Statistical Science
1986, Vol. 1, No. 3, 297–318

Generalized Additive Models

Trevor Hastie and Robert Tibshirani

Abstract. Likelihood-based regression models such as the normal linear regression model and the linear logistic model, assume a linear (or some other parametric) form for the covariates $X_1, X_2, \cdots, X_p$. We introduce the class of *generalized additive models* which replaces the linear form $\sum \beta_j X_j$ by a sum of smooth functions $\sum s_j(X_j)$. The $s_j(\cdot)$'s are unspecified functions that are estimated using a scatterplot smoother, in an iterative procedure we call the *local scoring* algorithm. The technique is applicable to any likelihood-based regression model: the class of *generalized linear models* contains many of these. In this class the linear predictor $\eta = \sum \beta_j X_j$ is replaced by the additive predictor $\sum s_j(X_j)$; hence, the name generalized additive models. We illustrate the technique with binary response and survival data. In both cases, the method proves to be useful in uncovering nonlinear covariate effects. It has the advantage of being completely automatic, i.e., no "detective work" is needed on the part of the statistician. As a theoretical underpinning, the technique is viewed as an empirical method of maximizing the *expected log likelihood*, or equivalently, of minimizing the *Kullback–Leibler distance* to the true model.

Key words and phrases: Generalized linear models, smoothing, nonparametric regression, partial residuals, nonlinearity.

1. INTRODUCTION

Likelihood-based regression models are important tools in data analysis. A typical scenario is the following. A likelihood is assumed for a response variable Y, and the mean or some other parameter is modeled as a linear function $\sum_1^p \beta_j X_j$ of a set of covariates $X_1, X_2, \cdots, X_p$. The parameters of the linear function are then estimated by maximum likelihood. Examples of this are the normal linear regression model, the logistic regression model for binary data, and Cox's proportional hazards model for survival data. These models all assume a linear (or some parametric) form for the covariate effects.

A trend in the past few years has been to move away from linear functions and model the dependence of Y on $X_1, X_2, \cdots, X_p$ in a more nonparametric fashion. For a single covariate, such a model would be $Y = s(X) +$ error where $s(X)$ is an unspecified smooth function. This function can be estimated by any so-called *scatterplot smoother*, for example a running mean, running median, running least squares line, kernel estimate, or a spline (see Reinsch (1967), Wahba and Wold (1975), Cleveland (1979), and Silverman (1985) for discussions of smoothing techniques). For the p covariates $\mathbf{X} = (X_1, X_2, \cdots, X_p)$, one can use a p-dimensional scatterplot smoother to estimate $s(\mathbf{X})$, or else assume a less general model such as the additive model $s(\mathbf{X}) = \sum_1^p s_j(X_j)$ and estimate it in an iterative manner.

In this paper we propose a class of models that extends the usual collection of likelihood-based regression models and a method for its estimation. This new class replaces the linear function $\sum_1^p \beta_j X_j$ by an additive function $\sum_1^p s_j(X_j)$; we call it the class of *generalized additive models*. The technique for estimating the $s_j(\cdot)$'s, called the *local scoring* algorithm, uses scatterplot smoothers to generalize the usual Fisher scoring procedure for computing maximum likelihood estimates. For example, the linear logistic model for binary data specifies $\log[p(\mathbf{X})/(1 - p(\mathbf{X}))] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$, where $p(\mathbf{X}) = \text{Prob}(Y = 1 \mid \mathbf{X})$. This is generalized to $\log[p(\mathbf{X})/(1 - p(\mathbf{X}))] = \sum_1^p s_j(X_j)$, and the *local scoring* procedure provides nonparametric, smooth estimates of the $s_j(\cdot)$'s. The smooth functions produced by the local scoring procedure can be used as a data description, for prediction, or to suggest covariate transformations. One can

Trevor Hastie is a member of the technical staff at AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, New Jersey 07974. Robert Tibshirani is Assistant Professor and NSERC University Research Fellow, Department of Preventive Medicine and Biostatistics and Department of Statistics, University of Toronto, Toronto, Ontario M5S 1A8, Canada.

allow a smooth estimate for all of the covariates or force a linear fit for some of them. Such a semiparametric model would naturally arise if categorical covariates were present, but would also be useful if, for reasons specific to the data at hand, a linear fit was desired for certain covariates.

The Gaussian and logistic models are members of the class of *generalized linear models* (GLM) (Nelder and Wedderburn, 1972). This comprehensive class restricts Y to be in the exponential family (with an unspecified scale parameter); the statistical package GLIM (generalized linear interactive modeling) performs estimation and diagnostic checking for these models. Generalized additive models extend GLM by replacing the linear predictor $\eta = \sum_1^p \beta_j X_j$ with an additive predictor of the form $\eta = \sum_1^p s_j(X_j)$. The third example, the proportional hazards model mentioned earlier, is not in the exponential family, and the likelihood it uses is not in fact a true likelihood at all. Nevertheless, we still think of it as a "likelihood-based" regression model and the techniques can be applied. The usual form for the relative risk $\exp(\sum_1^p \beta_j X_j)$ is replaced by the more general form $\exp(\sum_1^p s_j(X_j))$.

The local scoring procedure is similar to another method for estimating generalized additive models, *local likelihood* estimation (Hastie, 1984a; Tibshirani, 1984, and references therein). In this paper we compare the two techniques in some examples and find that the estimated functions are very similar. The advantage of the local scoring method is that it is considerably faster.

Generalized additive models provide one way to extend the additive model $E(Y \mid \mathbf{X}) = \sum_1^p s_j(X_j)$. At least two other extensions have been proposed. Friedman and Stuetzle (1981) introduced the *projection pursuit regression model* $E(Y \mid \mathbf{X}) = \sum_1^p s_j(a_j' \mathbf{X})$. The direction vectors $\mathbf{a}_j$ are found by a numerical search, while the $s_j(\cdot)$'s are estimated by smoothers. The *ACE* (alternating conditional expectation) model of Breiman and Friedman (1985) generalizes the additive model by estimating a transformation of the response: $E(\theta(Y) \mid \mathbf{X}) = \sum_1^p s_j(X_j)$. Breiman and Friedman discuss other extensions in their article.

This paper is nontechnical for the most part, with an emphasis on the techniques and their illustration through examples. In Section 2, we review the linear regression model and its generalization (the additive model). Section 3 reviews generalized linear models. In Section 4, we link smoothing and generalized linear models to produce a more general model. The two techniques for estimation are introduced and illustrated.

In Section 5, we present a unified framework in which to view the estimation procedures, and a general form of local scoring applicable to any likelihood-based regression model. Section 6 contains examples of the procedures, including the logistic model and Cox's model for censored data. In Section 7 we discuss multiple covariate models and backfitting procedures. Section 8 compares the local scoring and local likelihood procedures, and finally in Section 9 we discuss extensions of the models and related work.

2. THE LINEAR REGRESSION MODEL AND ITS SMOOTH EXTENSION

Our discussion will center on a response random variable Y, and a set of predictor random variables $X_1, X_2, \cdots, X_p$. A set of n independent realizations of these random variables will be denoted by $(y_1, x_{11}, \cdots, x_{1p}), \cdots, (y_n, x_{n1}, \cdots, x_{np})$. When working with a single predictor ($p = 1$), we'll denote it by X and its realizations by $x_1, x_2, \cdots, x_n$.

A regression procedure can be viewed as a method for estimating $E(Y \mid X_1, X_2, \cdots, X_p)$. The standard linear regression model assumes a simple form for this conditional expectation:

$$E(Y \mid X_1, X_2, \cdots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p. \tag{1}$$

Given a sample, estimates of $\beta_0, \beta_1, \cdots, \beta_p$ are usually obtained by least squares.

The additive model generalizes the linear regression model. In place of (1), we assume

$$E(Y \mid X_1, X_2, \cdots, X_p) = s_0 + \sum_{j=1}^{p} s_j(X_j), \tag{2}$$

where the $s_j(\cdot)$'s are smooth functions standardized so that $Es_j(X_j) = 0$. These functions are estimated one at a time, in a forward stepwise manner. Estimation of each $s_j(\cdot)$ is achieved through a *scatterplot smoother.*

2.1 Scatterplot Smoothers

Let's look first at the case of a single predictor. Our model is

$$E(Y \mid X) = s(X). \tag{3}$$

(If there is only one smooth function, we suppress the constant term s_0 and absorb it into the function.) To estimate $s(x)$ from data, we can use any reasonable estimate of $E(Y \mid X = x)$. One class of estimates are the *local average estimates:*

$$\hat{s}(x_i) = \text{Ave}_{j \in N_i}\{y_j\}, \tag{4}$$

where Ave represents some averaging operator like the mean and N_i is a *neighborhood* of x_i (a set of indices of points whose x values are *close* to x_i). The only type of neighborhoods we'll consider in this paper are *symmetric nearest neighborhoods.* Associated with a neighborhood is the *span* or *window* size w; this is

the proportion of the total points contained in each neighborhood. Let $[x]$ represent the integer part of x and assume that $[wn]$ is odd. Then a span w symmetric nearest neighborhood will contain $[wn]$ 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:

$$N_i = \left\{\max\left(i - \frac{[wn] - 1}{2}, 1\right), \cdots, i - 1, i, i + 1, \cdots, \min\left(i + \frac{[wn] - 1}{2}, n\right)\right\}. \tag{5}$$

We see that the neighborhoods are truncated near the end points if $([wn] - 1)/2$ points are not available. The span w controls the smoothness of the resulting estimate, and is usually chosen in some way from the data.

If Ave stands for arithmetic mean, then $\hat{s}(\cdot)$ is the *running mean*, a very simple scatterplot smoother. The running mean is not a satisfactory smoother because it creates large biases at the end points and doesn't generally 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 slight refinement of the running average, *the running lines smoother* alleviates these problems. The running lines estimate is defined by

$$\hat{s}(x_i) = \hat{\beta}_{0i} + \hat{\beta}_{1i}x_i, \tag{6}$$

where $\hat{\beta}_{0i}$ and $\hat{\beta}_{1i}$ are the least squares estimates for the data points in N_i:

$$\hat{\beta}_{1i} = \frac{\sum_{j\in N_i}(x_j - \bar{x}_i)y_j}{\sum_{j\in N_i}(x_j - \bar{x}_i)^2}, \qquad \hat{\beta}_{0i} = \bar{y}_i - \hat{\beta}_{1i}\bar{x}_i, \tag{7}$$

and $\bar{x}_i = (1/n)\sum_{j\in N_i} x_j, \bar{y}_i = (1/n)\sum_{j\in N_i} y_j$. An estimate of $s(a)$ for a not equal to one of the x_i's can be obtained by interpolation.

The running lines smoother is the most obvious generalization of the least squares line. If $w = 2$ (that is every neighborhood contains all the data points), the smooth agrees exactly with least squares regression line (note that with $w = 1$ a neighborhood at the end points would only contain about half of the data points). Although very simple in nature, the running lines smoother produces reasonable results and has the advantage that the estimate in a neighborhood can be found by updating the estimate of the previous neighborhood. As a result, a running lines smoother can be implemented in an $O(n)$ algorithm (an algorithm having number of computations proportional to n), a fact that will become important when we use it as a primitive in other procedures. For the rest of this paper, a "**smooth**$[\cdot]$" operation will refer to a running lines smoother for some fixed span.

It is important to note, however, that the running lines smoother plays no special role in the algorithms that are described in this paper. Other estimates of $E(Y \mid X)$ could be used, such as a kernel or spline smoother. Except for the increased computational cost, these smoothers could be expected to work as well or better than the running lines smoother.

Finally, using smooth as a building block, the full model (2) can be estimated in a forward stepwise manner. This is discussed in Section 7.

2.2 Span Selection and the Bias-Variance Tradeoff

The running lines smoother requires a choice of span size w. Let's look at the extreme choices first. When $w = 1/n$, $\hat{s}(x_i)$ is just y_i. This is not a good estimate because it has a high variance and is not smooth. If $w = 2$, $\hat{s}(\cdot)$ is the global least squares regression line. This estimate is *too* smooth and will not pick up curvature in the underlying function, i.e., it might be biased. Hence, the span size should be chosen between $1/n$ and 2 to tradeoff the bias and variability of the estimate.

A data-based criterion can be derived for this purpose if we consider the estimates of $E(Y \mid X)$ as empirical minimizers of the (integrated) prediction squared error (PSE)

$$\text{PSE} = E(Y - s(X))^2 \tag{8}$$

or equivalently the integrated mean squared error (MSE)

$$\text{MSE} = E(E(Y \mid X) - s(X))^2. \tag{9}$$

Let $\hat{s}_w^{-i}(x_i)$ be the running lines smooth of span w, at x_i, having removed (x_i, y_i) from the sample. Then the *cross-validation* sum of squares (CVSS) is defined by

$$\text{CVSS}(w) = (1/n)\sum_1^n (y_i - \hat{s}_w^{-i}(x_i))^2. \tag{10}$$

One can show that $E(\text{CVSS}(w))$ is approximately PSE, using the fact that $\hat{s}_w^{-i}(x_i)$ is independent of y_i. Thus it is reasonable to choose the span w that produces the smallest value of CVSS(w). This criterion effectively weighs bias and variance based on the sample. Cross-validation for span selection is discussed in Friedman and Stuetzle (1982). Note that if we used the observed residual error RSS $= (1/n)\sum_1^n (y_i - \hat{s}_w(x_i))^2$ to choose w, ($\hat{s}_w(x_i)$ being the fit at x_i with span w) we would get $w = 1/n$ and hence $\hat{s}(x_i) = y_i$. Not surprisingly, residual sum of squares (RSS) is not a good estimate of PSE. The point is that by choosing the span to minimize an estimate of *expected* squared error, we get a useful estimate.

3. A REVIEW OF GENERALIZED LINEAR MODELS (GLMs)

Generalized linear models (Nelder and Wedderburn, 1972) consist of *a random component*, *a systematic component*, and *a link function* linking the two components. The response Y is assumed to have exponential family density

$$(11)\quad f_Y(y;\theta;\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\},$$

where θ is called the natural parameter and ϕ is the scale parameter. This is the random component of the model. We also assume that the expectation of Y, denoted by μ, is related to the set of covariates $X_1, X_2, \cdots, X_p$ by $g(\mu) = \eta$, where $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$. η is the systematic component and $g(\cdot)$ is the link function. Note that the mean μ is related to the natural parameter θ by $\mu = b'(\theta)$; also, the most commonly used link for a given f is called the *canonical link*, for which $\eta = \theta$. It is customary, however, to define the model in terms of μ and $\eta = g(\mu)$ and thus θ does not play a role. Hence, when convenient we'll write $f_Y(y, \theta, \phi)$ as $f_Y(y, \mu, \phi)$.

Estimation of μ won't involve the scale parameter ϕ, so for simplicity this will be assumed known.

Given specific choices for the random and systematic components, a link function, and a set of n observations, the maximum likelihood estimate of $\hat\beta = (\hat\beta_0, \hat\beta_1, \cdots, \hat\beta_p)$ can be found by a Fisher scoring procedure. GLIM uses an equivalent algorithm called *adjusted dependent variable regression.* Given $\hat\eta$ (a current estimate of the linear predictor), with corresponding fitted value $\hat\mu$, we form the adjusted dependent variable

$$(12)\quad z = \hat\eta + (y - \hat\mu)\left(\frac{d\eta}{d\mu}\right).$$

Define weights W by

$$(13)\quad (W)^{-1} = \left(\frac{d\eta}{d\mu}\right)^2 V,$$

where V is the variance of Y at $\mu = \hat\mu$. The algorithm proceeds by regressing z on $1, x_1, \cdots, x_p$ with weights W to obtain an estimate $\hat\beta$. Using $\hat\beta$, a new $\hat\mu$ and $\hat\eta$ are computed. A new z is computed and the process is repeated, until the change in the deviance

$$(14)\quad \text{dev}(y, \hat\mu) = 2[l(y) - l(\hat\mu)]$$

is sufficiently small. In the above, $l(\mu)$ is the log likelihood $\sum \log f_Y(y_i, \mu_i, \phi)$. Nelder and Wedderburn show that the adjusted dependent variable algorithm is equivalent to the Fisher scoring procedure, that is the sequence of estimates is identical. It is attractive because no special optimization software is required, just a subroutine that computes weighted least squares estimates. Green (1984) gives an excellent discussion of iteratively reweighted least squares methods for maximum likelihood estimation.

A comprehensive description of generalized linear models is given by McCullagh and Nelder (1983).

4. SMOOTH EXTENSIONS OF GENERALIZED LINEAR MODELS

4.1 Specification of the Model

The linear predictor $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ specifies that $X_1, X_2, \cdots, X_p$ act in a linear fashion. A more general model is

$$(15)\quad \eta = s_0 + \sum_1^p s_j(X_j),$$

where $s_1(\cdot), \cdots, s_p(\cdot)$ are smooth functions. These functions will not be given a parametric form but instead will be estimated in a nonparametric fashion.

4.2 Estimation of the Model—Local Scoring

We require an estimate of the $s_j(\cdot)$'s in (15). For the linear model $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$, the estimates were found by repeatedly regressing the adjusted dependent variable z on $1, X_1, \cdots, X_p$. Since smoothing generalizes linear regression, in the smooth model $\eta = s(X)$, we can estimate $s(\cdot)$ by repeatedly smoothing the adjusted dependent variable on X. We call this procedure *local scoring* because the Fisher scoring update is computed using a local estimate of the score. This intuitive idea can be justified on firm grounds (see Section 5). For the full model (15), the smooths can be estimated one at a time in an iterative fashion. This idea is discussed in detail in Section 7.

In Figure 1 (Section 6) we display the results of local scoring smoothing (solid curve), $\exp(\hat s(x))/(1 + \exp(\hat s(x)))$, along with the usual linear estimate (almost straight curve) $\exp(\hat\beta_0 + \hat\beta_1 x)/(1 + \exp(\hat\beta_0 + \hat\beta_1 x))$, for some binary response data. This is one of the smooths from the analysis of Haberman's breast cancer data discussed in detail in Sections 6 and 7.

This procedure requires a choice of span. In the Gaussian or ordinary additive regression models we use the CVSS (10) to guide us in selecting spans. CVSS is approximately unbiased for the expected prediction squared error, whereas RSS is not and would lead us to pick spans of $1/n$. In the exponential family, the deviance is the analogue of RSS. It is a sample estimate of the expected Kullback-Leibler distance between a model and future observations. Just like the RSS it will be biased for this quantity. For span selection, one can think of cross-validating the deviance in order to get an approximately unbiased estimate for the Kullback-Leibler distance. This

however would be very expensive due to the nonlinear nature of the estimation procedure. In ordinary additive regression, simple deletion formula allow one to calculate cross-validated fits in $O(n)$ computations. In the local scoring algorithm, however, the entire estimation procedure would have to be repeated n times, and so cross-validation would be very expensive.

Instead, we find the span at each iteration by cross-validation as described in Section 2. Recall that $E(\text{CVSS}(w)) \approx \text{PSE}$ for a scatterplot smoother; the derivation of this rests on the fact that the fitted value $\hat{s}_w^{-i}(x_i)$ does not involve y_i, and thus is independent of y_i. In this setting, the response is the adjusted dependent variable z_i which is a function of y_i. The cross-validated fit for z_i is a function of z_j, $j \neq i$. Since z_j is a function of y_i from previous iterations, z_i is not independent of its cross-validated fit. However, if k_n is the number of points in the neighborhood, then one can show that under reasonable conditions the correlation is only $O(1/k_n)$.

To obtain smoother estimates, we use a slight modification of this criterion. We choose a larger span than the cross-validatory choice if it produces less than a 1% increase in CVSS(w).

4.3 Estimation of the Model—Local Likelihood

Hastie (1984a) and Tibshirani (1984) discuss another method for estimating smooth covariate functions called *local likelihood* estimation. For a single covariate, the usual (linear) procedure fits a line across the entire range of X, i.e., $\eta = \beta_0 + \beta_1 X$. To estimate the model $\eta = s(X)$, the local likelihood procedure generalizes this by assuming that locally $s(x)$ is linear and fits a line in a neighborhood around each X value. In the exponential family with canonical link, the local likelihood estimate of $s(x_i)$ is defined as

$$\hat{s}(x_i) = \hat{\beta}_{0i} + \hat{\beta}_{1i} x_i, \tag{16}$$

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

$$\log L_i = \sum_{j \in N_i} \left\{ \frac{y_j \theta_{ij} - b(\theta_{ij})}{a(\phi)} + c(y_j, \phi) \right\}, \tag{17}$$

and $\theta_{ij} = \beta_{0i} + \beta_{1i} x_j$. The local likelihood smooth applied to the Haberman data is shown in Figure 1 (dotted line). They are very similar, a fact that seems to be a general phenomenon. We discuss the relationship between the two procedures in Section 8.

Local scoring and local likelihood estimation provide two methods for estimating the covariate functions of a generalized linear model. In the next section, we introduce a theoretical framework in which to view both of these techniques. Besides providing a justification for the methods, this framework also produces a general form of local scoring that can be used in any likelihood-based regression model.

5. JUSTIFICATION OF THE SMOOTHING PROCEDURES

5.1 The Expected Log Likelihood Criterion

In Section 2 we discussed scatterplot smoothers as estimates of $E(Y \mid X)$. There we saw that by choosing the span to minimize an estimate of *expected* squared error (as opposed to residual sum of squares), we obtained a useful estimate. In this section, we will use this idea in a likelihood setting, basing the estimation procedures on *expected* log likelihood.

Consider a likelihood-based regression model with one covariate. We assume that the data pairs $(x_1, y_1), \cdots, (x_n, y_n)$ are independent realizations of random variables X and Y. Assume also that given $X = x$, Y has conditional density $h(y, \eta)$. Since η is a function of x, we will sometimes write $\eta(x)$ for emphasis. Denote the corresponding log likelihood for a single observation by $l(\eta, Y)$ or l for short. Now to estimate $\eta(\cdot)$, we could simply maximize $\sum_1^n l(\eta(x_i), y_i)$ over $\{\eta(x_1), \eta(x_2), \cdots, \eta(x_n)\}$. This is unsatisfactory, however, because it doesn't force the estimate to be smooth. In the logistic model, for example, it produces $\hat{\eta}(x_i) = +\infty$ if $y_i = 1$ and $-\infty$ if $y_i = 0$, and the estimated probabilities are just the observed y_i's. Looking back at the scatterplot smoothing discussion, we see that a remedy in the random variable case is to choose $\hat{\eta}(\cdot)$ to maximize the *expected* log likelihood:

$$E(l(\hat{\eta}(X), Y)) = \max_{\eta} E(l(\eta(X), Y)), \tag{18}$$

the expectation being over the joint distribution of X and Y. This has intuitive appeal since we are choosing the model to maximize the likelihood of all possible future observations.

In the exponential family the motivation is strengthened if we use the Kullback-Leibler distance as the generalization of squared error. This measures the distance between densities; the distance between a model with true parameter η^* and one with parameter η is defined as $K(\eta^*, \eta) = E_{\eta *} \log h(Y, \eta^*)/h(Y, \eta)$. We regard this equivalently as a measure of distance between the two parameters η^* and η, or even the associated means μ^* and μ. The following decompositions, one for squared error, the other for Kullback-Leibler distance, are easily derived:

$$\begin{aligned} &E(Y - \mu(X))^2 \\ &\quad = E(Y - \mu^*(X))^2 + E(\mu^*(X) - \mu(X))^2, \\ &EK(Y, \mu(X)) \\ &\quad = EK(Y, \mu^*(X)) + EK(\mu^*(X), \mu(X)), \end{aligned} \tag{19}$$

where $\mu^*(X)$ is the true conditional mean. From (19) we see that if we minimize the expected Kullback-

Leibler distance from future observations $EK(Y, \mu(X))$, then we get the model $\mu(X)$ closest to $\mu^*(X)$. If $\mu(x)$ is unrestricted, the minimum is achieved at $\mu(x) = \mu^*(x)$. If the distribution is Gaussian, the Kullback-Leibler distance becomes squared error (*times* ½). Since $EK(Y, \mu(X)) = El(Y, Y) - El(Y, \mu)$, we see that this is equivalent to maximizing the expected log likelihood.

The use of expected log likelihood has also been suggested by Brillinger (1977) and Owen (1983). In what follows, we show that standard maximum likelihood estimation for generalized linear models, local scoring, and local likelihood estimation can all be viewed as methods for empirically maximizing the expected log likelihood.

5.2 Derivation of the Estimation Techniques Via Expected Log Likelihood

One way to use (18) for estimation of $\eta(x)$ would be to assume a simple form for $\eta(x)$, like $\eta(x) = \beta_0 + \beta_1 x$. We would then be finding the *linear* $\eta(x)$ closest in Kullback-Leibler distance to $\eta^*(x)$. The expectation in (18) could then be replaced by its sample analogue, and the resultant expression maximized over β_0 and β_1. This is nothing more than standard maximum likelihood estimation.

Now suppose (as is the point of this paper) that we don't want to assume a parametric form for $\eta(x)$. Differentiating (18) with respect to (the number) η we get

$$E(dl/d\eta \mid x)_{\hat{\eta}(x)} = 0, \tag{20}$$

assuming expectation and differentiation can be interchanged. Given some initial estimate $\eta(x)$, a first order Taylor series expansion gives the improved estimate

$$\eta^1(x) = \eta(x) - E(dl/d\eta \mid x)/E(d^2l/d\eta^2 \mid x) \tag{21}$$

or

$$\eta^1(x) = E\left[\eta(x) - \frac{dl/d\eta}{E(d^2l/d\eta^2 \mid x)} \,\middle|\, x\right]. \tag{22}$$

This provides a recipe for estimating $\eta(\cdot)$ in practice. Starting with some initial estimate $\eta(x)$, a new estimate is obtained using formula (22), replacing the conditional expectations by scatterplot smoothers. The data algorithm analogue is thus

$$\eta^1(x) = \mathbf{smooth}\left[\eta(x) - \frac{dl/d\eta}{\mathbf{smooth}[d^2l/d\eta^2]}\right]. \tag{23}$$

Since the variance of each of the terms in the brackets is approximately $\propto E(d^2l/d\eta^2)$, the *smoother* could use weights $\propto \mathbf{smooth}(d^2l/d\eta^2)^{-1}$ for efficient estimation. The data algorithm consists of repeated iterations of (23), stopping when the deviance fails to change by a small amount.

In the exponential family case, we can simplify (22) before replacing $E(\cdot \mid x)$ by **smooth**. We compute $dl/d\eta = (y - \mu)V^{-1}(d\mu/d\eta)$, $d^2l/d\eta^2 = (y - \mu)\cdot(d/d\eta)\,[V^{-1}(d\mu/d\eta)] - (d\mu/d\eta)^2V^{-1}$, and

$$E((d^2l/d\eta^2) \mid x) = -(d\mu/d\eta)^2V^{-1}.$$

Hence the update simplifies to

$$\eta^1(x) = E[\eta(x) + (Y - \mu)\,(d\eta/d\mu) \mid x]. \tag{24}$$

The data analogue is

$$\eta^1(x) = \mathbf{smooth}[\eta(x) + (y - \mu)\,(d\eta/d\mu)] \tag{25}$$

with weights $(d\mu/d\eta)^2V^{-1}$. This is exactly a smooth of the adjusted dependent variable, suggested on intuitive grounds in Section 4.

Note that we chose the form (22) instead of (21). In the case of distributions, they are the same because conditional expectation is a projection operator. Most smoothers are not projections and thus the two forms are not equivalent in the data case. We chose (22) because in the Gaussian case it simplifies to $\hat{\eta}(x) = \mathbf{smooth}[y]$ without any iteration, whereas (21) would require iteration even in this simple case.

The local likelihood procedure can also be viewed as an empirical method of maximizing $El(\eta(X), Y)$. Instead of differentiating this expression (as above), we write $El(\eta(X), Y) = E(E(l(\eta(X), Y) \mid X = x))$. Hence it is sufficient to maximize $E(l(\eta(X), Y) \mid X = x)$ for each x. The corresponding data recipe can be derived as follows. Consider estimating $\eta(x)$ at some point $x = x_i$. An estimate of $E(l(\eta(X), Y) \mid X = x_i)$ is

$$E(l(\eta(X), Y) \mid X = x_i) = (1/k_n) \sum_{j \in N_i} l(\eta(x_j), y_j), \tag{26}$$

where k_n is the number of data points in N_i. Assuming $\eta(x) \approx \beta_{0i} + \beta_{1i}x$ for points in N_i, (26) is then maximized over β_{0i} and β_{1i}. The resulting estimate, $\hat{\eta}(x_i) = \hat{\beta}_{0i} + \hat{\beta}_{1i}x_i$, is the local likelihood estimate as defined in Section 4.

The algorithms described here can be used in any likelihood-based regression model. As a technical point, note that in the exponential family, we linked the additive predictor $\eta = \sum_1^p s_j(X_j)$ to the distribution of Y via $\eta = g(\mu)$. In some nonexponential family models, μ is a complicated function of the model parameters or may not exist at all. It would then be desirable to link η to some other parameter of the distribution. This is true in the Cox model (see the next section). In any case, there is little difficulty—however, η is linked to the distribution of Y, the likelihood is some function of η and its derivatives are used in the updating formula.

To summarize so far, maximization of the expected log likelihood has led to a general technique for estimating a smooth covariate function: the local scoring procedure. In the case of the exponential family likelihood this procedure corresponds to smoothing of the adjusted dependent variable. Standard (linear) maximum likelihood estimation and local likelihood estimation can also be viewed as empirical maximizers of expected log likelihood. Equivalently they can all be viewed as empirical minimizers of the expected Kullback-Leibler distance between the model and the estimate.

So far we have not addressed the problem of multiple covariates—this will be done in Section 7.

6. SOME EXAMPLES

6.1. The Gaussian Model

For the Gaussian model with identity link, (25) simplifies to $\eta^1(x) = \mathbf{smooth}[y]$, and the local scoring algorithm reduces to a running lines smooth of y and x. The local likelihood procedure also gives the running lines smooth of y on x, since the local maximum likelihood estimate is $\hat{\eta}(x_i) = \hat{\beta}_{0i} + \hat{\beta}_{1i}x_i$, $\hat{\beta}_{0i}$ and $\hat{\beta}_{1i}$ being the least squares estimates for the points in N_i. The Gaussian model is applied to a large meterological data set in Hastie and Tibshirani (1984).

6.2. The Linear Logistic Model

A binomial response model assumes that the proportion of successes Y is such that $n(x)Y \mid x \sim \text{Bin}(n(x), p(x))$, where $\text{Bin}(n(x), p(x))$ refers to the binomial distribution with parameters $n(x)$ and $p(x)$. Often the data is binary in which case $n(x) \equiv 1$. The binomial distribution is a member of the exponential family with canonical link

$$g(p(x)) = \log \frac{p(x)}{(1 - p(x))} = \eta(x).$$

In the *linear logistic model* we assume $\eta(x) = \beta_0 + \beta_1 x$, and the parameters are estimated by maximum likelihood using Fisher scoring or equivalently by using adjusted dependent variable regression. The smooth extension of this model generalizes the link relation to $\log[p(x)/(1 - p(x))] = \eta(x)$. The local scoring step is

$$(27) \quad \eta^1(x) = \mathbf{smooth}\left[\eta(x) + \frac{y - p(x)}{p(x)(1 - p(x))}\right]$$

with weights $n(x)p(x)(1 - p(x))$. We now demonstrate the procedure 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 four variables.

$$y_i = \begin{cases} 1 & \text{if patient } i \text{ survived 5 years or longer,} \\ 0 & \text{otherwise,} \end{cases}$$

x_{i1} = age of patient i at time of operation,

x_{i2} = year of operation i (minus 1900),

x_{i3} = number of positive axillary nodes detected in patient i.

Figure 1 shows the response variable plotted against the covariate *age*. The solid nonlinear function was estimated using the local scoring method, with a span of .6. Now for a single covariate one could simply average the 0-1 response directly—this produced the dashed curve in the figure. It is identical with the function found using the local likelihood method fitting local constants to the logits. The local likelihood smooth fitting local straight lines (the more usual approach) is the dotted curve. The three nonparametric estimates are all similar, with bias affecting the running mean near the end points, and all give a different qualitative description of the data than the linear fit (almost straight curve). One can compare the linear logistic fit to any of the smooth estimates by examining the corresponding drops in deviance. For example, the local scoring estimate produced a deviance of 5.6 less than the linear logistic fit, while using only 1.6 more degrees of freedom (see Section 9), and hence the linear logistic fit is not adequate for these data.

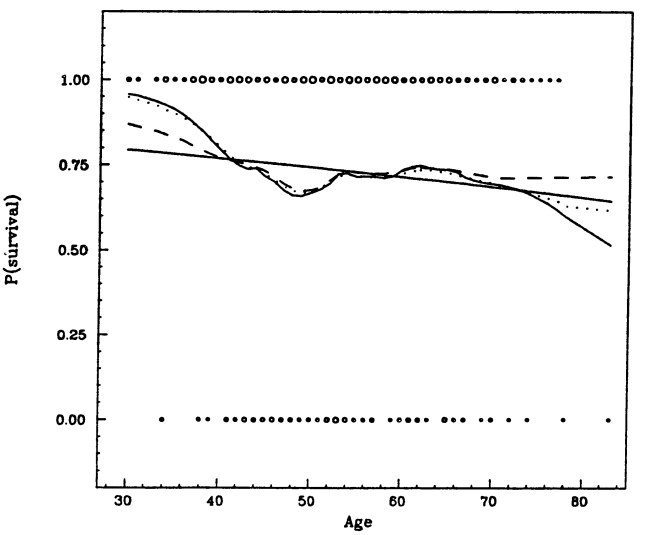

FIG. 1. *Survival of patients who underwent surgery versus age of the patient. The local scoring function is the solid curve, the local likelihood function is dotted, the running mean of the y's is dashed, and the linear logistic function is the almost straight curve. The area of the circles is proportional to the number of observations.*

We will see in the Section 7 that in fitting multiple covariate models, the approach of smoothing the response variable directly breaks down, whereas the local scoring and local likelihood techniques generalize easily. We will pursue this example in Section 7.

6.3. The Cox Model

The proportional hazards model of Cox (1972) is an example of a nonexponential family regression model. This model is used to relate a covariate to a possibly censored survival time. The data available are of the form $(y_1, x_1, \delta_1), \cdots, (y_n, x_n, \delta_n)$, the survival time y_i being complete if $\delta_1 = 1$ and censored if $\delta_i = 0$. We assume there are no ties in the survival times. The proportional hazards model assumes the hazard relation

$$\lambda(t \mid x) = \lambda_0(t)e^{\beta x}. \tag{28}$$

The parameter β can be estimated without specification of $\lambda_0(t)$ by choosing $\hat{\beta}$ to maximize the *partial likelihood* (PL)

$$\text{PL} = \prod_{i \in D} \frac{e^{\beta x_i}}{\sum_{j \in R_i} e^{\beta x_j}}. \tag{29}$$

In the above, D is the set of indices of the failures and $R_i = \{j \mid y_j \leq y_i\}$ the risk set just before the failure at y_i.

A more general model is

$$\lambda(t \mid x) = \lambda_0(t)e^{\eta(x)} \tag{30}$$

where $\eta(x)$ is a smooth function of x. One way to estimate $\eta(x)$ would be to apply the local scoring formula (23). Letting l equal the log partial likelihood and $C_i = \{k : i \in R_k\}$, (the risk sets containing individual i), straightforward calculations yield

$$\frac{\partial l}{\partial \eta(x_i)} = \delta_i - e^{\eta(x_i)} \sum_{k \in C_i} \frac{1}{\sum_{j \in R_k} e^{\eta(x_j)}} \tag{31}$$

and

$$\frac{\partial^2 l}{\partial \eta(x_i)^2} = -e^{\eta(x_i)} \sum_{k \in C_i} \frac{1}{\sum_{j \in R_k} e^{\eta(x_j)}} + e^{2\eta(x_i)} \sum_{k \in C_i} \frac{1}{(\sum_{j \in R_k} e^{\eta(x_j)})^2}. \tag{32}$$

Starting with say $\eta(x) = \hat{\beta}x$, smooths are applied to these quantities, as in (23), and the process is iterated.

The local likelihood technique can also be applied to the Cox model—this is described in Tibshirani (1984). We won't give details here. Instead, we'll illustrate the two estimation techniques with a real data example.

Miller and Halpern (1982) provide a number of analyses of the Stanford heart transplant data. The data, listed in their paper, consist of 157 observations of time to failure (months) and two covariates, age (years) and T5 mismatch score. Here we will consider only the age variable.

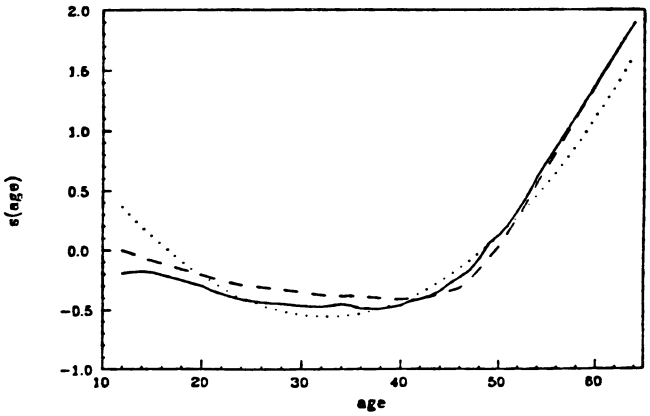

FIG. 2. *The Stanford heart transplant data. The solid curve is the local scoring function, the dashed line is the local likelihood function, and the dotted line is the proportional hazards quadratic fit.*

TABLE 1
Analysis of Stanford heart transplant data—age

Model	−2 log likelihood	Degrees of freedom
Null	902.40	0
Linear	894.82	1
Linear + quadratic	886.24	2
Local likelihood (span .5)	884.65	2.95
Local scoring (span .5)	884.66	2.95

Figure 2 shows the smooth obtained by local scoring (solid line) and local likelihood (broken line). Also shown is the fit obtained using a linear and quadratic term for age in a standard Cox analysis (dotted line). The smooths suggest that the relative risk stays about constant up to age 45, then rises sharply. The quadratic model forces a parametric shape on the function, and suggests (perhaps misleadingly) that the relative risk drops then rises. Table 1 summarizes the results of the various fitting procedures.

The approximate degrees of freedom or *number of parameters* of the model are discussed in Section 9. The table suggests that there is insufficient data to distinguish between the quadratic and smooth fits. This data set is analyzed more thoroughly in Tibshirani (1984).

7. MULTIPLE COVARIATES

When we have p covariates, represented by the vector $\mathbf{X} = (X_1, X_2, \cdots, X_p)$, a general model specifies $E(Y \mid \mathbf{X}) = \mu$ and $g(\mu) = \eta(\mathbf{X})$, where η is a function of p variables. We will first discuss the Gaussian case and show why it is necessary to restrict attention to an additive model.

We assume

$$Y = \eta(\mathbf{X}) + \varepsilon \tag{33}$$

where $\eta(\mathbf{X}) = E(Y \mid \mathbf{X})$, $\mathrm{Var}(Y \mid \mathbf{X}) = \sigma^2$, and the errors ε are independent of $\mathbf{X}$. The goal is to estimate $\eta(\mathbf{X})$. If we use the least squares criterion $E(Y - \eta(\mathbf{X}))^2$, the best choice for $\eta(\mathbf{X})$ is $E(Y \mid \mathbf{X})$. In the case of a single covariate, we estimated $E(Y \mid X)$ by a scatterplot smoother which in its crudest form is the average of those y_i in the sample for which x_i is close to x.

We could think of doing the same thing for multiple covariates: average the y_i for which x_i is close to x. However, it is well known that smoothers break down in higher dimensions (Friedman and Stuetzle, 1981); the *curse of dimensionality* takes its toll. The variance of an estimate depends on the number of points in the neighborhood. You have to look further for near neighbors in high dimensions, and consequently the estimate is no longer local and can be severely biased. This is the chief motivation for the additive model $\eta(\mathbf{X}) = s_0 + \sum_{j=1}^{p} s_j(X_j)$. Each function is estimated by smoothing on a single co-ordinate; we can thus include sufficient points in the neighborhoods to keep the variance of the estimates down and yet remain local in each co-ordinate. Of course, the additive model itself may be a biased estimate of the true regression surface, but hopefully this bias is much lower than that produced by high dimensional smoothers. The additive model is an obvious generalization of the standard linear model, and it allows easier interpretations of the contributions of each variable. In practice a mixture of the two will often be used:

$$\eta(\mathbf{X}) = s_0 + \sum_{j=1}^{q} s_j(X_j) + \sum_{j=q+1}^{p} \beta_j X_j. \tag{34}$$

7.1. Estimation—The Additive Regression Model

We now turn to the estimation of s_0, $s_1(\cdot)$, $\cdots$, $s_p(\cdot)$ in the additive regression model

$$E(Y \mid \mathbf{X}) = s_0 + \sum_{j=1}^{p} s_j(X_j), \tag{35}$$

where $Es_j(X_j) = 0$ for every j.

In order to motivate the algorithm, suppose the model $Y = s_0 + \sum_{j=1}^{p} s_j(X_j) + \varepsilon$ is in fact correct, and assume we know s_0, $s_1(\cdot)$, $\cdots$, $s_{j-1}(\cdot)$, $s_{j+1}(\cdot)$, $\cdots$, $s_p(\cdot)$. If we define the partial residual:

$$R_j = Y - s_0 - \sum_{k \neq j} s_k(X_k),$$

then $E(R_j \mid X_j) = s_j(X_j)$ and minimizes $E(Y - s_0 - \sum_{k=1}^{p} s_k(X_k))^2$. Of course we don't know the $s_k(\cdot)$'s, but this provides a way for estimating each $\hat{s}_j(\cdot)$ given estimates $\{\hat{s}_i(\cdot),\ i \neq j\}$. The resulting iterative procedure is known as the *backfitting* algorithm (Friedman and Stuetzle, 1981):

Backfitting Algorithm

Initialization: $s_0 = E(Y)$, $s_1^1(\cdot) \equiv s_2^1(\cdot) \equiv \cdots \equiv s_p^1(\cdot) \equiv 0$, $m = 0$.

Iterate: $m = m + 1$

for $j = 1$ to p do:

$$R_j = Y - s_0 - \sum_{k=1}^{j-1} s_k^m(X_k) - \sum_{k=j+1}^{p} s_k^{n-1}(X_k)$$

$$s_j^m(X_j) = E(R_j \mid X_j).$$

Until: $\mathrm{RSS} = E\left(Y - s_0 - \sum_{j=1}^{p} s_j^m(X_j)\right)^2$ fails to decrease.

In the above $s_j^m(\cdot)$ denotes the estimate of $s_j(\cdot)$ at the mth iteration. Notice that by effectively centering Y at the start, we guarantee that $Es_j^m(X_j) = 0$ at every stage. It is clear that RSS does not increase at any step of the algorithm and therefore converges. Breiman and Friedman (1985, Theorem 5.19) show in the more general context of the ACE algorithm that the solution $\sum s_j^\infty(X_j)$ is unique and is therefore the best additive approximation to $E(Y \mid \mathbf{X})$. This does not mean that the individual functions are unique, since dependence among the covariates can lead to more than one representation for the same fitted surface. These results do not depend on the validity of either the additive model for $E(Y \mid \mathbf{X})$ or the additive error assumption as in (33).

If we return to the world of finite samples, we replace the conditional expectations in the backfitting algorithm by their estimates, the scatterplot smooths. Brieman and Friedman have proved:

- For a restrictive (impractical) class of smoothers, the algorithm converges.
- For a less restrictive class, the procedure is mean square consistent in a special sense. Suppose that the mth iteration estimate of s_j is $\hat{s}_j^m$, where the *hat* implies it is a function of the sample size n. Let s_j^m be the estimate of s_j at the mth iteration of the algorithm applied to the distributions. Then $E(\hat{s}_j^m(X) - s_j^m(X))^2 \to 0$ as $n \to \infty$, with m fixed.

A special case arises if we use the least squares regression $\hat{a} + \hat{b}X_j$ to estimate $E(\cdot \mid X_j)$ at every stage of the algorithm. We can once again invoke the Breiman and Friedman results for this projection operator, which show that the algorithm converges to the usual

least squares estimate of the *multiple* regression of Y and $\mathbf{X}$. This is true for both the usual data estimate or the estimate in distribution space as in Section 5. Hastie and Tibshirani (1984) give an elementary proof of this fact due to Werner Stuetzle.

Although these results are encouraging, much work is yet to be done to investigate the properties of additive models. In multiple regression we need to worry about collinearity of covariates when interpreting regression coefficients; perhaps *cocurvity* has even worse implications when trying to interpret the individual functions in additive models. This would call for nonparametric analogues of linear principal components analysis—a standard device for determining lower dimensional *linear* manifolds in the data. Some work in this direction has been done (Hastie, 1984b; Young, Takane, and de Leuuw, 1978).

If the purpose of our analysis is prediction, these problems are less important. We proceed in an exploratory spirit, and hopefully a sound bed of theory will develop around these as yet unanswered questions.

7.2. Backfitting in the Local Scoring Algorithm

For multiple covariates the local scoring update (22) is given by

$$\eta^1(\mathbf{x}) = E\left[\eta(\mathbf{x}) - \frac{\partial l/\partial \eta}{E[\partial^2 l/\partial \eta^2 \mid \mathbf{x}]} \,\middle|\, \mathbf{x}\right] \tag{36}$$

and in exponential family case (24) is

$$\begin{aligned} \eta^1(\mathbf{x}) &= E[\eta(\mathbf{x}) + (Y - \mu)\ (\partial\eta/\partial\mu) \mid \mathbf{x}] \\ &= E(Z \mid \mathbf{x}) \end{aligned} \tag{37}$$

where $g(\mu) = \eta$ and Z is the adjusted dependent variable. For the reasons described in the previous section, we will restrict attention to an additive model:

$$\eta(\mathbf{X}) = s_0 + \sum_{j=1}^{p} s_j(X_j).$$

We see that (37) is of the same form as equation (35), with Z playing the role of Y. Thus to estimate the $s_j(\cdot)$'s, we fit an additive regression model to Z, treating it as the response variable Y in (35). The sum of the fitted functions is η^0 of the next iteration. This is the motivation for the *general local scoring algorithm* which we give for the exponential family case as in (37).

General Local Scoring Algorithm

Initialization: $s_0 = g(E(y))$, $s_1^0(\cdot) \equiv s_2^0(\cdot) \equiv \cdots \equiv s_p^0(\cdot) \equiv 0$, $m = 0$.

Iterate: $m = m + 1$

1. Form the adjusted dependent variable

$$Z = \eta^{m-1} + (Y - \mu^{m-1})(\partial\eta/\partial\mu^{m-1}),$$

where

$$\eta^{m-1} = s_0 + \sum_{j=1}^{p} s_j^{m-1}(X_j) \quad \text{and}$$

$$\eta^{m-1} = g(\mu^{m-1}).$$

2. Form the weights $W = (\partial\mu/\partial\eta^{m-1})^2 V^{-1}$.
3. Fit an additive model to Z using the backfitting algorithm with weights W, we get estimated functions $s_j^m(\cdot)$ and model η^m.

Until: $E\ \mathrm{dev}(Y, \mu^m)$ fails to decrease.

Step 3 of the algorithm is simply the additive regression backfitting algorithm (Section 7.1) with weights. Hastie and Tibshirani (1984, Appendix B) show why weights are required even in the distribution version of the algorithm. To incorporate them, the data is first transformed using the weights, and the backfitting algorithm is then applied to the transformed data.

From the results of the previous section, we see that the inner loop converges. In particular, if each smoother is replaced by the simple regression on the corresponding covariate (for data or distributions), the backfitting algorithm converges to the usual (weighted) multiple regression. This shows that in this case, the algorithm is identical with the usual GLM estimation procedure using Fisher scoring as in (12) and (13). Once again the data analogue of the algorithm replaces weighted conditional expectations by weighted smoothers. The span for each smoother is chosen by cross-validation as described in Section 4.2. Note that for nonexponential family models an additional backfitting step is required to compute the denominator of the second term in (36).

Stone (1986) has shown that under mild regularity conditions, a unique best additive approximation (in terms of Kullback-Leibler distance) exists for any exponential family model. We conjecture that the general local scoring algorithm converges to this best additive approximation.

It is important to stress the generality of the procedure. First, note that in either the backfitting algorithm or its generalization, different smoothers may be used for different covariates. As a simple example, a linear least squares fit would be used to "smooth" a binary covariate or a continuous covariate for which a linear fit was desired. Other possibilities might include a periodic smoother for a covariate like day of

the week, or a smoother that forces monotonicity (Friedman and Tibshirani, 1984). Secondly, interactions can be incorporated by defining a new covariate to be the product of two or more covariates, then smoothing on the new covariate. Interactions involving categorical covariates can be handled with dummy variables in the usual way.

The backfitting idea is also used in the local likelihood estimation procedure to incorporate multiple covariates. To estimate a new $s_j(\cdot)$, or to adjust $s_j(\cdot)$ for other $s_k(\cdot)$'s in the model, $s_j(\cdot)$ is re-estimated holding all others fixed. The algorithm cycles through the functions until convergence. The details can be found in Tibshirani (1984).

7.3. The Breast Cancer Example Continued

We continue our analysis of the breast cancer data using all three covariates. The model is now $\log [p(\mathbf{x})/(1-p(\mathbf{x}))] = s_0 + \sum_{j=1}^{3} s_j(x_j)$. This is preferable to modeling $p(\mathbf{x})$ by an additive sum, since we would have to check that the estimated probabilities are positive and add to 1; the logit transform allows our estimates to be unrestricted. There are other reasons for using the logit transform; on the logit scale prior probabilities appear only as an additive constant (McCullagh and Nelder, 1983, page 78). This is useful in biomedical problems where there is often some established population risk, and the problem is to see what factors modify this risk for the sample under study.

Table 2 summarizes the various models fitted (by local scoring). The approximate degrees of freedom (dof) or *number of parameters* of the model are discussed in Section 9. *Auto* in the column labeled spans indicates that each time a smooth was computed, the span was selected by cross-validation. The entry D^2 refers to the percentage of deviance explained and is in direct analogy to the more familiar R^2 in regression.

TABLE 2
The analysis of deviance (ANODEV) table for the breast cancer data

Model	Spans	Degrees of freedom	Deviance	D^2
Constant	1		353.67	
x_1, x_2, x_3	All linear	4	328.75	.07
x_1, x_2, x_3	All .5	8.8	307.89	.13
x_1, x_2, x_3	Auto	8.0	308.22	.13
x_1, x_3	Auto	5.9	317.66	.10
x_2, x_3	Auto	5.0	312.68	.12
x_2, x_2	Auto	4.1	346.71	.02
Parametric		7	302.30	.15

Figures 3, 4, and 5 show the estimated functions for our model with deviance 308.22 and dof = 8.8.

Landwehr, Pregibon, and Shoemaker (1984) analyzed this data set and in particular considered partial residual plots in order to identify the functional form

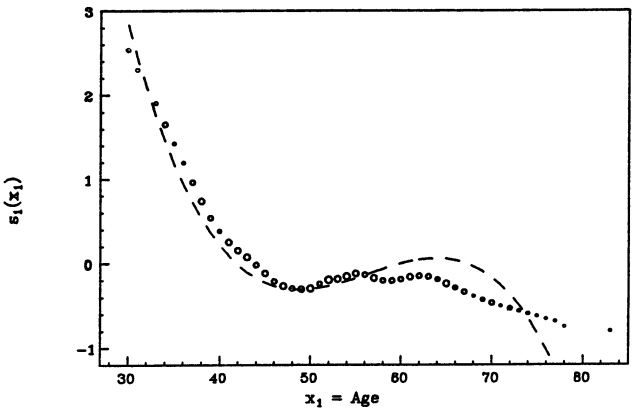

FIG. 3. *The circles represent* $\hat{s}$(age), *where the area of the circles is proportional to the number of points. The dashed term is the cubic polynomial term in* (38).

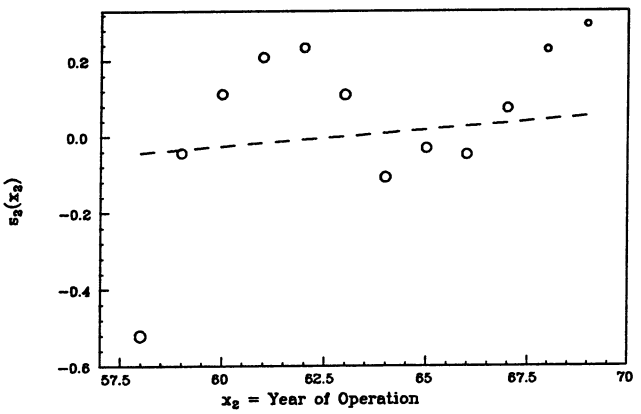

FIG. 4. *The circles represent* $\hat{s}$(year of operation). *The dashed term is the linear term in* (38).

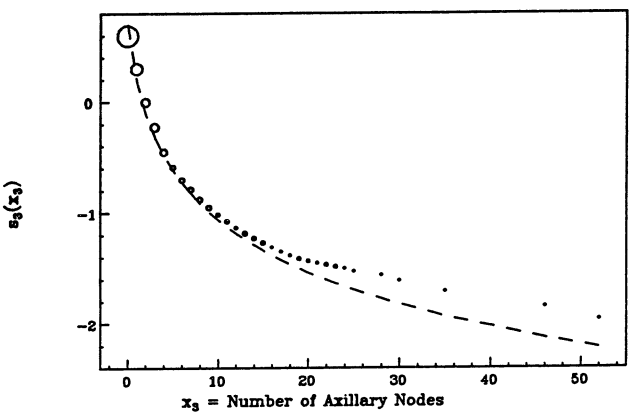

FIG. 5. *The circles represent* $\hat{s}$(number of positive axillary nodes). *The dashed term is the log term in* (38).

for each covariate. Their final model was

$$\text{(38)}\quad \begin{aligned}\operatorname{logit} p(\mathbf{x}) &= \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \beta_4 x_2 \\ &\quad + \beta_5 x_1 x_2 + \beta_6(\log(1 + x_3))\end{aligned}$$

with a deviance of 302.3 on 299 degrees of freedom. We fit this model in two ways: using a) GLIM and b) the backfitting procedure with linear fits for the transformed variables. As expected, the results agreed (up to four significant figures). This model is labeled *parametric* in the table. We have superimposed the parametric model terms in the figures, and note that the functions are very similar. If $\hat{\boldsymbol{\beta}}\mathbf{x}_i$ is the estimated linear model, and $\hat{p}_i$ the corresponding probability estimate, the partial residual for variable j and observation i is defined by

$$\text{(39)}\quad r(x_{ij}) = \hat{\beta}_j x_{ij} + \frac{y_i - \hat{p}_i}{\hat{p}_i(1 - \hat{p}_i)}.$$

Landwehr, Pregibon, and Shoemaker (1984) show that if the true model is

$$\log[p(\mathbf{x})/(1 - p(\mathbf{x}))] = \beta_0 + \sum_{k \neq j} \beta_k x_k + s_j(x_j),$$

and $s_j(\cdot)$ in linear, then $E[r(X_j) \mid X_j = x] \approx s_j(x)$. Thus they use the smooth of the partial residuals to suggest the functional form. This result breaks down if the other terms are not linear (Hastie, 1984a; Fienberg and Gong, 1984). One can see from the previous section that smoothing the partial residual corresponds to the first step of the general local scoring procedure in the local scoring algorithm, if our starting guess is the linear model. The local scoring procedure continues, however, by simultaneously estimating and adjusting nonparametric functions for all the covariates.

8. COMPARISON OF LOCAL SCORING TO LOCAL LIKELIHOOD ESTIMATION

In a number of examples that we have tried, the local scoring and local likelihood procedures give very similar results. This is not surprising in light of the discussion of Section 5, where we saw that both techniques are based on empirical estimates of the expected log likelihood. The difference seems to be in computational speed: local scoring is $O(n)$ while local likelihood, if the span increases like n^c, is $O(n^{c+1})$. For large data sets, the local scoring procedure is considerably faster. This leads us to ask: will the two procedures always give similar estimates? Artificially, they could be made very different. The reason for this is as follows. For a single covariate, the local likelihood procedure is completely local; that is if x_j is not in the neighborhood for estimating $s(x_i)$, then (x_j, y_j) has absolutely no effect on the estimate $\hat{s}(x_i)$. This is not true in the local scoring procedure, for as the smooth operation is iterated, the estimates $\hat{s}(x_j)$ enter into the computation of $\hat{s}(x_i)$. Thus sending y_j off to $+\infty$ would have a large effect on the estimate of $\hat{s}(x_i)$ in the smooth updating procedure, but no effect in the local likelihood procedure.

Given the theoretical basis of Section 5, it seems eminently reasonable that the two procedures be asymptotically equivalent in some sense. In Hastie and Tibshirani (1984) we sketch a proof of this fact for exponential families.

For finite samples, we can describe operationally the difference as follows, using logistic regression as an example. Suppose we start with $p(x_i) = \bar{y}$, the overall proportion of 1's. Then the first iteration for both procedures is identical:

- Local scoring regresses $z_j = \log[p(x_j)/(1 - p(x_j))] + (y_j - p(x_j)/(p(x_j)(1 - p(x_j)))$ on x_j for $j \in N_i$, with weights $p(x_j)(1 - p(x_j))$, to obtain the estimate $\eta^1(x_i)$; this is the local linear smoother operation in this neighborhood.
- Local likelihood does exactly the same operation in computing the maximum likelihood estimate (MLE) in the neighborhood, since this is the first step in the adjusted variable regression procedure used to compute the MLE.

The second iterations are very similar:

- Local scoring regresses

$$z_j = \eta^1(x_j) + (y_j - p^1(x_j))/(p^1(x_j)(1 - p^1(x_j)))$$

with weights $p^1(x_j)(1 - p^1(x_j))$ against x_j for $j \in N_i$ to obtain the estimate $\eta^2(x_i)$.

- Local likelihood, however, regresses

$$z_j = \eta_i^1(x_j) + (y_j - p_i^1(x_j))/(p_i^1(x_j)(1 - p_i^1(x_j)))$$

against x_j with weights $p^1(x_j)(1 - p^1(x_j))$, where $\eta_i^1(x_j)$ refers to the extrapolated value of η^1 at x_j derived from the linear estimate $\eta^1(x_j) = \hat{\beta}_{0i} + \hat{\beta}_{1i} x_j$.

If the function is fairly linear in the neighborhood then these two steps will yield similar estimates. For a given point x_j, the local scoring algorithm uses its latest estimate of $p(x_j)$ for every neighborhood in which x_j appears. The local likelihood procedure, however, uses a linear approximation (on the η scale) for $p(x_j)$ based on its estimate $p(x_i)$ for $j \in N_i$.

9. DISCUSSION

Generalized additive models provide a flexible method for identifying nonlinear covariate effects in exponential family models and other likelihood-based regression models. In the two data examples given in this paper, we utilized a degrees of freedom estimate to assess the importance of covariates. This is based on the expected decrease in the deviance due to

smoothing, computable from the trace of the appropriate "smoother matrix." We give details in Tibshirani (1984) and Hastie and Tibshirani (1984). In Hastie and Tibshirani (1985a and references therein) we also provide a method for computing confidence bands for the smooths.

There are a number of ways that the setup can be further generalized. For example, the local scoring algorithm can be extended to provide nonparametric estimation of the link function $g(\mu)$. Usually, $g(\mu)$ is assumed to be known; for example, in the linear logistic model, $g(\mu) = \log[\mu/(1 - \mu)]$. The generalization allows $g(\mu)$ to be estimated nonparametrically and hence provides a check of two of the assumptions inherent in linear logistic modeling: the linear form for the covariates and the logit link. Details may be found in Hastie and Tibshirani (1984). The local scoring procedure can also be generalized to fit a smooth version of McCullagh's (1980) model for ordinal data, analogous to the extension of the linear logistic model described here. In its most general form, the algorithm can be applied to any situation in which a criterion is optimized involving one or more smooth functions. We discuss this in Hastie and Tibshirani (1985b).

In the local scoring procedure we have used a running lines smoother, but we noted that other smoothers could be used. Cubic splines are a popular technique for smoothing and would be an interesting alternative. Wahba (1980) discusses the use of two-dimensional "thin-plate splines" for estimating response surfaces. These are more general than the additive model but are more difficult to interpret. O'Sullivan, Yandell, and Raynor (1984) look at splines for general exponential family models. Analogous to the Gaussian case, they emerge as the solution to a penalized likelihood problem. Again, an additive model is not considered; instead, a general surface is fitted. Green and Yandell (1985) propose similar techniques, with an emphasis on semiparametric models. Stone and Koo (1986a) investigate the use of additive B-splines for exponential family models.

The computations in this paper were performed using the GAIM (generalized additive interactive modeling) package, available upon request from either author. Also available from the authors are a GAIM function for the *S* statistical language and a special version of GAIM for the IBM PC.

ACKNOWLEDGMENTS

We would like to thank Arthur Owen for his ideas on expected log likelihood and degrees of freedom, and the Editor and Associate Editor for comments that improved the presentation. Support for this work from the Department of Energy, Office of Naval Research, and the United States Army Research Office is gratefully acknowledged. R. Tibshirani would also like to thank the Natural Sciences and Engineering Research Council of Canada for its support.

REFERENCES

BREIMAN, L. and FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80** 580–597.

BRILLINGER, D. (1977). Discussion of Consistent nonparametric regression, by C. J. Stone. *Ann. Statist.* **5** 622–623.

CLEVELAND, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74** 829–836.

COX, D. R. (1972). Regression models and life tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–202.

FIENBERG, S. and GONG, G. (1984). Discussion of Graphical methods for assessing logistic regression models, by J. M. Landwehr, D. Pregibon, and A. C. Shoemaker. *J. Amer. Statist. Assoc.* **79** 72–77.

FRIEDMAN, J. H. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823.

FRIEDMAN, J. H. and STUETZLE, W. (1982). Smoothing of scatterplots. Tech. Rept. Orion 3, Dept. of Statistics, Stanford Univ.

FRIEDMAN, J. H. and TIBSHIRANI, R. (1984). The monotone smoothing of scatterplots. *Technometrics* **26** 243–250.

GREEN, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation and some robust and resistant alternatives (with discussion). *J. Roy. Statist. Soc. Ser. B* **46** 149–192.

GREEN, P. J. and YANDELL, B. (1985). Semi-parametric generalized linear models. In *Generalized Linear Models. Lecture Notes in Statist.* (R. Gilchrist, B. Francis, and J. Whittaker, eds.) **32** 44–55. Springer, New York.

HABERMAN, S. J. (1976). Generalized residuals for log-linear models. *Proc. 9th Int. Biometrics Conference, Boston,* 104–122, Biometric Soc., Raleigh, N.C.

HASTIE, T. (1984a). Discussion of Graphical methods for assessing logistic regression models, by J. M. Landwehr, D. Pregibon, and A. C. Shoemaker. *J. Amer. Statist. Assoc.* **79** 77–78.

HASTIE, T. (1984b). Principal curves and surfaces. Tech. Rept. Orion 24 and unpublished Ph.D. dissertation, Dept. of Statistics, Stanford Univ.

HASTIE, T. and TIBSHIRANI, R. (1984). Generalized additive models. Tech. Rept. 98, Dept. of Statistics, Stanford Univ.

HASTIE, T. and TIBSHIRANI, R. (1985a). Generalized additive models: some applications. In *Generalized Linear Models. Lecture Notes in Statist.* (R. Gilchrist, B. Francis, and J. Whittaker, eds.) **32** 66–81. Springer, New York.

HASTIE, T. and TIBSHIRANI, R. (1985b). Smoothing in likelihood-based regression models. Tech. Rept. 1985-002, Dept. of Preventive Medicine and Biostatistics, Univ. of Toronto.

LANDWEHR, J. M., PREGIBON, D. and SHOEMAKER, A. C. (1984). Graphical methods for assessing logistic regression models (with discussion). *J. Amer. Statist. Assoc.* **79** 61–81.

MCCULLAGH, P. (1980). Regression models for ordinal data. *J. Roy. Statist. Soc. Ser. B* **42** 109–142.

MCCULLAGH, P. and NELDER, J. (1983). *Generalized Linear Models.* Chapman and Hall, London.

MILLER, R. G. and HALPERN, J. (1982). Regression with censored data. *Biometrika* **69** 521–531.

NELDER, J. A. and WEDDERBURN, R. W. M. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A* **135** 370–384.

O'SULLIVAN, F., YANDELL, B. and RAYNOR, W. (1984). Automatic smoothing of regression functions in generalized linear models.

Tech. Rept. 734, Dept. of Statistics, Univ. of Wisconsin, Madison.

OWEN, A. (1983). The estimation of smooth curves. Unpublished manuscript.

REINSCH, C. (1967). Smoothing by spline functions. *Numer. Math.* **10** 177–183.

SILVERMAN, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *J. Roy. Statist. Soc. Ser. B* **47** 1–52.

STONE, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14** 590–606.

STONE, C. J. and KOO, C.-Y. (1986a). Additive splines in statistics. *Proc. Statist. Comp. Sect. Amer. Statist. Assoc.*

TIBSHIRANI, R. (1984). Local likelihood estimation. Stanford technical report and unpublished Ph.D. dissertation, Dept. of Statistics, Stanford Univ.

WAHBA, G. (1980). Spline bases, regularization, and generalized cross-validation for solving approximation problems with large quantities of noisy data. *Proc. Conf. on Approximation Theory in Honour of George Lorenz, Jan. 8–10, Austin, Texas* (W. Chaney, ed.). Academic, New York.

WAHBA, G. and WOLD, S. (1975). A completely automatic French curve: Fitting spline functions by cross-validation. *Comm. Statist.* **4** 1–7.

YOUNG, F. W., TAKANE, Y. and DE LEUUW, J. (1978). The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. *Psychometrika* **43** 279–282.

Comment

David R. Brillinger

> "All considered, it is conceivable that in a minor way, nonparametric regression might, like linear regression, become an object treasured for both its artistic merit as well as usefulness."
>
> L. Breiman (1977)

This paper by Hastie and Tibshirani lays bare the insight of the above remark of Leo Breiman made in the course of the discussion of a seminal work on regression with smooth functions (Stone, 1977). Here Hastie and Tibshirani increase the store of both artistic merit and usefulness by plugging nonparametric regression into the generalized linear model and by alluding to a variety of possible further extensions. It all makes being a statistician these days a joy—it seems approaches are now available to attack most any applied problem that comes to hand. (Understanding the operational performance of those approaches is clearly another matter however.)

It was nice to be asked to comment on such a stimulating paper. I have divided my comments into several sections, striving to focus on individual strains present in the paper, believing that future research on those strains will proceed at different rates.

1. STRUCTURE OF A BASIC PROBLEM

One has data $(Y_i, \mathbf{X}_i)$, $i = 1, \cdots, n$, with n moderately large. One is willing to consider a model for the individual Ys wherein: i) the conditional distribution of Y given $\mathbf{X}$ belongs to an exponential family, ii) it involves $\mathbf{X}$ only through $\eta = \sum s_j(X_j)$ with the $s_j(\cdot)$ unknown, but smooth, and iii) $E\{Y \mid \mathbf{X}\} = h(\sum s_j(X_j))$, with $h(\cdot)$ known. The parameter of the model is $\theta = \{s_j(\cdot), j = 1, \cdots, p\}$, and possibly a scale. The two key elements of the model are a) that the $s_j(\cdot)$ are smooth and b) that $\sum s_j(X_j)$ is additive.

It is to be noted that this model continues the contemporary statistical trend to eliminate distinctions between the cases of finite and infinite dimensional θ or between discrete and continuous data.

The problem is of interest, for one may wish to make inferences from the data via the model or one may wish to validate a model with a low dimensional parameter by imbedding it in a broader model, for example.

2. CONSTRUCTION OF ESTIMATES

To begin, focus on estimating $\eta = \eta(\mathbf{X})$, via a relationship that characterizes the true value η_0. Suppose one has a function $\rho(Y \mid \eta)$ such that $E_0\{\rho(Y \mid \eta) \mid \mathbf{X}\}$ is maximized at $\eta = \eta_0$. An example would be $\log f(Y \mid \eta)$, $f(\cdot)$ denoting the conditional density of Y. Alternately, suppose one has a function $\psi(Y \mid \eta)$ such that $E_0\{\psi(Y \mid \eta) \mid \mathbf{X}\} = \mathbf{0}$ at $\eta = \eta_0$. An example would be $\partial \log f(Y \mid \eta)/\partial\eta$. Estimates of the true η_0 may be constructed by paralleling these relations on the data. For example, given weights $W_{ni}(\mathbf{X})$ such as in Stone (1977) one might take $\hat{\eta}$ to maximize

$$\sum_i \rho(Y_i \mid \hat{\eta}) W_{ni}(\mathbf{X})$$

or to satisfy

$$\sum_i \psi(Y_i \mid \hat{\eta}) W_{ni}(\mathbf{X}) = \mathbf{0}.$$

The estimate of Hastie and Tibshirani based on (26) takes this form. One can expect such estimates to be

David R. Brillinger is Professor of Statistics, University of California, Berkeley, California 94720.

consistent under regularity conditions. Stone (1977, page 643) gives some simple conditions. Such estimates were called conditional M-estimates by Brillinger (1977) and it was remarked there that one could form robust estimates directly (by limiting the influence of individual observations for example). It is further clear that partial likelihood estimates, censored data estimates, and unequal probability of selection estimates are particular cases.

The critical advance of Hastie and Tibshirani is to look for extrema with η of the form $\sum s_j(X_j)$. They limit consideration to likelihood- and partial likelihood-based estimates, but it is clear that they could go on to form for example robust-resistant ones by choice of ρ or ψ.

It is further apparent that were the dimension of $\mathbf{X}$, p, unclear one could add an Akaike type term in p and estimate p as well. Continuing, this makes it apparent that penalized maximum likelihood estimates also may be fit into this general setup. We have here a type of inverse unstable problem. These are often solved by forms of regularization (smoothing). It is perhaps worth remarking that the first approach above is a form of Courant regularization, while penalized likelihood would correspond to Tihonov regularization. (These techniques are discussed in Allison (1979).)

There is much insight in Hastie and Tibshirani's remark that because of the additivity of η in the $s_j(\cdot)$, the smoothing need not be local (in the $\mathbf{X}$-space).

3. COMPUTATIONS

In the next few years, the structure set out in the preceding section may not be expected to change too much. This is probably not true for the algorithms numerically determining the extrema.

Hastie and Tibshirani propose an iteratively reweighted least squares solution, as in GLIM, interwoven with a stepwise selection procedure as in Breiman and Friedman (1985). My experience with such algorithms is that they are troubled by initial values, precision/round-off, convergence criterion, underflow/overflow, and instability among other things. Nonlinear iterations can do strange things. In particular I expect better algorithms for determining the components of $\sum s_j(\cdot)$ to be developed.

4. SOME QUIBBLES

I do have some disagreements with the paper. In the abstract, it is stated: "It has the advantage of being completely automatic . . ." I see this as both a disadvantage and not true. A disadvantage because surely one wants flexible analyses. Not true because someone (the programmer?) has made many choices: machine precision, convergence criterion, smoother, ... The analyst will not know these choices at his peril.

In Section 7 it is stated: "This is the chief motivation for the additive model." The reason given is a statistical one. To my mind the motivation is substantive. Additivity is basic to science (see Luce and Tukey, 1964, particularly the references therein).

Two medical data sets are analyzed, but no inferences are made. Can the authors not set down some (biological) insight or understanding that has been gained from the analyses? Otherwise they might have just as well presented the results of simulations.

5. FURTHER ISSUES AND PROBLEMS

In this section I am not complaining about possible omissions from the present paper, rather I am interested in the authors' thoughts regarding future directions of work. The paper certainly stands on its own.

The statistical properties of the estimates need to be understood. What are they actually estimating in the case of a finite sample? In time series we know that the conventional spectrum estimate is estimating an average of the power spectrum, albeit concentrated near the frequency of concern. Is that the case here or are remote values influential? The time series case further suggests the possible utility of pretransforming the Xs to reduce bias.

The sampling variability of the estimates need to be assessed. Could the authors indicate their preferred technique. Mine would be a jackknife variant, because of its bias reducing properties and nonmodel dependence. There is a need for goodness of fit/validation procedures, diagnostics, measures of influence.

In power spectrum estimation, I do not generally take the same bandwidth for all frequencies in the conventional estimate (and more complex estimators have a similar effect). Here the span is taken to be the same. Have the authors thought of making it variable?

Smoothness is essential in the development in the paper. Yet many natural relationships are discontinuous and even multivalued. It would seem appropriate to develop techniques for such situations. For the former, perhaps one would smooth only when an estimate of the derivative is small.

6. A QUESTION

In Section 1, the authors refer to the ACE procedure of Breiman and Friedman (1985) as a means of determining a transform of the *dependent* variable. This involves maximizing a correlation. In Section 9, they refer to the use of local scoring (i.e., a likelihood-based technique) for the analogous problem of determining the link function. The two criteria are quite different seemingly. Can the authors comment? I wonder about yet another alternative, namely picking the transformations to maximize a nonparametric estimate of the

mean information in $\eta(\mathbf{X})$ about $\theta(Y)$. (This does not involve a Jacobian.)

ADDITIONAL REFERENCES

ALLISON, H. (1979). Inverse unstable problems and some of their applications. *Math. Sci.* **4** 9–30.

BREIMAN, L. (1977). Discussion of Consistent nonparametric regression, by C. J. Stone. *Ann. Statist.* **5** 621–622.

LUCE, R. D. and TUKEY, J. W. (1964). Simultaneous conjoint measurement: a new type of fundamental measurement. *J. Math. Psych.* **1** 1–27.

STONE, C. J. (1977). Consistent nonparametric regression (with discussion). *Ann. Statist.* **5** 595–645.

Comment

J. A. Nelder

I congratulate the authors on a fascinating piece of work and offer three comments.

1. In order to make smoothing work it is necessary to restrict it to one-dimensional covariate spaces, hence the strong assumption of additivity. In principle one could introduce cross-terms, e.g., have $x_{12} = x_1 x_2$, as well as x_1 and x_2, in the model; however, I suspect the convergence of the algorithm might now become immensely slow or even nonexistent because of the functional relations between the covariates. An alternative might be to include a term of the form $s_1(x_1) \cdot s_2(x_2)$, with coefficient to be estimated. Have the authors any comments on this problem?

J. A. Nelder is Visiting Professor, Department of Mathematics, Imperial College, 180 Queen's Gate, London SW7, England.

2. To me it seemed intuitively surprising that the figures in Table 2 show the generalized additive model to have one parameter more than the original parametric one, but a deviance nearly 6 higher. I then realized that the latter has a cross-term in it, and this appears to be important. What would be the effect of adding a term in $s_1(x_1) \cdot s_2(x_2)$ to the former? Also it would help interpretation if the difference in deviance were given when each term in their model was replaced by a parametric form. This would give summary statistics for differences visible in Figures 3, 4, and 5.

3. The new version of GLIM (3-77) now available has a facility for inserting new code. I very much hope that the authors can be persuaded to exploit this in order to make available the fitting of generalized additive models in GLIM.

Comment

Charles J. Stone

Hastie and Tibshirani deserve commendation for the originality, significance, and interest of their approach and the excellent expository review in the present paper.

Recently I have been working on a different approach to fitting more or less the same class of models, but using polynomial cubic splines to model the component functions $s_j(\cdot)$ and the Newton–Raphson method to calculate the ordinary maximum likelihood estimate. In order to avoid artificial end effects of polynomial fits such as those shown in Figures 2 and 3, the splines are constrained to be linear to the left of the first knot and to the right of the last knot. To avoid multiple representations of the constant term, zero sum constraints are imposed on the individual terms (when $p \geq 2$), as is done in this paper. Thus, if there are N knots, there are $N + 4$ degrees of freedom for the unconstrained spline and $N - 1$ degrees of freedom for the constrained spline. There is also 1 degree of freedom for the constant term; so there are $(N - 1)p + 1$ degrees of freedom in total. This approach will be referred to as the parametric spline approach to distinguish it from the smoothing spline approach favored by Wahba and others in which smoothing is achieved by a roughness penalty instead of by confining attention to spline models with a modest number of degrees of freedom. In theory, N should tend to infinity as the sample size n tends to

Charles J. Stone is Professor of Statistics, University of California, Berkeley, California 94720.

infinity so as to achieve the optimal rate of convergence (see Stone (1985, 1986)). Asymptotically optimal rules for selecting N based on the data have been obtained by Burman (1985). In practice $N = 5$ has proven sufficient. This is not surprising, since the standard linear approach allows only 1 degree of freedom per component function. Allowing 4 degrees of freedom should provide enough flexibility to fit the regular departures from nonlinearity that are likely to occur in practice, especially when linear constraints are used in the tails. For the flexibility is then highest in that portion of the axis that contains the bulk of the data. Linear restrictions on splines lead to tail behavior very similar to that of the linear smoothers (local linear regression) recommended by Stone (1975, 1977), Cleveland (1979), Friedman and Stuetzle (1981), and this paper. As Hastie and Tibshirani and others have pointed out, it is desirable to have a reasonable automatic default rule. The rule that has emerged from Stone and Koo (1986b) is this: given a specific covariate, order its observed values as $x_{(1)}, \cdots, x_{(n)}$; put knots at the minimum value $x_{(1)}$ and maximum value $x_{(n)}$; put additional knots at $x_{(i)}$, $i = i_2, i_3, i_4$, chosen so that the logits of $1/(n + 1)$, $i_2/(n + 1)$, $i_3/(n + 1)$, $i_4/(n + 1)$, $n/(n + 1)$ are approximately equally spaced.

In the few cases where the two approaches have been applied to the same data, the resulting curves appeared visually to be quite similar (see Stone and Koo, 1986a, and Devlin and Weeks, 1986), except that the approach of Hastie and Tibshirani leads to small scale roughness not present in curve estimates obtained by the parametric spline approach. The approaches seem equally feasible numerically and equally automatic. But the parametric spline approach has several conceptual advantages. In particular, the standard maximum likelihood method can be used to estimate the parameters and obtain confidence intervals that are asymptotically valid, at least when N is fixed. The χ^2 approximation to the asymptotic distribution of the logarithm of likelihood ratio statistics is also asymptotically valid with an integral number of degrees of freedom. The theory is analytically tractable even when $N \to \infty$ as $n \to \infty$, provided that the covariates are restricted to a compact set. Undoubtedly, Hastie and Tibshirani could site advantages for their approach.

In order to carry out the asymptotics for the parametric spline approach when $N \to \infty$ as $n \to \infty$, it seems necessary that the log likelihood function be strictly concave. Such concavity holds in generalized additive models when $\eta = \theta$ and for some other choices of the link function (such as that corresponding to probit models), but it is not true for an arbitrary link function. Strict concavity is desirable even when N is fixed, for it guarantees that the log likelihood function have at most one local maximum and that a local maximum, if it exists, be the unique global maximum. Hastie and Tibshirani do not explicitly mention strict concavity, which in their notation amounts to the requirement that $d^2l/d\eta^2 < 0$. Even without this requirement, it is true that $E(d^2l/d\eta^2 \mid x) < 0$, but perhaps the algorithm in (23) of this paper is more reliable when the log likelihood function is strictly concave.

Generalized additive modeling as studied by Hastie and Tibshirani, by Burman, and by myself is an extension of the generalized linear models (GLMs) introduced by Nelder and Wedderburn (1972). But, so far at least, one limitation of GLMs has been preserved; namely, the restriction to exponential models that involve a one-dimensional parameter θ. The most obvious practical advantage of considering a multidimensional parameter θ is that the setup would then include multinomial models for conditional distributions and thereby allow for categorical response variables Y having more than two possible categories. Once covariates are included we have a natural setup for developing reasonable and flexible multiple classification procedures. In the linear form of the model, each coordinate of θ would be a linear function of the covariates. In the additive extension, each coordinate would be an additive function of the covariates. Ideally, the fitting procedure should be such that the estimated conditional probabilities of the various categories are positive and sum to one. This can undoubtedly be done with the parametric spline approach. Can it also be done with the approach of Hastie and Tibshirani?

In the present paper, Hastie and Tibshirani treat Cox's proportional hazards model as being outside the framework of GLMs. However, the logarithm of the partial likelihood is of the form log-PL $= \sum_{i \in D} [\beta x_i - \log(\sum_{j \in R_i} e^{\beta x_j})]$. For each i, the expression enclosed by brackets is exactly in the form of a multinomial model, there being as many categories as there are elements in the risk set R_i. Thus the setup is essentially that of independent but not identically distributed multinomial response experiments. In particular, log-PL is a strictly concave function of the unknown parameters, so the parametric spline approach should also be viable. But the asymptotics, especially when $N \to \infty$ as $n \to \infty$, have yet to be worked out.

ADDITIONAL REFERENCES

BURMAN, P. (1985). Estimation of generalized additive models. To appear in *J. Multivariate Anal.*

DEVLIN, T. F. and WEEKS, B. J. (1986). Spline functions for logistic regression modeling. In *Proceedings of the Eleventh Annual SAS Users Group International Conference*, 646–651. SAS Institute, Inc., Cary, N. C.

STONE, C. J. (1975). Nearest neighbor estimators of a nonlinear regression function. In *Proceedings of Computer Science and*

Statistics: 8th Annual Symposium on the Interface, 413–418. Health Sciences Computer Facility, UCLA.

STONE, C. J. (1977). Consistent nonparametric regression (with discussion). *Ann. Statist.* **5** 595–645.

STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705.

STONE, C. J. and KOO, C.-Y. (1986b). *Function Estimates.* AMS Contemporary Math. Ser., Amer. Math. Soc., Providence, R. I.

Comment

Peter McCullagh

Hastie and Tibshirani are to be congratulated for presenting the theory and methodology of generalized additive models in a form that keeps incidental mathematical details at an acceptably low level. I have little to add and my single comment is therefore brief.

The whole thrust of the authors' development seems to be based implicitly on the following assumption, here reduced to the bare essentials: zero interaction is fundamentally more plausible than componentwise linearity in the covariates. Has there been any attempt to justify this point of view, either philosophically or empirically by examining a large number of examples or by any other means? A closely related question concerning statistical strategy is the following: at what stage of analysis does the assumption of zero interaction come under scrutiny?

Peter McCullagh is Professor of Statistics, Department of Statistics, The University of Chicago, 5734 University Avenue, Chicago, Illinois 60637.

Rejoinder

Trevor Hastie and Robert Tibshirani

1. THE GENERAL PROBLEM

In Section 5 of the paper, we motivated the local scoring and local likelihood estimation procedures as empirical methods for maximizing $E(l(\eta(X), Y))$. In the two procedures, the maximization problem is approached in different ways. In the *local likelihood* method, an estimate of $E((l(\eta(X), Y) \mid X = x))$ is constructed (for each x) and this has the form $(1/k_n) \cdot \sum_{j \in N_i} l(\eta(x_j), y_j)$ given in (26) of the paper. As Brillinger notes (his Section 2), one can generalize this and hence include robust estimates and many others.

On the other hand, the local scoring procedure maximizes $E(l(\eta(X), Y))$ by estimating the quantities in the update expressions (22) and (36). Note, however, that this procedure is not expressible as a maximation of the kind that Brillinger describes, i.e., a maximization of a function of the form $\sum_i \rho(Y_i \mid \hat{\eta}) W_{ni}(X)$. However, it is possible to write down a finite sample justification of local scoring (to answer a question of Brillinger's) based on the notion of penalized likelihood. This justification applies only in the special case in which the local scoring algorithm uses linear smoothers. Recall that a linear smoother is one for which the result of smoothing a vector $\mathbf{z}$ can be written simply as $\hat{\mathbf{z}} = S\mathbf{z}$, for some matrix S, called a "smoother matrix." Now suppose we have data $(y_1, x_{11}, x_{12}, \cdots, x_{1p}), \cdots, (y_n, x_{n1}, x_{n2}, \cdots, x_{np})$ and let S_j be the smoother matrix for the jth variable. Let $\mathbf{s}_j = (s_1(x_{1j}), s_2(x_{2j}), \cdots, s_n(x_{nj}))^t$, $j = 1, 2, \cdots, p$ and consider the following problem. Find $\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_p$ to maximize

$$l(\boldsymbol{\eta}) - \frac{1}{2} \sum_1^p \mathbf{s}_j^t (S_j^- - I) \mathbf{s}_j \tag{1}$$

where $\boldsymbol{\eta} = \alpha + \sum_1^p \mathbf{s}_j$ and S_j^- is a generalized inverse of S_j. Then it is easy to show that the local scoring procedure is a Fisher scoring step for maximizing (1) (see Hastie and Tibshirani, 1986a, for details). Now a typical smoother matrix is close to symmetric, has eigenvectors that are close to polynomials, and has eigenvalues that tend to decrease with increasing order of the eigenvector. Hence, the penalty term in (1) puts greater penalty on the higher order polynomial components of each $\mathbf{s}_j$. There is also a close tie here to smoothing splines. If we start with a penalty of the form $\sum_1^p \lambda_j \mathbf{s}_j^t K_j \mathbf{s}_j$, where K_j is an appropriate quadratic penalty matrix, we derive a local scoring procedure that uses cubic spline smoothers. Hence, there is close relation of local scoring to the work of O'Sullivan, Yandell, and Raynor (1986), Green (1985), and Green and Yandell (1985). These authors consider a

penalized likelihood approach, with emphasis on quadratic penalties leading to spline smoothing, as above. None of these authors use backfitting type algorithms, however, because their models contain only a single smooth function or surface (in addition to parametric terms) and hence backfitting is not required.

Following Brillinger's comment, we note that the local scoring procedure can also be used for robust estimation. For a general ψ function, the local scoring step equivalent to (22) is

$$\eta^1(x) = E\left[\eta(x) - \frac{\psi(Y \mid \eta)}{E(d\psi/d\eta \mid x)} \,\middle|\, x\right]. \tag{2}$$

As before, the conditional expectation is estimated by a smoother, and for multiple covariates, η might be an additive function.

Stone's parametric splines can be cast in the same setting. The spline fit on each covariate can be written as a linear operation and hence his model could be fit with a local scoring algorithm performing the appropriate linear fit at each smoothing step. By the results quoted in Section 7, this procedure would converge to the maximum likelihood estimates of the functions; this would, however, be a very inefficient way to fit the model, since it can be solved efficiently via the usual iterative methods.

A disadvantage of this global minimization framework is that it doesn't incorporate nonlinear smoothers. These include variable span smoothers (e.g., "supersmoother," Friedman and Stuetzle, 1982) and the "split-linear smoother" (MacDonald and Owen, 1984) for capturing discontinuities. These and other nonlinear smoothers would be useful for capturing the irregular or discontinuous functional behavior that Brillinger mentions, but we've had only a limited experience with them so far. We have incorporated a cubic spline smoother into the lastest version of GAIM (thanks to Finbarr O'Sullivan for his code) and have been happy with the results. As a final point, we reiterate another nice feature of local scoring: one is free to choose different smoothers for different covariates. Hence, one could use a spline smoother for one covariate, a parametric spline for a second covariate, a variable span smoother for a third covariate, and so on. Of course, straight line fits and categorical variables can also be used, resulting in a rich class of models.

2. ADDITIVITY AND INTERACTION

Drs. Brillinger, McCullagh, and Nelder bring up the question of additivity and interaction. It is difficult to come up with a clear definition for the latter; one possibility is to define interaction as being the lack of fit in a standard (componentwise) linear model. Phrased in this way, nonlinearity in a covariate is a kind of interaction. We suspect that McCullagh is referring to a more restrictive form of interaction, something like product interactions of two or more variables. We don't feel that zero interaction (of this latter sort) is "fundamentally more plausible" than componentwise linearity; instead, we view the method that we have presented for estimating nonlinearities as just another tool for detecting departures from the linear model. McCullagh's question concerning when (in an analysis) interaction should come under scrutiny is a deep one that we don't know how to answer. To further stress the difficulty of this question, we note that transformations of Y are another way to model certain kinds of product interactions on the original Y scale. The overall goal of all these tools is to find simple departures from a componentwise linear model; developing an effective strategy for this is a challenging and important problem.

We do want to emphasize that simple interactions can be incorporated in a generalized additive model. These include interactions of the form $\beta x_1 x_2$, $s(x_1 x_2)$, and $\beta \hat{s}_1(x_1) \cdot \hat{s}_2(x_2)$ (suggested by Nelder), where $\hat{s}_1(x_1)$ and $\hat{s}_2(x_2)$ are known functions, possibly obtained from an additive fit. We don't feel (as Nelder does) that convergence of local scoring will be a problem in these cases; it will simply take longer to converge as the constructed variables become more correlated.

More recently, we have experimented with the use of two-dimensional smoothers to fit surfaces more general than an additive one. Figure 1B illustrates such a surface.

The variables income (of the head of household) and age are two of a number of variables used to model the proportion of families having a telephone at home (the data is part of a telephone survey kindly furnished by Ed Fowlkes). The terms $s_1(\text{Inc}) + s_2(\text{Age})$ were included in an additive logistic model, together with several other variables. Figure 1A gives the additive surface defined by these two fitted functions. Figure 1B shows the estimated interaction surface $s_{1,2}(\text{Inc}, \text{Age})$. This was estimated by using a (kernel) surface smoother within the local scoring algorithm for this pair. The single function for income was quadratic, whereas in Figure 1B we see that that the income effect appears mostly monotone (except for a dip around the middle ages) and levels off at higher ages (and thus higher incomes). This leveling off goes unnoticed in the additive function model; rather it simply dampens the overall effect. This example illustrates the fact that an additive model can give us a reasonable idea of what is going on, while finer details can be discovered by fitting more general models.

3. COMPUTATIONAL CONSIDERATIONS

Brillinger reports many problems with iteratively reweighted least squares algorithms, and while we don't doubt that better procedures will be developed,

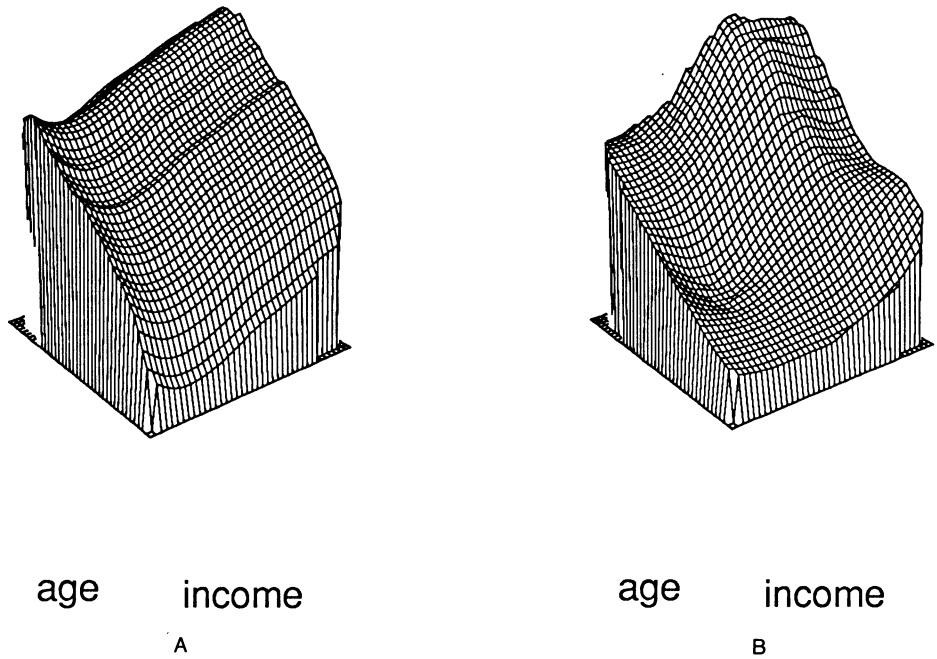

FIG. 1. *(A) Additive surface defined by* $s_1(\text{Inc}) + s_2(\text{Age})$. *This gives an idea of the types of surfaces produced by additive models. (B) Interactions surface* $s_{1,2}(\text{Inc, Age})$ *reveals two-dimensional features not captured by the additive model. The surface was estimated using a two-dimensional kernel smoother within the local scoring algorithm.*

we have had few difficulties with the present algorithm. Brillinger later told us that this problems occurred in special models that incorporated random effects, and perhaps this added complexity caused some of the difficulty.

In the local scoring algorithm, many variations are possible, in terms of the order of the smoothing and updating operations. In some early experimentation we had convergence problems with one variant. We chose the present method because it converged well in practice and because it reduces to Fisher scoring when linear fitting is used.

More recently, we have developed (with Andreas Buja, 1986) a new version of backfitting in which the linear components for all the variables are all fit in a separate projection. We have been able to prove convergence of this modified algorithm when a practical smoother like a cubic spline is used, something that neither we nor Breiman and Friedman (1985) have been able to do for the present algorithm. This algorithm should also be much more efficient computationally.

We note that in general, one has a choice of observed or expected information in the local scoring procedure; these correspond to Newton-Raphson and Fisher scoring, respectively. (In the exponential family with canonical link, they are the same.) In the paper, we used observed information for the general case, but we haven't yet studied this choice. Nor have we thought enough about concavity of the log likelihood, as mentioned by Stone.

Brillinger's points about an "automatic" algorithm are well taken. We were referring to the fact that our procedures eliminate some of the detective work necessary for finding nonlinearities from partial residual plots.

Finally, in response to Nelder's request for a GLIM version of the algorithms, we think we have a simple method for implementation via the new PASS facility that he is alluding to, but neither of us have the 3-77 version of GLIM and are looking forward to receiving it (for UNIX machines).

4. DIAGNOSTICS AND TOOLS FOR INFERENCE

Brillinger mentions the need for fit/validation procedures, diagnostics, and measures of influence. As mentioned in Section 9 and demonstrated in the examples, we have developed a notion degrees of freedom or "effective number of parameters," following that of Cleveland (1979). This is useful for assessing the importance of model terms. We also have a fairly simple way of estimating pointwise confidence bands for the estimated functions, if the smoothers used are linear. These are based on $\pm$ twice a local measure of standard deviation. See Hastie and Tibshirani (1984, 1985c) for further details of both the above techniques. Resampling methods, as suggested by Brillinger, would be another approach. An example of this is given in Efron and Tibshirani (1986), but we haven't yet studied this problem in detail.

The local scoring algorithm is not very robust to

outliers and making the smoothers robust would not solve the problem completely, if more than one covariate is present. What is needed is another outer loop in which points are downweighed based on the current fit; however, this may be computationally formidable. As far as diagnostics are concerned, Buja, Donnell, and Stuetzle (1986) have studied the analogous problem to collinearity in additive models (they call it "cocurvity"). This and much more work is needed to develop for additive models a diagnostic "black bag" like the one available for linear models.

5. THE DATA ANALYSES

In the two examples of our paper, we are guilty, as Brillinger points out, of brushing over the scientific aspects of the problem at hand. We will briefly try to make amends here. In the first example, the smooth in Figure 3 is interesting because it shows a plateau around age 50, something oncologists call the "Clemmenson hook." In the Cox model example, an interesting result was the disagreement, between the parametric and nonparametric analyses, as to whether the relative risk dropped or was about constant between ages 10 and 40. Further investigation (see Efron and Tibshirani, 1986) suggested that there was insufficient data in this age range to decide the issue.

To answer Nelder's questions on the first example, the addition of a term $\beta x_1 x_2$ to the generalized additive model did not significantly reduce the deviance, although it was significant when added to the parametric model. On Nelder's suggestion, we tried adding the term $\beta \hat{s}_1(x_1) \cdot \hat{s}_2(x_2)$ to the model, $\hat{s}_1(x_1)$ and $\hat{s}_2(x_2)$ being the functions from the generalized additive fit. This produced a drop in deviance of only 1.4. We also tried replacing each smooth by the corresponding parametric fit, as suggested by Nelder. The drops in deviance were 5.7, 3.6, and .01 on 1.7, 1.5, and 1.4, respectively. Hence, only the function for age is significantly better than its parametric fit.

For a more thorough data analysis using generalized additive models, we refer the reader to Hastie and Tibshirani (1985a and 1985c).

6. RELATED WORK AND EXTENSIONS

Stone discusses another approach to generalized additive model estimation, namely the use of fixed knot "parametric splines." His method does have the conceptual and mathematical advantages that he mentions, but practically speaking, we worry about the task of picking the number and position of the knots. How much does this choice effect the appearance of the final estimate? When many covariates are present, should the knots be chosen in some way to account for the other variables in the model? Another closely related approach, is that of smoothing splines, mentioned in Section 3 of this discussion. A comparative study of all these methods would be very useful.

Stone mentioned multiparameter models. We have, in fact, generalized the logistic model to incorporate ordered categorical responses (Hastie and Tibshirani, 1986b). We adapted the proportional odds model of McCullagh (1980):

$$\text{logit}[P(Y \le k \mid \mathbf{x})] = \alpha_k - \sum_{j=1}^{p} f_j(x_j), \quad k = 1, 2, \cdots, K-1, \tag{3}$$

where the response Y has K categories. The model essentially says that the histogram for the response categories shifts with the covariates according to $\eta(\mathbf{x}) = \sum_{j=1}^{p} f_j(x_j)$. We use the multinomial likelihood for estimation. The appropriate local scoring algorithm has an additional loop; we alternate between estimating the $K-1$ constants by weighted averages of $K-1$ adjusted dependent variates, and the additive functions by backfitting on a *scalar* linear combination of adjusted dependent variates. The model (3) can also be used when a continuous response has been categorized, and thus fills the gap between the extreme 0–1 response logistic regression model and the *continuous* response ordinary regression model.

Brillinger's final question concerning ACE and generalized additive models is a fascinating one. We would like to take this opportunity to clarify the relationship between the methods and report some current research. First note that, as alluded to in Section 9 of the paper, the local scoring algorithm can be used to estimate any function that appears in a model, not just a function of a covariate. We simply add a step like (22) to the algorithm for that function. Thus for example, we can estimate a link function (see Hastie and Tibshirani, 1984) or a variance function (Hastie and Pregibon, 1986). This fact will become important below.

Now consider the Gaussian additive model $E(Y \mid \mathbf{X}) = \alpha + \sum_1^p s_j(X_j)$. (We'll relate our comments to nonGaussian generalized additive models as we go along.) Two ways to extend this model are to allow a transformation of the mean, i.e., $E(Y \mid \mathbf{X}) = f(\alpha + \sum_1^p s_j(X_j))$ or a transformation of the response, i.e., $E(\theta(Y) \mid \mathbf{X}) = \alpha \sum_1^p s_j(X_j)$. The former has been looked at by Friedman and Owen (1986) and is a special case of link function estimation for generalized additive models via local scoring. The second model is the transformation model, for which Breiman and Friedman's (1985) ACE algorithm provides a method for estimation. The two models are not the same, even if $\theta(\cdot)$ is forced to be monotone. That is, we should not expect that $\hat{\theta}^{-1}(\cdot)$ will be close to $\hat{f}(\cdot)$ for a given data set.

Brillinger's question concerns two possible methods for estimating the functions of the transformation model. The ACE algorithm maximizes the correlation of the transformed variables. Brillinger's suggestion is to instead maximize the likelihood of the untransformed variables, by direct analogy to the parametric method of Box and Cox (1964). This likelihood would include a Jacobian, as Brillinger states, to account for the transformation $\theta(\cdot)$. One can carry through Brillinger's suggestion using the local scoring algorithm: unfortunately, the resultant algorithm requires estimates of the second and third derivatives of $\theta(\cdot)$. While we haven't tried it yet, our guess is that the algorithm might be unstable because of this.

Another approach to this problem, similar to Brillinger's suggestion, is given by Tibshirani (1986). He proposes an algorithm in which a (nonparametric) variance stabilizing transformation is used to estimate $\theta(\cdot)$. The procedure is called "RACE" for regression ACE. In both simulated and real data examples, he demonstrates that RACE eliminates many of the anomalies of ACE, in particular, sensitivity to the marginal distribution of the X's. RACE is likely to produce similar results (qualitatively) to Brillinger's suggestion, because the effect of the Jacobian is mainly to force $\theta(Y)$ to have constant variance (see Box and Cox (1964) and Tibshirani (1984, Remark F)).

A transformation of the response might also be useful in other generalized additive models, such as a Poisson model for categorical data. Marhoul (1984) looks at a related technique.

ACKNOWLEDGMENTS

We were fortunate to receive the comments of four such highly respected statisticans as Drs. Brillinger, McCullagh, Nelder, and Stone. They raise a number of important issues, some that we've thought about since the writing of the paper, but many others that we haven't yet resolved. We would like to thank them for their efforts and we would also like to thank Morris DeGroot for his editorial work. In our rejoinder we have tried to clarify and expand on some of these questions.

ADDITIONAL REFERENCES

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations. *J. Roy. Statist. Soc. Ser. B* **26** 211–252.

BUJA, A., DONNELL, D. and STUETZLE, W. (1986). Additive principal components. Manuscript in preparation.

EFRON, B. and TIBSHIRANI, R. (1986). Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy (with discussion). *Statist. Sci.* **1** 54–77.

FRIEDMAN, J. H. and OWEN, A. (1986). Predictive ACE. Unpublished manuscript.

GREEN, P. J. (1985). Penalized likelihood for general semi-parametric regression models. Tech. Rept. 2819, Dept. Statistics, Univ. Wisconsin, Madison.

HASTIE, T. and PREGIBON, D. (1986). Manuscript in preparation.

HASTIE, T. and TIBSHIRANI, R. (1985c). Generalized additive models: some applications. Tech. Rept. 14, Dept. Statistics, Univ. Toronto.

HASTIE, T. and TIBSHIRANI, R. (1986a). Generalized additive models, cubic splines and penalized likelihood. Tech. Rept., Biostatistics Group, Univ. Toronto.

HASTIE, T. and TIBSHIRANI, R. (1986b). Nonparametric logistic and proportional odds regression. To appear in *Appl. Statist.*

HASTIE, T., TIBSHIRANI, R., and BUJA, A. (1986). Manuscript in preparation.

MACDONALD, J. and OWEN, A. (1984). Smoothing with split linear fits. Report LCS07, Dept. Statistics, Stanford Univ.

MARHOUL, J. (1984). A model for large sparse contingency tables. Report LCS13, Dept. Statistics, Stanford Univ.

O'SULLIVAN, F., YANDELL, B. and RAYNOR, W. (1986). Automatic smoothing of regression functions in generalized linear models. *J. Amer. Statist. Assoc.* **81** 96–103.

TIBSHIRANI, R. (1986). Estimating optimal transformations for regression: a variation on ACE. Tech. Rept. 1986-001, Biostatistics Group, Univ. Toronto.