The Annals of Statistics
1998, Vol. 26, No. 1, 215–241

ON A SEMIPARAMETRIC SURVIVAL MODEL WITH FLEXIBLE COVARIATE EFFECT[1]

BY JENS P. NIELSEN, OLIVER LINTON AND PETER J. BICKEL

PFA Pension, *Yale University and University of California*, *Berkeley*

A semiparametric hazard model with parametrized time but general covariate dependency is formulated and analyzed inside the framework of counting process theory. A profile likelihood principle is introduced for estimation of the parameters: the resulting estimator is $n^{1/2}$-consistent, asymptotically normal and achieves the semiparametric efficiency bound. An estimation procedure for the nonparametric part is also given and its asymptotic properties are derived. We provide an application to mortality data.

1. Introduction. The Cox regression model specifies the stochastic hazard rate at exposure time t for covariate z as

$$\lambda(z, t) = \alpha(t) \exp(\beta z), \tag{1}$$

where α is nonparametric, while the dependency on the marker or covariate z is parametric. This model has been popular in survival analysis because β can conveniently be used to assess the impact of the marker or covariate, while no parametric restrictions on α are necessary: the partial likelihood principle provides efficient estimates of β in the semiparametric model (1) [see Andersen and Gill (1982)]. This in turn gives a powerful and easily applicable basis for the construction of likelihood ratio tests and confidence bands; for details see Cox (1972) and Andersen and Gill (1982). The proportionality assumption is justifiable in certain circumstances [see Cox (1972), pages 200–201] and certainly provides a clear interpretation; in any case, it can be tested [see Andersen, Borgan, Gill and Keiding (1993), pages 539–562, and McKeague and Utikal (1991)].

Despite the advantages of the Cox model, there have been many recent developments in survival analysis that have broken with this tradition. For example, Beran (1981) initiated work on the fully nonparametric hazard model where $\lambda(z, t)$ is unspecified; see also Dabrowska (1987). Nielsen (1996) treats the multiplicative nonparametric model where $\lambda(z, t) = \alpha(t)g(z)$ in which neither $\alpha(\cdot)$ nor $g(\cdot)$ is parametrically specified. Sasieni (1992a) examines a partially linear covariate effect, that is, $\lambda(z, t) = \alpha(t) \exp(\beta z_1 + g(z_2))$. Lin and Ying (1995) consider a general class of parametric covariate effects that are both additive and multiplicative. Andersen, Borgan, Gill and Keiding (1993) provides some discussion about modelling issues.

Received January 1996; revised April 1997.

[1]Research supported in part by an NSF grant.

AMS 1991 *subject classifications*. Primary 62G05; secondary 62M09.

Key words and phrases. Counting process theory, kernel estimation, predictability, semiparametric survival analysis.

In this paper, we suppose that

$$\lambda(z, t) = \alpha(t; \theta)g(z), \tag{2}$$

where $\{\alpha(\cdot\,; \theta)\}_{\theta \in \Theta}$ is a parametric class of hazard functions, while g is of unknown functional form. In many applications the shape of the baseline hazard is thought to be well understood: for example, in insurance problems the Gompertz–Makeham hazard has a long tradition of successful application, [Jordan (1975), page 21]. Meshalkin and Kagan (1972) claimed that the logarithm of baseline hazard is approximately linear for a number of chronic diseases. The covariate effect, however, is rarely specified precisely by behavioral models. For this reason (2) may provide a useful general starting point for model building for these data.

Unfortunately, as in frailty models [Clayton and Cuzick (1985)] and censored regression [Buckley and James (1979)] the partial likelihood principle does not extend to estimation of the parametric part in (2). A more generally applicable approach for semiparametric models is that of profile likelihood, which is discussed extensively in Bickel, Klaassen, Ritov and Wellner [(1993), Section 7.7]. This requires the use of a smoothing procedure for estimation of g: we use the kernel method proposed in Nielsen and Linton (1995) for this purpose. We then construct a profiled pseudolikelihood estimator of θ; see Section 3. We adopt the full counting process framework, introduced by Aalen (1978), allowing for multiple exits, censoring and time-varying covariates. The asymptotic results here make use of certain uniform convergence results obtained in Nielsen and Linton (1995). An additional technical problem that we must solve here is what we call the predictability problem. A number of terms we encounter are of the general form

$$\overline{M}_t = \sum_{i=1}^{n} \int_0^t h_i(u)\, dM_i(u), \tag{3}$$

where the M_i's are martingales but $h_i(u)$ is not a predictable process according to the usual definition. Since the integrands are not predictable, results like Rebolledo's martingale central limit theorem cannot be used to show $\overline{M}_t \to 0$ in probability [see Andersen, Borgan, Gill and Keiding (1993)]. We provide a general result in the Appendix establishing this convergence for integrands with certain additional structure found in our work. The general strategy we use is similar to that followed by Schick (1987) in an independent and identically distributed (i.i.d.) setting without censoring.

This paper is organized as follows. In Section 2 we outline the counting process framework, while in Section 3 we define estimators of θ and g in (2) based on the profile likelihood principle. In Section 4 we provide the asymptotic properties of the parametric and nonparametric estimates outlining the general approach to the asymptotics. We supply an empirical illustration in Section 5 and discuss some extensions in Section 6. Proofs are given in the Appendix.

2. A counting process formulation of the model. We observe n units, $i = 1, \ldots, n$. Let N_i count observed failures for the ith unit in the time interval $[0, 1]$. We assume that N_i is a one-dimensional counting process with respect to an increasing, right continuous, complete filtration $\mathscr{F}_{t,i}$, $t \in [0, 1]$, that is, one that obeys *les conditions habituelles* [see Andersen, Borgan, Gill and Keiding (1993), page 60]. The random intensity process λ_i of N_i is modelled as depending on marker values:

$$\lambda_i(t) = \alpha(t; \theta_0) g\{Z_i(t)\} Y_i(t), \tag{4}$$

where the function $g(\cdot)$ is positive and smooth, but otherwise unspecified, while $\alpha(t; \theta_0)$ is a member of a parametric family of hazard functions $\{\alpha(\,\cdot\,; \theta)\}_{\theta \in \Theta}$, where $\Theta \subseteq \mathbb{R}^p$. For identification, an arbitrary scale normalization is imposed on $\{\alpha(\,\cdot\,; \theta)\}_{\theta \in \Theta}$. Here, Y_i is a predictable process taking values in $\{0, 1\}$, indicating (by the value 1) when the ith individual is under risk, while the scalar Z_i is a predictable cadlag covariate or marker process. The marker $Z_i(u)$ is only observed for those u such that $Y_i(u) = 1$. Let

$$Z_i^*(u) = \begin{cases} Z_i(u), & \text{when } Y_i(u) = 1, \\ -\infty, & \text{when } Y_i(u) = 0. \end{cases}$$

We call Z_i^* the observed marker process.

The hazard function (2) is only a partial model specification. When important statistical issues such as efficiency [Bickel, Klaassen, Ritov and Wellner (1993)] or prediction schemes [Jewell and Nielsen (1993)] are involved a full model specification is necessary. For example, we specify some joint process for Y_i and Z_i. Then, conditional on $\{Y_i(u), Z_i(u)\colon 0 \le u \le t\}$, we say N_i has hazard function λ_i, where

$$\lambda_i(t) = \lim_{\delta \to 0} \frac{1}{\delta} \Pr\big\{N_i(t + \delta) - N_i(t) = 1 \mid Y_i(u), Z_i(u),\ 0 \le u \le t\big\}.$$

In this paper we use the following specification. We assume that the stochastic processes $(N_1, Z_1^*, Y_1), \ldots, (N_n, Z_n^*, Y_n)$ are independent and identically distributed for the n individuals. Let $F(z, u) = \Pr\{Z_i(u) \le z | Y_i(u) = 1\}$ be the conditional distribution function of the covariate process at time s, and let $f(z, u)$ be its density with respect to the d-dimensional Lebesgue measure. We assume that the covariate process is supported on the unit cube and that $E\{Y_i(u)\} = y(u)$, where $y(\cdot)$ is continuous. Finally, we take $\mathscr{F}_{t,i} = \sigma\{N_i(u), Z_i(u), Y_i(u);\ u \le t\}$ and $\mathscr{F}_t = \bigvee_{i=1}^n \mathscr{F}_{t,i}$. With these definitions, λ_i is predictable with respect to $\mathscr{F}_{t,i}$ and hence $\mathscr{F}_t$, while the processes $M_i(t) = N_i(t) - \Lambda_i(t)$, $i = 1, \ldots, n$, with compensators $\Lambda_i(t) = \int_0^t \lambda_i(u)\, du$, are square integrable local martingales with respect to $\mathscr{F}_t$ on the time interval $[0, 1]$. More precisely, $\Lambda_i(t)$ is the compensator of $N_i(t)$ with respect to the filtration $\mathscr{F}_{t,i} \vee \mathscr{H}_{t,\infty}$, where $\mathscr{H}_{t,\infty}$ is the σ-field generated by the entire future of Y_i and Z_i.

An important special case here is where we observe a cross-sectional sample $\{Z_i, V_i, \Delta_i\}_{i=1}^n$, where $\Delta_i = \mathbb{I}(T_i < C_i)$ and $V_i = T_i \wedge C_i$. Here, T_i, C_i are i.i.d. failure and right censoring times, respectively.

3. Definition of estimators of θ and *g*. To estimate θ we use a semiparametric profile likelihood method, called method 2 in Wellner [(1985), page 23.1–13], which gives the following three-step procedure:

STEP (i). First, g is estimated under the assumption that the true θ is known. This estimator of g depends on θ and on a smoothing parameter b. We denote the estimator by $\widehat{g}_\theta(z)$.

STEP (ii). Second, we derive the likelihood function for the observable data assuming that the true g is known. The true θ is now estimated from the pseudolikelihood that arises when g is replaced by $\widehat{g}_\theta(z)$. We denote the estimator by $\widehat{\theta}$.

STEP (iii). The final estimator of g is now calculated by assuming that $\widehat{\theta}$ is the true parameter and kernel smoothing using a bandwidth h. Therefore, the final estimator of g is of the form $\widehat{g}_{\hat{\theta}}(z)$.

Different amounts of smoothing may be appropriate when estimating g in Step (i) and in Step (iii), reflecting the common finding that in (nonadaptive) semiparametric procedures one should undersmooth the nonparametric estimation in order to achieve root-n consistency; essentially we need the bias to be $o(n^{-1/4})$, which will be the case if one assumes bounded second derivatives and takes bandwidth and kernel appropriately [see Bickel, Klaassen, Ritov and Wellner (1993)].

3.1. *Definition of* $\widehat{g}$. For any θ, we use the following leave-one-out procedure:

$$\widehat{g}_\theta(z) = \frac{\sum_{j\neq i} \int K_b\{z - Z_j(u)\}\, dN_j(u)}{\sum_{j\neq i} \int K_b\{z - Z_j(u)\}\alpha(u;\theta)Y_j(u)\, du}, \tag{5}$$

where K is a kernel function with $K_b(\cdot) = b^{-1}K(\cdot/b)$ for any b. Under our regularity conditions, $\widehat{g}_{\theta_0}(z)$ consistently estimates $g(z)$ [Nielsen and Linton (1995)], and furthermore, away from the true parameter value,

$$\widehat{g}_\theta(z) \to_p g_\theta(z) \equiv \frac{g(z)e_{\theta_0}(z)}{e_\theta(z)} \quad \text{where } e_\theta(z) = \int \alpha(u;\theta)f(z,u)y(u)\, du, \tag{6}$$

as we show in the Appendix. Let

$$g^*_\theta(z) = \frac{\sum_{j\neq i} \int K_b\{z - Z_j(u)\}\lambda_j(u)\, du}{\sum_{j\neq i} \int K_b\{z - Z_j(u)\}\alpha(u;\theta)Y_j(u)\, du},$$

and note that

$$\widehat{g}_\theta(z) - g^*_\theta(z) = \frac{\sum_{j\neq i}^n \int K_b\{z - Z_j(u)\}\, dM_j(u)}{\sum_{j\neq i} \int K_b\{z - Z_j(u)\}\alpha(u;\theta)Y_j(u)\, du}; \tag{7}$$

this quantity can be analyzed by martingale methods. We call $g^*_\theta(z) - g_\theta(z)$ the stable and $\widehat{g}_\theta(z) - g^*_\theta(z)$ the variable part of $\widehat{g}_\theta(z)$.

REMARK. It may appear that any reasonable nonparametric estimator would work here in the sense of providing the same asymptotic distribution for $\widehat{\theta}$; this is not correct. Consider using the estimator

$$\ddot{g}_\theta(z) = \frac{\sum_{j\neq i}\int K_b\{z - Z_j(u)\}(1/\alpha(u;\theta))\,dN_j(u)}{\sum_{j\neq i}\int K_b\{z - Z_j(u)\}Y_j(u)\,du}$$

in place of $\widehat{g}_\theta(z)$. The corresponding estimator of θ although consistent is inefficient relative to the estimate constructed from (5). In fact, $\ddot{g}_\theta(z)$ estimates $\bar{g}_\theta(z) = g(z)\omega_\theta(z)/\omega_{\theta_0}(z)$, where $\omega_\theta(z) = \int(\alpha(r;\theta_0)/\alpha(r;\theta))f(z,r)y(r)\,dr$ and $\omega_{\theta_0}(z) = \int f(z,r)y(r)\,dr$, so that $\bar{g}_\theta(z) \neq g_\theta(z)$, except when $\theta = \theta_0$. Heuristically, the advantage of (5) arises from the fact that it is a maximum likelihood estimator in the following special case. Suppose that time is discrete and the time-invariant covariate also takes only a finite number of integer values, for example, $T_i, Z_i \in \{1,\ldots,J\}$, where T_i is the survival time of the ith individual. Thus, $\lambda(j|Z_i = k) = \alpha(j)g(k)Y_i$. The maximum likelihood estimator of $g(k)$ is

$$\widehat{g}(k) = \frac{\sum_{i=1}^n \mathbb{I}(Z_i = k)}{\sum_{i=1}^n \alpha(T_i)Y_i\mathbb{I}(Z_i = k)},$$

which corresponds to (5).

3.2. *Definition of* $\widehat{\theta}$. The standard (conditional on Y and Z) log-likelihood for a counting process is $\sum_{i=1}^n\int \ln\lambda_i(u)\,dN_i(u) - \sum_{i=1}^n\int\lambda_i(u)\,du$ [see Aalen (1978)]. If g were known, this would be

$$\ell(\theta) = \sum_{i=1}^n\int \mu_\theta\{u, Z_i(u)\}\,dN_i(u) - \sum_{i=1}^n\int \exp\{\mu_\theta(u, Z_i(u))\}Y_i(u)\,du, \tag{8}$$

where $\mu_\theta(u,z) = \ln[\alpha(u;\theta)g(z)]$ is the logarithmic hazard. Since g is not known, we substitute $\widehat{g}_\theta\{Z_i(u)\}$ in (8) and estimate θ from the following profile pseudolikelihood:

$$\widehat{\ell}(\theta) = \sum_{i=1}^n\int \widehat{\mu}_\theta\{u, Z_i(u)\}\,dN_i(u) - \sum_{i=1}^n\int \exp\{\widehat{\mu}_\theta(u, Z_i(u))\}Y_i(u)\,du, \tag{9}$$

where $\widehat{\mu}_\theta(u,z) = \ln[\alpha(u;\theta)\widehat{g}_\theta(z)]$. Let $\widehat{\theta}$ be any maximizer of $\widehat{\ell}(\theta)$.

3.3. *Definition of the final estimator of* g. We define the final estimator of g by

$$\widetilde{g}_{\hat{\theta}}(z) = \frac{\sum_{i=1}^n\int K_h\{z - Z_i(u)\}\,dN_i(u)}{\sum_{i=1}^n\int K_h\{z - Z_i(u)\}\alpha(u;\widehat{\theta})Y_i(u)\,du}, \tag{10}$$

where h is a second smoothing parameter.

4. Asymptotic properties of $\widehat{\theta}$. In this section we give our main results and outline our general strategy for establishing the asymptotic properties of our procedure. Proofs are to be found in the Appendix. We show, using the approach of Cramér (1946), that under mild local conditions there exists a consistent solution of the profile likelihood equation which is asymptotically normal. As usual this does not guarantee that maximizing the profile likelihood yields a consistent solution of the equation. For that one needs global conditions such as those of Wald (1949) and Bahadur (1967). However, it does ensure that if we start a standard maximum seeking algorithm (such as steepest ascent or Newton–Raphson) sufficiently close to θ_0, then we will end up close to a consistent maximizer of the profile likelihood.

4.1. *Consistency.* We use the following conditions for consistency.

(A1) We have $\Pr\{Z_i(u) \in \mathscr{Z}\} = 1$ for all $u \in [0,1]$, where $\mathscr{Z} \subseteq \mathbb{R}$ is a compact set, and $g(z), \alpha(u;\theta), f(z,u)$ are positive and continuous on their domains of definition $\{\mathscr{Z}, [0,1] \times \mathscr{N}_0, \mathscr{Z} \times [0,1]\}$, where $\mathscr{N}_0 = \{\theta\colon |\theta - \theta_0| \le \delta\}$ is a neighborhood of θ_0 with $\delta < \infty$.

(A2) The functions g and α are Lipschitz continuous with respect to z and θ, respectively; that is, there exists a finite constant c such that $|g(z_1) - g(z_2)| \le c|z_1 - z_2|$ for all $z_1, z_2 \in \mathscr{Z}$ and $\sup_{u\in[0,1]}|\alpha(u;\theta_1) - \alpha(u;\theta_2)| \le c|\theta_1 - \theta_2|$ for all $\theta_1, \theta_2 \in \mathscr{N}_0$.

(A3) The kernel K is a probability density function supported on $[-1,1]$, symmetric about zero, and Lipschitz continuous on its support.

(A4) $nb^3 \to \infty$ and $b \to 0$.

We will show that $Q_n(\theta) = n^{-1}\{\widehat{\ell}(\theta) - \widehat{\ell}(\theta_0)\}$ converges in probability, uniformly in a neighborhood $\mathscr{N}_0$ of θ_0, to a nonrandom function $Q(\theta)$ that is locally uniquely maximized at θ_0. In fact, we will first show that $Q_n(\theta)$ can be approximated by $\bar{Q}_n(\theta) = n^{-1}\{\bar{\ell}(\theta) - \bar{\ell}(\theta_0)\}$, where

$$\bar{\ell}(\theta) = \sum_{i=1}^{n} \int \bar{\mu}_\theta\{u, Z_i(u)\} dN_i(u) - \sum_{i=1}^{n} \int \exp\big[\bar{\mu}_\theta\{u, Z_i(u)\}\big] Y_i(u)\, du$$

with $\bar{\mu}_\theta(u,z) = \ln\{\alpha(u;\theta) g_\theta(z)\}$. By a uniform law of large numbers, which holds under our conditions, $\bar{Q}_n(\theta)$ approaches

$$Q(\theta) = \iint \left[\ln\left\{\frac{\alpha(u;\theta) e_{\theta_0}(z)}{\alpha(u;\theta_0) e_\theta(z)}\right\} - \frac{\alpha(u;\theta) e_{\theta_0}(z)}{\alpha(u;\theta_0) e_\theta(z)} + 1\right] \alpha(u;\theta_0) g(z) f(z,u) y(u)\, du\, dz$$

in probability, uniformly over any compact neighborhood of θ_0. Since $Q(\theta)$ is continuous in θ at θ_0 by (A1) and (A2), and $Q(\theta) \le 0 = Q(\theta_0)$, because $\ln(x) - x + 1 \le 0$ for all x, the result is established. It follows that at least one consistent solution $\widehat{\theta}$ exists to the pseudolikelihood equation.

THEOREM 1 (Consistency of $\widehat{\theta}$). *Suppose that conditions* (A1)–(A4) *hold. Then, with a probability tending to* 1, *the pseudolikelihood has at least one consistent solution in* $\mathcal{N}_0$.

4.2. *Asymptotic normality.* We assume that the following hold:

(A5) The functions g, α and f are twice differentiable with respect to z and θ on their domains of definition with Lipschitz continuous second derivatives.

(A6) The semiparametric information matrix $\mathcal{I}_0$ is positive definite and finite, where

$$\mathcal{I}_0 = \iint \frac{\partial \bar{\mu}_{\theta_0}}{\partial \theta} \frac{\partial \bar{\mu}_{\theta_0}}{\partial \theta^T}(u, z)\alpha(u; \theta_0)g(z)f(z, u)y(u)\,du\,dz,$$

$$\frac{\partial \bar{\mu}_{\theta_0}}{\partial \theta}(u, z) = \frac{\partial \ln \alpha}{\partial \theta}(u, \theta_0) - \frac{\partial \ln e_{\theta_0}}{\partial \theta}(z).$$

(A7) θ_0 is an interior point of Θ.

(A8) $nb^4 \to 0$ and $nb^3 \to \infty$.

Let $\widehat{s}_\theta$ (the score vector) and $\widehat{H}_{\theta\theta}$ (the Hessian matrix) be the first and second derivatives of the pseudolikelihood $\widehat{\ell}$ standardized by sample size. By the mean value theorem

$$0 = n^{1/2}\widehat{s}_\theta(\theta_0) + \widehat{H}_{\theta\theta}(\check{\theta})n^{1/2}(\widehat{\theta} - \theta_0), \tag{11}$$

where $\check{\theta}$ lies between θ_0 and $\widehat{\theta}$. We first analyze the pseudoscore vector evaluated at the true θ_0,

$$\begin{aligned}
\widehat{s}_\theta(\theta_0) &= \frac{1}{n}\sum_{i=1}^{n}\int \frac{\partial \widehat{\mu}_{\theta_0}}{\partial \theta}\{u, Z_i(u)\}\,dN_i(u) \\
&\quad - \frac{1}{n}\sum_{i=1}^{n}\int \frac{\partial \widehat{\mu}_{\theta_0}}{\partial \theta}\{u, Z_i(u)\}\alpha(u; \theta_0)\widehat{g}_{\theta_0}\{Z_i(u)\}Y_i(u)\,du \\
&= \frac{1}{n}\sum_{i=1}^{n}\int \frac{\partial \widehat{\mu}_{\theta_0}}{\partial \theta}\{u, Z_i(u)\}\,dM_i(u) + \frac{1}{n}\sum_{i=1}^{n}\int \frac{\partial \widehat{\mu}_{\theta_0}}{\partial \theta}\{u, Z_i(u)\}\,d\Lambda_i(u) \qquad (12) \\
&\quad - \frac{1}{n}\sum_{i=1}^{n}\int \frac{\partial \widehat{\mu}_{\theta_0}}{\partial \theta}\{u, Z_i(u)\}\alpha(u; \theta_0)g\{Z_i(u)\}Y_i(u)\,du \\
&\quad - \frac{1}{n}\sum_{i=1}^{n}\int \frac{\partial \widehat{\mu}_{\theta_0}}{\partial \theta}\{u, Z_i(u)\}\alpha(u; \theta_0)\big[\widehat{g}_{\theta_0}\{Z_i(u)\} - g\{Z_i(u)\}\big]Y_i(u)\,du,
\end{aligned}$$

by substituting N by $M + \Lambda$ and $\widehat{g}_{\theta_0}$ by $g + \widehat{g}_{\theta_0} - g$. On substituting for Λ, we find that (12) $= 0$. We then break $\widehat{g}_{\theta_0} - g$ into stable and variable terms

[as in Nielsen and Linton (1995)]. Using the decomposition (7), we find, after interchanging the order of summation and integration, that

$$\sum_{i=1}^{n}\int \frac{\partial \widehat{\mu}_{\theta_0}}{\partial\theta}\{u, Z_i(u)\}\alpha(u;\theta_0)\{\widehat{g}_{\theta_0} - g^*_{\theta_0}\}\{Z_i(u)\}Y_i(u)\,du$$
$$= \sum_{i=1}^{n}\int \frac{\partial \widehat{\mu}^*_{\theta_0}}{\partial\theta}\{u, Z_i(u)\}\,dM_i(u),$$

where

$$\frac{\partial \widehat{\mu}^*_{\theta_0}}{\partial\theta}\{u, Z_i(u)\} = \sum_{j\neq i}^{n}\int \frac{(\partial \widehat{\mu}_{\theta_0}/\partial\theta)\{t, Z_j(t)\}\alpha(t;\theta_0)Y_j(t)K_b\{Z_j(t) - Z_i(u)\}\,dt}{\sum_k \int K_b\{Z_j(t) - Z_k(r)\}\alpha(r;\theta_0)Y_k(r)\,dr}.$$

Now substitute $\partial\bar{\mu}_\theta/\partial\theta + \partial \ln \widehat{g}_\theta/\partial\theta - \partial \ln g_\theta/\partial\theta$ for $\partial\widehat{\mu}_\theta/\partial\theta$ in the first term of (12). Collecting everything together we obtain

$$\widehat{s}_\theta(\theta_0) = n^{-1}\sum_{i=1}^{n}\int \frac{\partial\bar{\mu}_{\theta_0}}{\partial\theta}\{u, Z_i(u)\}\,dM_i(u) \tag{13}$$

$$- n^{-1}\sum_{i=1}^{n}\int \frac{\partial \widehat{\mu}^*_{\theta_0}}{\partial\theta}\{u, Z_i(u)\}\,dM_i(u)$$
$$+ n^{-1}\sum_{i=1}^{n}\int\left\{\frac{\partial \ln \widehat{g}_\theta}{\partial\theta} - \frac{\partial \ln g_\theta}{\partial\theta}\right\}(u,\theta_0)\,dM_i(u) \tag{14}$$

$$- n^{-1}\sum_{i=1}^{n}\int \frac{\partial \widehat{\mu}_{\theta_0}}{\partial\theta}\{u, Z_i(u)\}\alpha(u;\theta_0)\{g^*_{\theta_0} - g\}\{Z_i(u)\}Y_i(u)du. \tag{15}$$

We have written $\widehat{s}_\theta$ as a sum of four terms: the last term (15) is a stochastic average of $g^*_{\theta_0} - g$ that arises from the bias obtained in the estimation of g: it is asymptotically negligible, when a sufficiently small bandwidth is chosen. Undersmoothing is necessary in many semiparametric estimation problems; see Bickel, Klaassen, Ritov and Wellner [(1993), Section 7] for a discussion. The second and third terms (14) are of the form (3) and are also negligible (the formal proof, given in the Appendix, requires the solution to the predictability problem). We therefore have that

$$n^{1/2}\widehat{s}_\theta(\theta_0) = n^{1/2}s^e_\theta(\theta_0) + o_p(1) \quad \text{where}$$
$$s^e_\theta(\theta_0) = n^{-1}\sum_{i=1}^{n}\int \frac{\partial\bar{\mu}_{\theta_0}}{\partial\theta}\{u, Z_i(u)\}\,dM_i(u). \tag{16}$$

Then, since $\partial\bar{\mu}_{\theta_0}\{u, Z_i(u)\}/\partial\theta$ is a predictable process, we can apply Rebolledo's martingale central limit theorem to $s^e_\theta(\theta_0)$ [see Ramlau-Hansen (1983)]. In particular, under the conditions given below,

$$n^{1/2}s^e_\theta(\theta_0) \Rightarrow N(0, \mathscr{I}_0). \tag{17}$$

We also show that the Hessian matrix

$$\widehat{H}_{\theta\theta}(\theta) = n^{-1}\sum_{i=1}^{n}\int \frac{\partial^2\widehat{\mu}_\theta}{\partial\theta\partial\theta^T}\{u, Z_i(u)\}\,dN_i(u)$$
$$- n^{-1}\sum_{i=1}^{n}\int\left\{\frac{\partial^2\widehat{\mu}_\theta}{\partial\theta\partial\theta^T} + \frac{\partial\widehat{\mu}_\theta}{\partial\theta}\frac{\partial\widehat{\mu}_\theta}{\partial\theta^T}\right\}\{u, Z_i(u)\}\alpha(u;\theta)\widehat{g}_\theta\{Z_i(u)\}Y_i(u)\,du$$

satisfies

$$\sup_{\theta\in\mathscr{N}_n}\left|\widehat{H}_{\theta\theta}(\theta) - \mathscr{I}_0\right| \to_p 0, \tag{18}$$

where $\mathscr{N}_n = \{\theta\colon |\theta-\theta_0| \le \delta_n,\ \delta_n \to 0\}$ is a shrinking neighborhood of θ_0. In conclusion, the asymptotic distribution of $n^{1/2}(\widehat{\theta}-\theta_0)$ follows from (16), (17) and (18).

THEOREM 2 (Asymptotic distribution of $\widehat{\theta}$). *Let $\widehat{\theta}$ be a consistent solution to the pseudoscore equations. Suppose that* (A1)–(A8) *hold. Then*

$$n^{1/2}(\widehat{\theta}-\theta_0) \Rightarrow N(0, \mathscr{I}_0^{-1}).$$

Furthermore, $\mathscr{I}_0^{-1}$ can be consistently estimated by $\widehat{H}_{\theta\theta}^{-1}(\widehat{\theta})$. The cumulative baseline hazard $A(t) = \int_0^t \alpha(s;\theta_0)\,ds$ can be estimated by $\widehat{A}(t) = \int_0^t \alpha(s;\widehat{\theta})\,ds$, which satisfies

$$n^{1/2}\{\widehat{A}(t) - A(t)\} \Rightarrow \int_0^t \frac{\partial\alpha(s;\theta_0)}{\partial\theta^T}\,ds \times N(0, \mathscr{I}_0^{-1}).$$

REMARK. Condition (A8) excludes $b = O(n^{-1/5})$. In other words, the bandwidth for estimation of θ should be smaller than the magnitude which is optimal for estimating a one-dimensional function, that is, one should undersmooth inside $\widehat{\ell}$. To derive an optimal bandwidth for estimating θ requires higher-order expansions, as has been done for regression problems by Härdle, Hart, Marron and Tsybakov (1992) and Linton (1995), and is beyond the scope of this paper.

4.3. *Asymptotic distribution of $\widehat{g}$.* By the mean value theorem

$$n^{2/5}\{\widetilde{g}_{\hat\theta}(z) - g(z)\} = n^{2/5}\{\widetilde{g}_{\theta_0}(z) - g(z)\} + n^{-1/10}\frac{\partial\widetilde{g}_{\check\theta}(z)}{\partial\theta^T}n^{1/2}(\widehat{\theta}-\theta), \tag{19}$$

where $\check\theta$ lies between $\widehat{\theta}$ and θ. Under our conditions, $\partial\widetilde{g}_\theta(z)/\partial\theta$ is stochastically bounded uniformly in $\mathscr{N}_0$, and the second term on the right-hand side of (19) is negligible compared to the first term. The asymptotic distribution of $n^{2/5}\{\widetilde{g}_{\theta_0}(z) - g(z)\}$ was given in Nielsen and Linton (1995). We have the following theorem.

THEOREM 3 (Asymptotic distribution of $\widetilde{g}_{\hat\theta}$). *Suppose that assumptions* (A1)–(A8) *hold, and let* $0 < \lim_{n\to\infty} nh^5 = \gamma < \infty$. *Then*

$$n^{2/5}\{\widetilde{g}_{\hat\theta}(z) - g(z)\} \Rightarrow N[m(z), v(z)],$$

where

$$m(z) = \frac{\gamma^2}{2}\mu_2(K)\left\{2\frac{\partial g}{\partial z}(z)\frac{\partial \ln e_{\theta_0}}{\partial z}(z) + \frac{\partial^2 g}{\partial z^2}(z)\right\}, \qquad v(z) = \gamma^{-1}\|K\|^2\frac{g(z)}{e_{\theta_0}(z)}$$

with $\mu_2(K) = \int t^2 K(t)\,dt$ *and* $\|K\|^2 = \int K(t)^2\,dt$. *Finally*,

$$\tilde{v}(z) = \frac{nh\sum_{i=1}^n \int K_h\{z - Z_i(u)\}^2 dN_i(u)}{\left[\sum_{i=1}^n \int K_h\{z - Z_i(u)\}\alpha(u;\hat{\theta})Y_i(u)\,du\right]^2} \to_p v(z).$$

The uncertainty caused by estimating θ is of smaller order than that caused by the estimation of $g(\cdot)$ itself. Therefore, one can essentially ignore the presence of $\hat{\theta}$ and b as far as the properties of $\tilde{g}_{\hat{\theta}}(z)$ are concerned: bandwidth selection methods developed for kernel estimation of hazard functions, such as proposed in Nielsen and Linton (1995), can be used to choose h without modification; likewise, pointwise confidence intervals for $\lambda(z,t) = \alpha(t;\theta)g(z)$ would only involve the variability from $\tilde{g}_{\theta_0}(z)$.

4.4. *Efficiency of* $\hat{\theta}$. The information in the semiparametric model, $\mathscr{I}_0$, is smaller than in two relevant parametric submodels; this is not an adaptive situation in the sense of Bickel, Klaassen, Ritov and Wellner [(1993), page 2]. Compare first with the situation where g is known. In this case, the likelihood score function is

$$s_\theta = n^{-1}\sum_{i=1}^n \int \frac{\partial \ln \alpha}{\partial \theta}(u;\theta_0)\,dM_i(u),$$

and the information is

$$\mathscr{I}_{\theta\theta} = \iint \frac{\partial \ln \alpha}{\partial \theta}\frac{\partial \ln \alpha}{\partial \theta^T}(u;\theta_0)g(z)\alpha(u;\theta_0)f(z,u)y(u)\,du\,dz,$$

and clearly $\mathscr{I}_{\theta\theta} \geq \mathscr{I}_0$. Now suppose that the covariate effect depends on an unknown Euclidean parameter β, that is, $g(z;\beta)$, and the resulting model satisfies the conditions of Andersen and Gill (1982). In this case, the score function for β is

$$s_\beta = n^{-1}\sum_{i=1}^n \int \frac{\partial \ln g}{\partial \beta}\{Z_i(u);\beta\}\,dM_i(u),$$

and the information about θ is

$$\mathscr{I}^*_{\theta\theta} = \mathscr{I}_{\theta\theta} - \mathscr{I}_{\beta\theta}\mathscr{I}_{\beta\beta}^{-1}\mathscr{I}_{\theta\beta},$$

where

$$\mathscr{I}_{\beta\beta} = \iint \frac{\partial \ln g}{\partial \beta}\frac{\partial \ln g}{\partial \beta^T}(z;\beta)g(z;\beta)\alpha(u;\theta_0)f(z,u)y(u)\,du\,dz,$$

$$\mathscr{I}_{\theta\beta} = \iint \frac{\partial \ln \alpha}{\partial \theta}(u;\theta_0)\frac{\partial \ln g}{\partial \beta^T}(z;\beta)g(z;\beta)\alpha(u;\theta_0)f(z,u)y(u)\,du\,dz.$$

For any such parametric models,

$$\mathscr{I}_0 \leq \mathscr{I}^*_{\theta\theta} \leq \mathscr{I}_{\theta\theta},$$

and there is an information loss due to not knowing the function g.

We now show that $\mathscr{I}_0$ is the semiparametric efficiency bound. By definition, the information for a semiparametric model is the infimum of the information among all regular parametric submodels [Bickel, Klaassen, Ritov and Wellner (1993), page 46]. We formally exhibit a parametric model of g for which $\mathscr{I}_0 = \mathscr{I}^*_{\theta\theta}$; this shows that $\mathscr{I}_0$ is maximal. Let

$$g(z;\beta) = g_\theta(z) = g(z)\frac{e_{\theta_0}(z)}{e_\theta(z)}$$

with $\beta = \theta$. Then, by Fubini's theorem,

$$\begin{aligned}
\mathscr{I}_{\beta\beta} &= \iint \frac{\partial e}{\partial\theta}\frac{\partial e}{\partial\theta^T}(z)\frac{1}{e^2_\theta(z)}g(z)\alpha(u)f(z,u)y(u)\,du\,dz \\
&= \int\left\{\frac{\partial e}{\partial\theta}\frac{\partial e}{\partial\theta^T}(z)\frac{1}{e^2_\theta(z)}g(z)\int\alpha(u)f(z,u)y(u)\,du\right\}dz \\
&= \int\frac{\partial e}{\partial\theta}\frac{\partial e}{\partial\theta^T}(z)\frac{1}{e_\theta(z)}g(z)\,dz \\
&= \int\frac{\partial e}{\partial\theta^T}(z)\frac{1}{e_\theta(z)}g(z)\left\{\int\frac{\partial\alpha}{\partial\theta}(u)f(z,u)y(u)\,du\right\}dz \\
&= \iint\frac{\partial e}{\partial\theta^T}(z)\frac{1}{e_\theta(z)}\frac{\partial\alpha}{\partial\theta}(u)\frac{1}{\alpha(u)}g(z)\alpha(u)f(z,u)y(u)\,du\,dz \\
&= \mathscr{I}_{\theta\beta}.
\end{aligned}$$

After a similar computation, we see that $\mathscr{I}_0 = \mathscr{I}^*_{\theta\theta}$. To make this computation rigorous one needs to check that the model partially specified by $\alpha(\cdot,\theta)$ and $g(\cdot)$ satisfies the regularity conditions given in Andersen and Gill (1982). Similar calculations are given in Lin and Ying (1995) and Sasieni (1992a, b).

5. Numerical results. We illustrate the semiparametric procedure in a study of mortality data. Gompertz (1825) suggested the parametric model $\alpha(t;\theta) = \beta c^t$ as describing the force of pure mortality for British data. This was extended by Makeham (1860) to $\alpha(t;\theta) = \alpha + \beta c^t$. Both the Gompertz and Makeham laws simplify many actuarial expressions concerning annuities, and are thus widely used. They both have also been used successfully in numerous actuarial applications. For the above reasons, these laws are the most celebrated actuarial laws to this day. The Danish tariff system G82 is an example of a sophisticated modern technical tariff system exclusively employing the Gompertz–Makeham mortality law or slight modifications of it.

Our data are a sample of 2306 disabled Danish males observed from January 1, 1986, to January 1, 1988, and were provided by the Danish Committee for Assessment of Substandard Lives. It includes the single covariate time since disability along with the age of the individual. Out of a total of 3178 man-years there were 185 deaths. Before estimating the semiparametric model, we provide a graphical test of proportionality. In Figure 1 we plot

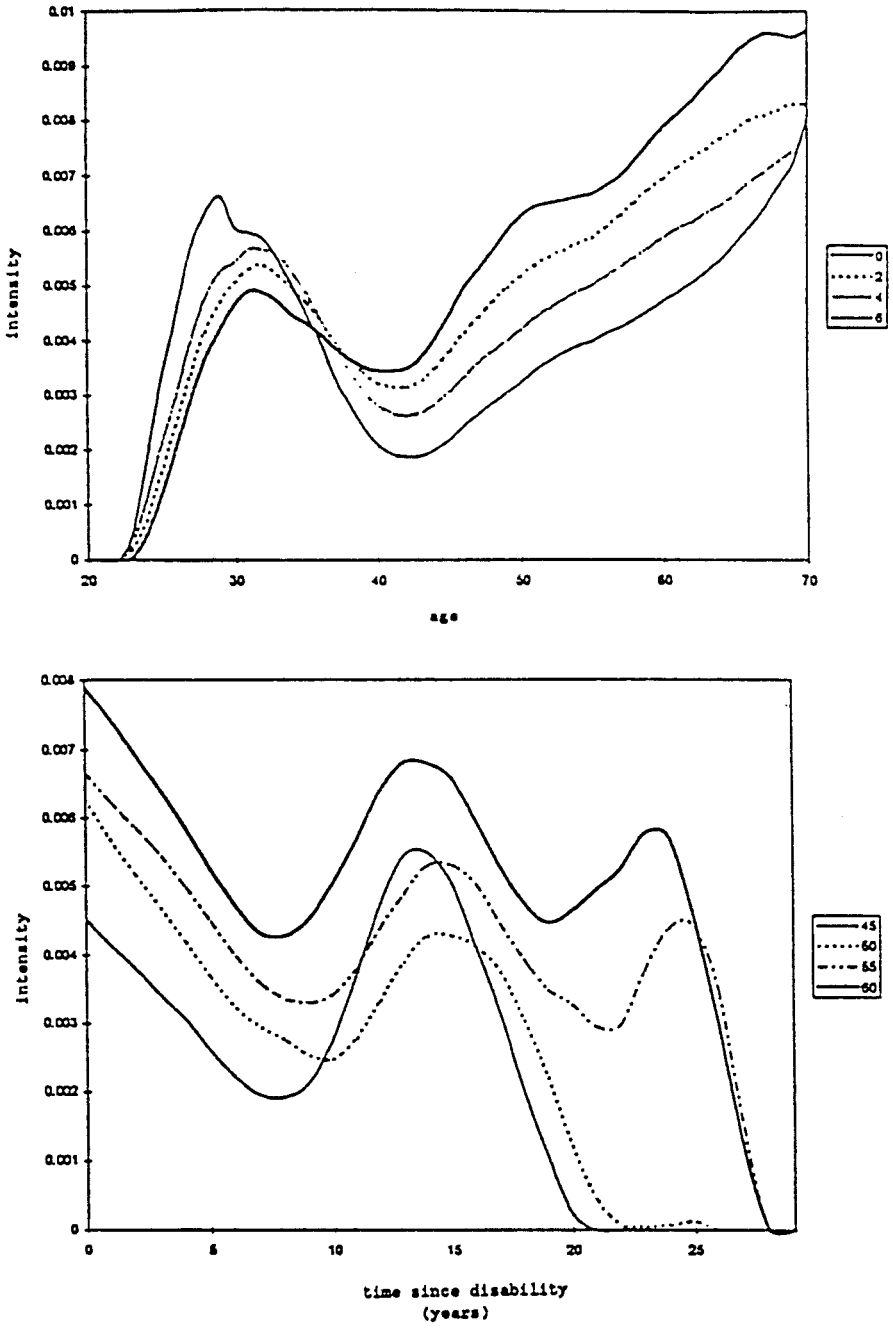

FIG. 1. *Cross-sectional effects of age and disability for alternative values of the second variable.*

the bivariate Nielsen and Linton (1995) estimates $\widehat{\lambda}(\cdot, z)$ [and $\widehat{\lambda}(t, \cdot)$] for a number of different values of t (or z). In the range where most of the data lie, (e.g., 40–60 years of age), proportionality seems quite reasonable.

We adopted the Gompertz–Makeham model for the baseline hazard, adjusted to satisfy identifiability, that is,

$$\alpha(t; \theta) = \alpha + c^t,$$

and $\lambda(t, z; \theta) = \alpha(t; \theta)g(z)$. In Table 1 we report our estimates of $\widehat{\alpha}$ and $\widehat{c}$ for a range of bandwidths, reporting also the maximized pseudolikelihood criterion $\widehat{\ell}(\widehat{\theta}(b))$, the maximized Nielsen–Linton [Nielsen and Linton (1995)] cross-validation criterion $\widehat{Q}(b; \widehat{\theta}(b))$ and the parameter standard errors. There is some variation in both parameter estimate and standard error with respect to bandwidth; but in all cases $\widehat{c}$ was significantly different from 1, while $\widehat{\alpha}$ was insignificantly different from zero. Figure 2 shows the estimated $\widehat{c}$ as it varies with b. Note that the minimum of $\widehat{\ell}(\widehat{\theta}(b))$ is attained at $b = 56$, while the minimum of $\widehat{Q}(b; \widehat{\theta}(b))$ is attained at $b = 60$. The latter criterion function is certainly better suited to the estimation of $g(\cdot)$ rather than $\widehat{\theta}$ and can be expected to select oversmoothed bandwidths. This suggests that the relevant range of bandwidths as far as $\widehat{\theta}$ is concerned are those less than $b = 60$ for which there is much less variation.

We now turn to the estimation of g. For this purpose the bandwidth $h = 80$ was chosen by eyeball for $\widetilde{g}_h$. In Figure 3 we give our final estimate of g along with 95% multiplicative pointwise confidence bands. It is not surprising that the general trend over the six available years is downward: this is due to the fact that the seriously disabled have very high initial mortality, while individuals with, say, minor back injuries have only minor additional mortality risk. The mixture of different disabilities in the sample is also perhaps responsible for the apparent nonmonotonicity in the middle.

TABLE 1

Parameter estimates, standard errors and the pseudolikelihood criterion for a range of bandwidths; Gompertz–Makeham baseline hazard

b	$\widehat{\ell}$	$\widehat{Q}$	$\widehat{\alpha}$	s.e.	$\widehat{c}$	s.e.
32	−1148.55	−99,164	0.0047276	0.0222	1.083961	0.06207
36	−1146.72	−75,291	0.0047552	0.0224	1.083950	0.06257
40	−1145.90	−69,424	0.0047688	0.0226	1.083869	0.06287
52	−1144.34	−59,807	0.0047557	0.0228	1.083490	0.06367
56	−1144.22	−58,682	0.0047577	0.0229	1.083415	0.06388
60	−1144.27	−58,369	0.0047798	0.0230	1.083409	0.06408
72	−1145.02	−61,485	0.0049886	0.0242	1.083803	0.06480
80	−1145.56	−64,485	0.0051747	0.0252	1.084163	0.06534
120	−1148.44	−87,778	0.0063541	0.0316	1.086306	0.06819
160	−1149.52	−121,690	0.0078201	0.0395	1.088160	0.07041

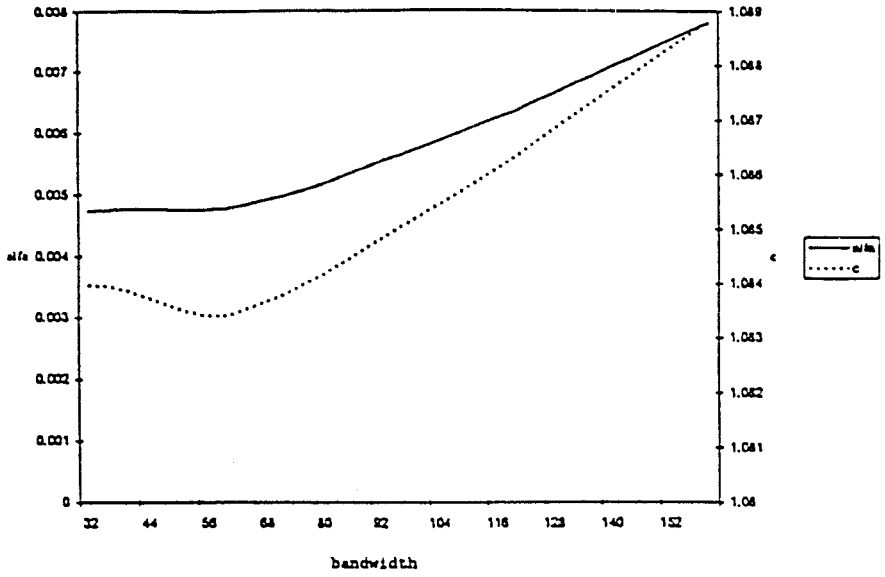

FIG. 2. *Parameter estimates $\widehat{\alpha}$ and $\widehat{c}$ by bandwidth b.*

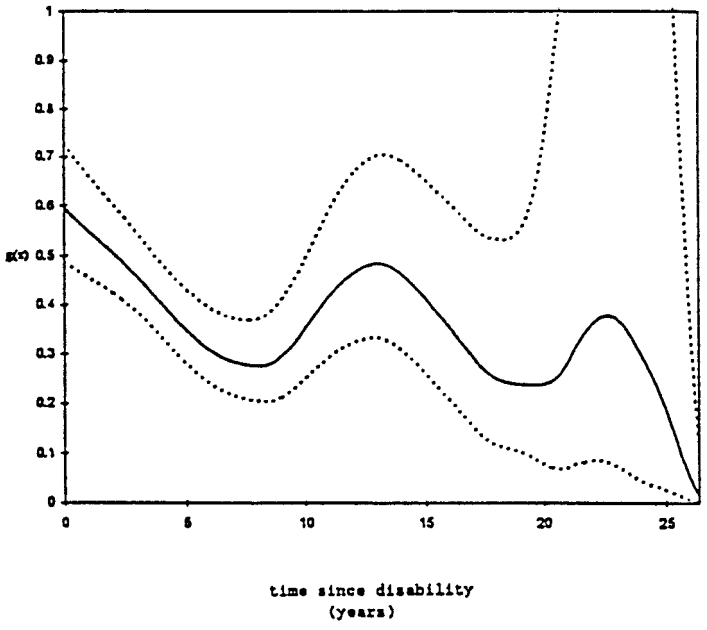

FIG. 3. *Covariate effect g against time since disability z with* 95% *multiplicative pointwise confidence bands.*

Additional computations can be found in Sandqvist (1995). This includes the same analysis for females and other subgroups for which qualitatively the same results are obtained.

6. Extensions. The extension of our procedure to multiple covariates is formally trivial: we just replace the univariate kernel K in (5) by a multivariate one. However, when there are many covariates, the nonparametric procedure on which our estimation is based may suffer from the curse of dimensionality. Although the main result Theorem 2 remains valid for any dimensions, to obtain convergence at rate $n^{1/2}$ one must use bias reduction (higher-order kernels) when dimensionality $d \geq 2$, requiring therefore some additional smoothness on α, g and f. Even if these conditions are satisfied, the small sample properties of $\widehat{\theta}$ are likely to be poor. Therefore, when there are many covariates, some dimensionality reducing strategy such as additive, multiplicative or even index structures, would seem desirable; see Andersen, Borgan, Gill and Keiding (1993) for an extensive discussion of such models in hazard estimation. We consider one particular dimensionality reducing model that has been much applied in regression problems, the index model

$$\lambda(z, t) = \alpha(t; \theta) g(\beta^T z),$$

where g is of unknown form [this is a special type of "projection pursuit" model; see Friedman and Stuetzle (1981)]. There are a number of plausible estimation methods here. One can use our previous method (which ignores the restriction on the covariate effect) to get estimates of θ and of the unrestricted covariate effect we call $g_U(z) = g(\beta^T z)$. Then, note that

$$E\left[\frac{\partial g_U(z)}{\partial z}\right] = E[g'(\beta^T z)]\beta,$$

where expectation is taken with respect to the distribution of z. Thus the average derivative is proportional to the parameters β. This fact has been used in the regression context to estimate parameters of index models [see Härdle and Stoker (1989)] by substituting nonparametric estimates of the derivatives of $g_U(\cdot)$ and averaging over the sample evaluation points. It has the advantage of simplicity, but is likely to be inefficient because it takes no account of the additional structure (beyond the mean) given in (8), and to suffer from the curse of dimensionality since partial derivatives of multidimensional smoothers must be computed. We therefore suggest the following alternative method based on profiling both β and θ. Let $\psi = (\theta, \beta)$ and

$$\widehat{g}_\psi(z) = \frac{n^{-1} \sum_{j \neq i} \int K_b\{\beta^T z - \beta^T Z_j(u)\}\, dN_j(u)}{n^{-1} \sum_{j \neq i} \int K_b\{\beta^T z - \beta^T Z_j(u)\}\alpha(u; \theta) Y_j(u)\, du} \equiv \frac{\widehat{r}_\beta(z)}{\widehat{e}_\psi(z)},$$

where as before $K(\cdot)$ is a one-dimensional kernel. Now substitute $\widehat{g}_\psi\{\beta^T Z_i(u)\}$ into (9) to obtain

$$\widehat{\ell}(\psi) = \sum_{i=1}^{n} \int \widehat{\mu}_\psi(u, Z_i(u))\, dN_i(u) - \sum_{i=1}^{n} \int \exp\{\widehat{\mu}_\psi(u, Z_i(u))\} Y_i(u)\, du, \tag{20}$$

where $\widehat{\mu}_\psi(u,z) = \ln[\alpha(u;\theta)\widehat{g}_\psi(\beta^T z)]$, and let $\widehat{\psi}$ be any maximizer of $\widehat{\ell}(\psi)$. The average derivative estimator could be used to provide starting values.

The asymptotic properties of $\widehat{\psi}$ follow by similar calculations to those used in establishing Theorem 2 above; specifically, under additional regularity conditions, one expects that

$$n^{1/2}(\widehat{\psi} - \psi) \Rightarrow N(0, \mathscr{I}_{00}^{-1}),$$

where

$$\mathscr{I}_{00} = \iint \frac{\partial \bar{\mu}_{\psi_0}}{\partial \psi} \frac{\partial \bar{\mu}_{\psi_0}}{\partial \psi^T}(u,z)\alpha(u;\theta_0)g(\beta^T z)f(z,u)y(u)\,du\,dz,$$

$$\frac{\partial \bar{\mu}_\psi}{\partial \theta}(u;\psi_0) = \frac{\partial \ln \alpha(u;\theta)}{\partial \theta} + p\lim_{n\to\infty} \frac{\partial \ln \widehat{g}_\psi\{\beta^T Z_i(u)\}}{\partial \theta},$$

$$\frac{\partial \bar{\mu}_\psi}{\partial \beta}(u;\psi_0) = p\lim_{n\to\infty} \frac{\partial \ln \widehat{g}_\psi\{\beta^T Z_i(u)\}}{\partial \beta},$$

while $\widehat{H}_{\psi\psi}^{-1}(\widehat{\psi}) \to_p \mathscr{I}_{00}^{-1}$, where $\widehat{H}_{\psi\psi}(\psi)$ is the Hessian matrix of $\widehat{\ell}(\psi)$. The asymptotic variance matrix has a form similar to that found by Klein and Spady (1991) in a related semiparametric index problem and depends on the conditional distributions $Z_j(u)|\beta^T Z_j(u)$: specifically,

$$p\lim_{n\to\infty} \frac{\partial \widehat{g}_\psi}{\partial \beta}(\beta^T z) = \{z - R(\beta^T z)\}g'(\beta^T z),$$

where $R(x) = E[Z_j(u)|\beta^T Z_j(u) = x]$.

APPENDIX

A.1. The predictability issue. The purpose of the present section is to provide a useful method for proving convergence in probability of certain key quantities. Consider

$$\overline{M}_t = \sum_{i=1}^n \int_0^t h_i^{(n)}(u)\,dM_i(u),$$

where M_i is the martingale defined in Section 2, but $h_i^{(n)}$ is not a predictable process according to the usual definition. Basically what is needed is to perform an approximation of the type

$$\overline{M}_t = \sum_{i=1}^n \int_0^t h_i^{(n)}(u)dM_i(u) \simeq \sum_{i=1}^n \int_0^t \widetilde{h}_i^{(n)}(u)dM_i(u), \tag{21}$$

where the $\widetilde{h}_i^{(n)}$'s are predictable processes, and then to apply standard martingale theory to the right-hand side. If the $h_i^{(n)}$'s and the $\widetilde{h}_i^{(n)}$'s were cadlag,

one way to make this approximation would be to employ partial integration for cadlag functions, that is,

$$\begin{aligned}(22)\qquad &\sum_{i=1}^{n}\int_0^t h_i^{(n)}(u)\,dM_i(u)-\sum_{i=1}^{n}\int_0^t \widetilde{h}_i^{(n)}(u)\,dM_i(u)\\ &\qquad=\sum_{i=1}^{n}\int_0^t M_i(u)\,d(h_i^{(n)}-\widetilde{h}_i^{(n)})(u),\end{aligned}$$

provided $(h_i^{(n)}-\widetilde{h}_i^{(n)})(0), (h_i^{(n)}-\widetilde{h}_i^{(n)})(t)=0$. If the differentials $d(h_1^{(n)}-\widetilde{h}_1^{(n)}),\dots,d(h_n^{(n)}-\widetilde{h}_n^{(n)})$ have a sufficiently simple structure, then (22) can be used to carry out the approximation needed to solve the predictability problem. While partial integration is a useful solution in some simple cases, it is unfortunately not widely applicable since it is excessively crude when applied to high-dimensional kernels.

The method presented below is based on simple approximations of the integral (21) by integrals of the same form, but where the $h_i^{(n)}$'s obey certain leave-one-out properties. To this end we need the following definition.

DEFINITION A.1. The sequence of processes $\{h_{i_1,\dots,i_k}^{(n)}\}$ is of the leave-k-out type if $h_{i_1,\dots,i_k}^{(n)}$ is predictable with respect to the filtration given by

$$\mathscr{F}_t^{(i_1,\dots,i_k;n)}=\bigvee_{j\notin\{i_1,\dots,i_k\}}\mathscr{F}_{j,1}^{(n)}\bigvee_{l=1}^{k}\mathscr{F}_{i_l,t}^{(n)}.$$

We use below the facts that $h_j^{(n)}(t)$ is predictable with respect to $\mathscr{F}_t^{(j;n)}$, that $h_j^{(n)}(t)-h_{i,j}^{(n)}(t)$ is predictable with respect to $\mathscr{F}_t^{(i,j;n)}$ and that $M_j=N_j-\Lambda_j$ is a martingale with respect to both filtrations. These are consequences of the i.i.d. setup we adopted.

LEMMA 1. *Suppose that the processes* $\{h_i^{(n)}(u)\}$ *and* $\{h_{i,j}^{(n)}(u)\}$, $i,j=1,\dots,n$, *are cadlag and of the leave-one-out and leave-two-out types, respectively. Let* $\alpha_i=\{E\int_0^t h_i^{(n)}(u)^2\,d\Lambda_i(u)\}^{1/2}$ *and* $\delta_i=\max_{j\le n}[E\int_0^t\{h_i^{(n)}(u)-h_{i,j}^{(n)}(u)\}^2\,d\Lambda_i(u)]^{1/2}$. *Then*

$$E(\overline{M}_t^2)\le n\sum_{i=1}^{n}\delta_i^2+2\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\delta_j+\sum_{i=1}^{n}\alpha_i^2.$$

PROOF. We divide into diagonal out and diagonal in terms,

$$\begin{aligned}E(\overline{M}_t^2)&=E\Bigg[\sum_{i=1}^{n}\sum_{j=1}^{n}\int_0^t\int_0^t h_i^{(n)}(u)h_j^{(n)}(v)\mathbf{1}(u=v)\,dM_i(u)\,dM_j(v)\Bigg]\\ &\quad+E\Bigg[\sum_{i=1}^{n}\sum_{j=1}^{n}\int_0^t\int_0^t h_i^{(n)}(u)h_j^{(n)}(v)\mathbf{1}(u\ne v)\,dM_i(u)\,dM_j(v)\Bigg]\\ &=\mathrm{I}+\mathrm{II}.\end{aligned}$$

We first deal with II. Writing $h_i^{(n)} = h_{i,j}^{(n)} + [h_i^{(n)} - h_{i,j}^{(n)}]$,

$$\begin{aligned}
\mathrm{II} &= E\Bigg[\sum_{i=1}^n \sum_{j=1}^n \int_0^t \int_0^t h_{i,j}^{(n)}(u) h_{i,j}^{(n)}(v) \mathbf{1}(u \neq v)\, dM_i(u)\, dM_j(v)\Bigg] \\
&\quad + 2E\Bigg[\sum_{i=1}^n \sum_{j=1}^n \int_0^t \int_0^t h_i^{(n)}(u)\big[h_j^{(n)}(v) - h_{i,j}^{(n)}(v)\big] \mathbf{1}(u \neq v)\, dM_i(u)\, dM_j(v)\Bigg] \\
&\quad + E\Bigg[\sum_{i=1}^n \sum_{j=1}^n \int_0^t \int_0^t \big[h_i^{(n)}(u) - h_{i,j}^{(n)}(u)\big]\big[h_j^{(n)}(v) - h_{i,j}^{(n)}(v)\big] \\
&\qquad\qquad \times \mathbf{1}(u \neq v)\, dM_i(u)\, dM_j(v)\Bigg] \\
&= \mathrm{IIa} + \mathrm{IIb} + \mathrm{IIc}.
\end{aligned}$$

The first term is zero, by the following argument. First, replace $\mathbf{1}(u \neq v)$ by $\mathbf{1}(u < v)$ and multiply by a factor of 2. Then note that $k_j(v) = \sum_{i=1}^n h_{i,j}^{(n)}(v) \int_0^v h_{i,j}^{(n)}(u)\, dM_i(u)$ is predictable with respect to $\mathscr{F}_v^{(j;n)}$, so that

$$IIa = E\Bigg[\sum_{j=1}^n \int_0^t k_j(v)\, dM_j(v)\Bigg] = 0,$$

since M_j is a martingale with respect to $\mathscr{F}_v^{(j;n)}$.

Now for the second term:

$$\begin{aligned}
\mathrm{IIb} &= 2E\Bigg[\sum_{i=1}^n \sum_{j=1}^n \int_0^t h_i^{(n)}(u)\, dM_i(u) \int_0^t \big[h_j^{(n)}(v) - h_{i,j}^{(n)}(v)\big]\, dM_j(v)\Bigg] \\
&\le 2\sum_{i=1}^n \sum_{j=1}^n \Bigg[E\Bigg\{\int_0^t h_i^{(n)}(u)^2\, d\Lambda_i(u)\Bigg\}\Bigg]^{1/2} \\
&\qquad \times \Bigg[E\Bigg\{\int_0^t \big[h_j^{(n)}(v) - h_{i,j}^{(n)}(v)\big]^2\, d\Lambda_j(v)\Bigg\}\Bigg]^{1/2} \\
&= 2\sum_{i=1}^n \sum_{j=1}^n \alpha_i \delta_j,
\end{aligned}$$

where the second line follows by the Cauchy–Schwarz inequality and the martingale property.

For the third term, we apply the Cauchy–Schwarz inequality to obtain

$$\mathrm{IIc} \le nE\Bigg[\sum_{i=1}^n \int_0^t \big\{h_i^{(n)}(u) - h_{i,j}^{(n)}(u)\big\}^2 d\Lambda_i(u)\Bigg] = n\sum_{i=1}^n \delta_i^2.$$

As for the diagonal term I,

$$\begin{aligned}\mathrm{I} &= E\Bigg[\sum_{i=1}^{n}\sum_{j=1}^{n}\int_0^t\int_0^t h_i^{(n)}(v)h_j^{(n)}(v)\Delta N_i(v)\,dM_j(v)\Bigg]\\ &= E\Bigg[\sum_{j=1}^{n}\int_0^t h_i^{(n)}(v)^2\,dN_j(v)\Bigg]\\ &= E\Bigg[\sum_{j=1}^{n}\int_0^t h_i^{(n)}(v)^2\,d\Lambda_j(v)\Bigg] = \sum_{i=1}^{n}\alpha_i^2. \qquad \square\end{aligned}$$

A.2. Proofs of theorems. Without loss of generality, suppose that $\mathscr{Z} = [\underline{a}, \bar{a}]$ for $-\infty < \underline{a} < \bar{a} < \infty$, and define the interior region $\mathscr{Z}_n^0 = [\underline{a} + b, \bar{a} - b]$ and boundary region $\partial\mathscr{Z}_n = \mathscr{Z} \setminus \mathscr{Z}_n^0$.

The following lemma establishes global convergence for a number of useful quantities: including the standardized denominator of (5) and its partial derivatives with respect to θ:

$$\widehat{e}_\theta(z) = n^{-1}\sum_{j\neq i}\int K_b\{z - Z_j(u)\}\alpha(u;\theta)Y_j(u)\,du;$$

$$\frac{\partial\widehat{e}_\theta}{\partial\theta}(z) = n^{-1}\sum_{j\neq i}\int K_b\{z - Z_j(u)\}\frac{\partial\alpha(u;\theta)}{\partial\theta}Y_j(u)\,du;$$

$$\frac{\partial^2\widehat{e}_\theta}{\partial\theta\partial\theta^T}(z) = n^{-1}\sum_{j\neq i}\int K_b\{z - Z_j(u)\}\frac{\partial^2\alpha(u;\theta)}{\partial\theta\partial\theta^T}Y_j(u)\,du.$$

LEMMA 2 (Uniform convergence). *Suppose that* (A1)–(A3) *hold. Suppose further that* $nb^3 \to 0$ *and* $b \to 0$. *Then, as* $n \to \infty$, *we have the following*:

(a) $$\sup|\widehat{e}_\theta(z) - e_\theta(z)| \to_p 0,$$

(b) $$\sup|\widehat{g}_\theta(z) - g_\theta(z)| \to_p 0,$$

where the suprema are taken over $(z, \theta) \in \mathscr{Z} \times \mathscr{N}_0$. *Furthermore, if also* (A5) *holds, then*:

(c) $$\sup\left|\frac{\partial\widehat{e}_\theta}{\partial\theta_\gamma}(z) - \frac{\partial e_\theta}{\partial\theta_\gamma}(z)\right| \to_p 0,$$

(d) $$\sup\left|\frac{\partial^2\widehat{e}_\theta}{\partial\theta_\gamma\partial\theta_\delta}(z) - \frac{\partial^2 e_\theta}{\partial\theta_\gamma\partial\theta_\delta}(z)\right| \to_p 0,$$

(e) $$\sup\left|\frac{\partial\ln\widehat{e}_\theta}{\partial\theta_\gamma}(z) - \frac{\partial\ln e_\theta}{\partial\theta_\gamma}(z)\right| \to_p 0,$$

(f) $$\sup\left|\frac{\partial^2\ln\widehat{e}_\theta}{\partial\theta_\gamma\partial\theta_\delta}(z) - \frac{\partial^2\ln e_\theta}{\partial\theta_\gamma\partial\theta_\delta}(z)\right| \to_p 0,$$

for all $\gamma, \delta = 1, \ldots, p$, *where the suprema are taken over* $(z, \theta) \in \mathscr{D} \times \mathscr{N}_0$. *Finally, we have the following*:

$$\text{(g)} \qquad \begin{aligned} \sup_{(z,\theta)\in\mathscr{D}_n^0\times\mathscr{N}_0} \left|\widehat{g}_\theta^*(z) - g_\theta(z)\right| &= O_p(b^2); \\ \sup_{(z,\theta)\in\partial\mathscr{D}_n\times\mathscr{N}_0} \left|\widehat{g}_\theta^*(z) - g_\theta(z)\right| &= O_p(b). \end{aligned}$$

PROOF. [(a) and (b)] Nielsen and Linton [(1995), Theorem 2] established uniform consistency for $\widehat{g}_{\theta_0}(z)$ over $\mathscr{D}$ using the global convergence criteria developed by Bickel and Wichura (1971). Because of the Lipschitz continuity condition on α, uniform convergence over $\mathscr{N}_0$ follows too.

[(c) and (d)] Likewise (c) and (d) follow from the Lipschitz continuity of $\partial\alpha(u;\theta)/\partial\theta$ and $\partial^2\alpha(u;\theta)/\partial\theta\,\partial\theta^T$.

[(e) and (f)] Statements (e) and (f) are consequences of (a), (c) and (d), because

$$\frac{\partial \ln \widehat{e}_\theta}{\partial\theta_\gamma}(z) - \frac{\partial \ln e_\theta}{\partial\theta_\gamma}(z) = \frac{1}{\widehat{e}_\theta(z)}\left[\left\{\frac{\partial \widehat{e}_\theta}{\partial\theta_\gamma}(z) - \frac{\partial e_\theta}{\partial\theta_\gamma}(z)\right\} - \frac{\partial \ln e_\theta}{\partial\theta_\gamma}(z)\{\widehat{e}_\theta(z) - e_\theta(z)\}\right],$$

and $\{\inf \widehat{e}_\theta(z)\}^{-1} = O_p(1)$, which is itself implied by the inequality

$$\inf e_\theta(z) \le \inf \widehat{e}_\theta(z) + \sup|\widehat{e}_\theta(z) - e_\theta(z)| = \inf \widehat{e}_\theta(z) + o_p(1),$$

which follows by (a) and (A1). Similarly, we can write $\partial^2 \ln \widehat{e}_\theta(z)/\partial\theta\,\partial\theta^T - \partial^2 \ln e_\theta(z)/\partial\theta\,\partial\theta^T$ in terms of $\widehat{e}_\theta(z) - e_\theta(z)$, $\partial \widehat{e}_\theta(z)/\partial\theta - \partial e_\theta(z)/\partial\theta$ and $\partial^2\widehat{e}_\theta(z)/\partial\theta\,\partial\theta^T - \partial^2 e_\theta(z)/\partial\theta\,\partial\theta^T$, and apply the results (a), (c) and (d).

(g) Statement (g) follows by a Taylor expansion using (A2) and (A3); the boundary effect is due to our use of an uncorrected kernel estimator throughout and can be removed by appropriate techniques [see Andersen, Borgan, Gill and Keiding (1993), page 251]. An alternative approach here is to trim out boundary observations. □

PROOF OF THEOREM 1. It suffices to show that the following hold:

$$\text{(C1)} \qquad \sup_{\theta\in\mathscr{N}_0}\left|n^{-1}\sum_{i=1}^n \int [\ln \widehat{g}_\theta\{Z_i(u)\} - \ln g_\theta\{Z_i(u)\}]\, dN_i(u)\right| \to_p 0;$$

$$\text{(C2)} \qquad \sup_{\theta\in\mathscr{N}_0}\left|n^{-1}\sum_{i=1}^n \int \alpha(u;\theta)[\widehat{g}_\theta\{Z_i(u)\} - g_\theta\{Z_i(u)\}]Y_i(u)\, du\right| \to_p 0.$$

By the Cauchy–Schwarz inequality, the left-hand side of (C2) is majorized by

$$\sup_{\theta\in\mathscr{N}_0,\, z\in\mathscr{D}} |\widehat{g}_\theta(z) - g_\theta(z)|\left\{\sup_{\theta\in\mathscr{N}_0} n^{-1}\sum_{i=1}^n \int \alpha^2(u;\theta)Y_i(u)\, du\right\}^{1/2} \to_p 0,$$

where convergence to zero follows from Lemma 2(b) and the fact that the second random variable is tight due to (A1)–(A2).

By the mean value theorem and positivity of $\widehat{g}_\theta(z)$,

$$|\ln \widehat{g}_\theta(z) - \ln g_\theta(z)| \le \{\inf \widehat{g}_\theta(z)\}^{-1} |\widehat{g}_\theta(z) - g_\theta(z)|,$$

where $\{\inf \widehat{g}_\theta(z)\}^{-1} = O_p(1)$ by Lemma 2(b) and (A1) [(A1) implies that $\inf g_\theta(z) > 0$]. Therefore, (C1) is satisfied. □

PROOF OF THEOREM 2. We must show that

$$n^{1/2}\{\widehat{s}_\theta(\theta_0) - s^e_\theta(\theta_0)\} = o_p(1) \tag{S1}$$

and that

$$\sup \left| \widehat{H}_1(\theta) + \iint \frac{\partial \bar{\mu}_\theta}{\partial \theta} \frac{\partial \bar{\mu}_\theta}{\partial \theta^T}(u, z) \alpha(u; \theta) g_\theta(z) f(z, u) y(u)\, dz\, du \right| = o_p(1), \tag{H1}$$

$$\sup \left| \widehat{H}_2(\theta) - \iint \frac{\partial^2 \bar{\mu}_\theta}{\partial \theta \partial \theta^T}(u, z) \{\alpha(u; \theta_0) g(z) - \alpha(u; \theta) g_\theta(z)\} \times f(z, u) y(u)\, dz\, du \right| = o_p(1), \tag{H2}$$

$$\sup |\widehat{H}_3(\theta)| = o_p(1), \tag{H3}$$

$$\sup |\widehat{H}_4(\theta)| = o_p(1), \tag{H4}$$

where the suprema are taken over $\mathcal{N}_n = \{\theta: |\theta - \theta_0| \le \delta_n\}$ for some sequence $\delta_n \to 0$. Here, $\widehat{H}_{\theta\theta}(\theta) = \sum_{j=1}^4 \widehat{H}_j(\theta)$, where

$$\widehat{H}_1(\theta) = -n^{-1} \sum_{i=1}^n \int \frac{\partial \widehat{\mu}_\theta}{\partial \theta} \frac{\partial \widehat{\mu}_\theta}{\partial \theta^T} \{u, Z_i(u)\} \alpha(u; \theta) g_\theta\{Z_i(u)\} Y_i(u)\, du,$$

$$\widehat{H}_2(\theta) = n^{-1} \sum_{i=1}^n \int \frac{\partial^2 \widehat{\mu}_\theta}{\partial \theta \partial \theta^T} \{u, Z_i(u)\} \times \big[\alpha(u; \theta_0) g\{Z_i(u)\} - \alpha(u; \theta) g_\theta\{Z_i(u)\}\big] Y_i(u)\, du,$$

$$\widehat{H}_3(\theta) = n^{-1} \sum_{i=1}^n \int \frac{\partial^2 \widehat{\mu}_\theta}{\partial \theta \partial \theta^T} \{u, Z_i(u)\}\, dM_i(u),$$

$$\widehat{H}_4(\theta) = -n^{-1} \sum_{i=1}^n \int \left\{ \frac{\partial^2 \widehat{\mu}_\theta}{\partial \theta \partial \theta^T} + \frac{\partial \widehat{\mu}_\theta}{\partial \theta} \frac{\partial \widehat{\mu}_\theta}{\partial \theta^T} \right\} \{u, Z_i(u)\} \times \alpha(u; \theta) \{\widehat{g}_\theta - g_\theta\}\{Z_i(u)\} Y_i(u)\, du.$$

We begin with the Hessian results (H1)–(H4) because they only require the uniform convergence results of Lemma 2.

PROOF OF (H1). By the triangle inequality, we can bound the left-hand side of (H1) by

$$\begin{aligned}
&\left|\widehat{H}_1(\theta) + n^{-1}\sum_{i=1}^{n}\int \frac{\partial\bar{\mu}_\theta}{\partial\theta}\frac{\partial\bar{\mu}_\theta}{\partial\theta^T}\{u; Z_i(u)\}\alpha(u;\theta)g_\theta\{Z_i(u)\}Y_i(u)\,du\right| \\
&\quad + \left|n^{-1}\sum_{i=1}^{n}\int \frac{\partial\bar{\mu}_\theta}{\partial\theta}\frac{\partial\bar{\mu}_\theta}{\partial\theta^T}\{u; Z_i(u)\}\alpha(u;\theta)g_\theta\{Z_i(u)\}Y_i(u)\,du\right. \\
&\qquad \left. - \iint \frac{\partial\bar{\mu}_\theta}{\partial\theta}\frac{\partial\bar{\mu}_\theta}{\partial\theta^T}(u,z)\alpha(u;\theta)g_\theta(z)f(z,u)y(u)\,dz\,du\right| \\
&\quad = \mathrm{I}(\theta) + \mathrm{II}(\theta), \quad \text{say},
\end{aligned}$$

where $\sup_{\theta\in\mathscr{N}_n} II(\theta) = o_p(1)$ by a standard uniform law of large numbers, while by the Cauchy–Schwarz inequality,

$$\begin{aligned}
I(\theta) &\le \sup_z\left|\frac{\partial\ln\widehat{e}_\theta}{\partial\theta}\frac{\partial\ln\widehat{e}_\theta}{\partial\theta^T}(z) - \frac{\partial\ln e_\theta}{\partial\theta}\frac{\partial\ln e_\theta}{\partial\theta^T}(z)\right| \\
&\qquad \times n^{-1}\sum_{i=1}^{n}\int \alpha^2(u;\theta)g_\theta^2\{Z_i(u)\}Y_i(u)\,du \\
&= o_p(1),
\end{aligned}$$

where $o_p(1)$ is uniform in θ by Lemma 2(e), the inequality $|a-b| \le 2|b|\,|a-b| + |a-b|^2$ and the tightness (in $\theta\in\mathscr{N}_n$) of $n^{-1}\sum_{i=1}^{n}\int\alpha^2(u;\theta)g_\theta^2\{Z_i(u)\}Y_i(u)\,du$. □

PROOF OF (H2). This follows from the results

$$\begin{aligned}
(23)\qquad & n^{-1}\sum_{i=1}^{n}\int \frac{\partial^2\bar{\mu}_\theta}{\partial\theta\,\partial\theta^T}\{u, Z_i(u)\} \\
&\qquad \times\big[\alpha(u;\theta_0)g\{Z_i(u)\} - \alpha(u;\theta)g_\theta\{Z_i(u)\}\big]Y_i(u)\,du = o_p(1)
\end{aligned}$$

uniformly over $\theta\in\mathscr{N}_n$, where

$$\frac{\partial^2\bar{\mu}_\theta}{\partial\theta\,\partial\theta^T}(u,z) = \frac{\partial^2\ln\alpha}{\partial\theta\,\partial\theta^T}(u,\theta) - \frac{\partial^2\ln e_\theta}{\partial\theta\,\partial\theta^T}(z)$$

and

$$\begin{aligned}
(24)\qquad & n^{-1}\sum_{i=1}^{n}\int\left\{\frac{\partial^2\widehat{\mu}_\theta}{\partial\theta\partial\theta^T} - \frac{\partial^2\bar{\mu}_\theta}{\partial\theta\partial\theta^T}\right\}\{u, Z_i(u)\} \\
&\qquad \times\big[\alpha(u;\theta_0)g\{Z_i(u)\} - \alpha(u;\theta)g_\theta\{Z_i(u)\}\big]Y_i(u)\,du = o_p(1)
\end{aligned}$$

uniformly over $\theta \in \mathscr{N}_n$. A uniform law of large numbers implies (23), while (24) follows by Lemma 2(f) and the Cauchy–Schwarz inequality. □

PROOF OF (H3) AND (H4). Statements (H3) and (H4) follow by applying Cauchy–Schwarz and Lemma 2. □

We now turn to the properties of the score function, (S1). We must show that

$$n^{-1/2}\sum_{i=1}^{n}\int\left\{\frac{\partial\widehat{\mu}_{\theta_0}}{\partial\theta}-\frac{\partial\bar{\mu}_{\theta}}{\partial\theta}\right\}\{u, Z_i(u)\}\,dM_i(u)\to_p 0, \tag{25}$$

$$n^{-1/2}\sum_{i=1}^{n}\int\frac{\partial\widehat{\mu}^{*}_{\theta_0}}{\partial\theta}\{u, Z_i(u)\}\,dM_i(u)\to_p 0, \tag{26}$$

$$n^{-1/2}\sum_{i=1}^{n}\int\frac{\partial\widehat{\mu}_{\theta_0}}{\partial\theta}\{u, Z_i(u)\}\alpha(u;\theta_0)(g^{*}_{\theta_0}-g)\{Z_i(u)\}Y_i(u)\,du\to_p 0. \tag{27}$$

PROOF OF (27). First, we use the triangle inequality to bound the left-hand side of (27) by $T_{n1}+T_{n2}$, where

$$T_{n1}=\left|n^{-1/2}\sum_{i=1}^{n}\int\frac{\partial\widehat{\mu}_{\theta_0}}{\partial\theta}\{u, Z_i(u)\}\alpha(u;\theta_0)(g^{*}_{\theta_0}-g)\{Z_i(u)\}\times\mathbf{1}\{Z_i(u)\in\mathscr{D}_n^0\}Y_i(u)\,du\right|,$$

$$T_{n2}=\left|n^{-1/2}\sum_{i=1}^{n}\int\frac{\partial\widehat{\mu}_{\theta_0}}{\partial\theta}\{u, Z_i(u)\}\alpha(u;\theta_0)(g^{*}_{\theta_0}-g)\{Z_i(u)\}\times\mathbf{1}\{Z_i(u)\in\partial\mathscr{D}_n\}Y_i(u)\,du\right|.$$

Then use Cauchy–Schwarz to obtain the bound

$$T_{n1}\le\left[n^{-1}\sum_{i=1}^{n}\int\left\{\frac{\partial\widehat{\mu}_{\theta_0}}{\partial\theta}\{u, Z_i(u)\}\right\}^2\alpha^2(u;\theta_0)Y_i(u)\,du\right]^{1/2}\times n^{1/2}\sup_{z\in\mathscr{D}_n^0}|g^{*}_{\theta_0}(z)-g(z)|.$$

The first term on the right-hand side is $O_p(1)$, while the second is $O_p(n^{1/2}b^2)$ [$=o_p(1)$ by the bandwidth conditions] by Lemma 2(g). As for T_{n2}, note that

$$n^{-1}\sum_{i=1}^{n}\int\mathbf{1}\{Z_i(u)\in\partial\mathscr{D}_n\}Y_i(u)\,du=O_p(b),$$

so that $T_{n2}=O_p(n^{1/2}b^2)$ too. □

PROOF OF (25). Note that

$$\begin{aligned}\left\{\frac{\partial\widehat{\mu}_{\theta_0}}{\partial\theta}-\frac{\partial\overline{\mu}_{\theta}}{\partial\theta}\right\}\{u,Z_i(u)\}&=-\left\{\frac{\partial\ln\widehat{e}_{\theta_0}}{\partial\theta}-\frac{\partial\ln e_{\theta_0}}{\partial\theta}\right\}\{Z_i(u)\}\\&=-\widehat{e}_{\theta_0}^{-1}\left\{\frac{\partial\widehat{e}_{\theta_0}}{\partial\theta}-\frac{\partial e_{\theta_0}}{\partial\theta}\right\}\{Z_i(u)\}\\&\quad+\frac{\partial e_{\theta_0}}{\partial\theta}\left\{\frac{\widehat{e}_{\theta_0}-e_{\theta_0}}{\widehat{e}_{\theta_0}e_{\theta_0}}\right\}\{Z_i(u)\}.\end{aligned}$$

Note that

$$\overline{M}_t=n^{-1/2}\sum_{i=1}^{n}\int\widehat{e}_{\theta_0}^{-1}\left\{\frac{\partial\widehat{e}_{\theta_0}}{\partial\theta}-\frac{\partial e_{\theta_0}}{\partial\theta}\right\}\{Z_i(u)\}\,dM_i(u)$$

is of the general form $\overline{M}_t=\sum_{i=1}^{n}\int_0^t h_i^{(n)}(u)\,dM_i(u)$, where the M_i is a martingale, but $\{h_i^{(n)}(u)\}$ is not a predictable process according to the usual definition. Furthermore, the random denominator $\widehat{e}_{\theta_0}$ can take very small values and even negative values. Define $\mathscr{A}_n=\{\inf\widehat{e}_{\theta_0}(z)>c\}$, where $c=\inf e_{\theta_0}(z)/2$, and note that $\Pr[\mathscr{A}_n^c]\to 0$ as $n\to\infty$. Therefore, for any $\varepsilon>0$,

$$\begin{aligned}\Pr\bigl[|\overline{M}_t|>\varepsilon\bigr]&\le\Pr\bigl[\{|\overline{M}_t|>\varepsilon\}\cap\mathscr{A}_n\bigr]+\Pr[\mathscr{A}_n^c]\\&=\Pr\bigl[\{|\overline{M}_t|>\varepsilon\}\cap\mathscr{A}_n\bigr]+o(1)\\&\le\Pr\bigl[\{|\overline{M}_t^*|>\varepsilon\}\bigr]+o(1),\end{aligned}$$

where

$$\overline{M}_t^*=n^{-1/2}\sum_{i=1}^{n}\int\mathbf{1}[\widehat{e}_{\theta_0}\{Z_i(u)\}>c]\,\widehat{e}_{\theta_0}^{-1}\left\{\frac{\partial\widehat{e}_{\theta_0}}{\partial\theta}-\frac{\partial e_{\theta_0}}{\partial\theta}\right\}\{Z_i(u)\}\,dM_i(u).$$

We can now apply Lemma 1 with

$$h_i^{(n)}(u)=n^{-1/2}\,\widehat{e}_{\theta_0}^{-1}\mathbf{1}[\widehat{e}_{\theta_0}\{Z_i(u)\}>c]\left\{\frac{\partial\widehat{e}_{\theta_0}}{\partial\theta_\pi}-\frac{\partial e_{\theta_0}}{\partial\theta_\pi}\right\}\{Z_i(u)\},$$

for any $\pi=1,\dots,p$, and $h_{i,j}^{(n)}(u)$ the same quantity but with the jth term left out of the definition of $\widehat{e}_\theta\{Z_i(u)\}$ and $\partial\widehat{e}_{\theta_0}\{Z_i(u)\}/\partial\theta_\pi$. First, note that

$$\int E[h_i^{(n)}(u)^2]d\Lambda_i(u)=O(n^{-2}b^{-1})+O(n^{-1}b^4)$$

[see Nielsen and Linton (1995) and the argument above for (27)]. The expression for $h_i^{(n)}(u)-h_{i,j}^{(n)}(u)$ is quite complicated and is omitted for brevity; clearly it suffices to work with the following leading components thereof:

$$\begin{aligned}\theta_{i,j}^{(n)}(u)&=\frac{1}{n^{3/2}}\,\widehat{e}_{\theta_0}^{-1}\{Z_i(u)\}\mathbf{1}[\widehat{e}_{\theta_0}\{Z_i(u)\}>c]\\&\quad\times\int K_b\{Z_i(u)-Z_j(t)\}\frac{\partial\alpha(t;\theta_0)}{\partial\theta_\pi}Y_j(t)\,dt;\\\rho_{i,j}^{(n)}(u)&=n^{-1/2}\left\{\frac{\partial\widehat{e}_{\theta_0}}{\partial\theta_\pi}-\frac{\partial e_{\theta_0}}{\partial\theta_\pi}\right\}\{Z_i(u)\}\frac{\int K_b\{Z_i(u)-Z_j(t)\}\alpha(t;\theta_0)Y_j(t)\,dt}{\widehat{e}_{\theta_0}e_{\theta_0}\{Z_i(u)\}}\\&\quad\times\mathbf{1}[\widehat{e}_{\theta_0}\{Z_i(u)\}>c];\end{aligned}$$

$$\phi_{i,j}^{(n)}(u) = n^{-1/2}\bigl(\mathbf{1}[\widehat{e}_{\theta_0}\{Z_i(u)\} > c] - \mathbf{1}\bigl[\widehat{e}_{\theta_0}^{(j)}\{Z_i(u)\} > c\bigr]\bigr)$$
$$\times \widehat{e}_{\theta_0}^{-1}\left\{\frac{\partial \widehat{e}_{\theta_0}}{\partial \theta_\pi} - \frac{\partial e_{\theta_0}}{\partial \theta_\pi}\right\}\{Z_i(u)\},$$

where the superscript j is used to denote the fact that the jth observation was left out. We have, for example,

$$\begin{aligned} &E\left[\int \{\theta_{i,j}^{(n)}(u)\}^2\, d\Lambda_i(u)\right] \\ &\quad \le c^2 n^{-3} \int E\left[\left\{\int K_b\{Z_i(u) - Z_j(t)\}\frac{\partial \alpha}{\partial \theta_\pi}(t;\theta_0) Y_j(t)\, dt\right\}^2\right] d\Lambda_i(u) \\ &\quad \le c^2 n^{-3} \int \left\{\frac{\partial \alpha}{\partial \theta_\pi}(t;\theta_0)\right\}^2 dt \int E\left[\int K_b^2\{Z_i(u) - Z_j(t)\} Y_j(t)\, dt\right] d\Lambda_i(u) \\ &\quad = O(n^{-3} b^{-1}), \end{aligned}$$

by the Cauchy–Schwarz inequality and standard change of variables argument for kernels. Similar arguments apply to $E[\int \{\rho_{i,j}^{(n)}(u)\}^2\, d\Lambda_i(u)]$ and $E[\int \{\phi_{i,j}^{(n)}(u)\}^2\, d\Lambda_i(u)]$. Therefore, $E[|\overline{M}_t|^2] \to 0$, which by Chebyshev's inequality implies that $\overline{M}_t \to_p 0$. Similar techniques can be applied to show that $\overline{\overline{M}}_t \to_p 0$, where

$$\overline{\overline{M}}_t = n^{-1/2} \sum_{i=1}^n \int \frac{\partial e_{\theta_0}}{\partial \theta}\left\{\frac{\widehat{e}_{\theta_0} - e_{\theta_0}}{\widehat{e}_{\theta_0} e_{\theta_0}}\right\}\{Z_i(u)\}\, d\overline{M}_i(u). \qquad \square$$

PROOF OF (26). We show that

$$n^{-1/2} \sum_{i=1}^n \int \frac{\partial \mu_{\theta_0}^*}{\partial \theta}\{u, Z_i(u)\}\, d\overline{M}_i(u) \to_p 0, \tag{28}$$

$$n^{-1/2} \sum_{i=1}^n \int \left\{\frac{\partial \widehat{\mu}_{\theta_0}^*}{\partial \theta} - \frac{\partial \mu_{\theta_0}^*}{\partial \theta}\right\}\{u, Z_i(u)\}\, d\overline{M}_i(u) \to_p 0, \tag{29}$$

where

$$\frac{\partial \mu_{\theta_0}^*}{\partial \theta}\{u, Z_i(u)\} = e_{\theta_0}^{-1}\{Z_i(u)\} \int \frac{\partial \overline{\mu}_{\theta_0}}{\partial \theta}\{t, Z_i(u)\}\alpha(t;\theta_0) f\{Z_i(u), t\} y(t)\, dt \tag{30}$$

is the pointwise probability limit of $\partial \widehat{\mu}_{\theta_0}^*\{u, Z_i(u)\}/\partial\theta$. In fact, since

$$\frac{\partial \overline{\mu}_\theta}{\partial \theta}(u,z) = \frac{\partial \ln \alpha}{\partial \theta}(u;\theta) - \frac{\partial \ln e_\theta}{\partial \theta}(z), \qquad \frac{\partial e_\theta}{\partial \theta}(z) = \int \frac{\partial \alpha}{\partial \theta}(u;\theta) f(z,u) y(u)\, du,$$

we have, on substituting into (30) and using

$$\int \frac{\partial \alpha}{\partial \theta}(t;\theta_0) f\{Z_i(u), t\} y(t)\, dt = e_{\theta_0}^{-1} \frac{\partial e_{\theta_0}}{\partial \theta}\{Z_i(u)\} \int \alpha(t;\theta_0) f\{Z_i(u), t\} y(t)\, dt,$$

that $\partial \widehat{\mu}_{\theta_0}^*\{u, Z_i(u)\}/\partial\theta = 0$ and (28) holds immediately. The proof of (29) is very similar to that used in (25) above, and we omit the details. $\square$

That $\widehat{H}_{\widehat{\theta}\widehat{\theta}}$ consistently estimates $\mathscr{I}_0$ follows from the previous results (H1)–(H4) about uniform convergence of the Hessian. The weak convergence result follows by a standard application of the delta method. □

Acknowledgments. We thank the Danish Committee for Assessment of Substandard Lives, for providing the data, and Jette Sandqvist, for help with the computations. A computer program in Turbo Pascal is available from the first author upon request.

REFERENCES

AALEN, O. O. (1978). Nonparametric inference for a family of counting processes. *Ann. Statist.* **6** 701–726.

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.

BAHADUR, R. R. (1967). Rates of convergence of estimates and test statistics. *Ann. Math. Statist.* **38** 303–324.

BERAN, R. J. (1981). Nonparametric regression with randomly censored survival data. Technical report, Univ. California, Berkeley.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Inference in Semiparametric Models*. Johns Hopkins Univ. Press.

BICKEL, P. J. and WICHURA, M. J. (1971). Convergence criteria for multiparameter stochastic processes and some applications. *Ann. Math. Statist.* **42** 1656–1670.

BUCKLEY, J. and JAMES, I. (1979). Linear regression with censored data. *Biometrika* **66** 429–436.

CLAYTON, D. and Cuzick, J. (1985). Multivariate generalizations of the proportional hazard model (with discussion). *J. Roy. Statist. Soc. Ser. A* **148** 82–117.

COX, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34** 187–220.

COX, D. R. (1974). Partial likelihood. *Biometrika* **62** 269–276.

CRAMÉR, H. (1946). *Mathematical Methods in Statistics*. Princeton Univ. Press.

DABROWSKA, D. M. (1987). Non-parametric regression with censored survival time data. *Scand. J. Statist.* **14** 181–192.

DABROWSKA, D. M. (1989). Uniform consistency of the kernel conditional Kaplan–Meier estimator. *Ann. Statist.* **17** 1157–1167.

DABROWSKA, D. M. (1992). Variable bandwidth conditional Kaplan–Meier estimate. *Scand. J. Statist.* **19** 351–361.

DELLACHERIE, ?. and MEYER, ?. (1980). *Probabilities and Potential B*. North-Holland, Amsterdam.

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

GOMPERTZ, B. (1825). On the nature of the function expressive of the law of human mortality. *Philos. Trans. Roy. Soc. London*.

HÄRDLE, W., HART, J., MARRON, J. S. and TSYBAKOV, A. B. (1992). Bandwidth choice for average derivative estimation. *J. Amer. Statist. Assoc.* **87** 218–226.

HÄRDLE, W. and STOKER, T. (1989). Estimating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84** 986–995.

JEWELL, N. P. and NIELSEN, J. P. (1993). A framework for consistent prediction rules based on markers. *Biometrika* **80** 153–164.

JORDAN, C. W. (1975). *Life Contingencies*. The Society of Actuaries, Chicago.

KLEIN, R. W. and SPADY, R. H. (1991). An efficient semiparametric estimator for binary choice models. *Econometrica* **61** 387–421.

LIN, D. Y. and YING, Z. (1995). Semiparametric analysis of general additive–multiplicative hazard models for counting processes. *Ann. Statist.* **23** 1712–1734.

LINTON, O. B. (1995). Second order approximation in the partially linear regression model. *Econometrica* **63** 1079–1112.

LINTON, O. B. and NIELSEN, J. P. (1995). A marginal integration approach to estimating structured nonparametric regression. *Biometrika* **82** 93–101.

MAKEHAM, W. M. (1860). On the law of mortality, and the construction of mortality tables. *Journal of the Institute of Actuaries* **8**.

MCKEAGUE, I. W. and UTIKAL, K. J. (1991). Goodness of fit tests for additive hazards and proportional hazards models. *Scand. J. Statist.* **18** 177–195.

MESHALKIN, L. D. and KAGAN, A. R. (1972). Discussion of "Regression models and life tables," by D. R. Cox. *J. Roy. Statist. Soc. Ser. B* **34** 213.

NIELSEN, J. P. (1990). Kernel estimation of densities and hazards: a counting process approach. Ph.D. dissertation, Biostatistics, Univ. California, Berkeley.

NIELSEN, J. P. (1996). Multiplicative and additive marker dependent hazard estimation based on marginal integration. Unpublished manuscript, PFA Pension.

NIELSEN, J. P. and LINTON, O. B. (1995). Kernel estimation in a nonparametric marker dependent hazard model. *Ann. Statist.* **23** 1735–1748.

RAMLAU-HANSEN, H. (1983). Smoothing counting process intensities by means of kernel functions. *Ann. Statist.* **11** 453–466.

SANDQVIST, J. L. (1995). Invalidedødelighed—en semiparametrisk produktmodel. M.Sc. dissertation, Laboratory of Actuarial Science, Univ. Copenhagen.

SASIENI, P. (1992a). Non-orthogonal projections and their application to calculating the information in a partly linear Cox model. *Scand. J. Statist.* **19** 215–234.

SASIENI, P. (1992b). Information bounds for the conditional hazard ratio in a nested family of regression models. *J. Roy. Statist. Soc. Ser. B* **54** 617–635.

SCHICK, A. (1987). A note on the construction of asymptotically linear estimators. *J. Statist. Plann. Inference* **16** 89–105.

SHORACK, G. R. and WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics*. Wiley, New York.

SILVERMAN, B. W. (1978). Weak and strong uniform consistency of the kernel estimate of a density function and its derivatives. *Ann. Statist.* **6** 177–184.

WALD, A. (1949). A note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* **20** 595–601.

WELLNER, J. A. (1985). Semiparametric models: progress and problems. *Bull. Internat. Statist. Inst.* **51**(4) 23.1.1–23.1.20.

JENS P. NIELSEN
PFA PENSION
SUNDKROGSGADE 4
DK-2100 COPENHAGEN
DENMARK
E-MAIL: jp@pfa.dk

OLIVER LINTON
COWLES FOUNDATION FOR RESEARCH
IN ECONOMICS
YALE UNIVERSITY
30 HILLHOUSE AVENUE
NEW HAVEN, CONNECTICUT 06520-8281
E-MAIL: oliver.linton@yale.edu

PETER J. BICKEL
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720
E-MAIL: bickel@stat.berkeley.edu