The Annals of Statistics
1997, Vol. 25, No. 4, 1661–1690

LOCAL LIKELIHOOD AND LOCAL PARTIAL LIKELIHOOD IN HAZARD REGRESSION

BY JIANQING FAN,[1] IRÈNE GIJBELS[2] AND MARTIN KING

Hong Kong Chinese University, Université Catholique de Louvain and University of North Carolina

In survival analysis, the relationship between a survival time and a covariate is conveniently modeled with the proportional hazards regression model. This model usually assumes that the covariate has a log-linear effect on the hazard function. In this paper we consider the proportional hazards regression model with a nonparametric risk effect. We discuss estimation of the risk function and its derivatives in two cases: when the baseline hazard function is parametrized and when it is not parametrized. In the case of a parametric baseline hazard function, inference is based on a local version of the likelihood function, while in the case of a nonparametric baseline hazard, we use a local version of the partial likelihood. This results in maximum local likelihood estimators and maximum local partial likelihood estimators, respectively. We establish the asymptotic normality of the estimators. It turns out that both methods have the same asymptotic bias and variance in a common situation, even though the local likelihood method uses information about the baseline hazard function.

1. The proportional hazards regression model. In survival analysis, one is interested in exploring the possible relationship between a survival time T and a covariate X. It is often convenient to work with the *conditional hazard function* of T given $X = x$, defined as

$$\lambda(t|x) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} P\{t \le T < t + \Delta t | T \ge t, X = x\},$$

which is the instantaneous rate of failure at time t, given a particular value x for the covariate X. The *proportional hazards model* is often used to describe the covariate effect on the survival time. The model is given by

$$\lambda(t|x) = \lambda_0(t)\Psi(x). \tag{1.1}$$

When $\Psi(0) = 1$, the function $\lambda_0(t)$ is the conditional hazard function of T given $X = 0$, and it is called the *baseline hazard function*. Under model (1.1), the conditional failure rates associated with any two values of the covariate

Received November 1995; revised November 1996.
[1]On leave from the University of North Carolina at Chapel Hill and supported by NSF Grant DMS-95-04414 and NSA Grant 96-1-0015.
[2]Supported by "Projet d'Actions de Recherche Concertées" (No. 93/98-164) and Research Grant 15.001.95F of the National Science Foundation (FNRS), Belgium.
AMS 1991 *subject classifications*. Primary 62G05; secondary 62E20, 60G44.
Key words and phrases. Asymptotic normality, censored data, local likelihood, local partial likelihood, proportional hazards.

X are proportional. The most common form of the proportional hazards regression model is obtained by taking the reparametrization

$$\Psi(x) = \exp\{\psi(x)\}, \tag{1.2}$$

as is done in this paper. Model (1.2) was introduced by Cox (1972). See Fleming and Harrington (1991), Andersen, Borgan, Gill and Keiding (1993), and references therein for the literature concerning this model.

In many applications, the survival times of some subjects are not fully observed; instead they are censored. Consider the bivariate data $\{(X_i, T_i): i = 1, \ldots, n\}$, which form an i.i.d. sample from the population (X, T). For a variety of reasons, including, for example, termination of the study or early withdrawal from the study, not all of the survival times $T_1, \ldots, T_n$ may be fully observable. Those incomplete observations are right censored. The observed data can then be formulated as follows. Suppose we have an *independent censoring scheme*, in which i.i.d. censoring times $C_1, \ldots, C_n$ are independent of the survival times given the covariates. We observe for the ith participant an *event-time* $Z_i = \min(T_i, C_i)$, a censoring indicator $\delta_i = I\{T_i \leq C_i\}$, as well as an associated covariate X_i. Then we denote the observed data

$$\{(X_i, Z_i, \delta_i): i = 1, \ldots, n\},$$

which are an i.i.d. sample from the population

$$(X, \min(T, C), I\{T \leq C\}).$$

For convenience, we assume throughout this paper that the random variables T and C are positive and continuous. The covariate X is assumed to remain constant over time.

It can be shown that under the proportional hazards model (1.1),

$$\Psi(x) = \frac{E\{\delta | X = x\}}{E\{\Lambda_0(Z) | X = x\}}, \tag{1.3}$$

where $\Lambda_0(t) = \int_0^t \lambda_0(u)\, du$ denotes the *cumulative baseline hazard function*. This implies that $\Psi(x)$ can easily be estimated if the baseline hazard function is known.

If a parametric form is assumed for both the baseline hazard function $\lambda_0(\cdot)$ and the risk factor $\psi(\cdot)$, inference is often based on the likelihood function. See for example Aitkin and Clayton (1980). If the function $\psi(\cdot)$ is not parametrized, we work with a local version of this likelihood function. Gentleman and Crowley (1991) propose such a version, with the uniform kernel function, when $\lambda_0(\cdot)$ is known and use an iterative algorithm when $\lambda_0(\cdot)$ is unknown. In the first part of this paper, we modify the estimator of Gentleman and Crowley (1991) (hereafter designated as G–C) to allow for a general kernel function and to permit derivative estimation. Our approach is noniterative when $\lambda_0(\cdot)$ is parametrized. The asymptotic distribution of the resulting local likelihood estimators is established.

When the baseline hazard function is completely unknown and the form of the function $\psi(x)$ is given, then inference can be based on the *partial likelihood*, as in Cox (1972). If $\psi(x)$ is not parametrized, which is the main

interest in this paper, we use a local version of the partial likelihood function. This local partial likelihood is introduced in Tibshirani and Hastie (1987) with nearest neighbor type of uniform windows. The generalization to nonuniform kernels requires some careful thoughts. To this end, we derive a general form of the local partial likelihood from a profile point of view. This sheds some new light on the concept of partial likelihood, and elaborates some connections between the local likelihood and the local partial likelihood. One of the main objectives in this paper is to derive the sampling properties of the proposed local likelihood and local partial likelihood estimators. These properties have not been studied so far, not even for the case of a uniform kernel.

When the baseline hazard function is parametrized, both the local likelihood and the local partial likelihood estimator can be employed. The former uses the knowledge of the baseline hazard function while the latter ignores this knowledge. To gain some insight into the efficiency of the local partial likelihood estimators, we compare the asymptotic biases and variances of both estimators. A bit surprisingly, we find that the methods have the same asymptotic bias and variance for the most common situation, where the derivative curve is estimated with a local quadratic fit. In other words, asymptotically, knowledge of the baseline hazard function does not provide any extra information about the derivative curve. Hence, it is preferable to use the local partial likelihood method since it is robust against misspecification of a parametric form of the baseline hazard function and has the same asymptotic efficiency as the local full likelihood method.

We focus in this paper on estimation of the risk factor $\psi(\cdot)$, describing the effect of the covariate on the hazard function. Estimation of the baseline hazard function is of secondary interest here, and procedures for this estimation task will be discussed rather briefly.

There are a vast number of references on nonparametric regression techniques and it is impossible to mention them all. See the references in Hastie and Tibshirani (1990a) and Fan and Gijbels (1996) for starting points. Much closely related work has been done on censored data. Marron and Padgett (1987), Müller and Wang (1990, 1994), Stute and Wang (1993), and Hjort (1996) among others, have studied the estimation of hazard rates via kernel methods. Nonparametric estimation of the conditional hazard and distribution function using local linear fits was investigated in Li and Doss (1995). Tibshirani and Hastie (1987) and G–C use a local modeling principle to study the nonparametric proportional hazards model, while O'Sullivan (1988), Hastie and Tibshirani (1990b) and Kooperberg, Stone and Truong (1995a, b), among others, use spline methods to study the model. These papers give convincing examples that demonstrate the usefulness of nonparametric techniques and inspire our current work.

Section 2 of this paper deals with the situation where the baseline hazard function is parametrized and discusses inference based on the local likelihood. Section 3 considers a nonparametric baseline hazard function, introduces the local partial likelihood estimators and studies their asymptotic

behavior. Section 4 briefly discusses the relative efficiency between the local likelihood method and the local partial likelihood method. In Section 5 we investigate the finite sample behavior of the local partial likelihood method. The proofs of the presented results are given in Section 6.

2. Parametric baseline hazard function.

2.1. *Likelihood function.* Assume temporarily that the baseline hazard function $\lambda_0(\cdot)$ has been parametrized as $\lambda_0(t) = \lambda_0(t;\theta)$ and that $\psi(x)$ has been parametrized as $\psi(x) = \psi(x;\beta)$. Let $f(t|x)$ denote the conditional density function of T given $X = x$, and let $S(t|x) = P\{T > t|X = x\}$ be its conditional survivor function. The conditional distribution function of C given $X = x$ is denoted by $G(t|x)$. Under independent and *noninformative censoring*, which is when $G(t|x)$ does not involve the parameters θ and β, it can be shown that the conditional likelihood function is given by

$$L = \prod_u f(Z_i|X_i)\prod_c S(Z_i|X_i),$$

where $\prod_u$ and $\prod_c$ denote, respectively, the product involving the uncensored and the censored observations. This kind of likelihood appears often in the literature [c.f. Aitkin and Clayton (1980)]. See also Section 5.3.4 of Fan and Gijbels (1996) for a simple derivation of this likelihood function. Under the proportional hazards model (1.1), we have

$$\log L = \sum_{i=1}^{n} \big[\delta_i\{\log \lambda_0(Z_i;\theta) + \psi(X_i;\beta)\} - \Lambda_0(Z_i;\theta)\exp\{\psi(X_i;\beta)\}\big]. \tag{2.1}$$

Maximization of (2.1) leads to the maximum likelihood estimators of θ and β.

2.2. *Local likelihood.* Suppose now that the form of $\psi(x)$ is not specified, and that the pth order derivative of $\psi(x)$ at the point x exists. Then, by Taylor's expansion,

$$\psi(X) \approx \psi(x) + \psi'(x)(X - x) + \cdots + \frac{\psi^{(p)}(x)}{p!}(X - x)^p,$$

for X in a neighborhood of x. Let h be the bandwidth parameter that controls the size of the local neighborhood and let K be a kernel function that smoothly weighs down the contribution of remote data points. Set

$$\mathbf{X} = \big\{1, X - x, \ldots, (X - x)^p\big\}^T \quad \text{and} \quad \mathbf{X}_i = \big\{1, X_i - x, \ldots, (X_i - x)^p\big\}^T,$$

where T denotes the transpose of a vector. Then, locally around x, $\psi(X)$ can be modeled as

$$\psi(X) \approx \mathbf{X}^T\beta, \tag{2.2}$$

where $\beta = (\beta_0, \ldots, \beta_p)^T = \{\psi(x), \ldots, \psi^{(p)}(x)/p!\}^T$. Using the local model (2.2), and incorporating the localizing weights, we obtain the local (log) likelihood

$$l_n(\beta, \theta) = n^{-1} \sum_{i=1}^{n} \big[\delta_i \{\log \lambda_0(Z_i; \theta) + \mathbf{X}_i^T \beta\} - \Lambda_0(Z_i; \theta) \exp(\mathbf{X}_i^T \beta) \big] K_h(X_i - x), \tag{2.3}$$

where $K_h(t) = h^{-1}K(t/h)$. Compare with (2.1). This local likelihood, with a uniform kernel and a given λ_0, was used also by G–C. It can be easily seen that $l_n(\beta, \theta)$ is strictly concave with respect to β. Hence, this local log-likelihood has a unique maximizer with respect to β. However, $l_n(\beta, \theta)$ is not always concave with respect to β and θ. See Section 2.4 for further discussion on this issue.

Let $\hat{\beta}$ and $\hat{\theta}$ be the maximizers of (2.3). Then, according to our parametrization, a natural estimator of $\psi^{(\nu)}(x)$, for $\nu = 0, \ldots, p$, is

$$\hat{\psi}_\nu(x) = \nu! \hat{\beta}_\nu. \tag{2.4}$$

As x varies across the range of a set, we obtain an estimated (derivative) curve on that set.

2.3. *Sampling properties*. Throughout this section we assume that the censoring scheme is independent and noninformative. We will use the superscripts $'$ and $''$ to indicate the gradient vector and the Hessian matrix, respectively, of a function. For example, $\lambda_0'(t; \theta)$ denotes the gradient vector of $\lambda_0(t; \theta)$ with respect to the parameter vector θ. For purposes of identifiability, we assume that $\lambda_0(1; \theta) = 1$. The local maximum likelihood estimators $\hat{\beta}$ and $\hat{\theta}$ solve the local likelihood equation $l_n'(\beta, \theta) = 0$. This equation is given by

$$\begin{aligned} \frac{\partial l_n(\beta, \theta)}{\partial \beta} &= n^{-1} \sum_{i=1}^{n} \big[\delta_i - \Lambda_0(Z_i; \theta) \exp(\mathbf{X}_i^T \beta) \big] \\ &\qquad \times \mathbf{X}_i K_h(X_i - x) = 0, \\ \frac{\partial l_n(\beta, \theta)}{\partial \theta} &= n^{-1} \sum_{i=1}^{n} \big\{ \delta_i \xi_0'(Z_i; \theta) - \Lambda_0'(Z_i; \theta) \exp(\mathbf{X}_i^T \beta) \big\} \\ &\qquad \times K_h(X_i - x) = 0, \end{aligned} \tag{2.5}$$

where we introduce the notation $\xi_0(t; \theta) = \log \lambda_0(t; \theta)$. Let θ_0 be the true parameter of $\lambda_0(t; \theta)$, and similarly let

$$\beta^0 = \{\psi(x), \ldots, \psi^{(p)}(x)/p!\}^T,$$

be the true parameter in the local model (2.2). In order to have a consistency property for $\hat{\beta}$ and $\hat{\theta}$, the true parameters β^0 and θ_0 must solve the asymptotic counterpart of (2.5). That this holds for the first local likelihood equation follows from (1.3). For the second local likelihood equation, this is justified by the following proposition.

PROPOSITION 1. *If θ_0 is an interior point of the parameter space, and*

$$E\{\|\delta\xi_0'(Z;\theta_0)\|\,|X=x\}<\infty \quad \text{and} \quad E\{\|\Lambda_0(Z;\theta_0)\|\,|X=x\}<\infty,$$

then

$$E\{\delta\xi_0'(Z;\theta_0)-\Lambda_0'(Z;\theta_0)\Psi(x)\,|X=x\}=0, \tag{2.6}$$

where $\Psi(x)=\exp\{\psi(x)\}$.

PROOF. Let $N(t)=I\{Z\le t,\ \delta=1\}$, $Y(t)=I\{Z\ge t\}$ and let

$$\mathscr{F}_t=\sigma\{X,N(u),Y(u),0\le u\le t\}$$

be the history up to time t. Set

$$M(t)=N(t)-\int_0^t Y(u)\Psi(X)\lambda_0(u;\theta_0)\,du.$$

Then, $M(t)$ is an $\mathscr{F}_t$-martingale. Note that

$$\delta\xi_0'(Z;\theta_0)-\Lambda_0'(Z;\theta_0)\Psi(X)=\int_0^\infty \xi_0'(t;\theta_0)\,dM(t).$$

Since $\xi_0'(t;\theta_0)$ is $\mathscr{F}_t$-measurable, (2.6) follows by taking the conditional expectation of the above equality with respect to Z given $X=x$. This completes the proof. □

Equation (2.6) gives the first-order Bartlett identity of the local likelihood. We also need the following second-order Bartlett identity.

PROPOSITION 2. *If θ_0 is an interior point of the parameter space, and*

$$E\Big[\big\|\delta\{\xi_0'(Z;\theta_0)\}^{\otimes 2}\big\|\,|X=x\Big]<\infty \quad \text{and} \quad E\{\|\Lambda_0''(Z;\theta_0)\|\,|X=x\}<\infty,$$

then

$$\begin{aligned}&E\{\delta\xi_0''(Z;\theta_0)-\Lambda_0''(Z;\theta_0)\Psi(x)\,|X=x\}\\&\quad=-E\Big[\delta\{\xi_0'(Z;\theta_0)\}^{\otimes 2}\,|X=x\Big],\end{aligned} \tag{2.7}$$

where $A^{\otimes 2}$ denotes AA^T for a vector or matrix A.

PROOF. Using the notation of the proof of Proposition 1, we have

$$\delta\frac{\lambda_0''(Z;\theta_0)}{\lambda_0(Z;\theta_0)}-\Lambda_0''(Z;\theta_0)\Psi(X)=\int_0^\infty \frac{\lambda_0''(u;\theta_0)}{\lambda_0(u;\theta_0)}\,dM(u).$$

Hence,

$$E\left\{\delta\frac{\lambda_0''(Z;\theta_0)}{\lambda_0(Z;\theta_0)}-\Lambda_0''(Z;\theta_0)\Psi(x)\,|X=x\right\}=0.$$

Using this and

$$\xi_0''(Z;\theta_0)=\frac{\lambda_0''(Z;\theta_0)}{\lambda_0(Z;\theta_0)}-\left\{\frac{\lambda_0'(Z;\theta_0)}{\lambda_0(Z;\theta_0)}\right\}^{\otimes 2}$$

we obtain (2.7). □

Before we state the main results of this section, we first introduce some notation. Let

$$H = \operatorname{diag}\{1, h, \ldots, h^p\}^T, \qquad \mathbf{u} = \{1, u, \ldots, u^p\}^T$$

and put

$$S_0(x;\theta_0) = \int_{-\infty}^{+\infty} E\left\{\delta\begin{pmatrix}\mathbf{u}\\ \xi_0'(Z;\theta_0)\end{pmatrix}^{\otimes 2}\middle| X = x\right\}K(u)\,du,$$

$$S_1(x;\theta_0) = \int_{-\infty}^{+\infty} E\left\{\delta\begin{pmatrix}\mathbf{u}\\ \xi_0'(Z;\theta_0)\end{pmatrix}^{\otimes 2}\middle| X = x\right\}K^2(u)\,du \tag{2.8}$$

and

$$b_n(x) = \frac{\psi^{(p+1)}(x)}{(p+1)!}h^{p+1}\left\{\int \mathbf{u}\mathbf{u}^T K(u)\,du\right\}^{-1}\int u^{p+1}\mathbf{u}K(u)\,du. \tag{2.9}$$

We now impose some convenient conditions.

CONDITION A. (i) θ_0 is an interior point of the parameter space.

(ii) There exists an $\eta > 0$ such that

$$E\{\Lambda_0(Z;\theta_0)^{2+\eta}\,|X\}, \qquad E\{|\xi_0'(Z;\theta_0)|^{2+\eta}\,|X\} \quad \text{and} \quad E\{|\Lambda_0'(Z;\theta_0)|^{2+\eta}\,|X\}$$

are finite and continuous at the point $X = x$.

(iii) The functions

$$E(\delta|X), \quad E\{\Lambda_0(Z;\theta_0)\,|X\}, \quad E\{\Lambda_0'(Z;\theta_0)\,|X\}, \quad E\{\Lambda_0''(Z;\theta_0)\,|X\},$$
$$E\{\delta\xi_0'(Z;\theta_0)\,|X\} \quad \text{and} \quad E\{\delta\xi_0''(Z;\theta_0)\,|X\}$$

are continuous at the point $X = x$.

(iv) There exists a function $M(z)$, with $EM(Z) < \infty$, such that

$$\left|\frac{\partial^3}{\partial\theta_j\,\partial\theta_k\,\partial\theta_l}\xi_0(z;\theta)\right| < M(z), \qquad \left|\frac{\partial^3}{\partial\theta_j\,\partial\theta_k\,\partial\theta_l}\Lambda_0(z;\theta)\right| < M(z),$$

for all z, and for all θ in a neighborhood of θ_0.

(v) The kernel function $K \geq 0$ is a bounded density with a compact support.

(vi) The function $\psi(\cdot)$ has a continuous $(p+1)$th derivative around the point x.

(vii) The density $f(\cdot)$ of X is continuous at the point x and $f(x) > 0$.

(viii) $nh \to \infty$ and nh^{2p+3} is bounded.

THEOREM 1. *Under Condition* A, *there exists a solution* $\hat\beta$ *and* $\hat\theta$ *to the local likelihood equation* (2.5) *such that*

$$H(\hat\beta - \beta^0) \to_P 0 \quad \textit{and} \quad \hat\theta - \theta_0 \to_P 0.$$

THEOREM 2. *Under Condition* A, *the solution given in Theorem* 1 *is asymptotically normal*:

$$\sqrt{nh}\begin{pmatrix} H(\hat{\beta}-\beta^0)-b_n(x) \\ \hat{\theta}-\theta_0 \end{pmatrix} \to_D N\{0, f^{-1}(x)S_0(x;\theta_0)^{-1}S_1(x;\theta_0)S_0(x;\theta_0)^{-1}\}. \tag{2.10}$$

REMARK 1. Note that the bias term

$$b_n(x) = \frac{\psi^{(p+1)}(x)}{(p+1)!}h^{p+1}\left\{\int \mathbf{u}\mathbf{u}^T K(u)\,du\right\}^{-1}\int u^{p+1}\mathbf{u}K(u)\,du,$$

of $\hat{\beta}$ admits the same expression as that of the least-squares nonparametric regression estimator. See expression (3.8) of Fan and Gijbels (1996). The explanation for this is that the bias comes from the approximation error and hence is independent of the model.

REMARK 2. For the parametric linear model $\psi(X;\beta) = \mathbf{X}^T\beta$, one would directly maximize the log-likelihood (2.1). In that case, our proofs of Theorems 1 and 2 show that

$$\sqrt{n}\begin{pmatrix} \hat{\beta}-\beta^0 \\ \hat{\theta}-\theta_0 \end{pmatrix} \to_D N(0, \Sigma^{-1}),$$

where $\Sigma = E\left\{\delta\begin{pmatrix} \mathbf{X} \\ \xi_0'(Z;\theta_0) \end{pmatrix}^{\otimes 2}\right\}$.

REMARK 3. The likelihood equations (2.5) depend on the point x in which we would like to estimate the function ψ and its derivatives. Only data points falling in a certain neighborhood of x are involved in determining $\hat{\beta} = \hat{\beta}(x)$ and $\hat{\theta} = \hat{\theta}(x)$. Using simply the estimator $\hat{\theta}(x)$ to estimate θ_0, resulting from solving equations (2.5), is not satisfactory and does not reflect our model assumption. All data points should be used in order to efficiently estimate the global parameter θ_0. A possible approach is to maximize (2.3), with θ fixed, over a range of x-values to obtain an estimate $\hat{\psi}(\cdot,\theta)$ for each θ, and then to maximize (2.1) with $\psi(X_i;\beta)$ replaced by $\hat{\psi}(X_i,\theta)$. This is essentially a profile likelihood method and can be implemented by the following simple iterative algorithm: given θ, obtain $\hat{\psi}(\cdot)$ and given $\hat{\psi}(\cdot)$, update θ and so on. This idea is similar to that given in G–C.

As an illustration of Theorem 2, we now consider the particular situation that we estimate $\psi'(\cdot)$ using a local quadratic fit. In this case, $p = 2$ and $\nu = 1$. The bias and variance of the local likelihood estimator for $\psi'(x)$ is then given by the second marginal component of (2.10). For this special case we obtain the following asymptotic normality result.

COROLLARY 1. *Under the conditions of Theorem 2 with* $p = 2$, *and if* K *is symmetric, then*

$$\sqrt{nh^3}\left\{\hat{\psi}_1(x) - \psi'(x) - \frac{1}{6}\int t^3 K_1^*(t)\,dt\,\psi^{(3)}(x)h^2\right\} \to_D N\left\{0, \frac{\sigma^2(x)}{f(x)}\int K_1^*(t)^2\,dt\right\},$$

where $\sigma^2(x) = E\{\delta|X = x\}^{-1}$ *and* $K_1^*(t) = tK(t)/\int t^2K(t)\,dt$.

Note that when θ_0 is known, one would maximize (2.3) with respect to β. The resulting derivative estimator has the same asymptotic bias and variance as those given in Corollary 1. In other words, under the conditions given in Corollary 1, $\hat{\psi}_1(x)$ is adaptive in the sense that it estimates $\psi'(x)$ as well as in the case that θ_0 is given.

2.4. *Concavity of the local likelihood.* As in most of the parametric likelihood theory [see, e.g., Sections 6.3 and 6.4 of Lehmann (1983)], we only know that there exists a consistent solution to the local likelihood equation. But if there are multiple roots, we don't know which solution is consistent. However, if $l_n(\beta, \theta)$ is strictly concave, then the solution to (2.5) is unique and must be consistent.

The Hessian matrix of $l_n(\beta, \theta)$ is given by

$$\begin{aligned} l_n''(\beta,\theta) &= n^{-1}\sum_{i=1}^n K_h(X_i - x) \\ &\quad\times\begin{pmatrix} -\Lambda_0(Z_i;\theta)\exp(\mathbf{X}_i^T\beta)\mathbf{X}_i\mathbf{X}_i^T & -\exp(\mathbf{X}_i^T\beta)\mathbf{X}_i\Lambda_0'(Z_i;\theta)^T \\ -\exp(\mathbf{X}_i^T\beta)\Lambda_0'(Z_i;\theta)\mathbf{X}_i^T & -\exp(\mathbf{X}_i^T\beta)\Lambda_0'(Z_i;\theta) + \delta_i\xi_0''(Z_i;\theta) \end{pmatrix} \\ &= -n^{-1}\sum_{i=1}^n K_h(X_i - x)\Lambda_0(Z_i;\theta)\exp(\mathbf{X}_i^T\beta)\begin{pmatrix} X_i \\ \Lambda_0'(Z_i;\theta)/\Lambda_0(Z_i;\theta) \end{pmatrix}^{\otimes 2} \\ &\quad -n^{-1}\sum_{i=1}^n K_h(X_i - x) \\ &\quad\times\begin{pmatrix} 0 & 0 \\ 0 & \exp(\mathbf{X}_i^T\beta)\Lambda_0(Z_i;\theta)\{\log\Lambda_0(Z_i;\theta)\}'' - \delta_i\xi_0''(Z_i;\theta) \end{pmatrix}. \end{aligned} \tag{2.11}$$

Clearly the first term on the right-hand side of (2.11) is negatively definite and if

$$\{\log\Lambda_0(Z_i;\theta)\}'' \geq 0 \quad\text{and}\quad \xi_0''(Z_i;\theta) \leq 0, \tag{2.12}$$

then $l_n(\beta, \theta)$ is strictly concave. We summarize the result as follows.

THEOREM 3. *A necessary condition for the strict concavity of* $l_n(\beta, \theta)$ *is condition* (2.12). *For this special case, the maximum local likelihood estimator is unique and possesses the property stated in Theorem* 2.

EXAMPLE (Weibull distribution). For the Weibull baseline distribution with scale parameter ρ and index κ, the hazard and cumulative hazard function are given by

$$\lambda_0(t; \rho, \kappa) = \kappa\rho(\rho t)^{\kappa-1} \quad \text{and} \quad \Lambda_0(t; \rho, \kappa) = (\rho t)^{\kappa}.$$

Model (1.1) can be written as

$$\lambda(t|x) = t^{\kappa-1}\exp\{\psi_0(x)\} \quad \text{where } \psi_0(x) = \psi(x) + \log\kappa + \kappa\log\rho,$$

and our estimation equations (2.5) estimate κ and $\psi_0(\cdot)$. Note that ρ is not identifiable. The above expression can also be seen from the identifiability condition $\lambda_0(1; \theta) = 1$, which implies that $\kappa\rho^{\kappa} = 1$ and hence $\lambda_0(t; \rho, \kappa) = t^{\kappa-1}$. Clearly, $\lambda_0(t; \rho, \kappa)$ satisfies condition (2.12).

3. Nonparametric baseline hazard function.

3.1. *From likelihood to partial likelihood.* Let $t_1 < \cdots < t_N$ denote the ordered failure times and let (j) denote the label of the item failing at t_j. Denote by $\mathscr{R}_j$ the risk set at time t_j- that is, $\mathscr{R}_j = \{i\colon Z_i \geq t_j\}$. Consider the "least informative" nonparametric modeling for $\Lambda_0(\cdot)$; that is, $\Lambda_0(t)$ has a jump θ_j at t_j. More precisely,

$$\Lambda_0(t; \theta) = \sum_{j=1}^{N} \theta_j I\{t_j \leq t\}. \tag{3.1}$$

Then

$$\Lambda_0(Z_i; \theta) = \sum_{j=1}^{N} \theta_j I\{i \in \mathscr{R}_j\}.$$

Substituting these two expressions into (2.1), we obtain that

$$\log L = \sum_{j=1}^{N}\left[\log\theta_j + \psi\{X_{(j)}; \beta\}\right] - \sum_{i=1}^{n}\left[\sum_{j=1}^{N}\theta_j I\{i \in \mathscr{R}_j\}\exp\{\psi(X_i; \beta)\}\right]. \tag{3.2}$$

The maximum of $\log L$ with respect to θ_j $(j = 1, \ldots, N)$ is obtained at

$$\hat{\theta}_j = \left[\sum_{i \in \mathscr{R}_j}\exp\{\psi(X_i; \beta)\}\right]^{-1}. \tag{3.3}$$

This is the Breslow estimator of the baseline hazard function [see Breslow (1972, 1974)]. Substituting (3.3) into (3.2), we obtain

$$\max_{\lambda_0}\log L = \sum_{j=1}^{N}\left(\psi\{X_{(j)}; \beta\} - \log\left[\sum_{i \in \mathscr{R}_j}\exp\{\psi(X_i; \beta)\}\right]\right) - N. \tag{3.4}$$

Therefore, the maximum likelihood estimate of β under the nonparametric model (3.1) is the β that maximizes (3.4). The objective function in (3.4) is the same as the partial likelihood function in Cox (1975) and is a profile likelihood. This kind of derivation is due to Breslow (1972). Theory on partial likelihood can be found in Wong (1986).

In summary, when $\lambda_0(\cdot)$ is not specified, one should use the maximum partial likelihood estimator. This method is equivalent to the maximum likelihood estimator with the "least informative" baseline hazard function, namely with the cumulative hazard function parametrized as in (3.1) with a large number of parameters.

3.2. *Local partial likelihood.* When the forms of $\psi(x;\beta)$ and $\Lambda_0(t;\theta)$ are not specified, one should use the local model (2.2) along with a local version of the partial likelihood in (3.4). That is, find the β that maximizes the local partial likelihood

$$\sum_{j=1}^{N} K_h\{X_{(j)} - x\}\left[\mathbf{X}_{(j)}^T\beta - \log\left\{\sum_{i\in\mathscr{R}_j}\exp\left(\mathbf{X}_i^T\beta\right)K_h(X_i - x)\right\}\right]. \tag{3.5}$$

Clearly (3.5) is just a localized version of (3.4). It can also be derived from the local log-likelihood (2.3) using the "least-informative" nonparametric modeling (3.1). In other words, the maximum local partial likelihood estimator is the maximum local likelihood estimator when $\Lambda_0(\cdot;\theta)$ is parametrized with a large number of parameters, as in (3.1).

Let $\hat{\beta}$ maximize (3.5) with respect to $\beta = \{\beta_0, \ldots, \beta_p\}^T$. Then an obvious estimator of $\psi^{(\nu)}(x)$ is as in (2.4), namely,

$$\hat{\psi}_\nu(x) = \nu!\hat{\beta}_\nu,$$

where $\hat{\beta}_\nu$ estimates β_ν. Note that the function value $\psi(x)$ is not directly estimable; (3.5) does not involve the intercept $\beta_0 = \psi(x)$ since it cancels out. This is not surprising since from the proportional hazards model (1.1), it is already clear that $\psi(x)$ is only identifiable to within a constant factor. The identifiability of $\psi(x)$ is ensured by imposing the condition $\psi(0) = 0$. Then the function $\psi(x) = \int_0^x \psi'(t)\,dt$ can be estimated by

$$\hat{\psi}(x) = \int_0^x \hat{\psi}_1(t)\,dt.$$

For practical implementation, Tibshirani and Hastie (1987) suggested approximating the integration by the trapezoidal rule.

When there is more than one covariate, one could use a multivariate Taylor expansion to approximate $\psi(\cdot)$ locally with a polynomial of order p. This would lead to a straightforward generalization of the results presented here. However, a serious problem in multivariate situations is the curse of dimensionality. A possible approach to tackle this problem is to consider additive modeling [see Hastie and Tibshirani (1990b)] or low-order interaction models [see Kooperberg, Stone and Truong (1995a, b)]. The advantage of

such modeling is that one can use low-dimensional smoothing techniques such as the one proposed here as building blocks along with a backfitting type of algorithm [see, e.g., Hastie and Tibshirani (1990a)] to fit all low-dimensional functions involved. Another possible approach is to use the average of the estimated multivariate hazard regression surface to estimate each additive component. See, for example, Linton and Nielsen (1995) and Fan, Härdle and Mammen (1995) for details. We would anticipate that these two approaches for estimating the additive components would enjoy optimal rates of convergence as in Stone (1994). Formal theoretical derivations remain to be done.

3.3. *Estimation of the baseline hazard function.* As mentioned in the introduction, estimation of the baseline hazard function is not a primary goal of this paper. We therefore only briefly outline two possible approaches to this problem.

From (3.1) and (3.3), we suggest the following estimator for the cumulative hazard function:

$$\hat{\Lambda}_0(t) = \sum_{j=1}^{N} \left[\sum_{i \in \mathscr{R}_j} \exp\{\hat{\psi}(X_i)\} \right]^{-1} I\{t_j \le t\}. \tag{3.6}$$

A kernel smoothing technique can then be employed to obtain an estimate of $\lambda_0(t)$ via

$$\hat{\lambda}_0(t) = \int W_g(t - x)\, d\hat{\Lambda}_0(x), \tag{3.7}$$

where W is a given kernel function and g is a given bandwidth.

An alternative approach to estimating $\Lambda_0(\cdot)$ and $\lambda_0(\cdot)$ is to use a local polynomial fit. For simplicity, we use the local linear fit to illustrate the idea. Locally around a given point t_0, one can approximate

$$\Lambda_0(t) \approx \exp(\beta_0 + \beta_1(t - t_0)), \qquad \lambda_0(t) \approx \exp(\beta_0 + \beta_1(t - t_0))\beta_1$$

for t in a neighborhood of t_0. For a given estimator $\hat{\psi}(\cdot)$ of $\psi(\cdot)$, the local version of the likelihood (2.1) can be expressed as

$$\begin{aligned} \sum_{i=1}^{n} W_g(Z_i - t_0) \big[\delta_i \{ \beta_0 + \beta_1(Z_i - t_0) + \log \beta_1 + \hat{\psi}(X_i) \} \\ - \exp\{\beta_0 + \beta_1(Z_i - t_0)\} \exp\{\hat{\psi}(X_i)\} \big], \end{aligned} \tag{3.8}$$

where W is a kernel function and g is a bandwidth. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ maximize (3.8). Then,

$$\hat{\Lambda}_0(t_0) = \exp(\hat{\beta}_0) \quad \text{and} \quad \hat{\lambda}_0(t_0) = \exp(\hat{\beta}_0)\hat{\beta}_1$$

are smoothed type estimators for $\Lambda_0(t_0)$ and $\lambda_0(t_0)$, respectively.

The above approaches are quite different from the local full likelihood procedure proposed by G–C. In order to estimate both, the covariate effect and the baseline hazard function, G–C rely on the following iterative proce-

dure. For a given estimate of $\psi(\cdot)$, they estimate the baseline cumulative hazard function using all data points, and applying for example the Breslow estimator. Then, the obtained cumulative hazard function estimator is used to build up a local (full) likelihood as described in Section 2.2, and maximization of this local likelihood leads to an estimate of the function $\psi(\cdot)$. G–C suggest iterating between the two estimation steps until some convergence criterion is met.

As demonstrated in Corollary 2, our direct method of estimating $\psi'(x)$ is asymptotically already as good as the case where $\lambda_0(\cdot)$ is known. Thus, the iterative approach cannot improve asymptotically the efficiency of our local likelihood method. This is also demonstrated in our simulations.

3.4. *Asymptotic property of the maximum local partial likelihood estimator*. Since the local partial likelihood (3.5) does not involve $\beta_0 = \psi(x)$, we write

$$\beta^* = (\beta_1, \ldots, \beta_p)^T, \qquad \hat{\beta}^* = \left(\hat{\beta}_1, \ldots, \hat{\beta}_p\right)^T \quad \text{and}$$

$$\mathbf{X}_i^* = \left\{X_i - x, \ldots, (X_i - x)^p\right\}^T.$$

Correspondingly, let

$$\beta^{0*} = \left\{\psi'(x), \ldots, \psi^{(p)}(x)/p!\right\}^T, \qquad H^* = \operatorname{diag}(h, \ldots, h^p)^T,$$

$$\mathbf{u}^* = \{u, \ldots, u^p\}^T.$$

Set $\nu_1 = \int \mathbf{u}^* K(u)\, du$,

$$A = \int \mathbf{u}^* \mathbf{u}^{*T} K(u)\, du - \nu_1 \nu_1^T \quad \text{and} \quad B = \int K^2(u)(\mathbf{u}^* - \nu_1)^{\otimes 2}\, du. \tag{3.9}$$

Put

$$P(u|x) = P\{Z \geq u | X = x\} \quad \text{and} \quad \Lambda(t, x) = \int_0^t P(u|x)\, \lambda_0(u)\, du. \tag{3.10}$$

Writing the conditional probability as the conditional expectation of an indicator function, it can easily be seen by exchanging the integration with the conditional expectation that

$$\Lambda(t, x) = E\left[\Lambda_0\{\min(Z, t)\} \,|\, X = x\right],$$

where $\Lambda_0(\cdot)$ is the cumulative baseline hazard function. Finally, let $\hat{\beta}^*$ be the maximizer of the local partial likelihood

$$\sum_{j=1}^{N} K_h\{X_{(j)} - x\}\left[\mathbf{X}_{(j)}^{*T}\beta^* - \log\left\{\sum_{i \in \mathscr{R}_j} \exp\left(\mathbf{X}_i^{*T}\beta^*\right) K_h(X_i - x)\right\}\right] \tag{3.11}$$

with respect to β^*.

We now impose some convenient technical conditions for asymptotic normality.

CONDITION B. (i) The kernel function $K \geq 0$ is a bounded density function with compact support.

(ii) The function $\psi(\cdot)$ has a continuous $(p+1)$th derivative around the point x.

(iii) The density $f(\cdot)$ of X is continuous at the point x and $f(x) > 0$.

(iv) The conditional probability $P(u|\cdot)$ is equicontinuous at x.

(v) The sequence $nh \to \infty$ and nh^{2p+3} is bounded.

We now state the main result of this section.

THEOREM 4. *Under Condition* B, *we have*

$$\sqrt{nh}\left\{H^*(\hat{\beta}^* - \beta^{0*}) - \frac{\psi^{(p+1)}(x)}{(p+1)!}A^{-1}bh^{p+1}\right\} \to_D N\left\{0, \frac{\sigma^2(x)}{f(x)}A^{-1}BA^{-1}\right\}, \tag{3.12}$$

where $\sigma^2(x) = E\{\delta|X = x\}^{-1}$ *and* $b = \int u^{p+1}(\mathbf{u}^* - \nu_1)K(u)\,du$.

Note that in general the bias vector and variance matrix in (3.12) depend on K in a different way than in nonparametric regression [e.g., compare with (3.18) and (3.19) of Fan and Gijbels (1996)]. Theorem 4 gives the joint asymptotic normality of the derivative estimators. In particular, the bias and variance of $\hat{\psi}_\nu(x)$ can be obtained by taking the νth marginal component of (3.12). For example, if $p = 2$ and $\nu = 1$, we have the following result.

COROLLARY 2. *Under the conditions of Theorem* 4 *with* $p = 2$, *if* K *is symmetric*, *then*

$$\sqrt{nh^3}\left\{\hat{\psi}_1(x) - \psi'(x) - \frac{1}{6}\int t^3K_1^*(t)\,dt\,\psi^{(3)}(x)h^2\right\} \to_D N\left\{0, \frac{\sigma^2(x)}{f(x)}\int K_1^*(t)^2\,dt\right\}, \tag{3.13}$$

where $K_1^*(t) = tK(t)/\int t^2K(t)\,dt$.

As a consequence of (3.13), the theoretical optimal bandwidth, which minimizes the asymptotic weighted mean integrated squared error,

$$\int\left[\left\{\frac{1}{6}\int t^3K_1^*(t)\,dt\,\psi^{(3)}(x)h^2\right\}^2 + \frac{\sigma^2(x)}{nh^3f(x)}\int K_1^*(t)^2\,dt\right]w(x)\,dx,$$

is given by

$$h_{\text{opt}} = C(K)\left[\frac{\int \sigma^2(x)w(x)/f(x)\,dx}{\int\{\psi^{(3)}(x)\}^2 w(x)\,dx}\right]^{1/7} n^{-1/7},$$

with

$$C(K) = \left[\frac{27\int K_1^*(t)^2\,dt}{\{\int t^3 K_1^*(t)\,dt\}^2}\right]^{1/7} = \left[\frac{27\int t^2K^2(t)\,dt}{\{\int t^4K(t)\,dt\}^2}\right]^{1/7}.$$

4. Relative efficiency. In this section, we comment on the relative efficiency between the local likelihood and the local partial likelihood estimators when the baseline hazard function is correctly parametrized. We are interested in knowing how much efficiency is lost when the local partial likelihood method, which ignores the form of the baseline hazard function, is used.

For the convenience of the discussion, we assume that we use a local quadratic fit ($p = 2$) to estimate the derivative function $\psi'(x)$, and that we use a symmetric kernel. For this special case, the local likelihood and the local partial likelihood estimators have the same asymptotic bias and asymptotic variance (see Corollaries 1 and 2). Hence the relative efficiency of the two estimators, defined as the ratio of their asymptotic variances, is equal to 1. This result is somewhat surprising, because even though the information of the baseline hazard function was used in the local likelihood, it does not improve the sampling property of the local likelihood estimator. The reason for this asymptotic result is that we assume that $h \to 0$ and the kernel K is symmetric. Hence, in (2.3), the local parameter β_1 is asymptotically orthogonal to the other parameters β_0, β_2 and θ.

The above result is asymptotic in nature. We are not sure how small h should be in order for the orthogonality to exist, and hence it is possible for finite samples that the local likelihood estimator performs better than the local partial likelihood estimator.

We choose to compare derivative estimators instead of estimators for ψ itself because the local partial likelihood does not directly involve the local parameter for $\psi(x)$. We do not intend to compare the two methods for other cases, such as local cubic and quartic fits, since such a large order of local fit is rarely used in applications.

5. Simulation study. The proportional hazards model (1.1) is equivalent to the following transformed regression model:

$$\log \Lambda_0(T) = -\psi(X) + \log(\varepsilon), \tag{5.1}$$

where ε has the standard exponential distribution. Model (5.1) enables us to generate data easily from model (1.1) and to inspect visually the noise-to-signal ratio for a simulated model (see Figure 1). The following four simulated models are used in our simulations.

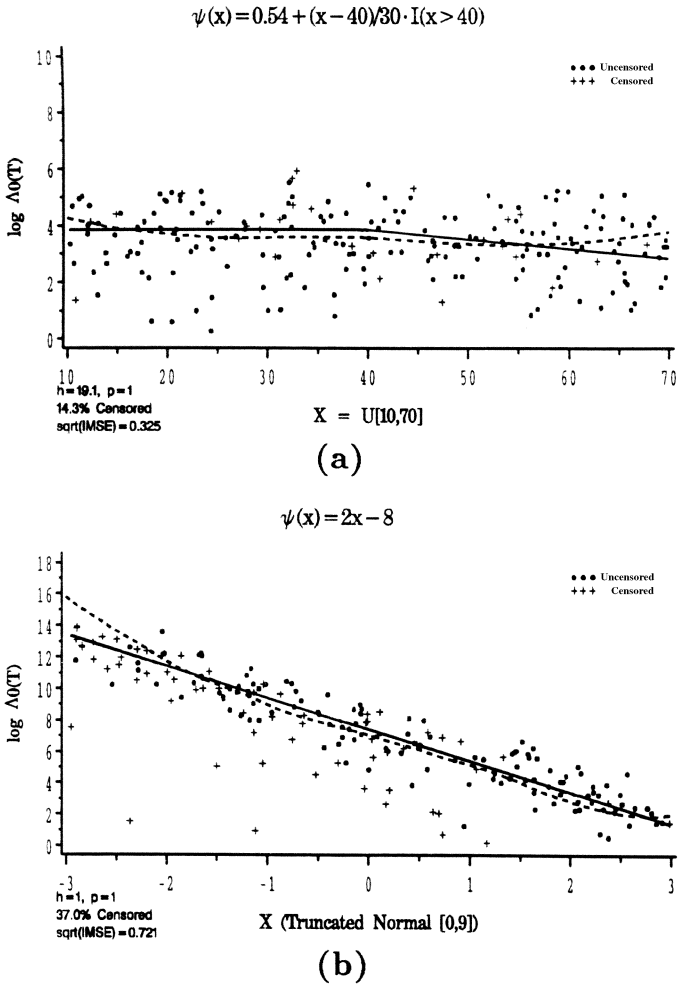
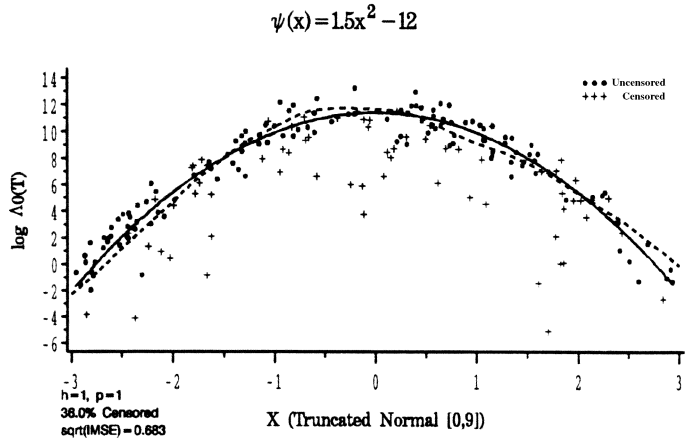


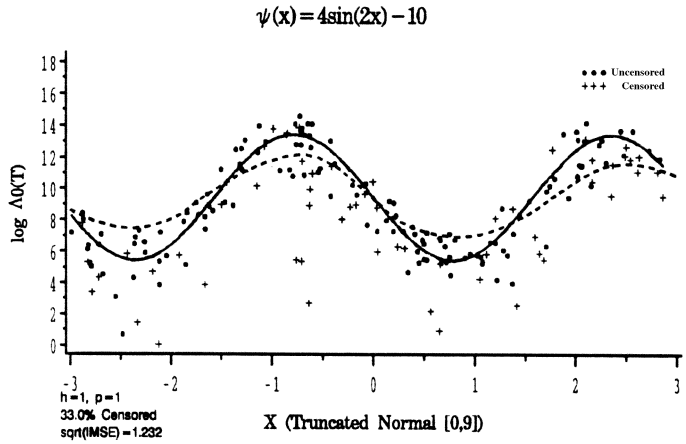
FIG. 1. *A typical simulated set of data and estimated curve for Models* 1–4. *The true curve* $\psi(\cdot)$ *is indicated as a solid line, and the estimated curve* $\hat{\psi}(\cdot)$ *as a dashed line. The estimated curves are obtained by using the local partial likelihood method with the Epanechnikov kernel. Presented in Figures* (a)–(d) *are, respectively, results for Models* 1–4.

MODEL 1. In this model, X is taken to be a uniform random variable distributed on $[10, 70]$, $\lambda_0(t) = 0.007$ and $\psi(x) = \psi_1(x) = 0.54 + (x - 40)/30 \cdot I\{x \geq 40\}$. The censoring random variable C is independent of X and T and its distribution is indicated in Table 1. The sample size for this example is 210. This model was suggested by G–C and is included here for purpose of comparison.

MODELS 2–4. In the next three models, we take $X \sim N(0, 3^2)$ but truncated at $[-3, 3]$, $\lambda_0(t) = 3t^2$. The censoring random variable C, given $X = x$,



(c)



(d)

FIG. 1. (*continued*)

is distributed uniformly on $[0, a(x)]$ where $a(x) = \exp[c_1 I\{\psi(x) > b\}/3 + c_2 I\{\psi(x) \le b\}/3]$, with b being the mean height of $\psi(x)$. Here c_1 and c_2 are chosen so that about 15%–20% of the total data are censored in each of the regions $\{x\colon \psi(x) > b\}$ and $\{x\colon \psi(x) \le b\}$. Thus, the total censoring rate is about 30%–40%. The three ψ functions in Models 2–4 are as follows:

$$\psi_2(x) = 2x - 8, \qquad \psi_3(x) = 1.5x^2 - 12, \qquad \psi_4(x) = 4\sin(2x) - 10.$$

The sample size is 200 for Models 2–4.

In the implementation of the local partial likelihood estimation, we use the local linear fit with the Epanechnikov kernel. Three different bandwidths are tried and results are reported in Tables 1 and 2.

TABLE 1
Comparison of two methods for estimating ψ_1

	Method*	Span 70	Span 105	Span 140
No censoring	1	0.357 ± 0.0082	0.354 ± 0.0076	0.362 ± 0.0076
	2	0.289 ± 0.0062	0.321 ± 0.0073	0.351 ± 0.0065
Censoring: $U[0, 400]$	1	0.371 ± 0.0086	0.357 ± 0.0085	0.356 ± 0.0074
	2	0.300 ± 0.0079	0.316 ± 0.0081	0.335 ± 0.0070
Censoring: $U[0, 200]$	1	0.366 ± 0.0079	0.359 ± 0.0079	0.368 ± 0.0086
	2	0.321 ± 0.0089	0.338 ± 0.0096	0.350 ± 0.0080

* Method 1 stands for the iterative local likelihood method of Gentleman and Crowley (1991). Method 2 is the local partial likelihood method in this article.

To compare our method with that of G–C, we need to adjust the bandwidth in order to have a comparable amount of smoothing. For example, G–C use nearest neighborhood uniform windows with a span 70 for the uniform [10, 70] design with $n = 210$. This is essentially equivalent to a uniform kernel with a bandwidth $h = 10$. To make the local linear fit with the Epanechnikov kernel have a comparable amount of smoothing as that used in G–C, we use the bandwidth $h = 1.27 \times 10$. The factor 1.27 is calculated based on the canonical kernel idea of Marron and Nolan (1988). Table 1 presents the average of $\sqrt{\text{MISE}}$ and its standard deviation based on 100 simulations. The result for the performance of Method 1 is adapted from G–C. Table 1 reveals the fact that the local partial likelihood method performs comparably to or even somewhat better than the iterative procedure of G–C. This seems consistent with our theoretical result that the noniterative local partial likelihood estimator is effective. A typical set of simulated data and the estimated curve are presented in Figure 1(a). The figure shown is the span 105 case with $C \sim$ Uniform[0, 400].

We now demonstrate the performance of our local partial likelihood estimator via simulated Models 2–4. Table 2 summarizes the result on $\sqrt{\text{MISE}}$ based on 100 simulations. In each cell, the first, second and third numbers represent the mean, standard deviation and the median of $\sqrt{\text{MISE}}$ based on 100 simulations. Note that for the small bandwidth $h = 0.5$, some local neighborhoods can contain very few data points and contribute a lot to the

TABLE 2
Performance of the local partial likelihood estimator

Model	Parameters b, c_1, c_2	$h = 0.5$	$h = 1.0$	$h = 1.5$
2	7.5, 14, 9	$1.698 \pm 5.034, 1.001$	$0.723 \pm 0.315, 0.683$	$0.594 \pm 0.300, 0.524$
3	7.5, 14, 8	$1.154 \pm 0.792, 0.992$	$0.796 \pm 0.238, 0.776$	$1.156 \pm 0.256, 1.148$
4	9.5, 16, 11	$48.19 \pm 228.5, 2.657$	$1.335 \pm 0.710, 1.219$	$1.916 \pm 0.062, 1.920$

average of $\sqrt{\text{MISE}}$. Figure 1(b)–(d) depict a typical estimated curve for Models 2–4. The bandwidth, $\sqrt{\text{MISE}}$ and the censoring rate are indicated in the lower left corner of each figure. The estimated curves have values of $\sqrt{\text{MISE}}$ close to the average of the 100 simulations. Figure 1 also indicates the satisfactory performance of the local partial likelihood method, even though the noise is quite large, the baseline hazard function is unknown and the data are censored.

6. Proofs.

6.1. *Proof of Theorem* 1. Let $\alpha = H(\beta - \beta^0)$, $\hat{\alpha} = H(\hat{\beta} - \beta^0)$ and $U_i = H^{-1}\mathbf{X}_i$. Put

$$l_n(\alpha, \theta) = n^{-1} \sum_{i=1}^{n} \Big[\delta_i \big\{ \xi_0(Z_i; \theta) + \mathbf{X}_i^T \beta^0 + U_i^T \alpha \big\} - \Lambda_0(Z_i; \theta) \exp\big(\mathbf{X}_i^T \beta^0 + U_i^T \alpha\big) \Big] K_h(X_i - x).$$

Then the problem is equivalent to showing that there exists a solution $\hat{\alpha}$ and $\hat{\theta}$ to the likelihood equation

$$\begin{aligned} \frac{\partial l_n(\alpha, \theta)}{\partial \alpha} &= n^{-1} \sum_{i=1}^{n} \big\{ \delta_i - \Lambda_0(Z_i; \theta) \exp\big(\mathbf{X}_i^T \beta^0 + U_i^T \alpha\big) \big\} \\ &\quad \times U_i K_h(X_i - x) = 0, \\ \frac{\partial l_n(\alpha, \theta)}{\partial \theta} &= n^{-1} \sum_{i=1}^{n} \big\{ \delta_i \xi_0'(Z_i; \theta) - \Lambda_0(Z_i; \theta) \exp\big(\mathbf{X}_i^T \beta^0 + U_i^T \alpha\big) \big\} \\ &\quad \times K_h(X_i - x) = 0, \end{aligned} \tag{6.1}$$

such that

$$\hat{\alpha} \to_P 0, \qquad \hat{\theta} - \theta_0 \to_P 0.$$

Let $\gamma = (\alpha^T, \theta^T)^T$ and $\gamma_0 = (0^T, \theta_0^T)^T$. Denote by S_ε the sphere centered at γ_0 with radius ε. We will show that for any sufficiently small ε, the probability that

$$\sup_{\gamma \in S_\varepsilon} l_n(\gamma) < l(\gamma_0) = l_n(0, \theta_0) \tag{6.2}$$

tends to 1. Hence $l_n(\gamma)$ has a local maximum in the interior of S_ε. Since at a local maximum the likelihood equation (6.1) must be satisfied, it follows that for any $\varepsilon > 0$, with probability tending to 1, the likelihood equation has a solution $\{\hat{\alpha}(\varepsilon), \hat{\theta}(\varepsilon)\}$ within S_ε. Let $(\hat{\alpha}, \hat{\theta})$ be the closest root to γ_0. Then

$$P\Big\{ \|\hat{\alpha}\|^2 + \|\hat{\theta} - \theta_0\|^2 \le \varepsilon \Big\} \to 1.$$

This in turn implies that

$$H\big(\hat{\beta} - \beta^0\big) \to_P 0 \quad \text{and} \quad \hat{\theta} - \theta_0 \to_P 0.$$

We now establish (6.2). Denote by γ_j and γ_{0j} the jth elements of γ and γ_0, respectively. By Taylor's expansion around the point γ_0,

$$
\begin{aligned}
l_n(\gamma) - l_n(\gamma_0) &= l_n'(\gamma_0)^T(\gamma - \gamma_0) \\
&\quad + \tfrac{1}{2}(\gamma - \gamma_0)^T l_n''(\gamma_0)(\gamma - \gamma_0) + R_n(\gamma^*),
\end{aligned} \tag{6.3}
$$

with γ^* lying between γ_0 and γ, and where

$$
R_n(\gamma) = \frac{1}{6}\sum_{j,k,l}(\gamma_j - \gamma_{0j})(\gamma_k - \gamma_{0k})(\gamma_l - \gamma_{0l})\frac{\partial^3}{\partial\gamma_j\,\partial\gamma_k\,\partial\gamma_l} l_n(\gamma).
$$

First of all, by recalling that $\beta_0 = \psi(x)$, we see that

$$
\begin{aligned}
l_n'(\gamma_0) &= n^{-1}\sum_{i=1}^n K_h(X_i - x)\begin{pmatrix} \{\delta_i - \Lambda_0(Z_i;\theta_0)\exp(\mathbf{X}_i^T\beta^0)\}U_i \\ \delta_i\,\xi_0'(Z_i;\theta_0) - \Lambda_0'(Z_i;\theta_0)\exp(\mathbf{X}_i^T\beta^0)\end{pmatrix} \\
&\to_P f(x)\begin{pmatrix} E\{\delta - \Lambda_0(Z;\theta_0)\Psi(x)\,|\,X = x\}\int \mathbf{u}K(u)\,du \\ E[\,\delta\xi_0'(Z;\theta_0) - \Lambda_0'(Z;\theta_0)\Psi(x)\,|\,X = x]\end{pmatrix}.
\end{aligned}
$$

By (1.3) and (2.6), we conclude that

$$
l_n'(\gamma_0) \to_P 0.
$$

Thus, with probability tending to 1,

$$
\left|l_n'(\gamma_0)^T(\gamma - \gamma_0)\right| \le \varepsilon^3. \tag{6.4}
$$

Analogously,

$$
\begin{aligned}
&l_n''(\gamma_0) \\
&\quad = -n^{-1}\sum_{i=1}^n K_h(X_i - x) \\
&\qquad \times\begin{pmatrix} \exp(\mathbf{X}_i^T\beta^0)\Lambda_0(Z_i;\theta_0)U_iU_i^T & \exp(\mathbf{X}_i^T\beta^0)U_i\Lambda_0'(Z_i;\theta_0)^T \\ \exp(\mathbf{X}_i^T\beta^0)\Lambda_0'(Z_i;\theta_0)U_i^T & -\delta_i\,\xi_0''(Z_i;\theta_0) + \Lambda_0''(Z_i;\theta_0)\exp(\mathbf{X}_i^T\beta^0)\end{pmatrix} \\
&\quad \to_P -f(x) \\
&\qquad \times\begin{pmatrix} \Psi(x)E\{\Lambda_0(Z;\theta_0)\,|\,X = x\}\int \mathbf{u}\mathbf{u}^TK(u)\,du & \Psi(x)\int \mathbf{u}K(u)\,du\,E\{\Lambda_0'(Z;\theta_0)\,|\,X = x\}^T \\ \Psi(x)E\{\Lambda_0'(Z;\theta_0)\,|\,X = x\}\int \mathbf{u}^TK(u)\,du & E\{-\delta\xi_0''(Z;\theta_0) + \Psi(x)\Lambda_0''(Z;\theta_0)\,|\,X = x\}\end{pmatrix}.
\end{aligned}
$$

By (1.3), (2.6) and (2.7)

$$
\begin{aligned}
l_n''(\gamma_0) &= -f(x) \\
&\qquad \times\begin{pmatrix} E\{\delta\,|\,X = x\}\int \mathbf{u}\mathbf{u}^TK(u)\,du & \int \mathbf{u}K(u)\,du\,E\{\delta\xi_0'(Z;\theta_0)\,|\,X = x\}^T \\ E\{\delta\xi_0'(Z;\theta_0)\,|\,X = x\}\int \mathbf{u}^TK(u)\,du & E[\,\delta\{\xi_0'(Z;\theta_0)\}^{\otimes 2}\,|\,X = x]\end{pmatrix} \\
&\qquad + o_P(1) \\
&\quad = -f(x)S_0(x;\theta_0) + o_P(1).
\end{aligned} \tag{6.5}
$$

Thus, with probability tending to 1,

$$
(\gamma - \gamma_0)^T l_n''(\gamma_0)(\gamma - \gamma_0) < -af(x)\varepsilon^2 \quad \text{for all } \gamma \in S_\varepsilon, \tag{6.6}
$$

where a is the smallest eigenvalue of $S_0(x;\theta_0)$.

By Condition A(iv),

$$|R_n(\gamma)| \le C\varepsilon^3 n^{-1} \sum_{i=1}^{n} M(Z_i) = C\varepsilon^3\{EM(Z) + o_P(1)\} \tag{6.7}$$

for some constant $C > 0$.

Substituting (6.4), (6.6) and (6.7) into (6.3), we conclude with probability tending to 1 that when ε is small enough,

$$l_n(\gamma) - l_n(\gamma_0) \le 0 \quad \text{for all } \gamma \in S_\varepsilon.$$

This completes the proof of Theorem 1. □

6.2. *Proof of Theorem* 2. We continue to use the notation introduced in the proof of Theorem 1. By Taylor's expansion and Condition A(iv), we have

$$0 = l_n'(\hat\gamma) = l_n'(\gamma_0) + l_n''(\gamma_0)(\hat\gamma - \gamma_0) + O_P\big(\|\hat\gamma - \gamma_0\|^2\big).$$

Hence, by (6.5),

$$\begin{aligned}(\hat\gamma - \gamma_0) &= \{l_n''(\gamma_0) + o_P(1)\}^{-1} l_n'(\gamma_0) \\ &= \{-f(x)S_0(x;\theta_0) + o_P(1)\}^{-1} l_n'(\gamma_0).\end{aligned} \tag{6.8}$$

Thus, we only need to establish the asymptotic normality of $l_n'(\gamma_0)$. We first compute the mean and the variance of $l_n'(\gamma_0)$.

First of all, by Taylor's expansion,

$$\begin{aligned}&\exp\{\psi(X_i)\} - \exp\big(\mathbf{X}_i^T\beta^0\big) \\ &\quad = \Psi(x)\frac{\psi^{(p+1)}(x)}{(p+1)!}(X_i - x)^{p+1}\{1 + o_P(1)\}.\end{aligned} \tag{6.9}$$

By (6.1), (1.3) and (2.6), we get

$$\begin{aligned}El_n'(\gamma_0) &= EK_h(X_i - x)\begin{pmatrix} \{\delta_i - \Lambda_0(Z_i;\theta_0)\exp\big(\mathbf{X}_i^T\beta^0\big)\}U_i \\ \delta_i\xi_0'(Z_i;\theta_0) - \Lambda_0'(Z_i;\theta_0)\exp\big(\mathbf{X}_i^T\beta^0\big)\end{pmatrix} \\ &= EK_h(X_i - x)\begin{pmatrix} \Lambda_0(Z_i;\theta_0)\big[\exp\{\psi(X_i)\} - \exp\big(\mathbf{X}_i^T\beta^0\big)\big]U_i \\ \Lambda_0'(Z_i;\theta_0)\big[\exp\{\psi(X_i)\} - \exp\big(\mathbf{X}_i^T\beta^0\big)\big]\end{pmatrix}.\end{aligned}$$

From (6.9), it can be calculated that

$$\begin{aligned}El_n'(\gamma_0) &= f(x)\Psi(x)\frac{\psi^{(p+1)}(x)}{(p+1)!}h^{p+1}\begin{pmatrix} E\{\Lambda_0(Z;\theta_0)\,|X = x\}\int \mathbf{u}u^{p+1}K(u)\,du \\ E\{\Lambda_0'(Z;\theta_0)\,|X = x\}\int u^{p+1}K(u)\,du\end{pmatrix} \\ &\quad + o(1) \\ &= f(x)\frac{\psi^{(p+1)}(x)}{(p+1)!}h^{p+1}\begin{pmatrix} E\{\delta\,|X = x\}\int \mathbf{u}u^{p+1}K(u)\,du \\ E\{\delta\xi_0'(Z;\theta_0)\,|X = x\}\int u^{p+1}K(u)\,du\end{pmatrix} + o(1) \\ &\equiv f(x)\bar b_n(x) + o(1).\end{aligned} \tag{6.10}$$

Similarly,

$$\begin{aligned}\operatorname{Var}\{l_n'(\gamma_0)\} &= n^{-1}EK_h^2(X-x)\begin{pmatrix}\{\delta-\Lambda_0(Z;\theta_0)\exp(\mathbf{X}^T\beta^0)\}U\\ \delta\xi_0'(Z;\theta_0)-\Lambda_0'(Z;\theta_0)\exp(\mathbf{X}^T\beta^0)\end{pmatrix}^{\otimes 2}\\ &\quad + O(n^{-1}h^{2p+2})\\ &= n^{-1}EK_h^2(X-x)\begin{pmatrix}\{\delta-\Lambda_0(Z;\theta_0)\Psi(X)\}U\\ \delta\xi_0'(Z;\theta_0)-\Lambda_0'(Z;\theta_0)\Psi(X)\end{pmatrix}^{\otimes 2}\\ &\quad + o(n^{-1}).\end{aligned} \tag{6.11}$$

We now use the counting process notation introduced in the proof of Proposition 1 to simplify the expected value in (6.11). Note that

$$\begin{pmatrix}\{\delta-\Lambda_0(Z;\theta_0)\Psi(X)\}U\\ \delta\xi_0'(Z;\theta_0)-\Lambda_0'(Z;\theta_0)\Psi(X)\end{pmatrix}=\int_0^\infty\begin{pmatrix}U\\ \xi_0'(Z;\theta_0)\end{pmatrix}dM(t).$$

By conditioning on X and using the fact that the predictable variation process

$$\langle M,M\rangle(t)=Y(t)\Psi(X)\lambda_0(t;\theta_0),$$

the expected value in (6.11) can be expressed as

$$\begin{aligned}&EK_h^2(X-x)\int_0^\infty\begin{pmatrix}U\\ \xi_0'(t;\theta_0)\end{pmatrix}^{\otimes 2}Y(t)\Psi(X)\lambda_0(t;\theta_0)\,dt\\ &\quad=EK_h^2(X-x)\int_0^\infty\begin{pmatrix}U\\ \xi_0'(t;\theta_0)\end{pmatrix}^{\otimes 2}dN(t)\\ &\quad=EK_h^2(X-x)\delta\begin{pmatrix}U\\ \xi_0'(Z;\theta_0)\end{pmatrix}^{\otimes 2}.\end{aligned}$$

Therefore, we have

$$\begin{aligned}&\operatorname{Var}\{l_n'(\gamma_0)\}\\ &\quad=n^{-1}h^{-1}f(x)\int_{-\infty}^{+\infty}E\left\{\delta\begin{pmatrix}\mathbf{u}\\ \xi_0'(Z;\theta_0)\end{pmatrix}^{\otimes 2}\middle|X=x\right\}K^2(u)\,du+o(n^{-1}h^{-1})\\ &\quad=n^{-1}h^{-1}f(x)S_1(x;\theta_0)+o(n^{-1}h^{-1}).\end{aligned}$$

To prove the asymptotic normality, we use the Cramér–Wold device. For any constant vector $b\neq 0$, we need to show

$$\sqrt{nh}\left\{b^Tl_n'(\gamma_0)-b^TEl_n'(\gamma_0)\right\}\to_D N\{0,f(x)b^TS_1(x;\theta_0)b\}. \tag{6.12}$$

Note that the left-hand side of (6.12) admits the form

$$\sqrt{nh}\,n^{-1}\sum_{i=1}^n\{K_h(X_i-x)Y_i-EK_h(X_i-x)Y_i\}.$$

To establish the asymptotic normality, we only need to verify the Lyapounov condition:

$$E\sum_{i=1}^{n}\left|\sqrt{nh}\,n^{-1}\{K_h(X_i-x)Y_i-EK_h(X_i-x)Y_i\}\right|^{2+\eta}=o(1),$$

for some $\eta>0$. By Condition A(ii), the left-hand side of the above expression is bounded by

$$2(n^{-1}h)^{1+\eta/2}nE\left|YK_h(X-x)\right|^{2+\eta}=O\{(nh)^{-\eta/2}\}\to 0.$$

This verifies (6.12). Consequently, by (6.8) and (6.10),

$$\begin{aligned}\sqrt{nh}\,&\{\hat\gamma-\gamma_0+S_0(x;\theta_0)^{-1}\bar b_n(x)\}\\ &=\{-f(x)S_0(x;\theta_0)+o_P(1)\}^{-1}\{l_n'(\gamma_0)-El_n'(\gamma_0)\}\\ &\quad+o_P(\sqrt{nh}\,h^{p+1})\\ &\to N\{0,f^{-1}(x)S_0(x;\theta_0)^{-1}S_1(x;\theta_0)S_0(x;\theta_0)^{-1}\}.\end{aligned}\tag{6.13}$$

It remains to simplify the bias expression $S_0(x;\theta_0)^{-1}\bar b_n(x)$, namely to show that

$$S_0(x;\theta_0)^{-1}\bar b_n(x)=\begin{pmatrix}b_n(x)\\0\end{pmatrix},\tag{6.14}$$

where $S_0(x;\theta_0)$, $\bar b_n(x)$ and $b_n(x)$ are given, respectively, by (2.8), (6.10) and (2.9).

Since $\int \mathbf{u}^TK(u)\,du$ is the first row of the matrix $\int \mathbf{u}\mathbf{u}^TK(u)\,du$, it follows that

$$\left\{\int\mathbf{u}^TK(u)\,du\right\}\left\{\int\mathbf{u}\mathbf{u}^TK(u)\,du\right\}^{-1}=(1,0,\ldots,0).$$

One can easily verify the above equality by multiplying with the matrix $\int \mathbf{u}\mathbf{u}^TK(u)\,du$ on both sides. Hence,

$$\left\{\int\mathbf{u}^TK(u)\,du\right\}\left\{\int\mathbf{u}\mathbf{u}^TK(u)\,du\right\}^{-1}\left(\int u^{p+1}\mathbf{u}K(u)\,du\right)=\int u^{p+1}K(u)\,du.$$

Using this, one can easily verify that

$$S_0(x;\theta_0)\begin{pmatrix}b_n(x)\\0\end{pmatrix}=\bar b_n(x),$$

and so (6.14) holds. Combining (6.13) and (6.14), we obtain Theorem 2. □

6.3. *Proof of Theorem* 4. We first state a simple lemma that will be used throughout this section.

LEMMA 1. *Suppose that K is bounded and compactly supported. If $g(\cdot)$ is continuous at the point x and $P(t|\cdot)$ is equicontinuous at the point x, then*

$$\sup_{0\le t\le\tau}|c_n(t)-c(t)|\to_P 0,$$

provided that $h \to 0$, $nh/\log n \to \infty$, $0 < \tau \le +\infty$, *where*

$$c_n(t) = n^{-1} \sum_{i=1}^{n} Y_i(t) g(X_i) K_h(X_i - x)$$

and

$$c(t) = f(x) g(x) P(t|x) \int K(u)\, du,$$

with $Y_i(t) = I\{Z_i \ge t\}$.

Upon conditioning on $X_1, \ldots, X_n$ one can apply Theorem 37 in Chapter 2 of Pollard (1984) to get the result.

Recall the notations of Section 3.3. Denote by $\hat{\alpha}^* = H^*(\hat{\beta}^* - \beta^{0*})$ and $U_i^* = (H^*)^{-1}\mathbf{X}_i^*$. Then, by (3.11), $\hat{\alpha}^*$ maximizes

$$l_n(\alpha) = n^{-1} \sum_{j=1}^{N} K_h\{X_{(j)} - x\} \Bigg[\mathbf{X}_{(j)}^{*T} \beta^{0*} + U_{(j)}^{*T} \alpha - \log\Bigg\{ \sum_{i \in \mathscr{R}_j} \exp\big(\mathbf{X}_i^{*T} \beta^{0*} + U_i^{*T} \alpha\big) K_h(X_i - x) \Bigg\} \Bigg]$$

with respect to α. We will prove a somewhat more general result. Let $\hat{\alpha}$ maximize

$$l_n(\alpha, \tau) = n^{-1} \sum_{j=1}^{N} K_h\{X_{(j)} - x\} I\{Z_{(j)} \le \tau\} \times \Bigg[\mathbf{X}_{(j)}^{*T} \beta^{0*} + U_{(j)}^{*T} \alpha - \log\Bigg\{ \sum_{i \in \mathscr{R}_j} \exp\big(\mathbf{X}_i^{*T} \beta^{0*} + U_i^{*T} \alpha\big) K_h(X_i - x) \Bigg\} \Bigg].$$

Then our case corresponds to that of $\tau = \infty$.

Let

$$N_i(t) = I\{Z_i \le t,\ \delta_i = 1\} \quad \text{and} \quad Y_i(t) = I\{Z_i \ge t\}.$$

Put

$$S_{n,0}(\alpha, u) = n^{-1} \sum_{i=1}^{n} Y_i(u) \exp\big(\mathbf{X}_i^{*T} \beta^{0*} + U_i^{*T} \alpha\big) K_h(X_i - x).$$

Then

$$l_n(\alpha, \tau) = \int_0^{\tau} n^{-1} \sum_{i=1}^{n} K_h(X_i - x) \big(\mathbf{X}_i^{*T} \beta^{0*} + U_i^{*T} \alpha\big)\, dN_i(u) - \int_0^{\tau} \log\{n S_{n,0}(\alpha, u)\} n^{-1} \sum_{i=1}^{n} K_h(X_i - x)\, dN_i(u).$$

The proof of Theorem 4 consists of the following main steps.

STEP 1. Consistency:

$$\hat{\alpha}^* \to_P 0. \tag{6.15}$$

STEP 2. Asymptotic normality of $l_n'(0,\tau)$:

$$\sqrt{nh}\,\{l_n'(0,\tau) - b_n(\tau)\} \to_D N\{0, \Sigma(\tau, x)\}, \tag{6.16}$$

where $l_n'(0,\tau) = (\partial l_n(\alpha,\tau)/\partial\alpha)|_{\alpha=0}$, and

$$b_n(\tau) = f(x)\Psi(x)\frac{\psi^{(p+1)}(x)}{(p+1)!}\Lambda(\tau, x)bh^{p+1}$$

with $\Lambda(\tau, x)$ defined by (3.10) and

$$\Sigma(\tau, x) = f(x)\Psi(x)\Lambda(\tau, x)B.$$

STEP 3. For any $\hat\alpha^{**} \to_P 0$,

$$l_n''(\hat\alpha^{**}, \tau) \to_P -f(x)\Psi(x)\Lambda(\tau, x)A \equiv \Sigma_1(\tau, x). \tag{6.17}$$

Once Steps 1–3 are established, we proceed as follows. Since $\hat\alpha^*$ maximizes $l_n(\alpha,\tau)$, it follows from a Taylor expansion around 0 that

$$0 = l_n'(\hat\alpha^*, \tau) = l_n'(0,\tau) + l_n''(\hat\alpha_n^{**}, \tau)\,\hat\alpha^*,$$

where $\hat\alpha_n^{**}$ lies between 0 and $\hat\alpha^*$, and hence $\hat\alpha_n^{**} \to_P 0$. Thus, by (6.17),

$$\hat\alpha^* + \Sigma_1(\tau, x)^{-1}b_n(\tau) = -l_n''(\hat\alpha_n^{**}, \tau)^{-1}\{l_n'(0,\tau) - b_n(\tau)\} + o_P(h^{p+1}).$$

By (6.16), (6.17) and Slutsky's theorem, we conclude that

$$\begin{aligned}&\sqrt{nh}\left\{\hat\alpha^* + \Sigma_1(\tau, x)^{-1}b_n(\tau)\right\}\\ &\quad\to_D N\left\{0, \Sigma_1(\tau, x)^{-1}\Sigma(\tau, x)\Sigma_1(\tau, x)^{-1}\right\}.\end{aligned} \tag{6.18}$$

From (6.18), we obtain (3.12). Thus it remains to prove (6.15), (6.16) and (6.17).

PROOF OF (6.15). Let the filtration $\mathscr{F}_{nt}$ be the statistical information accruing during the time $[0, t]$, namely,

$$\mathscr{F}_{nt} = \sigma\{X_i, N_i(u), Y_i(u), i = 1, \ldots, n, 0 \le u \le t\}.$$

Then, under the independent censoring scheme,

$$M_i(t) = N_i(t) - \int_0^t Y_i(u)\exp\{\psi(X_i)\}\lambda_0(u)\,du \tag{6.19}$$

is an $\mathscr{F}_{nt}$-martingale. Let

$$S_{n,0}^*(u) = n^{-1}\sum_{i=1}^n K_h(X_i - x)Y_i(u)\exp\{\psi(X_i)\}$$

and

$$S_{n,1}^*(u) = n^{-1}\sum_{i=1}^n K_h(X_i - x)Y_i(u)\exp\{\psi(X_i)\}U_i^*.$$

Then, by (6.19), we can write

$$l_n(\alpha, t) - l_n(0, t) = A_n(\alpha, t) + X_n(\alpha, t), \tag{6.20}$$

where

$$A_n(\alpha, t) = \int_0^t S_{n,1}^*(u)^T\alpha\lambda_0(u)\,du - \int_0^t \log\left\{\frac{S_{n,0}(\alpha, u)}{S_{n,0}(0, u)}\right\}S_{n,0}^*(u)\,\lambda_0(u)\,du,$$

and

$$X_n(\alpha, t) = \int_0^t n^{-1} \sum_{i=1}^n K_h(X_i - x)\left[U_i^{*T}\alpha - \log\left\{\frac{S_{n,0}(\alpha, u)}{S_{n,0}(0, u)}\right\}\right] dM_i(u).$$

By Lemma 1, we have

$$\begin{aligned} A_n(\alpha, t) &= f(x)\Psi(x)\Lambda(t, x) \\ &\quad \times\left[\nu_1^T\alpha - \log\left\{\int \exp(\mathbf{u}^{*T}\alpha)K(u)\, du\right\}\right] + o_P(1) \\ &\equiv A(\alpha, t) + o_P(1). \end{aligned} \tag{6.21}$$

The process $X_n(\alpha, \cdot)$ is a locally square integrable martingale with the predictable variation process

$$\begin{aligned} B_n(t) &\equiv \langle X_n(\alpha, \cdot), X_n(\alpha, \cdot)\rangle(t) \\ &= \sum_{i=1}^n \int_0^t n^{-2}K_h^2(X_i - x)\left[U_i^{*T}\alpha - \log\left\{\frac{S_{n,0}(\alpha, u)}{S_{n,0}(0, u)}\right\}\right]^2 \\ &\quad \times Y_i(u)\exp\{\psi(X_i)\}\lambda_0(u)\, du. \end{aligned}$$

By using Lemma 1, it can be calculated that

$$EX_n(\alpha, t)^2 = EB_n(t)^2 = O(n^{-1}h^{-1}) \to 0, \qquad 0 \le t \le \tau.$$

This result, with (6.20) and (6.21), leads to

$$l_n(\alpha, \tau) - l_n(0, \tau) = A(\alpha, \tau) + o_P(1).$$

Clearly, $A(\alpha, \tau)$ is strictly concave, with a maximum at the point $\alpha = 0$. Since $\hat{\alpha}^*$ maximizes the concave function $l_n(\alpha, \tau) - l_n(0, \tau)$, by the concavity lemma [see Andersen and Gill (1982)],

$$\hat{\alpha}^* \to_P 0.$$

PROOF OF (6.16). Let

$$S_{n,1}(\alpha, u) = n^{-1} \sum_{i=1}^n Y_i(u)\exp\left(\mathbf{X}_i^{*T}\beta^{0*} + U_i^{*T}\alpha\right)K_h(X_i - x)U_i^*.$$

Then, by Lemma 1, we have

$$\sup_{0 \le u \le \tau}\left\|\frac{S_{n,1}(0, u)}{S_{n,0}(0, u)} - \nu_1\right\| \to_P 0. \tag{6.22}$$

By (6.19), we can express $l_n'(0, \tau)$ as

$$l_n'(0, \tau) = U_n(\tau) + B_n(\tau), \tag{6.23}$$

where

$$U_n(\tau) = n^{-1} \sum_{i=1}^n \int_0^\tau K_h(X_i - x)\left\{U_i^* - \frac{S_{n,1}(0, u)}{S_{n,0}(0, u)}\right\} dM_i(u),$$

and

$$B_n(\tau) = n^{-1} \sum_{i=1}^n \int_0^\tau K_h(X_i - x)\left\{U_i^* - \frac{S_{n,1}(0, u)}{S_{n,0}(0, u)}\right\} Y_i(u)\exp\{\psi(X_i)\}\lambda_0(u)\, du.$$

Note that

$$B_n(\tau) = n^{-1}\sum_{i=1}^{n}\int_0^{\tau} K_h(X_i - x)\left\{U_i^* - \frac{S_{n,1}(0,u)}{S_{n,0}(0,u)}\right\} \times Y_i(u)\left[\exp\{\psi(X_i)\} - \exp\left\{\psi(x) + \mathbf{X}_i^{*T}\beta^{0*}\right\}\right]\lambda_0(u)\,du.$$

By Taylor's expansion we have

$$\exp\{\psi(X_i)\} - \exp\left\{\psi(x) + \mathbf{X}_i^{*T}\beta^{0*}\right\} = \Psi(x)\frac{\psi^{(p+1)}(x)}{(p+1)!}(X_i - x)^{p+1} + o_P(h^{p+1}). \tag{6.24}$$

Thus, by Lemma 1, (6.22) and (6.24), we obtain

$$\begin{aligned} B_n(\tau) &= f(x)\Psi(x)\frac{\psi^{(p+1)}(x)}{(p+1)!} \\ &\quad \times \Lambda(\tau, x)\int_{-\infty}^{+\infty} K(u)(\mathbf{u}^* - \nu_1)u^{p+1}\,du\,h^{p+1} + o_P(h^{p+1}) \\ &= b_n(\tau) + o_P(h^{p+1}). \end{aligned} \tag{6.25}$$

We now treat the process $U_n(t)$, using the martingale central limit theorem [see Theorem 5.3.5 of Fleming and Harrington (1991)]. The predictable variation process $U_n^*(t) = \sqrt{nh}\,U_n(t)$ is

$$\langle U_n^*, U_n^*\rangle(t) = n^{-1}h\sum_{i=1}^{n}\int_0^t K_h^2(X_i - x)\left\{U_i^* - \frac{S_{n,1}(0,u)}{S_{n,0}(0,u)}\right\}^{\otimes 2} \times Y_i(u)\exp\{\psi(X_i)\}\lambda_0(u)\,du.$$

By Lemma 1,

$$\begin{aligned} \langle U_n^*, U_n^*\rangle(t) &= f(x)\Psi(x)\Lambda(t,x)\int K^2(u)(\mathbf{u}^* - \nu_1)^{\otimes 2}\,du + o_P(1) \\ &= \Sigma(t,x) + o_P(1). \end{aligned}$$

Write the lth element of the vector $U_n^*(t)$ as

$$\frac{\sqrt{nh}}{n}\sum_{i=1}^{n}\int_0^t K_h(X_i - x)H_{n,l}(u)\,dM_i(u).$$

Then $H_{n,l}(u)$ is a bounded random vector. To prove the asymptotic normality, we need to check the Lindeberg condition:

$$\sum_{i=1}^{n}\int_0^t n^{-1}hK_h^2(X_i - x)H_{n,l}^2(u)I\left\{\sqrt{h/n}\,|K_h(X_i - x)H_{n,l}(u)| > \varepsilon\right\} \times Y_i(u)\exp\{\psi(X_i)\}\lambda_0(u)\,du \to_P 0 \quad \text{for all } \varepsilon > 0.$$

The last statement is valid since the random variable $K\{(X_i - x)/h\}H_{n,l}(x)$ is bounded and hence when n is large enough, the set indicator becomes

empty. This establishes that

$$\sqrt{nh}\, U_n(t) \to_D N\{0, \Sigma(t,x)\}, \qquad 0 < t \le \tau.$$

This result, with (6.23) and (6.25), proves (6.16).

PROOF OF (6.17). Let

$$S_{n,2}(\alpha, u) = n^{-1} \sum_{i=1}^{n} Y_i(u) \exp\bigl(\mathbf{X}_i^{*T}\beta^{0*} + U_i^{*T}\alpha\bigr) K_h(X_i - x) U_i^* U_i^{*T}.$$

Then

$$l_n''(\alpha, t) = -\int_0^t \frac{S_{n,2}(\alpha,u)S_{n,0}(\alpha,u) - S_{n,1}(\alpha,u)S_{n,1}(\alpha,u)^T}{S_{n,0}(\alpha,u)^2} \times n^{-1} \sum_{i=1}^{n} K_h(X_i - x)\, dN_i(u).$$

For any consistent estimator $\hat{\alpha}^{**} \to_P 0$, since the random vector U_i^* and the other involved random variables are bounded (by the continuity assumptions), it can be easily shown that

$$l_n''(\hat{\alpha}^{**}, \tau) = l_n''(0, \tau) + o_P(1). \tag{6.26}$$

By using Lemma 1,

$$\frac{S_{n,2}(0,u)S_{n,0}(0,u) - S_{n,1}(0,u)S_{n,1}(0,u)^T}{S_{n,0}(0,u)^2} - A \to_P 0$$

uniformly in $0 \le u \le \tau$, where $A = \int \mathbf{u}^* \mathbf{u}^{*T} K(u)\, du - \nu_1 \nu_1^T$ from (3.9). Thus, by Lemma 1,

$$\begin{aligned} l_n''(0,\tau) &= -A \int_0^\tau n^{-1} \sum_{i=1}^{n} K_h(X_i - x)\, dN_i(u) + o_P(1) \\ &= -An^{-1} \sum_{i=1}^{n} K_h(X_i - x) N_i(\tau) + o_P(1) \\ &= -Af(x) E\{N(\tau) \mid X = x\} + o_P(1), \end{aligned} \tag{6.27}$$

with $N(\tau) = I\{Z \le \tau,\ \delta = 1\}$, as in the proof of Proposition 1.

Recall that the compensator of $N(t)$ is $\Psi(X)\int_0^t Y(u)\lambda_0(u)\, du$. Hence,

$$E\{N(\tau) \mid X = x\} = \Psi(x)\Lambda(\tau, x). \tag{6.28}$$

Combining (6.26)–(6.28), we conclude that

$$l_n''(\hat{\alpha}^{**}, \tau) = -f(x)\Psi(x)\Lambda(\tau, x)A + o_P(1).$$

This proves (6.17) and completes the proof of Theorem 4. □

Acknowledgment. The authors thank the referees for their very valuable remarks which led to a considerable improvement of the paper.

REFERENCES

AITKIN, M. and CLAYTON, D. G. (1980). The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. *J. Roy. Statist. Soc. Ser. C* **29** 156–163.

ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. and KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.

ANDERSEN, P. K. and GILL, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10** 1100–1120.

BRESLOW, N. E. (1972). Comment on "Regression and life tables," by D. R. Cox. *J. Roy. Statist. Soc. Ser. B* **34** 216–217.

BRESLOW, N. E. (1974). Covariance analysis of censored survival data. *Biometrics* **30** 89–99.

COX, D. R. (1972). Regression models and life-tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **4** 187–220.

COX, D. R. (1975). Partial likelihood. *Biometrika* **62** 269–276.

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London.

FAN, J., HÄRDLE, W. and MAMMEN, E. (1995). Direct estimation of additive and linear components for high dimensional data. Inst. Statist. Mimeo Series #2339. Univ. North Carolina, Chapel Hill.

FLEMING, T. R. and HARRINGTON, D. P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

GENTLEMAN, R. and CROWLEY, J. (1991). Local full likelihood estimation for the proportional hazards model. *Biometrics* **47** 1283–1296.

HASTIE, T. and TIBSHIRANI, R. (1990a). *Generalized Additive Models*. Chapman and Hall, London.

HASTIE, T. and TIBSHIRANI, R. (1990b). Exploring the nature of covariate effects in the proportional hazards model. *Biometrics* **46** 1005–1016.

HJORT, N. L. (1996). Dynamic likelihood hazard rate estimation. *Biometrika*. To appear.

KOOPERBERG, C., STONE, C. J. and TRUONG, Y. (1995a). Hazard regression. *J. Amer. Statist. Assoc.* **90** 78–94.

KOOPERBERG, C., STONE, C. J. and TRUONG, Y. (1995b). The L_2 rate of convergence for hazard regression. *Scand. J. Statist.* **22** 143–157.

LEHMANN, E. L. (1983). *Theory of Point Estimation*. Wadsworth & Brooks/Cole, Pacific Grove, CA.

LI, G. and DOSS, H. (1995). An approach to nonparametric regression for life history data using local linear fitting. *Ann. Statist.* **23** 787–823.

LINTON, O. and NIELSEN, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* **82** 93–100.

MARRON, J. S. and NOLAN, D. (1988). Canonical kernels for density estimation. *Statist. Probab. Lett.* **7** 195–199.

MARRON, J. S. and PADGETT, W. J. (1987). Asymptotically optimal bandwidth selection from randomly right-censored samples. *Ann. Statist.* **15** 1520–1535.

MÜLLER, H. G. and WANG, J. L. (1990). Analyzing changes in hazard functions: an alternative to change-point models. *Biometrika* **77** 610–625.

MÜLLER, H. G. and WANG, J. L. (1994). Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics* **50** 61–76.

O'SULLIVAN, F. (1988). Nonparametric estimation of relative risk using splines and cross-validation. *SIAM J. Sci. Statist. Comput.* **9** 531–542.

POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

STONE, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). *Ann. Statist.* **22** 118–184.

STUTE, W. and WANG, J. L. (1993). A strong law under random censorship. *Ann. Statist.* **21** 1591–1607.

TIBSHIRANI, R. and HASTIE, T. (1987). Local likelihood estimation. *J. Amer. Statist. Assoc.* **82** 559–567.

WONG, W. H. (1986). Theory of partial likelihood. *Ann. Statist.* **14** 88–123.

J. FAN
DEPARTMENT OF STATISTICS
HONG KONG CHINESE UNIVERSITY
SHATIN
HONG KONG

I. GIJBELS
INSTITUT DE STATISTIQUE
UNIVERSITÉ CATHOLIQUE DE LOUVAIN
VOIE DU ROMAN PAYS 20
B-1348 LOUVAIN-LA-NEUVE
BELGIUM
E-MAIL: gijbels@stat.ucl.ac.be

M. KING
DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA 27599-3260