The Kernel Estimate of a Regression Function in Likelihood-Based Models

JOAN G. STANISWALIS*

Smoothing splines have a penalized likelihood motivation (Good and Gaskins 1971) allowing direct application to nonparametric regression in likelihood-based models. The notion of a weighted likelihood for the nonparametric kernel estimation of a regression function is proposed, generalizing the local likelihood theory of Tibshirani and Hastie (1987). Let the data be of the form (x_i, Y_i) $(i = 1, \ldots, n)$, where $x_i \in [0, 1]^d$ are lattice points and the Y_i are independent random variables from a family of distributions with parameter $\lambda_i = g(x_i)$, with g having continuous partial derivatives of order $k \geq 2$. The goal is to arrive at a nonparametric estimate $\hat{\lambda}_o$ of $\lambda_o = g(x_o)$ for a fixed point $x_o \in [0, 1]^d$. We consider the estimator $\hat{\lambda}_o$ that maximizes the weighted likelihood function $\mathbf{W}(\lambda) = \Sigma_{i=1}^n W[(x_o - x_i)/b] \log f(Y_i:\lambda)$, with f the density of Y_i, W a symmetric kernel with compact support, and b the bandwidth that controls the degree of smoothing. Sufficient conditions for consistency and asymptotic normality of $\hat{\lambda}_o$ are given. If the Y_i are normal random variables with mean λ_i and equal variance, then $\hat{\lambda}_o$ is the kernel estimator of Priestly–Chao (1972). It is a weighted average of Y_i corresponding to x_i in a neighborhood of x_o. The kernel governs the weights and the bandwidth controls the size of the neighborhood. The kernel estimator of the relative risk function is developed for censored survival times under the assumption of the Cox proportional hazards model. The weighted likelihood approach based on the full likelihood is illustrated with real and simulated data.

KEY WORDS: Local likelihood; Nonparametric regression; Proportional hazards model; Weighted likelihood.

1. INTRODUCTION

Consider the model $Y_i = g(x_i) + \varepsilon_i$ $(i = 1, \ldots, n)$, where $\varepsilon_1, \ldots, \varepsilon_n$ are iid random variates with mean 0 and variance σ^2. The assumption that g is a smooth function implies that if x_i is near x, then Y_i contains information about $g(x)$. The kernel estimator of g is

$$g_n(x, b) = \sum_{i=1}^{n} Y_i W\left(\frac{x - x_i}{b}\right) \Big/ (nb^d). \tag{1.1}$$

Here $b \in (0, .5)$ is the bandwidth and W is a direct product of a kernel w: $R \to R$ of order k with compact support on $[-1, 1]$; that is,

$$\begin{aligned} \int_{-1}^{1} v^j w(v)\, dv &= 1 && \text{if } j = 0 \\ &= 0 && \text{if } j < k \\ &= W_k \neq 0 && \text{if } j = k. \end{aligned}$$

The kernel estimator of $g(x)$ is a weighted average of observations Y_i that correspond to x_i in a neighborhood of x. The weights are governed by the kernel with the order selected to match the smoothness of g. The bandwidth controls the size of the neighborhood, that is, the degree of smoothing imposed on the noisy observations of g. If $b \to 0$ and $nb^d \to \infty$ as $n \to \infty$, then $g_n(x, b)$ is a consistent estimator of $g(x)$.

For small sample sizes,

$$g_n(x, b) = \frac{\sum_{i=1}^{n} Y_i W\left(\frac{(x - x_i)}{b}\right)}{\sum_{i=1}^{n} W\left(\frac{(x - x_i)}{b}\right)} \tag{1.2}$$

is preferred because it has smaller bias. The two estimators are asymptotically equivalent and have been studied extensively (Gasser and Müller 1979a; Müller 1984; Rice 1984a,b).

In classical least squares regression, a functional form is specified for the regression curve g. This functional form, say $g(x; \beta)$, depends on a finite number of unknown parameters β. The least squares estimate of g is $g(x, \hat{\beta})$, where $\hat{\beta}$ is chosen to maximize

$$-\sum_{i=1}^{n} [Y_i - g(x_i; \beta)]^2. \tag{1.3}$$

Compare (1.3) with the following weighted least squares criterion for the *nonparametric* estimation of $g(x)$:

$$-\sum_{i=1}^{n} W\left(\frac{x - x_i}{b}\right) [Y_i - g(x)]^2. \tag{1.4}$$

In (1.4), $g(x)$ replaces the $g(x; \beta)$ that appears in (1.3). If $g(x)$ is regarded as a single unknown parameter λ, then it may be estimated by solving for $\hat{\lambda}$, which maximizes

$$-\sum_{i=1}^{n} W\left(\frac{x - x_i}{b}\right) [Y_i - \lambda]^2. \tag{1.5}$$

The resulting estimate $\hat{\lambda}$ of $g(x)$ is precisely $g_n(x, b)$, given by (1.2). Equation (1.5) appears in Cleveland (1979) when

* Joan G. Staniswalis is Assistant Professor, Department of Biostatistics, Medical College of Virginia, Virginia Commonwealth University, Richmond, VA 23298-0032. This research was supported in part by the Grants-In-Aid Program for faculty of Virginia Commonwealth University. The author thanks John Ventre for his competent execution of the simulation studies, Trevor Hastie for providing the GAIM program, and Brian Yandell for reading a preliminary draft. Thanks are due to John A. Rice and Walter H. Carter, Jr. for helpful discussions and encouragement. The author is grateful to the referees and associate editors for asking pertinent questions and making suggestions that clarified the exposition.

Journal of the American Statistical Association
March 1989, Vol. 84, No. 405, Theory and Methods

he uses locally weighted least squares regression to locally fit a polynomial of degree 0 to the data.

Therefore, if the noise variables ε_i have a Gaussian density function denoted by $f(\varepsilon)$, then (1.2) is the maximizer with respect to λ of the weighted likelihood criterion

$$\mathbf{W}(\lambda) = \sum_{i=1}^{n} W\left(\frac{x - x_i}{b}\right) \log f(Y_i - \lambda). \quad (1.6)$$

In this article, the maximizer of $\mathbf{W}(\lambda)$ is studied for other densities f.

In Section 2, the asymptotic properties of the maximizer of the weighted likelihood $\mathbf{W}(\lambda)$ are presented. The choice of the bandwidth and kernel are discussed in Sections 3 and 4, respectively. In the remaining sections the use of the kernel estimators is demonstrated for censored survival data, following the Cox proportional hazards model.

2. WEIGHTED LIKELIHOODS

The log-likelihood of the sample (x_i, Y_i) for $i = 1, \ldots, n$ is $\Sigma_{i=1}^{n} \log f(Y_i; \lambda_i)$. The goal is to estimate $\lambda_o = g(x_o)$ for a fixed point $x_o \in R^d$. Here $x_o = (x_{1o}, \ldots, x_{do})$ such that $x_{jo} \in [b, 1 - b]$ for $j = 1, \ldots, d$ is assumed to avoid boundary modifications of the kernel (Rice 1984a). If x_i is close to x_o, then $\lambda_i = g(x_i)$ and hence Y_i contains information about $\lambda_o = g(x_o)$. Consider the estimator $\hat{\lambda}_o = g_n(x_o, b)$ of λ_o that maximizes with respect to λ the weighted likelihood function

$$\mathbf{W}(\lambda) = \sum_{i=1}^{n} W\left(\frac{x_o - x_i}{b}\right) \log f(Y_i; \lambda).$$

Set $W_o = \Sigma_{i=1}^{n} W((x_o - x_i)/b)$. The sum W_o is approximately equal to nb^d for large n and $\mathbf{W}(\lambda)/W_o$ is an empirical estimate of $E[\log f(Y; \lambda_o)|x_o]$. Sufficient conditions for the consistency of $\hat{\lambda}_o$ as an estimator of λ_o and for the asymptotic normality of $\hat{\lambda}_o$ are stated in the following theorem.

Theorem. Let f satisfy the following regularity conditions: $\log f(y; \lambda)$ has three continuous partial derivatives with respect to λ, there exist integrable $H_i(y)$ such that $E|H_i(Y)|^2 < \infty$, and $|\partial^i \log f(y; \lambda)/\partial\lambda^i| \leq H_i(y)$ for all λ and $i = 1, 2$. Then (a) if $nb^d \to \infty$ and $b \to 0$ as $n \to \infty$, $|g_n(x_o, b) - g(x_o)| \to 0$ in probability as $n \to \infty$; (b) if $nb^d \to \infty$ and $nb^{d+2k} \to 0$ as $n \to \infty$, $(nb^d I(\lambda_o)/W_2)^{1/2}$ $[g_n(x_o, b) - g(x_o)]$ converges in distribution to a standard normal random variable, where

$$W_2 = \left[\int_{-1}^{1} w^2(v)\, dv\right]^d$$

and

$$I(\lambda_o) = E\{[\partial \log f(Y; \lambda_o)/\partial\lambda_o]^2 | x_o\}.$$

The proofs appear in Staniswalis (1987).

This theorem is useful for deriving nonparametric confidence regions for the location of the optimum of a regression function, examples of which are given in Staniswalis and McCrady (1988) for count data and Staniswalis and Cooper (1988) for quantal data. When first investigating the use of kernel estimators in those applications, the effort seemed futile because of the small number of levels of the independent variables in the experimental design. Computer simulations supported the nonparametric inference procedures despite the small sample sizes.

3. BANDWIDTH SELECTION

3.1 $d = 1$

Cross-validation (Wahba and Wold 1975) may be used to select the bandwidth, in which case b is chosen to maximize

$$\sum_{j=1}^{n} \log f(Y_j; \hat{\lambda}_{(j)}).$$

Here $\hat{\lambda}_{(j)}$ is the maximizer with respect to λ of

$$\sum_{\substack{i=1 \\ i \neq j}}^{n} W\left(\frac{x_j - x_i}{b}\right) \log f(Y_i; \lambda).$$

This is a global bandwidth selection procedure, whose asymptotic optimality properties are not known.

For the setting given in Section 1, that is, $\varepsilon_1, \ldots, \varepsilon_n$ iid, Rice (1984b) proposed a procedure for the estimation of the global bandwidth that minimizes the integrated mean squared error of the kernel estimator. Müller and Stadtmüller (1987) proposed a procedure for the estimation of the local bandwidth that minimizes the asymptotic mean squared error. The author also proposed a data-adaptive procedure for local bandwidth selection (Staniswalis 1985). A consistent estimate of the optimal local bandwidth, optimal in the sense of minimizing the exact mean squared error of the kernel estimator, is derived. That procedure can be modified for estimating the surface $g(x)$ in the more general setting of Section 2.

3.2 $d > 1$

When $d > 1$, one must not only estimate the optimal size of the bandwidth, but also the shape and orientation of the neighborhood over which the kernel estimator is smoothing. The results in Staniswalis (1985) could be extended to estimate the size and orientation of an elliptical neighborhood that minimizes the mean squared error of the kernel estimator.

In the real data problems of Section 6, the kernel estimate is plotted for several values of global bandwidths for a square neighborhood about x. The cross-validation procedure could have been used to select a bandwidth, but was not deemed necessary in a setting where the kernel estimator was only used for exploratory purposes.

4. KERNEL SELECTION

Recall that $g(x)$ is assumed to have continuous partial derivatives of order k. The kernel $w(v)$ is a kernel of order k and is thus selected to match the smoothness of the dose–response curve (Gasser and Müller 1979a). The results are

presented in terms of a general kernel w, but all of the applications and simulations use $k = 2$ and Müller's (1984) quartic kernel:

$$w(v) = \frac{15}{16}(1 - v^2)^2, \qquad -1 \le v \le 1$$
$$= 0, \qquad \text{elsewhere.}$$

Some reasons for this choice of k and w are as follows.

The boundary modifications (Rice 1984a) for the kernel w become more complicated as the order of the kernel increases. This is because k kernel estimators are jackknifed to arrive at the bias-corrected kernel estimator in the boundary. The order $k = 2$ was fixed in the simulations because $k = 2$ is the easiest to implement.

Smooth estimates of derivatives of $g(x)$ are ultimately needed in the biomedical applications presented in Section 6, because confidence regions for the location of the global optimum of the regression function are ultimately desired. Those confidence regions may be calculated via the delta method using consistent estimates of the derivatives of g. The quartic kernel was used because it is a kernel of order $k = 2$ that satisfies the boundary conditions (Gasser and Müller 1979b) needed for the first and second derivatives of $g_n(x, b)$ in Section 1 to be consistent estimators of the first and second derivatives of $g(x)$, respectively.

5. CENSORED SURVIVAL DATA

5.1 The Cox Model

In what follows, x is a fixed design variable and the data set contains a control group. Similar results could be developed in a more general setting (see Sec. 5.3 and Staniswalis 1987).

Suppose that the observations are $(x_i, Y_{ij}, \delta_{ij})$, where x_i = ith combination dose level of d drugs; T_{ij} = survival time of the jth experimental unit at the ith combination dose level; C_{ij} = censorship time of the jth experimental unit at the ith combination dose level; $Y_{ij} = \min(T_{ij}, C_{ij})$; and

$$\delta_{ij} = 1 \quad \text{if } T_{ij} \le C_{ij}$$
$$= 0 \quad \text{if } T_{ij} > C_{ij},$$

for $j = 1, \ldots, m_i$ and $i = 1, \ldots, n$.

Let $p(t; g(x))$ denote the probability density function of the uncensored failure times at the combination dose level x. Assume the hazard functions of the uncensored survival times satisfy the Cox proportional hazards model $h(t; x) = h_o(t) \exp[-g(x)]$, where x equal to the 0 vector ϕ corresponds to the control group and $g(\phi) = 0$. Denote the survival function corresponding to $p(t; g(x))$ by $S(t; x)$ and $S_o(t) = S(t; x)|_{x=\phi}$. Then

$$p(t; g(x)) = h_o(t)\exp[-g(x)]\,[S_o(t)]^{\exp[-g(x)]}$$

and $S(t; x) = [S_o(t)]^{\exp[-g(x)]}$ (Kalbfleisch and Prentice 1980). The goal is to estimate $g(x)$ and $S_o(t)$ nonparametrically. Here $g(x)$ is the negative-log-relative risk. Exploratory diagnostic tools for determining whether the proportional hazards assumption is plausible are illustrated in Staniswalis (1987).

Hastie and Tibshirani (1986) used a nonparametric technique known as local scoring to estimate g under the generalized additive model assumption $g(x) = g_1(x_1) + \cdots + g_d(x_d)$. O'Sullivan (1986) used penalized likelihoods (Good and Gaskins 1971) to arrive at a smoothing spline estimate of the overall surface $g(x)$. They used the partial likelihood of Cox (1975), which does not require prior knowledge or estimation of $S_o(t)$.

In this article, the survival function of the control group $S_o(t)$ is assumed unknown, but the *full* likelihood, rather than the partial likelihood, function is used. An initial estimate of $S_o(t)$ is obtained from the data in the control group, then $g(x)$ is estimated. The estimate of $S_o(t)$ is updated, using the initial estimate of $g(x)$. Finally, $g(x)$ is updated. Given an estimate of $S_o(t)$, the kernel estimate of $g(x)$ is easy to calculate and is given explicitly without iteration, unlike Tibshirani and Hastie's (1986) procedure, which uses the partial likelihoods.

5.2 The Weighted Likelihood Formulation for the Cox Model

If the censoring mechanism is independent of the failure times and the dose levels, then the log-likelihood for the sample (for right-continuous survival functions), up to a constant, is

$$\sum_{i,j} \{\delta_{ij}\log p(Y_{ij}; g(x_i)) + (1 - \delta_{ij})\log S(Y_{ij}; g(x_i))\}$$
$$= \sum_{i,j} \{\delta_{ij}\log h_o(Y_{ij}) - \delta_{ij}\, g(x_i) + \exp[-g(x_i)]\log S_o(Y_{ij})\}.$$

The weighted likelihood for estimating $\lambda = g(x)$ is

$$\mathbf{W}(\lambda) = \sum_{i,j} W\left(\frac{x - x_i}{b}\right)\{\delta_{ij}\log h_o(Y_{ij}) - \delta_{ij}\lambda + \exp(-\lambda)\log S_o(Y_{ij})\}.$$

The maximizer of $\mathbf{W}$ with respect to λ is

$$g_n(x, b) = \log\left[\frac{-\sum_{i,j} W((x - x_i)/b)\log S_o(Y_{ij})}{\sum_{i,j} W((x - x_i)/b)\delta_{ij}}\right].$$

If $f(y; \lambda)$ is set equal to $p(y; \lambda)^{\delta}\, S(y; \lambda)^{1-\delta}$, then the theorem in Section 2 holds under the additional assumptions $\lim_{t\to\infty} S_o(t)\log S_o(t) = 0$ and $\lim_{t\to 0} S_o(t)\log S_o(t) = 0$.

The numerator of the kernel estimate of $g(x)$ makes sense because $-\log S_o(T_{ij})$ is exponentially distributed with mean $\exp[g(x_i)]$. This is verified by transforming the T_{ij} to a uniform random variable U_{ij} with the survival function $S(t; x_i)$:

$$U_{ij} = (S_o(T_{ij}))^{\exp[-g(x_i)]} \sim \text{uniform}[0, 1]. \qquad (5.1)$$

Finally, observe that $-\log(U_{ij}^{1/a})$ $(a > 0)$ is exponentially distributed with parameter a; set $a = \exp[-g(x_i)]$ and the result follows.

The denominator of the kernel estimate of $g(x)$ makes sense because if C_{ij} $(j = 1, \ldots, m_i;\ i = 1, \ldots, n)$ are iid random censoring times independent of T_{ij}, iid exponentially distributed survival times with parameter a, then the maximum likelihood estimator of $1/a$ is

$$\sum_{i,j} Y_{ij} \Big/ \sum_{i,j} \delta_{ij}, \tag{5.2}$$

where $Y_{ij} = \min(T_{ij}, C_{ij})$.

An algorithm for estimating $g(x)$ and $S_o(t)$ from the data is proposed. It exploits the fact that $S_o^{-1}(U_{ij})$, where U_{ij} is given by (5.1), has $S_o(t)$ as its survival function for $j = 1, \ldots, m_i$ and $i = 1, \ldots, n$. The survival function $S_o(t)$ of the control group is ultimately estimated with the nonnegative differentiable spline smooth estimate of Klotz (1982), which Whittemore and Keller (1986) claim is uniformly consistent.

1. Compute an initial estimate $\tilde{S}_o(t)$ of S_o(t) from the control group survival times.
2. Compute an initial estimate $\tilde{g}(x_i)$ of $g(x)$ according to

$$\tilde{g}(x_i) = \log\left[\frac{-\sum_{j=1}^{m_i} \log \tilde{S}_o(Y_{ij})}{\sum_{j=1}^{m_i} \delta_{ij}}\right].$$

3. Construct $(i = 1, \ldots, n;\ j = 1, \ldots, m_i)$ $\tilde{U}_{ij} = [\tilde{S}_o(Y_{ij})]^{\exp[-\tilde{g}(x_i)]}$.
4. Solve for $\tilde{Y}_{ij} = \tilde{S}_o^{-1}(\tilde{U}_{ij})$. It is assumed that $\tilde{S}_o(t)$ is invertible.
5. Compute the estimate $\hat{S}_o(t)$ of $S_o(t)$ with Klotz's spline smooth estimate based on $\tilde{Y}_{ij}$ $(j = 1, \ldots, m_i;\ i = 1, \ldots, n)$.
6. Compute

$$g_n(x, b) = \log\left[\frac{-\sum_{i,j} W((x - x_i)/b)\log \hat{S}_o(Y_{ij})}{\sum_{i,j} W((x - x_i)/b)\delta_{ij}}\right].$$

The following steps are optional:

7. Stop if $\hat{S}_o(t)$ and $g_n(x, b)$ have converged to a solution; otherwise, go to step 8.
8. Replace $\tilde{S}_o(t)$ with the current estimate $\hat{S}_o(t)$. Replace $\tilde{g}(x)$ with $g_n(x, b)$.
9. Go to step 3.

It is the subject of future research to show that this algorithm provides a consistent estimator of $g(x)$ and $S_o(t)$, given numerical convergence (as yet unspecified) in step 7. As in the backfitting algorithm (Friedman and Stuetzle 1981) used in the local scoring procedure of Tibshirani and Hastie (1986), a parametric estimate of $S_o(t)$ in step 1 could be used to initiate the algorithm and eventually arrive at nonparametric and distribution-free estimates of $g(x)$ and $S_o(t)$.

In Step 5, Klotz's (1982) spline estimate of $S_o(t)$ is preferred to the Kaplan–Meier (1958) estimate, because it is smooth and the transformed responses $\hat{S}_o(t)$ have the same resolution as the original responses. The Kaplan–Meier estimate of $S_o(t)$ is a sum of step functions. Therefore, transforming the responses by the Kaplan–Meier estimator in step 6 would discretize the responses, thereby throwing away information in the data.

5.3 Continuous Covariates in the Cox Model

The algorithm proposed in Section 5.2 is for censored survival data blocked according to fixed levels of a design variable. Here we consider the case where x is a continuous random variable and a control group is not available.

If the conditional density of $Y \mid x$ depends on x through $\lambda = g(x)$ and the marginal density of x does not depend on g, then the expected value of the weighted likelihood of (x, Y) may be empirically maximized by maximizing the conditional weighted likelihood for $Y \mid x$ with respect to λ for each x. The weighted likelihood formulation for the kernel estimator given in Section 5.2 carries over to this situation. The theorem in Section 2 does not claim that the kernel estimator is a consistent estimator of $g(x)$, given a uniformly consistent estimator of the baseline hazard function. But it is easily extended using the known results (Schuster 1972) for the Nadaraya (1964)–Watson (1964) kernel estimator in place of the asymptotic results for the Priestly–Chao (1972) kernel estimator.

We indicate how the algorithm in Section 5.2 may be modified to produce a nonparametric estimate of the relative risk and baseline hazard function when a control group is not available. Select an arbitrary value of the covariate x, say x_c, as the point of reference to which all other values of the covariate are compared. Experimental units with values of x in a neighborhood centered about x_c are the control group. The procedure developed in Sections 5.1 and 5.2 can be used here if $S(t; x_c)$ replaces $S_o(t)$. It is conjectured that this algorithm provides consistent estimators of $g(x)$ and $S(t; x_c)$ under appropriate assumptions on the size of the neighborhood about x_c defining the control group. Nevertheless, consistency of these estimators should be independent of x_c.

The estimators previously described for $g(x)$ and $S(t; x_c)$ are illustrated by Staniswalis (1987) on the Stanford heart transplant data reported by Miller and Halpern (1982). The kernel estimate has the same general shape and magnitude as the cubic smoothing spline calculated by O'Sullivan (1986). In that example it does not seem to matter whether one uses the full likelihood theory or the partial likelihood theory.

6. SIMULATIONS AND AN EXAMPLE WITH CENSORED SURVIVAL DATA

6.1 Simulations

This simulation illustrates that the proposed kernel estimator extracts the underlying structure in the censored data when $S_o(t)$ is known. Additional simulations, where

both $S_o(t)$ and g are estimated from the data, are needed to assess the procedure better. Survival times following an exponential proportional hazards model were simulated $h_o(t) = 1$ $(t > 0)$, because when $S_o(t)$ is known $-\log S_o(T_i)$ satisfy an exponential proportional hazards model. The chosen form for $g(x)$ was $g(x) = \sin(2\pi x)$ $(x \in [0, 1])$. The survival times were simulated by generating uniform random variates (SAS Institute 1982) and applying the appropriate inverse transformation. The simulated censoring random variable C was uniform on [0, 10], corresponding to about 10% censoring of the control group. The censoring mechanism was the same for each treatment group.

Data from an experimental design with 20 experimental units at each of 10 equally spaced dose levels were simulated. The kernel estimator of g was

$$g_n(x, b) = \log\left[\frac{\sum w((x - x_i)/b)Y_i}{\sum w((x - x_i)/b)\delta_i}\right],$$

where $w(v)$ is the quartic kernel. Rice's (1984a) boundary modification of the kernel estimator was used.

At each of the 10 dose levels, the median and quartiles of the simulated $g_n(x, b)$ were calculated. These are plotted in Figure 1 for $b = .35$ and are based on 500 iterations. The kernel estimator correctly suggests the presence of nonlinear trends in the data.

6.2 A Biomedical Example, $d = 2$

This example illustrates the need for a nonparametric estimate of the overall surface $g(x)$ for validating the selection of a parametric model and for providing qualitative information about the bias present in subsequent parametric inferences. The following data are taken from Carter, Wampler, and Stablein (1983). Combination treatments of 5-fluorouracil (5-FU) and cyclophosphamide (CTX) were given by single intraperitoneal injections to mice with leukemia. There were 25 combination dose groups, with 7 to 8 animals in each group. The hazard functions of the uncensored survival times were modeled with the Cox proportional hazards model. Of interest is the estimation of the combination level of CTX and 5-FU that corresponds to maximum survival, that is, the global optimum of $g(x)$ for x within the experimental region.

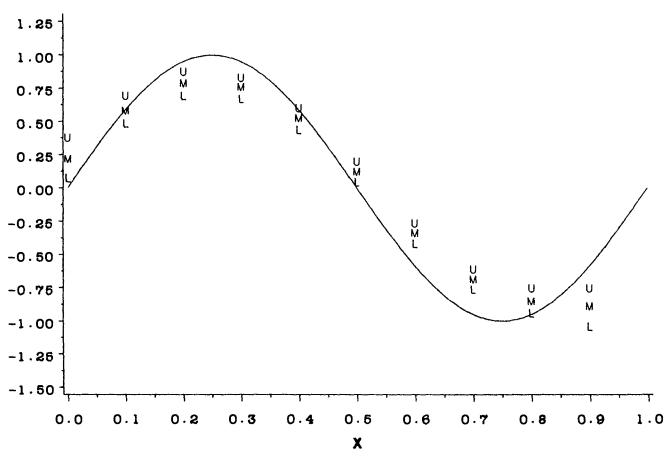

Figure 1. Simulated Quartiles and Medians of the Kernel Estimator, b = .35: U, 75th percentile; M, 50th percentile; L, 25th percentile; ——, g(x) used in the simulation.

A standard (linear) proportional hazards model using one dummy variable for each of the 24 dose groups was fit to the data with partial likelihoods, using PROC PHGLM in SAS. The parameter estimates (Table 1) indicate that an optimal treatment exists in the interior of the experimental region near 5-FU = 225 milligrams/kilogram (mg/kg), with CTX = 150 mg/kg. All of the animals in that dose group experienced some enhancement in survival times relative to the control group.

The initial estimate $\tilde{S}_o(t)$ of the survival function $S_o(t)$ computed in step 1 of the algorithm proposed in Section 5.2 is plotted in Figure 2. It was estimated with an exponential hazard rate $h_o(t) = c$ $(t > 0)$, where $c = 1/10.875$ according to Equation (5.2) for the control group. Included for comparison in Figure 2 is Klotz's (1982) spline smooth estimate, based on the data from the control group alone as well as in the final estimate of $S_o(t)$ computed in step 5. Note how the final estimate of $S_o(t)$ resembles Klotz's spline smooth estimate of $S_o(t)$ based on the control group alone rather than the initial parametric estimate.

Figures 3–5 have plots of the kernel estimates of the dose–response surface computed according to one pass through steps 1–6. The quartic kernel and several values of the bandwidth were used. The kernel estimator in the interior of the experimental region is averaging over approximately $2nb^2$ dose groups. The kernel was not modified in the boundary according to Rice (1984a); therefore, the kernel estimator in the boundary is averaging over between nb^2 and $2nb^2$ dose groups.

The dose levels were unequally spaced, thus Benedetti's (1977) form of the kernel estimator was used after transforming the doses to $[0, 1) \times [0, 1)$:

$$g_n(x, b) = \log\left[\frac{-\sum_{i=1}^{n} W\left[\frac{x - x_i}{b}\right] \log[\hat{S}_o(Y_i)](x_{1,i+1} - x_{1i})(x_{2,i+1} - x_{2i})}{\sum_{i=1}^{n} \delta_i W\left[\frac{x - x_i}{b}\right](x_{1,i+1} - x_{1i})(x_{2,i+1} - x_{2i})}\right],$$

where $x = (x_1, x_2)$, x_1 = dose level of 5-FU, x_2 = dose level of CTX, $x_{1,n+1} = 1$, and $x_{2,n+1} = 1$.

Table 1. Parameter Estimates for the Negative-Log-Relative Risk (relative to the control group) Using a Cox Proportional Hazards Model and a Dummy Variable for Each of the Dose Groups in the 5-FU and CTX Leukemia Experiment

	Negative-log-relative risk				
	5FU (mg/kg)				
CTX (mg/kg)	0	100	150	225	338
0	—	2.46	2.82	2.76	3.51
100	4.79	6.34	5.16	4.88	4.56
150	4.99	5.25	5.70	6.97	3.53
225	5.56	5.85	5.79	4.93	3.50
338	5.85	5.99	6.11	5.65	2.30

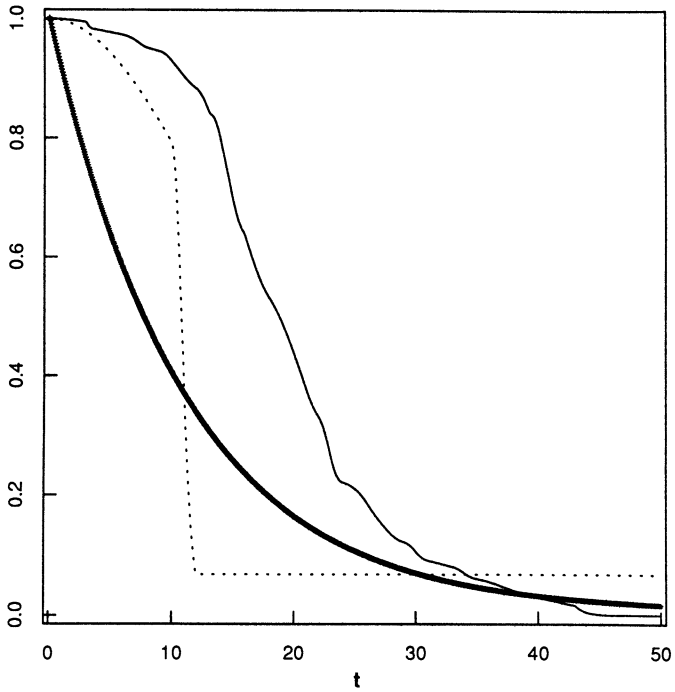

Figure 2. Estimates of $S_0(t)$: Leukemia Data. —— $\hat{S}_0(t)$, Klotz's spline smooth estimate based on all of the data and b = .20; ▬ $\tilde{S}_0(t)$, parametric estimate based on the control group data; ··· $S_0(t)$, Klotz's spline smooth estimate based on the control group data.

The maximum of the kernel estimate (Fig. 3, $b = .20$; Fig. 4, $b = .25$; Fig. 5, $b = .30$) reflects the prior observation that an optimal treatment exists in the interior of the experimental region in the neighborhood of 225 mg/kg of 5-FU and 150 mg/kg of CTX. Reflected in the local minimum of the kernel estimate is the fact that large doses of 5-FU in combination with large doses of CTX have toxic effects that negate the therapeutic effects of the drugs. For the purpose of estimating the global optimum of $g(x)$, the kernel estimate with a large value of the bandwidth (Fig. 5) suggests that a parametric model quadratic in the doses might be adequate.

Next, the data were analyzed with the GAIM package of Hastie and Tibshirani (1986), which uses local scoring, a running line smoother with a cross-validated value for the span, and the model $g(x) = g_1(x_1) + g_2(x_2)$. Figure 6 is a plot of the estimate of the overall surface $g(x)$. Reflected in $g_1(x_1)$ is the prior knowledge that 5-FU alone has therapeutic effects. The same is true for $g_2(x_2)$ and CTX. Furthermore, a comparison of $g_1(x_1)$ and $g_2(x_2)$ indicates that CTX alone has greater efficacy than 5-FU alone. This is substantiated by the data and consistent with the physician's prior experience. The surface estimated by GAIM also reflects the fact that large doses of CTX in combination with large doses of 5-FU are toxic to the animals. Nevertheless, the kernel estimate of $g(x)$ (Fig. 3) suggests that the combined effect of these drugs is not merely additive, but that an interaction exists between 5-FU and CTX.

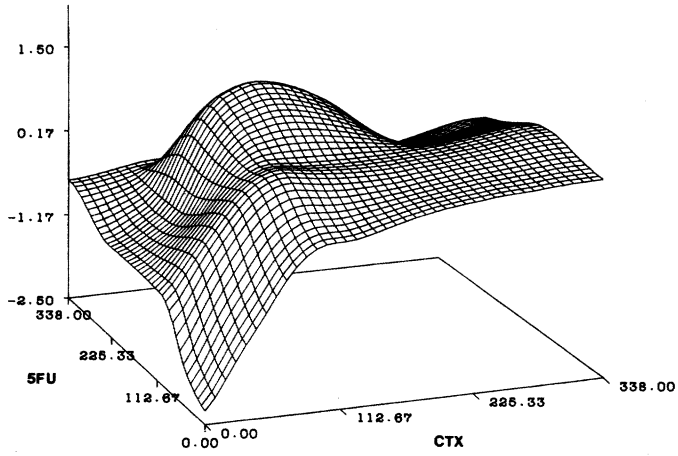

Figure 4. Kernel Estimate of g(x) for the Leukemia Data: b = .25.

Figure 7 is a plot of the quadratic response surface $g(x)$ estimated according to Carter et al. (1983) with partial likelihoods and Cox's proportional hazards model. Contrary to our prior expectations (Table 1), the confidence region for the optimal treatment is centered about the dose 95 mg/kg of 5-FU and 254 mg/kg of CTX. The reason appears to be that the parametric model masks the marked improvement in survival that was realized for a few of the animals in the *single* dose group administered 225 mg/kg of 5-FU and 150 mg/kg of CTX. It instead emphasizes the less dramatic improvement in the chances of survival that was realized at *several* dose groups given large doses of CTX in combination with small doses of 5-FU.

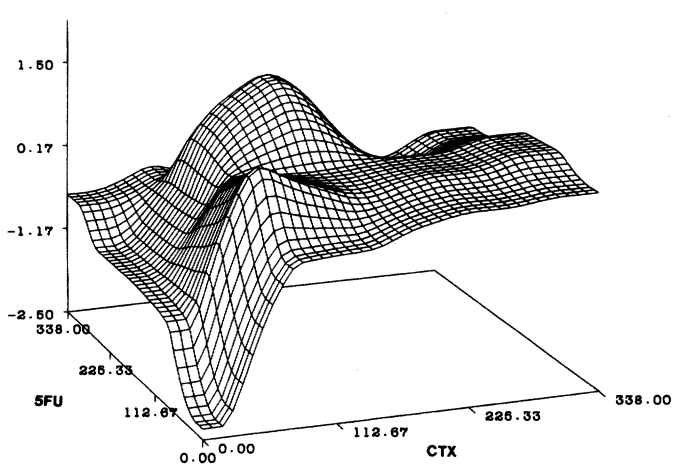

Figure 3. Kernel Estimate of g(x) for the Leukemia Data: b = .20.

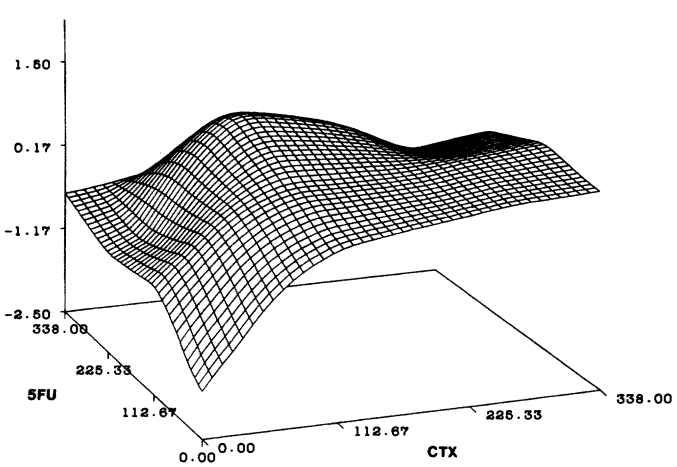

Figure 5. Kernel Estimate of g(x) for the Leukemia Data: b = .30.

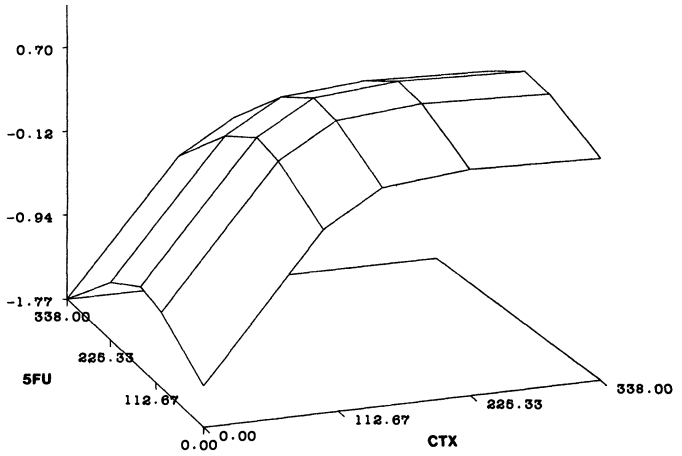

Figure 6. Generalized Additive Model of g(x) for the Leukemia Data: Estimate of the Overall Surface g(x).

7. CONCLUSIONS

We do not intend to argue that a nonparametric estimator of the response surface $g(x)$ replace the parametric estimator when the model is known to be correct. Rather, we suggest it be used when one is not prepared to adopt a parametric model, or that it be used to aid in the selection/validation of a parametric model.

If one is interested in the parametric modeling of a response surface when $d > 1$, then the generalized additive model *might* provide more insight than the overall surface fit with splines or kernel estimators. Nevertheless, the overall surface provided by kernel estimators and splines remains useful. The GAIM package for the generalized additive model is a wonderful tool for exploratory analysis and model building, but it is limited in its ability to detect structures in the data that are hidden by the projections onto $x_1, \ldots, x_d$, respectively.

As described in Section 3, existing local bandwidth selection procedures for kernel estimators can be easily extended to the more general models considered in this article. Kernel estimators with a local bandwidth selection procedure have the potential to reflect the local structure in the data more accurately than the smoothing splines that use a global smoothing parameter (O'Sullivan, Yandell, and Raynor 1986).

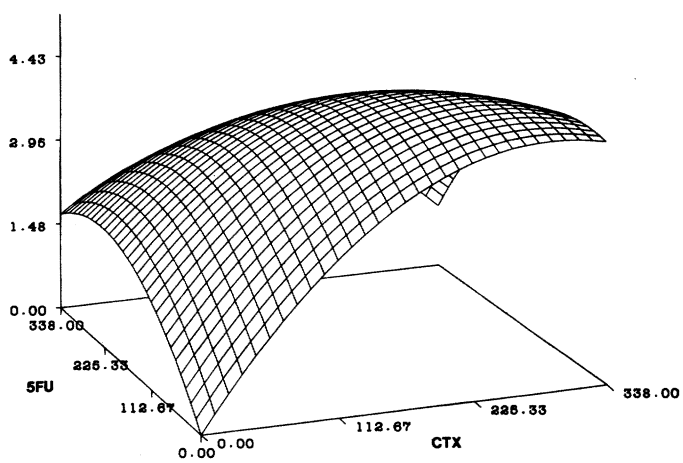

Figure 7. Parametric Estimate of g(x) for the Leukemia Data: Quadratic Model.

[*Received December 1986. Revised June 1988.*]

REFERENCES

Benedetti, J. K. (1977), "On the Nonparametric Estimation of Regression Functions," *Journal of the Royal Statistical Society*, Ser. B, 39, 248–253.

Carter, W. H., Jr., Wampler, G. L., and Stablein, D. M. (1983), *Regression Analysis Survival Data in Cancer Chemotherapy (Statistics: Textbooks and Monographs*, 44), New York: Marcel Dekker.

Cleveland, W. S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 823–836.

Cox, D. R. (1975), "Partial Likelihood," *Biometrika*, 62, 269–276.

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

Gasser, T., and Müller, H. G. (1979a), "Kernel Estimation of Regression Functions," in *Lecture Notes in Mathematics: Smoothing Techniques for Curve Estimation*, eds. T. Gasser and M. Rosenblatt, New York: Springer-Verlag, pp. 23–68.

——— (1979b), "Nonparametric Estimation of Regression Functions and Their Derivatives," Reprint 38 (Sonderforschungsbereich, 123), University of Heidelberg, West Germany, Dept. of Biostatistics.

Good, I. J., and Gaskins, R. A. (1971), "Nonparametric Roughness Penalties for Probability Densities," *Biometrika*, 58, 255–277.

Hastie, T., and Tibshirani, R. (1986), "Generalized Additive Models," *Statistical Science*, 1, 297–318.

Kalbfleisch, J. D., and Prentice, R. L. (1980), *The Statistical Analysis of Failure Time Data*, New York: John Wiley.

Kaplan, E. L., and Meier, P. (1958), "Nonparametric Estimation From Incomplete Observations," *Journal of the American Statistical Association*, 53, 457–481.

Klotz, J. (1982), "Spline Smooth Estimates of Survival," *IMS Lecture Notes: Survival Analysis*, 2, 14–19.

Miller, R., and Halpern, J. (1982), "Regression With Censored Data," *Biometrika*, 69, 521–531.

Müller, H. G. (1984), "Smooth Optimum Kernel Estimators of Densities, Regression Curves, and Modes," *The Annals of Statistics*, 12, 766–774.

Müller, H. G., and Stadtmüller, U. (1987), "Variable Bandwidth Kernel Estimators of Regression Curves," *The Annals of Statistics*, 15, 182–201.

Nadaraya, E. A. (1964), "On Estimating Regression," *Theory of Probability and Its Applications*, 9, 141–142.

O'Sullivan, F. (1986), "Relative Risk Estimation," Technical Report 76, University of California, Berkeley, Dept. of Statistics.

O'Sullivan, F., Yandell, B. S., and Raynor, W. J., Jr. (1986), "Automatic Smoothing of Regression Functions in Generalized Linear Models," *Journal of the American Statistical Association*, 81, 96–103.

Priestly, M. B., and Chao, M. T. (1972), "Nonparametric Function Fitting," *Journal of the Royal Statistical Society*, Ser. B, 34, 385–392.

Rice, J. (1984a), "Boundary Modification for Kernel Regression," *Communications in Statistics, Part A—Theory and Methods*, 13, 893–900.

——— (1984b), "Bandwidth Choice for Nonparametric Regression," *The Annals of Statistics*, 12, 1215–1230.

SAS Institute, Inc. (1982), *SAS User's Guide: Statistics*, Cary, NC: Author.

Schuster, E. F. (1972), "Joint Asymptotic Distribution of the Estimated Regression Function at a Finite Number of Points," *The Annals of Statistics*, 43, 84–88.

Staniswalis, J. G. (1985), "Local Bandwidth Selection for Kernel Estimates," unpublished Ph.D. dissertation, University of California, San Diego, Dept. of Mathematics.

——— (1987), "A Weighted Likelihood Formulation for Kernel Estimators of a Regression Function With Biomedical Applications," Technical Report 5, Virginia Commonwealth University, Medical College of Virginia, Dept. of Biostatistics.

Staniswalis, J. G., and Cooper, V. D. (1988), "Kernel Estimates of Dose–Response," *Biometrics*, 44, 1103–1119.

Staniswalis, J. G., and McCrady, C. W. (1988), "The Use of Kernel Estimators in Describing Human *T* Lymphocyte Proliferation Induced by Phorbol Esters and Ca^{2+} Ionophore," *Journal of the American College of Toxicology*, 7, 939–951.

Tibshirani, R., and Hastie, T. (1987), "Local Likelihood Estimation," *Journal of the American Statistical Association*, 82, 559–567.

Wahba, G., and Wold, S. (1975), "A Completely Automatic French Curve," *Communications in Statistics—Theory and Methods*, 4, 1–17.

Watson, G. S. (1964), "Smooth Regression Analysis, *Sankhyā*, Ser. A, 26, 359–372.

Whittemore, A., and Keller, J. (1986), "Survival Estimation Using Splines," *Biometrics*, 42, 495–506.