uncovering nonlinear dependencies. It is useful as a descriptive tool or to suggest transformations of the covariates. We also discuss some methods for inference.

KEY WORDS: Smoothing; Generalized linear models; Nonparametric regression.

1. INTRODUCTION

Figure 1 plots 100 data pairs along with the least squares line summarizing the relationship of a response (Y) and a covariate (X). Also shown in Figure 1 is a scatterplot smooth. This was computed by a type of local fitting—around each x value a window of 20 points was formed and a least squares line was fit to the points in the window. The value of the smooth at x is given by the value of the local line at x. As we can see, the smooth captures the trend of the data better than the least squares line. The reason is simple—the smooth does not make as rigid an assumption as the straight line about the form of the relationship between Y and X.

In recent years, there has been a great deal of interest in scatterplot smoothing by local fitting (see, e.g., Cleveland 1979; Friedman and Stuetzle 1981), and the availability of fast computers has been essential in this development. These smooths are useful as a descriptive tool (as we have seen above) and also as building blocks for nonparametric regression models. Important developments in the latter area can be found in Friedman and Stuetzle (1981) and Breiman and Friedman (1985).

In this article, we extend smoothing ideas to other kinds of data. In particular, we consider (X, Y) data whose relationship is expressible through a likelihood function. Our idea is to replace a simple parametric function like $\beta_0 + x\beta_1$ appearing in the likelihood with an unspecified smooth function $s(x)$ and to estimate $s(x)$ locally. Take, for example, the situation in which y is a 0–1 response and x is a covariate. The usual linear logistic model assumes that $\log(p(x)/(1 - p(x))) = \beta_0 + x\beta_1$, where $p(x) = \Pr(Y = 1 \mid X = x)$. For such a data set, Figure 2 shows the logistic regression line, estimated by maximum likelihood. On the same plot, the observed logits are shown. (Since we cannot take the logit of 0 or 1, the y's were grouped first.) Also appearing in Figure 2 is a smooth estimate, based on the more general model $\log(p(x)/(1 - p(x))) = s(x)$, with $s(x)$ an arbitrary smooth function. As was the case in the scatterplot example, the smooth does a better job of capturing the relationship between Y and X than the line does. The smooth in Figure 2 was produced by a technique we call *local likelihood estimation*. The basic idea is a simple extension of the local fitting technique used in scatterplot smoothing. Given a global method for estimating a linear response (e.g., maximum likelihood estimation in the linear logistic model), we apply it locally, estimating a separate line in a window around each x value. The value of the estimated line at x is the estimate of the smooth response function at x. Multiple covariates are incorporated in an additive model that is estimated iteratively.

The function estimates produced by local likelihood estimation are useful for descriptive, exploratory analysis, or to suggest a transformation of a covariate. By varying the window size, we can control the smoothness of the estimated function: the larger the windows, the smoother the estimated function. When each window contains 100% of the data, the local likelihood procedure corresponds exactly to the global linear method.

An outline of the article is as follows. In Section 2 we review scatterplot smoothing and introduce the local likelihood idea. The logistic and proportional hazard models are used for illustration. In Section 3 we discuss "degrees of freedom" approximations, and finally in Section 4 we discuss the relationship of this work to other techniques. Early development of this work can be found in Tibshirani (1982) and Hastie (1983), and the reader interested in further details may refer to Tibshirani (1984).

* Robert Tibshirani is NSERC University Fellow and Assistant Professor, Department of Preventive Medicine and Biostatistics and Department of Statistics, University of Toronto, Toronto, Canada M5S 1A8. Trevor Hastie is a member of the research staff in the Statistics and Data Analysis Research Group, AT&T Bell Laboratories, Murray Hill, NJ 07974. The authors thank Tom DiCiccio, Bradley Efron, Jerome Friedman, and Paul Switzer for their helpful comments, and Art Owen for his ideas on finding degrees of freedom by simulation. Suggestions by two editors and a referee improved this article substantially. A large part of this research was completed at Stanford University, where the first author was supported by the Natural Sciences and Engineering Research Council of Canada and both authors were supported by the Department of Energy, Office of Naval Research and by the U.S. Army Research Office.

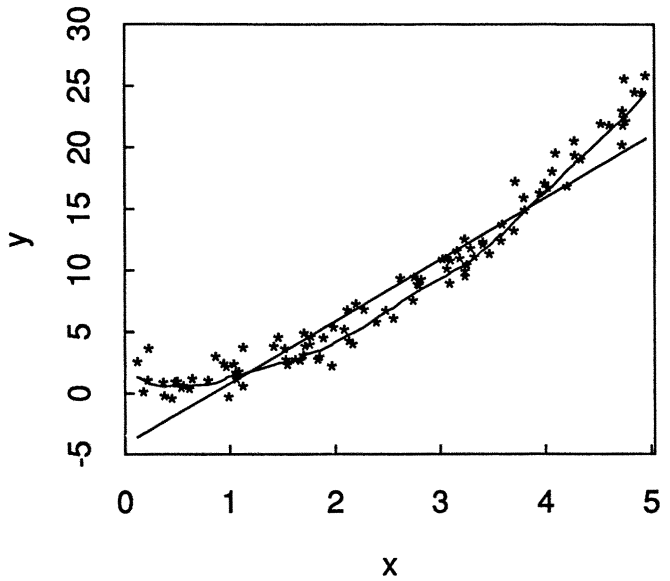

Figure 1. Least Squares Line and Scatterplot Smooth.

2. LOCAL LIKELIHOOD ESTIMATION—A DESCRIPTION

2.1 A Review of Scatterplot Smoothing

Given independent data pairs $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, assumed to be realizations of a response variable Y and a predictor X, a *scatterplot smoother* estimates

$$s(x) = \mathbf{E}(Y \mid X = x), \tag{1}$$

where $s(\cdot)$ is a smooth function. We will not define exactly what "smooth" means here; vaguely speaking, we are thinking of $s(\cdot)$ as a function less smooth than a straight line but smoother than an interpolating polynomial.

There are many ways to estimate $s(\cdot)$; we will concentrate here on the method of "local fitting." This is defined as follows. Let $\mathbf{Fit}(D, x)$ be some real-valued function of x, depending on the data D, representing the value at x of some function fitted to the data. For example, $\mathbf{Fit}(D, x)$ could be $\hat{\beta}_0 + x\hat{\beta}_1$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the regression coefficients based on D. A *local fit* estimate is defined as

$$\hat{s}(x_i) = \mathbf{Fit}(\{(x_j, y_j) : j \in N_i\}, x_i), \tag{2}$$

where N_i is a "neighborhood" of x_i (a set of indexes of points whose x values are "close" to x_i). The only neighborhoods we will consider in this article are *symmetric nearest neighborhoods*. Associated with a neighborhood is the *span* or window size w; this is the proportion of the total points that each neighborhood contains. Let $[x]$ represent the integer part of x and assume that $[wn]$ is odd. Then a span w symmetric nearest neighborhood contains $[wn]$ points: the ith point plus $([wn] - 1)/2$ points on either side of the ith point. Assuming that the data points are sorted by increasing x value, a formal definition is

$$N_i = \{\max(i - ([wn] - 1)/2, 1), \ldots, i - 1, i, i + 1, \ldots, \min(i + ([wn] - 1)/2, n)\}. \tag{3}$$

Note that the neighborhoods are truncated near the endpoints if $([wn] - 1)/2$ points are not available. The span controls the smoothness of the resulting estimate: larger spans will produce smoother (less variable) estimates but with possibly more bias. A span of $1/n$ corresponds to 1 point per neighborhood. The span is either fixed a priori or chosen adaptively from the data.

If **Fit** stands for arithmetic mean, then $\hat{s}(\cdot)$ is the *running mean*, a very simple scatterplot smoother. The running mean is not usually satisfactory, because it creates large biases at the endpoints. In addition, unless the abscissa values are equally spaced it does not reproduce straight lines (i.e., if the data lie exactly along a straight line, the smooth of the data will not be a straight line). A refinement of the running average, the *running lines smoother*, alleviates these problems. Instead of fitting a mean in a neighborhood, it fits a least squares line. The value of the least squares line at x_i is the estimated smooth there.

The running lines smoother is the most obvious generalization of the least squares line. When w is 2 (so that every neighborhood contains all of the data points), the smooth agrees exactly with the least squares line. Although very simple in nature, the running lines smoother produces reasonable results and has the advantage that the estimates can be updated. That is, to find $\hat{s}(x_{i+1})$ from $\hat{s}(x_i)$, only an $O(1)$ operation is needed. This makes the entire smoothing algorithm $O(n)$.

2.2 Local Likelihood Estimation: Definition

Suppose that we have n independent realizations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ of random variables X and Y with $Y \mid X = x \sim f(Y, \theta)$, where θ is a function of x. The likelihood is given by $L(\theta_1, \theta_2, \ldots, \theta_n) = \prod_1^n f(y_j, \theta_j)$. A standard modeling procedure would assume a parsimonious form for the θ_i's, say $\theta_i = \beta_0 + x_i\beta_1$. Then $L(\cdot)$ would be a function of β_0 and β_1; these parameters would be estimated by maximizing $L(\cdot)$. The *local likelihood method* assumes only that θ_i is a "smooth" function of x:

$$\theta_i = s(x_i). \tag{4}$$

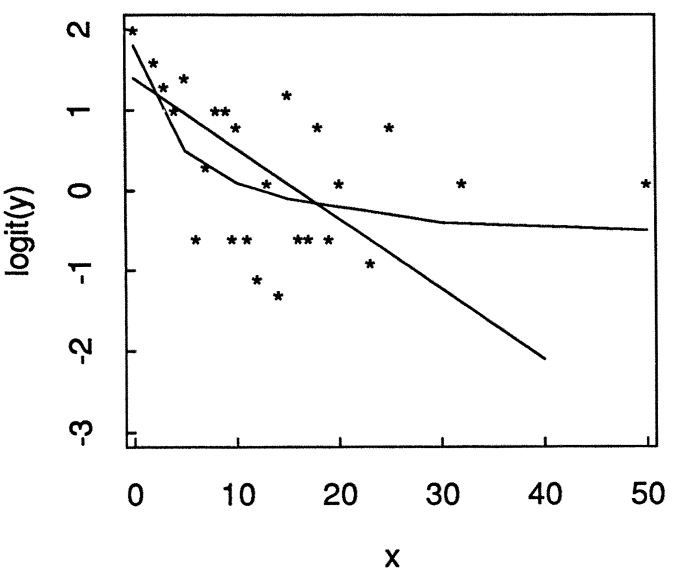

Figure 2. Logistic Regression Line and Local Likelihood Smooth.

To estimate $\{s(x_1), s(x_2), \ldots, s(x_n)\}$, we could try to maximize $L(s(x_1), s(x_2), \ldots, s(x_n))$; however, this would result in an unsatisfactory estimate due to overfitting. In many situations, it would simply reproduce the data. As an alternative, we define the *local likelihood estimate* of $s(x_i)$ as

$$\hat{s}(x_i) = \hat{\beta}_{0i} + x_i\hat{\beta}_{1i}, \tag{5}$$

where $\hat{\beta}_{0i}$ and $\hat{\beta}_{1i}$ maximize the local likelihood

$$L_i(\beta_{0i}, \beta_{1i}) = \prod_{j \in N_i} f(y_j, \beta_{0i} + x_j\beta_{1i}). \tag{6}$$

Note that i is fixed in (6), with j varying over the points of the neighborhood.

The local likelihood method produces a smooth estimate of the curve $s(\cdot)$ at the points $\{x_1, x_2, \ldots, x_n\}$. It avoids overfitting by averaging over neighborhoods. Note that values of $\hat{s}(x)$ for x not equal to one of the x_i's can be obtained by some sort of interpolation.

If we take $\beta_{1i} = 0$ for every i, what we call *local likelihood estimation with constants,* we have a procedure analogous to the running mean. This is not as useful, because it tends to produce large biases at the endpoints, but is more tractable theoretically.

More generally, suppose that we have n data tuples of the form $(y_i, x_i, \mathbf{c}_i)$, where y is a response variable, x is a covariate, and $\mathbf{c}$ is a vector containing any additional information. (In censored data problems, $\mathbf{c}$ would indicate whether y is censored; in many problems, like regression, $\mathbf{c}$ is empty.) Suppose that modeling considerations lead to maximization of a function of the form

$$L(\theta_1, \theta_2, \ldots, \theta_n) = g^n(y_1, y_2, \ldots, y_n, \theta_1, \theta_2, \ldots, \theta_n, \mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n), \tag{7}$$

with the superscript denoting that g is based on n observations. The local likelihood estimate of $s(\cdot)$ is $\hat{s}(x_i) = \hat{\beta}_{0i} + x_i\hat{\beta}_{1i}$, where $\hat{\beta}_{0i}$ and $\hat{\beta}_{1i}$ maximize the local likelihood

$$L_i(\beta_{0i}, \beta_{1i}) = g^{[N_i]}(\{y_j, \beta_{0i} + x_j\beta_{1i}, \mathbf{c}_j\}, j \in N_i), \tag{8}$$

$[N_i]$ denoting the number of observations in the neighborhood. The proportional hazards model of Cox (1972) is an example of a nonstandard model that fits into this framework (see Sec. 2.5).

A special case of model (6) occurs when Y has an exponential family density of the form

$$\exp[\{y_j\theta_j - b(\theta_j) - h(y_j, a(\phi))\}/a(\phi)] \tag{9}$$

with respect to some carrier measure. If the scale parameter ϕ is known, then (9) is an exponential family; if ϕ is unknown, (9) is not generally an exponential family, but the estimation procedure is unchanged because the local score function for θ does not involve $a(\phi)$. Letting $\mu = \mathbf{E}Y$, we assume that $\theta = g(\mu) = s(x)$. The function $g(\cdot)$ is called the *link function,* and the relation $\theta = g(\mu)$ is the *canonical* or *natural* link. We proceed by forming the local likelihood [as in (6)] and estimate the local slope and intercept β_{0i} and β_{1i}. Note that in the Gaussian case, $\hat{s}(\cdot)$ reduces to the running lines smooth defined earlier. This procedure can be viewed as an extension of the family of *generalized linear models* (Nelder and Wedderburn 1972). A generalized linear model is defined by $Y \mid x \sim f(Y, \boldsymbol{\theta})$ and $g(\mu) = \beta_0 + x\beta_1$. In the local likelihood procedure $\beta_0 + x\beta_1$ is generalized to $s(x)$.

Note that in this formulation we have modeled the natural parameter θ. We could just as well model some other parameter (like $\mathbf{E}Y$); in any specific problem, there may be reasons to prefer one parameterization to another. For example, in the binary response problem, it is more convenient to model the natural parameter $\log[p/(1-p)]$ than the expectation p, because the latter would require that the estimated smooth stay between 0 and 1.

Estimation of $\boldsymbol{\beta}_i = (\beta_{0i}, \beta_{1i})$ in the exponential family model or any local likelihood model of the form (5) and (6) is performed using a Newton–Raphson search in each neighborhood, going in order as i runs from 1 to n. The local likelihood estimate $\hat{\boldsymbol{\beta}}_i$ is used as a starting value for the maximization of $L_{i+1}(\cdot)$; because the estimates do not tend to differ much from one neighborhood to the next, convergence is typically achieved in two or three iterations.

If k_n is the number of points in a neighborhood, an $O(k_n)$ operation is required for computing each local likelihood estimate and thus the entire procedure is $O(k_n n)$. This is not a problem for moderate n (say $n \approx 200$), because of the small number of iterations required. For larger data sets, we speed up the procedure by calculating the fit at every mth point; this reduces the running time by about a factor of m. The smooths for the remaining x values are obtained by interpolation.

Some subtleties may arise in the estimation in the nonstandard case—see Example 2, the Cox model (Sec. 2.5).

2.3 Remarks

2.3.1. Asymptotic Properties. If the neighborhoods shrink in size but the number of points in each neighborhood goes to infinity at an appropriate rate, then it is reasonable to expect that the local likelihood estimate will be (pointwise) consistent and efficient. In Tibshirani and Hastie (1985) we proved this for a single covariate in exponential family. Note that when fitting local constants in that setting, no iteration is necessary for the fitted value because μ is simply the mean of the y's in the neighborhood and $\hat{s}(x_i)$ can be obtained by applying the link function to this mean. Thus we prove the result by (a) showing that the contribution of the slope parameter is asymptotically negligible and (b) applying a central limit theorem to the running mean estimate. As pointed out by a referee, this begs the question: Is it worthwhile fitting local lines instead of local constants? We believe that fitting lines is worthwhile because it reduces bias at the endpoints. The simulation described in Section 2.4 provides some evidence of this. It is important to note that if multiple covariates are present then iteration is required for local constants, even in the exponential family.

2.3.2. Number of Parameters—"Degrees of Freedom." Given a local likelihood fit based on span w, $1/n < w <$

2, we would expect the "number of parameters" it uses to be somewhere between 2 (the number for span = 2) and n (the number for span = $1/n$). In Section 4 we provide a definition of "number of parameters," or "degrees of freedom" (df), and a method for computing it. This can be used, in conjunction with the value of the overall likelihood $\prod_1^n f(y_i, \hat{\theta}_i)$, to assess the fit of the model.

2.3.3. Span Selection. The local likelihood procedure requires the choice of a span size w. One method is to try a range of spans and examine the resulting estimate and the value of the global likelihood that it produces. Sometimes, however, it may be desirable to have an automatic method for span selection. In scatterplot smoothing, one popular method for choosing the span is cross-validation (see, e.g., Friedman and Stuetzle 1982). In the local likelihood setting, cross-validation turns out to be very expensive computationally. As an alternative, we use, as a rough rule of thumb, a form of Akaike's information criterion (AIC) (Akaike 1973). Having fit a model with maximized likelihood L and p independent parameters, the AIC is defined by $2 \log L + 2p$. The first term measures the goodness of fit of the model, and the second term penalizes the number of parameters used. Hence the AIC attempts to trade off variability and bias. The AIC can be thought of as an extension of Mallows's (1973) C_p statistic to likelihood models. (The two criteria coincide in the linear regression context.) In this setting we select w to minimize AIC based on the value of the global likelihood and the number of parameters, as described in Section 4. We do not have any results, however, on the asymptotic correctness of this method of span selection.

2.3.4. Handling of Ties. For data with tied x values, two things are done. First, each neighborhood is expanded (if necessary) to ensure that if a point j is in a given neighborhood, so is any other point k having $x_k = x_j$. This makes the estimation procedure invariant to the incoming order of the data points. Second, the smooths for each of the tied values are averaged and each smooth value is assigned the average. That is, if $x_j = x_{j+1} = \cdots = x_{j+m}$, then for each $j \le i \le j + m$, $\hat{s}(x_i)$ is assigned the value $\sum_j^{j+m} \hat{s}(x_i)/(m + 1)$.

2.3.5. Multiple Covariates. The previous discussion shows how the local likelihood idea can be used to estimate the smooth for a single covariate. If p covariates are available, we assume a model of the form $\boldsymbol{\theta} = \sum_{j=1}^p s_j(\cdot)$. Tibshirani (1984) discussed a forward stepwise approach to the estimation of such a model, with readjustment through "backfitting." This procedure works by holding all but one smooth fixed and reestimating the remaining smooth. This process is iterated until convergence. Backfitting can be used for detecting nonlinearity in one covariate by forcing all of the other covariates to have a linear fit. Many theoretical details need to be worked out, including convergence of the algorithm and selection rules.

2.3.6. Other Fitting Procedures. The local likelihood procedure uses local linear estimation because it works well (especially in reducing bias at the endpoints) and is simple. More sophisticated kernel-type estimates could be used to make the procedure robust and increase the smoothness of the estimated function. Borrowing ideas from scatterplot smoothing (see, e.g., Cleveland 1979), we can downweight points based both on their distance from the center of the neighborhood and the size of the residual. Although we have not investigated downweighting in general, a robust algorithm for the proportional hazards model is discussed in Tibshirani (1984).

2.4 Example 1: The Logistic Model for Binary Data

Suppose that we have data of the form $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where the response y is 0 or 1, x is an explanatory variable, and the observations are assumed to be independent.

Let $\mathbf{x} = (1, x)$ and let $p(\mathbf{x}) = \Pr(y = 1 \mid \mathbf{x})$. The log-likelihood of the data is

$$\log L = \sum_{j=1}^{n} \{y_j \log p_j + (1 - y_j)\log(1 - p_j)\}, \quad (10)$$

where $p_j = p(\mathbf{x}_j)$. The linear logistic model assumes that logit $p(\mathbf{x}) = \mathbf{x}^t\boldsymbol{\beta}$.

The local likelihood method, on the other hand, assumes that logit $p(\mathbf{x}) = s(x)$ and $s(x)$ is estimated through the local likelihood corresponding to (6). With multiple covariates, the model takes the form logit $p(\mathbf{x}) = \alpha + \sum_1^p s_j(\cdot)$, where each $s_j(\cdot)$ is assumed to have mean 0 to ensure identifiability. We now illustrate this technique on some real data.

A study conducted between 1958 and 1970 at the University of Chicago's Billings Hospital concerned the survival of patients who had undergone surgery for breast cancer (Haberman 1976). There are 306 observations on 4 variables: $y = 1$ if patient survived ≥ 5 years, 0 other-

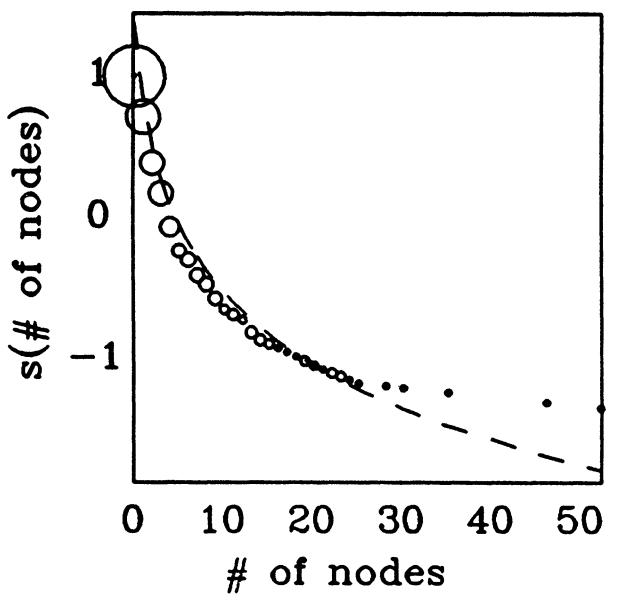

Figure 3. Estimate for Number of Nodes (circles: local likelihood smooth; broken curve: parametric function). The area of each circle is proportional to the number of data points.

wise; x_1 = age of patient at time of operation; x_2 = year of operation; and x_3 = number of positive auxillary nodes detected.

A local likelihood fit with the single variable number of nodes reduced the null "deviance" (i.e., twice log-likelihood ratio statistic) from 353.7 to 319.9 using an estimated 2.4 df. The smooth in Figure 2 is the estimate $\hat{s}$(number of nodes). By comparison, a linear logistic fit (straight line in Figure 2) reduced the deviance to 330.7 on 1 df. When all three variables were put into the model, the local likelihood procedure (with backfitting) produced the smooths shown in Figures 3 and 4 (year of operation had little effect and is not shown). The final model has a deviance of 307.74 on (306–2.4–2.5–2.4) = 298.7 df.

Landwehr, Pregibon, and Shoemaker (1984) analyzed this data set to explore the usefulness of partial residual plots in identifying parametric forms of covariate effects. Their final model was

$$\text{logit } p(\mathbf{x}) = \beta_0 + x_1\beta_1 + x_1^2\beta_2 + x_1^3\beta_3 + x_2\beta_4 + x_1x_2\beta_5 + (\log(1 + x_3))\beta_6. \quad (11)$$

The deviance of this model is 302.3 on 299 df. The fitted terms for each covariate are superimposed on Figures 3 and 4 (broken lines), ignoring the marginally significant x_1x_2 term. The functions are very similar.

Hastie (1984) and Hastie and Tibshirani (1986) discussed the relative merits of the local likelihood and partial residual plot procedures. They gave two reasons to suggest why the local likelihood procedure is preferable:

1. The partial residual technique, in suggesting the parametric form for a covariate effect, relies on the assumption that the covariate forms for the other effects are correct. Indeed, these effects are usually assumed to be linear. The local likelihood procedure finds the best functional form for all covariates simultaneously.

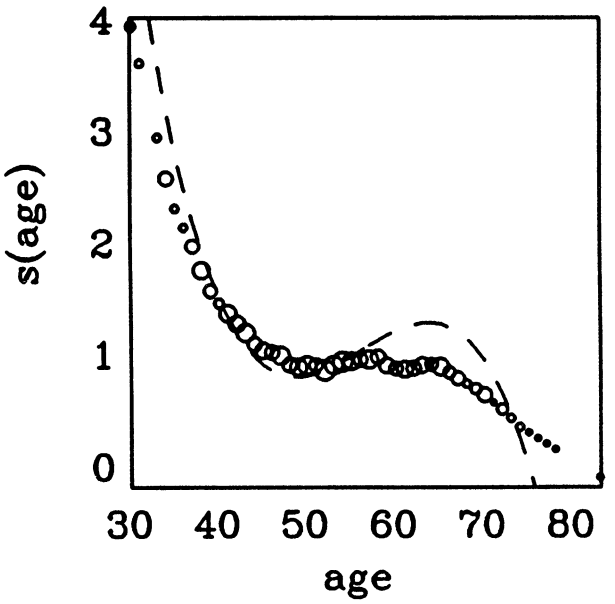

Figure 4. Estimate for Age (circles: local likelihood smooth; broken line: parametric function). The area of each circle is proportional to the number of data points.

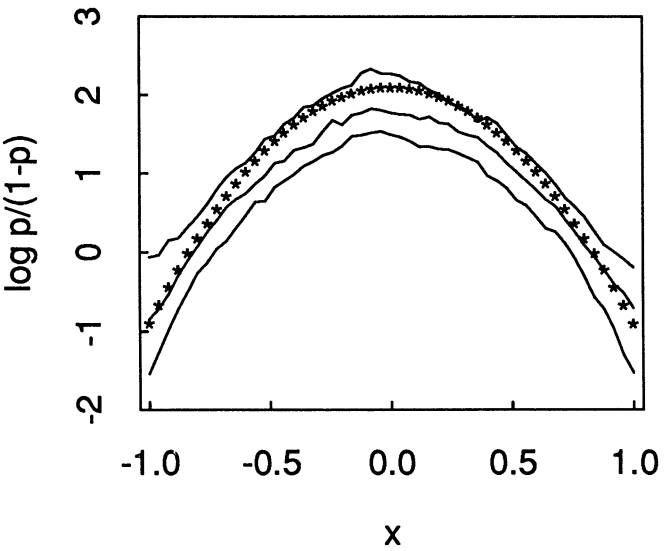

Figure 5. True Quadratic (asterisks) and Quartiles of Local Likelihood Estimates, Fitting Local Lines.

2. The partial residual technique requires quite a bit of ingenuity in identifying the various covariate effects. The local likelihood procedure, on the other hand, is automatic.

The local likelihood technique also has advantages over smoothing of the y's directly, that is, fitting a model of the form $y = \Sigma\, s_j(x_j)$. For more than one covariate, one would have to truncate the smooths to ensure that the fitted values stay between 0 and 1. By modeling the logit, the local likelihood technique avoids this problem. It also inherits (locally) the usual advantages of logistic regression.

To determine the value (if any) of fitting local lines over local constants, we ran a small simulation. We chose a sample size of 51 and fixed the x values equally spaced on $[-1, 1]$. Bernoulli random variables were generated from the model $\log(p(x)/(1 - p(x))) = 2 - 3x^2$. Figures 5 and 6 show the median and quartiles for 100 local likelihood estimates, fitting lines and constants, respectively. A span of .5 was used in each case.

The asterisks are the true quadratic function. We notice that the estimates of Figure 6 are considerably more biased near the endpoints, whereas the two methods are similar in the middle of the data. This confirms the usefulness of fitting local lines to reduce endpoint bias.

2.5 Example 2: Cox's Proportional Hazards Model

In the censored data problem we observe data triples (y_i, x_i, δ_i) $(i = 1, 2, \ldots, n)$, where δ_i indicates whether or not the response y_i is censored. The data are assumed to be sorted by the covariate x, that is, $x_1 \leq x_2 \leq \cdots \leq x_n$. The proportional hazards model of Cox (1972) models the relationship between y and x by assuming that x acts on the hazard function in a multiplicative way, that is, $\lambda(y \mid x) = \lambda_0(y)\exp(x\beta)$, where $\lambda_0(y)$ is an unspecified function and $\lambda(y \mid x)$ is the hazard function at covariate

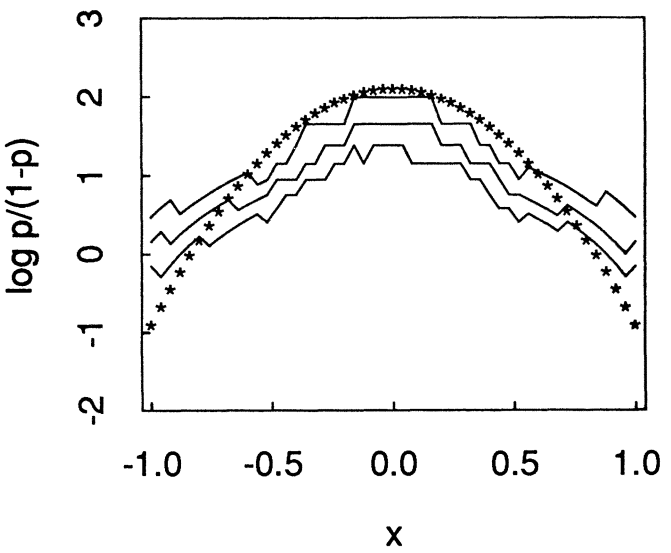

Figure 6. True Quadratic (asterisks) and Quartiles of Local Likelihood Estimates, Fitting Local Constants.

level x. This assumption allows β to be estimated independently of $\lambda_0(y)$ by maximizing the *partial likelihood* (PL):

$$\mathrm{PL} = \prod_{i \in D} \frac{e^{x_i\beta}}{\sum_{j \in R_i} e^{x_j\beta}}, \tag{12}$$

where D is the set of indexes of the uncensored y's and $R_i = \{j \mid y_j \geq y_i\}$, the risk set at time $y_i - 0$. We generalize this to $\lambda(y \mid x) = \lambda_0(y)\exp(s(x))$. Note that because of the arbitrary baseline hazard, the function $s(x)$ is only determined up to an additive constant, so for definitiveness we define $s(x_1) = 0$. To estimate $s(x_1), s(x_2), \ldots, s(x_n)$, we apply the local likelihood technique. The local PL for the data in N_i is

$$\mathrm{PL}_i = \prod_{l \in D \cap N_i} \frac{\exp(\alpha_i + x_l\beta_i)}{\sum_{j \in R_l \cap N_i} \exp(\alpha_i + x_j\beta_i)}. \tag{13}$$

Note, however, that α_i is not estimable from PL_i since the $\exp(\alpha_i)$ terms cancel one another, giving

$$\mathrm{PL}_i = \prod_{l \in D \cap N_i} \frac{\exp(x_l\beta_i)}{\sum_{j \in R_l \cap N_i} \exp(x_j\beta_i)}. \tag{14}$$

Let $\hat{\beta}_i$ maximize $L_i(\cdot)$. Although α_i [and thus $s(x_i)$] is not estimable locally, we can use the slope estimates $\{\hat{\beta}_1, \ldots, \hat{\beta}_n\}$ to estimate $\{s(x_1), \ldots, s(x_n)\}$, as follows. We have $s(x_i) = \int_c^{x_i} s'(z)dz$ and $s'(x) = \beta_i$ for $x \in N_i$; hence to estimate $s(x_i)$ we can use any estimate of $\int_c^{x_i} s'(z)dz$ based on $(x_1, \hat{\beta}_1), \ldots, (x_n, \hat{\beta}_n)$. We use the trapezoidal rule defined by

$$\hat{s}(x_i) = \sum_1^i (x_j - x_{j-1}) \frac{(\hat{\beta}_j + \hat{\beta}_{j-1})}{2}. \tag{15}$$

With more than one covariate, the model takes the form $\lambda(t \mid \mathbf{x}) = \lambda_0(t)\exp(\sum_{j=1}^p s_j(\cdot))$. The smooths are estimated in a forward stepwise manner, with backfitting, as discussed earlier.

Table 1. Mouse Leukemia Data

Model	−2 log PL	Number of parameters
Null	1189.06	0
Smooth, span = .7	1173.98	1.85
Linear	1183.16	1
Linear + quadratic	1183.07	2
Piecewise linear	1177.34	2

Kalbfleisch and Prentice (1980) analyzed the results of a study designed to examine the genetic and viral factors that may influence the development of spontaneous leukemia in AKR mice. The original data set contains 204 observations, with six covariates and both cancerous and noncancerous deaths recorded. Kalbfleisch and Prentice performed a number of analyses—we will consider any death as the endpoint and the single covariate antibody level (%). Antibody level took on continuous values, although about half of the mice had a value of 0.

Table 1 shows the results of the local likelihood procedure applied to these data, and a graph of the estimated smooth for antibody is shown in Figure 7. It is markedly nonlinear, changing slope at antibody level 7.5%. Also included in Table 1 are linear and quadratic terms for antibody. Even with a quadratic term, the fit of the parametric Cox model is significantly worse than the local likelihood smooth.

Based on Figure 7, a piecewise linear covariate was created by joining each of the leftmost and rightmost smooth values to the bending point by straight lines. For this covariate, −2 log PL was 1177.34, still significantly worse than the smooth model. This indicates that the bowed shape of the smooth between antibody levels 7.5% and 80% is supported by the data.

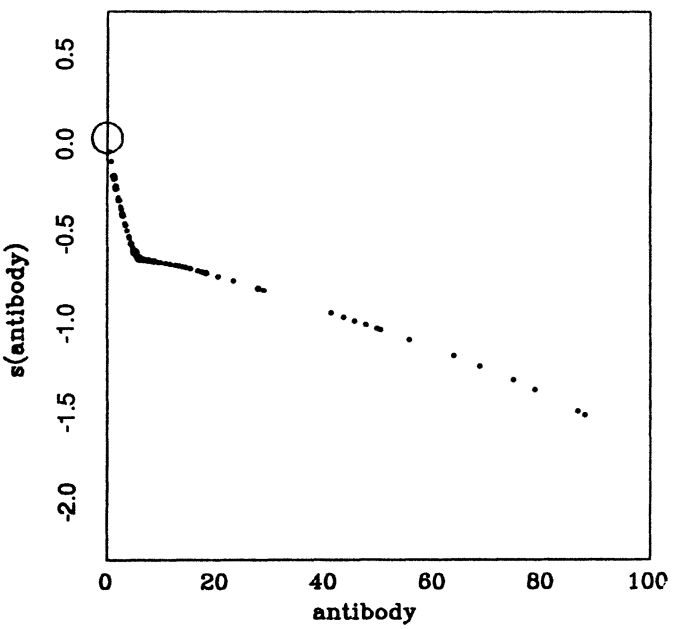

Figure 7. Local Likelihood Smooth for Mouse Leukemia Data (the area of each circle is proportional to the number of data points).

3. DEGREES OF FREEDOM APPROXIMATIONS

In generalized linear models, the goodness of fit of an estimate $\hat{\boldsymbol{\mu}}$ is measured by the deviance. Wilks's theorem tells us that, given two nested *linear* models and the hypothesis that the smaller model is correct, the deviance decrease in fitting the larger model is asymptotically $\phi\chi^2_{p_2-p_1}$, where p_1 and p_2 are the ranks of the two linear spaces. That is, the additional number of parameters fit give the number of *degrees of freedom* (df) of the corresponding deviance decrease.

This leads us to ask similar questions for the smooth estimates described in this article. We will restrict our discussion to the exponential family case. The question of interest is, *How many "parameters" are fit by a smooth?* This will depend on the span. With a span of 2 (i.e., every neighborhood contains all of the data points), 2 parameters are used. With a span of $1/n$ (i.e., 1 point per neighborhood), n independent parameters are used. Thus for spans in the range $1/n$ to 2, the number of parameters should be somewhere between 2 and n.

We first define the "number of parameters" or "degrees of freedom" of a scatterplot smoother, along the same lines as Cleveland (1979). A running lines smoother is a linear smoother; that is, the fit $\hat{\mathbf{y}}$ can be written as $\hat{\mathbf{y}} = S\mathbf{y}$, where S is called a *smoother matrix*. (This follows because least squares fitting is a linear operation.) We expand the expected residual sum of squares (RSS),

$$\begin{aligned}\mathbf{E}(\mathrm{RSS}(\mathbf{y}, \hat{\mathbf{y}})) &= \mathbf{E}(\mathbf{y} - \hat{\mathbf{y}})^t(\mathbf{y} - \hat{\mathbf{y}}) \\ &= (n - [2\,\mathrm{tr}(S) - \mathrm{tr}(S^tS)])\sigma^2 \\ &\quad + \mathbf{f}^t(I - S)^t(I - S)\mathbf{f}, \qquad (16)\end{aligned}$$

where $\mathbf{f} = \mathbf{E}\mathbf{y}$ and $\sigma^2 = \mathrm{var}(y_i)$. The first term in (16) relates to the variance of the fitted values, and the second term measures the bias. By analogy with linear regression, we define the number of df used in a fit $\hat{\mathbf{y}}$ by

$$\mathrm{df}(\hat{\mathbf{y}}) = 2\,\mathrm{tr}(S) - \mathrm{tr}(S^tS). \qquad (17)$$

In standard linear regression this expression reduces to p, the number of parameters in the model. A simplification of (17) occurs for a running lines smoother, because one can show that $\mathrm{tr}(S^tS) = \mathrm{tr}(S)$ because of the fact that each row of S is a row of a projection matrix, although S itself is not a projection matrix. [In particular, if $S = \{s_{ij}\}$, then $s_{ii} = \sum_j s_{ij}^2$, so $\mathrm{tr}(S^tS) = \mathrm{tr}(SS^t) = \mathrm{tr}(S)$]. Hence (df) simplifies to $\mathrm{tr}(S)$. Note that $\mathrm{tr}(SS^t)\sigma^2$ is also a relevant quantity; it is the sum of the variances of the fitted values $\sum \mathrm{var}(\hat{y}_i)$.

The quantity $\mathrm{df}(\hat{\mathbf{y}})$ is useful both for assessing a single fit and for comparing two fits. For example, suppose that we want to compare two fits $\hat{\mathbf{y}}_1 = S_1\mathbf{y}$ and $\hat{\mathbf{y}}_2 = S_2\mathbf{y}$. If we assume that the biases are the same, we have $\mathbf{E}(\mathrm{RSS}(\mathbf{y}, \hat{\mathbf{y}}_2) - \mathrm{RSS}(\mathbf{y}, \hat{\mathbf{y}}_1)) = [(2\,\mathrm{tr}(S_2) - \mathrm{tr}(S_2^tS_2)) - (2\,\mathrm{tr}(S_1) - \mathrm{tr}(S_1^tS_1))]\sigma^2$, or simply $[\mathrm{tr}(S_1) - \mathrm{tr}(S_2)]\sigma^2$ for running lines smoothers.

Analogous results also hold approximately for any local likelihood fit $\hat{\boldsymbol{\mu}}$ in the exponential family. Assume that the Y's are independently distributed with density of the form (9) with $\phi = 1$, and define the *deviance* as

$$\mathbf{Dev}(\mathbf{y}, \boldsymbol{\mu}) = 2[l(\mathbf{y}) - l(\boldsymbol{\mu})], \qquad (18)$$

where $l(\boldsymbol{\mu}) = \sum_1^n \log f_Y(y_i, \mu_i, \phi)$ is the log-likelihood, written for convenience as a function of $\boldsymbol{\mu}$ instead of $\boldsymbol{\theta}$. We generalize (17) and define (implicitly) the df of a fit $\hat{\boldsymbol{\mu}}$ by

$$\mathbf{E}\,\mathbf{Dev}(\mathbf{y}, \hat{\boldsymbol{\mu}}) = \mathbf{E}\,\mathbf{Dev}(\mathbf{y}, \boldsymbol{\mu}) - \mathrm{df}(\hat{\boldsymbol{\mu}}) + I(\boldsymbol{\mu}, \mathbf{h}), \qquad (19)$$

where $\mathbf{h} = \mathbf{E}\hat{\boldsymbol{\mu}}$ and $I(\boldsymbol{\mu}, \mathbf{h})$ is twice the Kullback–Leibler distance between $\boldsymbol{\mu}$ and $\mathbf{h}$ (see the Appendix). The first term on the right is independent of the fitting mechanism and compares with n in (17) and, in many cases, is asymptotically equal to n; the last term is a bias term. Now suppose that $\hat{\boldsymbol{\mu}}$ is based on a covariate vector $\mathbf{x}$ with span w. We show in the Appendix that $\mathrm{df}(\hat{\boldsymbol{\mu}})$ can be approximated by $\mathrm{tr}(S)$, if S is the running means smoother matrix based on $\mathbf{x}$ with span w.

This definition is once again useful for testing. Consider two fits $\hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\mu}}_2$ based on a covariate vector $\mathbf{x}$ and spans w_1 and w_2. Let S_1 and S_2 be the smoother matrices, based on $\mathbf{x}$, that produce running means smooths of spans w_1 and w_2, respectively. Then assuming that $\mathbf{E}\hat{\boldsymbol{\mu}}_1 \approx \mathbf{E}\hat{\boldsymbol{\mu}}_2$, we have that $\mathbf{E}((\mathbf{Dev}(\mathbf{y}, \hat{\boldsymbol{\mu}}_1) - \mathbf{Dev}(\mathbf{y}, \hat{\boldsymbol{\mu}}_2)) \approx \mathrm{df}(\hat{\boldsymbol{\mu}}_1) - \mathrm{df}(\hat{\boldsymbol{\mu}}_2) \approx \mathrm{tr}(S_2) - \mathrm{tr}(S_1)$. The derivation of these approximations, given in the Appendix, are rough, and we do not provide error bounds. Additional assumptions are also needed about the variation in the variance function.

Given a local likelihood fit, we can easily work out $\mathrm{tr}(S)$ and use it to determine the approximate significance of the smooth. As an example, for 200 equally spaced X values and a span of .5, $\mathrm{tr}(S)$ is about 3.6. Hence the smooth uses roughly 3.6 parameters.

So far, we have discussed only the expectation of the deviance, not its distribution. In Tibshirani (1984) we described a simulation study to assess the accuracy of the trace formula and to study the distribution of the deviance decrease. The formula turns out to be quite good for the Gaussian and logistic models but not very good for the Cox model. Hence we resort to simulation to estimate the df for the Cox model. As for the distribution of the deviance decrease, it is not χ^2, but is somewhat more spread out. Hence the χ^2 distribution with the appropriate df (or more specifically, the corresponding gamma distribution) should be used only as a rough reference.

4. DISCUSSION AND RELATED WORK

The local likelihood method extends nonparametric regression techniques to likelihood-based regression models. The literature on nonparametric regression is rich; see, for example, Rosenblatt (1971), Wahba and Wold (1975), Stone (1977), Cleveland (1979), Li (1984), and Silverman (1985). In the multiple covariate case, the local likelihood technique provides a method for estimating what Hastie and Tibshirani (1986) called a "generalized additive model," any model in which the linear term $\sum x_j\beta_j$ is replaced by

an additive term $\Sigma\, s(x_j)$. In that article, we discussed another closely related estimation technique, *local scoring*, and compared it with local likelihood estimation.

Generalized additive models provide one way of extending the additive model $\mathbf{E}(Y \mid \mathbf{x}) = \Sigma_1^p s_j(x_j)$ [see Hastie, Tibshirani, and Buja (1987) for a discussion of the additive model]. At least two other extensions have been proposed. Friedman and Stuetzle (1981) introduced the *projection pursuit regression model:*

$$\mathbf{E}(Y \mid \mathbf{x}) = \sum_1^p s_j(\mathbf{a}_j^t\mathbf{x}). \tag{20}$$

The directions $\mathbf{a}_j$ are found by a numerical search, and the $s_j(\cdot)$'s are estimated by smoothers.

The alternating conditional expectations (ACE) model (Breiman and Friedman 1985) generalizes the additive model by estimating a transformation of the response:

$$\mathbf{E}(\theta(Y) \mid \mathbf{x}) = \sum_1^p s_j(x_j). \tag{21}$$

The local likelihood idea could also be used to estimate a response transformation, or indeed any other function appearing in a model, for example, a link or variance function in a generalized linear model. We have not pursued this, however.

In the local likelihood procedure we have used local linear fitting. O'Sullivan, Yandell, and Raynor (1986) looked at splines for general exponential family models. They emerge as the solution to a penalized likelihood problem. An additive model is not considered; instead, a general surface is fitted. Green (1985) and Green and Yandell (1985) looked at similar techniques, with an emphasis on semiparametric models. Brant (1985) discussed a technique involving local likelihood estimation with constants.

APPENDIX: DERIVATION OF THE TRACE FORMULA FOR DEGREES OF FREEDOM

Given a fit $\hat{\boldsymbol{\mu}}$ based on the local likelihood fit of a response vector $\mathbf{y}$, a covariate $\mathbf{x}$, and span w, let S be the smoother matrix, based on $\mathbf{x}$ that produces a running means smooth of span w. As before, we assume that $\phi = 1$. Let $\boldsymbol{\mu} = \mathbf{E}\mathbf{y}$, $\mathbf{h} = \mathbf{E}\hat{\boldsymbol{\mu}}$, and $I(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = 2\mathbf{E}_{\mu_1} \Sigma_1^n \log f_Y(Y, u_{1i}, \phi)/f_Y(Y, \mu_{2i}, \phi)$, or twice the *Kullback–Leibler distance* between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. We wish to establish that

$$\mathrm{df}(\hat{\boldsymbol{\mu}}) \approx \mathrm{tr}(S), \tag{A.1}$$

where $\mathrm{df}(\hat{\boldsymbol{\mu}})$ is implicitly defined by

$$\mathbf{E}\ \mathbf{Dev}(\mathbf{y}, \hat{\boldsymbol{\mu}}) = \mathbf{E}\ \mathbf{Dev}(\mathbf{y}, \boldsymbol{\mu}) - \mathrm{df}(\hat{\boldsymbol{\mu}}) + I(\boldsymbol{\mu}, \mathbf{h}).$$

We can replace the terms $\mathbf{Dev}(\mathbf{y}, \cdot)$ by $I(\mathbf{y}, \cdot)$ (Hoeffding's lemma; see Efron 1975). Now by the definition of Kullback–Leibler distance in the exponential family, $\mathbf{E}I(\mathbf{y}, \mathbf{h}) = \mathbf{E}I(\mathbf{y}, \mathbf{u}) + I(\boldsymbol{\mu}, \mathbf{h})$, and we can expand $I(\mathbf{y}, \mathbf{h})$ as

$$I(\mathbf{y}, \mathbf{h}) = I(\mathbf{y}, \hat{\boldsymbol{\mu}}) + I(\hat{\boldsymbol{\mu}}, \mathbf{h}) + \Delta, \tag{A.2}$$

where $\Delta = 2(g(\hat{\boldsymbol{\mu}}) - g(\mathbf{h}))^t(\mathbf{y} - \hat{\boldsymbol{\mu}})$. By using these relationships we get

$$\mathrm{df}(\hat{\boldsymbol{\mu}}) = [\mathbf{E}(I(\hat{\boldsymbol{\mu}}, \mathbf{h})) + \mathbf{E}\,\Delta].$$

First we concentrate on the term $\mathbf{E}(I(\hat{\boldsymbol{\mu}}, \mathbf{h}))$. This can be thought of as the analog of the sum of the variances of the fits, using I as a "metric." We show now that it is approximately a weighted sum of variances and that $\mathbf{E}(I(\hat{\boldsymbol{\mu}}, \mathbf{h})) \approx \mathrm{tr}(S)$.

Denoting by $V(\cdot)$ the variance function, let $\sigma_i = V(\mu_i)$ and $a_i = V(h_i)$. In addition, let $D(a_i)$ be a diagonal matrix with iith entry a_i and denote by v_i the variance of $\hat{\mu}_i$. Then a standard Taylor series approximation gives

$$\begin{aligned}\mathbf{E}(I(\hat{\boldsymbol{\mu}}, \mathbf{h})) &\approx \mathbf{E}(\hat{\boldsymbol{\mu}} - \mathbf{h})^tD(a_i)^{-1}(\hat{\boldsymbol{\mu}} - \mathbf{h}) \\ &= \sum_1^n v_i a_i^{-1}. \end{aligned}\tag{A.3}$$

We assume that $\hat{\boldsymbol{\mu}}$ was obtained by local likelihood fits with constants; this means that $\hat{\boldsymbol{\mu}} = S\mathbf{y}$, where S is the running means smoother matrix. Then $v_i = \Sigma_{j\in N_i} \sigma_j/[N_i]^2$, and if we assume that $ave_{[N_i]}(\sigma_j) \approx \sigma_i$, then $v_i \approx \sigma_i/[N_i]$. Finally, assuming that $a_i \approx \sigma_i$, we have

$$\mathbf{E}(I(\hat{\boldsymbol{\mu}}, \mathbf{h})) \approx \sum_1^n \frac{1}{[N_i]} = \mathrm{tr}(S). \tag{A.4}$$

Finally, we show that $\mathbf{E}\,\Delta \approx 0$. Note that in the case of generalized linear models, we have

$$\Delta = 2(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^tX^t(\mathbf{y} - \hat{\boldsymbol{\mu}}) = 0. \tag{A.5}$$

Then (A.2) is a special case of Simon's theorem (Simon 1973), a Pythagorean relation for Kullback–Leibler distance. For running lines fits $\hat{\boldsymbol{\mu}} = S\mathbf{y}$, $\Delta = 2(S\mathbf{y} - \mathbf{h})^t(\mathbf{y} - \hat{\boldsymbol{\mu}}) = 2(\mathbf{y} - \mathbf{f})^tS^t(I - S)\mathbf{y}$, with expectation $\sigma^2\,\mathrm{tr}(S^t(I - S)) = 0$. For local likelihood fits in the exponential family we can write $g(\hat{\boldsymbol{\mu}}) - g(\mathbf{h}) \approx D^{-1}(a_i)(\hat{\boldsymbol{\mu}} - \mathbf{h})$. Since once again $\hat{\boldsymbol{\mu}} \approx S\mathbf{y}$, we have

$$\begin{aligned}\mathbf{E}(\Delta) &\approx 2\mathbf{E}(\mathbf{y} - \boldsymbol{\mu})^tS^tD^{-1}(a_i)(I - S)\mathbf{y} \\ &= \mathrm{tr}(S^tD^{-1}(a_i)(I - S)D(\sigma_i)) \\ &= 2(\mathrm{tr}(SD^{-1}(a_i)D(\sigma_i))) - \mathrm{tr}(S^tD^{-1}(a_i)SD(\sigma_i)).\end{aligned}$$

The term on the right, however, is another representation for $\mathbf{E}(\hat{\boldsymbol{\mu}} - \mathbf{h})^tD(a_i)^{-1}(\hat{\boldsymbol{\mu}} - \mathbf{h})$, and thus from the derivation leading to (A.4), $\mathbf{E}\,\Delta \approx 0$.

[*Received February 1983. Revised August 1986.*]

REFERENCES

Akaike, H. (1973), "Information Theory and an Extension of the Entropy Maximization Principle," in *2nd International Symposium on Information Theory*, eds. B. N. Petrov and F. Csak, Kiado: Akademia, pp. 267–281.

Brant, R. (1985), "Smooth Residual Plots for Generalized Linear Models," technical report, University of Minnesota, Dept. of Applied Statistics.

Breiman, L., and Friedman, J. H. (1985), "Estimating Optimal Correlations for Multiple Regression and Correlation," *Journal of the American Statistical Association*, 80, 580–597.

Cleveland, W. S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 828–836.

Cox, D. R. (1972), "Regression Models and Life Tables," *Journal of the Royal Statistical Society*, Ser. B, 34, 187–202.

Efron, B. (1975), "The Geometry of Exponential Families," *The Annals of Statistics*, 6, 362–376.

Friedman, J. H., and Stuetzle, W. (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association*, 76, 817–823.

——— (1982), "Smoothing of Scatterplots," technical report (Orion 003), Stanford University, Dept. of Statistics.

Green, P. J. (1985), "Penalized Likelihood for General Semi-parametric Regression Models," Technical Report 2819, University of Wisconsin–Madison, Dept. of Statistics.

Green, P., and Yandell, B. (1985), "Semiparametric Generalized Linear

Models," in *Proceedings of the 2nd International GLIM Conference* (Lecture Notes in Statistics 32), Berlin: Springer-Verlag.
Haberman, S. (1976), "Generalized Residuals for Log-linear Models," in *Proceedings of the 9th International Biostatistics Conference,* Boston, pp. 104–122.
Hastie, T. (1983), "Non-parametric Logistic Regression," technical report (Orion 016), Stanford University, Dept. of Statistics.
——— (1984), Comment on "Graphical Methods for Assessing Logistic Regression Models," by J. M. Landwehr, D. Pregibon, and A. Shoemaker, *Journal of the American Statistical Association,* 79, 77–78.
Hastie, T., and Tibshirani, R. (1986), "Generalized Additive Models" (with discussion), *Statistical Science,* 1, No. 3, 297–318.
Hastie, T., Tibshirani, R., and Buja, A. (1987), "Linear Smoothers and Additive Models," submitted for publication.
Kalbfleisch, J. D., and Prentice, R. L. (1980), *The Statistical Analysis of Failure Time Data,* New York: John Wiley.
Landwehr, J. M., Pregibon, D., and Shoemaker, A. (1984), "Graphical Methods for Assessing Logistic Regression Models," *Journal of the American Statistical Association,* 79, 61–71.
Li, K. C. (1984), "Regression Models With Infinitely Many Parameters: Consistency of Bounded Linear Functionals, *The Annals of Statistics,* 12, 601–611.
Mallows, C. L. (1973), "Some Comments on C_p," *Technometrics,* 15, 661–675.
Nelder, J. A., and Wedderburn, R. W. M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society,* Ser. A, 135, 370–384.
O'Sullivan, F., Yandell, B., and Raynor, W. (1986), "Automatic Smoothing of Regression Functions in Generalized Linear Models," *Journal of the American Statistical Association,* 81, 96–103.
Rosenblatt, M. (1971), "Curve Estimates," *Annals of Mathematical Statistics,* 42, 1815–1841.
Silverman, B. W. (1985), "Some Aspects of the Spline Smoothing Approach to Non-parametric Regression Curve Fitting" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 36, 111–147.
Simon, G. (1973), "Additivity of Information in Exponential Family Laws," *Journal of the American Statistical Association,* 68, 478–482.
Stone, C. (1977), "Consistent Non-parametric Regression," *The Annals of Statistics,* 5, 595–620.
Tibshirani, R. (1982), "Non-parametric Estimation of Relative Risk," technical report (Orion 22), Stanford University, Dept. of Statistics.
——— (1984), "Local Likelihood Estimation," unpublished Ph.D. dissertation, Stanford University, Dept. of Statistics.
Tibshirani, R., and Hastie, T. (1985), "Local Likelihood Estimation," unpublished manuscript, University of Toronto, Dept. of Statistics.
Wahba, G., and Wold, S. (1975), "A Completely Automatic French Curve: Fitting Spline Functions by Cross-Validation," *Communications in Statistics, Part A—Theory and Methods,* 4, 1–7.