A General Semiparametric Approach to Inference with Marker-Dependent Hazard Rate Models

Gerard.J. van den Berg[1], Lena Janys*[2], Enno Mammen[3], and Jens Perch Nielsen[4]

[1]*School of Economics, University of Bristol, IFAU, University of Groningen, IZA, ZEW and CEPR, The Priory Road Complex, Priory Road, Clifton, Bristol BS81TU, U.K., gerard.vandenberg@bristol.ac.uk*

[2]*Institute for Financial Economics and Statistics, University of Bonn, Adenauerallee 24-42, 53115 Bonn, Germany, ljanys@uni-bonn.de*

[3]*Institute for Applied Mathematics, Heidelberg University, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany, mammen@math.uni-heidelberg.de*

[4] *Cass Business School, City, University of London, 106 Bunhill Row, London EC1Y8TZ, U.K., jens.nielsen.1@city.ac.uk*

June 19, 2019

Abstract

We examine a new general class of hazard rate models for duration data, containing a parametric and a nonparametric component. Both can be a mix of a time effect and possibly time-dependent covariate effects. A number of well-known models are special cases. In a counting process framework, a general profile likelihood estimator is developed and the parametric component of the model is shown to be asymptotically normal and efficient. Finite sample properties are investigated in simulations. The estimator is applied to investigate the long-run relationship between birth weight and later-life mortality.

Keywords. Covariate effects; duration analysis; kernel estimation; mortality; semiparametric estimation.

JEL-Codes. C14, C41.

*Corresponding author

1 Introduction

The analysis of duration data using large samples is widespread in economics, actuarial science and finance, and also in biostatistics and engineering. In each of these fields it is of primary concern that the model to be estimated is not unduly restrictive. Semiparametric models provide a balance between flexibility and limited dimensionality. The most common semiparametric model is the Cox proportional hazard model for the hazard rate $\lambda(t)$,

$$\lambda(t) = \exp\{\beta' W(t))\}\alpha(t) \tag{1}$$

in which the covariate (or "marker") effects β are the parameters of interest and the dependence $\alpha(t)$ of the hazard rate on the elapsed duration or time is unspecified; see Cox (1972). The partial likelihood estimator of the parameter β does not depend on the functional form of α.

However, the estimator has some significant disadvantages. It requires the assumption that the covariates $W(t)$ affect the hazard rate by way of the parametric functional form $\exp\{\beta' W(t)\}$ and it does not include unobserved heterogeneity. The recent econometric literature has focused on the latter problem with major contributions from for example Bijwaard et al. (2013) and Wolter (2016), who provide extensions to the Cox model to accommodate unobserved heterogeneity, and Hausman and Woutersen (2014) who propose a semiparametric estimator for discrete duration data. In contrast, in this paper we focus on relaxing the assumptions associated with the covariate function and we concentrate on continuous data.

Perhaps ironically, there is often more prior knowledge or consensus about how the hazard rate varies with the elapsed duration or time t than about how it varies with the covariates. For example, in the study of adult mortality, it is natural to model the effect of age t on the mortality rate by way of the Gompertz specification $\exp(\theta t)$ (or by minor generalizations of it) especially if relatively homogeneous sub-populations are considered and extreme ages are not taken into consideration; see e.g. Wetterstrand (1981) and Gavrilov and Gavrilova (1991). At the same time, there is no well-established functional form for the dependence of the mortality rate on socio-economic class, level of education, and so on. Empirical studies sometimes discretize individual characteristics into a few categories and estimate effects of corresponding binary indicators using model (1), see e.g. Osler et al. (2003). If the true mortality rate is a smooth function of the individual characteristics then the estimated effects may be biased.

Other examples are provided by the literature on unemployment durations and job durations. Theoretical models based on job search theory make precise predictions on how individual hazard rates of the unemployment and job duration distributions depend on the timing of external events and on labor market fluctuations; see van den Berg (2001). This provides functional forms for the time-varying profile of these hazard rates. Similarly, theoretical models based on learning theory predict that the hazard rate of the job duration distribution depends on tenure in a specific inverse-gaussian fashion (Jovanovic, 1984). Conversely, it is more difficult to acquire theoretical guidance on how individual characteristics such as work experience and job complexity affect the hazard rates. Robust empirical guidance for how the unemployment duration hazard depends on the time spent in unemployment is provided by Hausman and Woutersen (2014)'s application of their flexible semiparametric estimator. Using US data they demonstrate that the dependence on time in unemployment is sufficiently regular for simple functional forms to capture $\alpha(t)$ over wide duration intervals.

If the functional form of the effects of the covariates $W(t)$ on the hazard rate is unknown then the partial likelihood method used for the estimation of model (1) does not apply. In the current paper we propose a general semiparametric model that does not specify the functional form of covariate effects on the hazard rate, and we develop an estimation method for this model. The model has the form

$$\lambda(t) = \alpha\{X(t); \theta\} g\{Z(t)\} \tag{2}$$

Here, $g(\cdot)$ is unspecified while $\alpha(\cdot; \theta)_{\theta \in \Theta}$ is a parametric class of functions. The vectors $X(t)$ and $Z(t)$ are covariate or marker processes, and their elements may include the elapsed duration (or time) t. Of course, if X and Z are time-invariant then λ is simply the conditional density of t given the covariates, divided by the conditional survivor function. We show that this model has many existing semiparametric models as special cases. Note that it also includes nonproportional hazard rate models. In applications, the researcher may be particularly interested in the function g, for example if $Z(t)$ includes a policy instrument or treatment regime or if it includes a marker used to predict future outcomes. However, in other applications θ may be the parameter of interest. In that case, if the functional form of g is unknown, the estimation of a model that assumes an incorrect functional form for g may result in biased estimates of θ.

The estimator that we develop is a three-step profile likelihood estimator inspired by a related approach by Nielsen et al. (1998) for a more restrictive semiparametric model. In our first stage, we estimate g best possible under the assumption that θ is actually known. In the second stage, we use this estimator $\widehat{g}_\theta$ of g in a profile likelihood, recognizing that the stochastic hazard $\alpha\{X(t); \theta\}\widehat{g}_\theta\{Z(t)\}$ has a parametric specification family of hazards, enabling the application of standard maximum likelihood methodology; see Borgan (1984). In the third estimation stage, we estimate g by $\widehat{g}_{\widehat{\theta}}$ using local linear kernel hazard regression. A major methodological contribution of our paper is that we improve on the asymptotic analysis in the existing literature for semiparametric hazard rate inference by using the improved asymptotic approximation theory of counting process martingales developed in Mammen and Nielsen (2007). In effect, our estimator of

θ is square-root-n consistent, asymptotically normal and efficient.[1] In addition to these desirable theoretical properties, the estimator is straightforward to use and is applicable as-is to important estimation problems, which is not universally true for flexible semiparametric estimators. We provide some simulations for further guidance.

We apply our newly devised estimation method to the study of the effect of birth weight on longevity. Longevity is an important economic variable, as it plays a role in savings decisions, pension and health insurance, and the costs and benefits of medical interventions. Recently, the interest in effects of conditions in utero on high-age health has been growing. It has been shown that a range of diseases and death causes at high ages have "developmental origins", i.e. can be affected by conditions in utero. The latter conditions are often summarized by birth weight. (See overviews in Poulter et al., 1999, Rasmussen, 2001, Kuh and Ben-Shlomo, 2004, Davey Smith, 2005, Huxley et al., 2007, and Almond and Currie, 2011.) Studies in this literature use simple parametric specifications for the effect of birth weight. For example, Leon et al. (1998) distinguish between four intervals for birth weight in its effect on mortality due to ischaemic heart disease. Others simply use a binary indicator for whether birth weight is "low" , i.e., below 2500 grams, or not. Alternatively, a linear relation is postulated between log birth weight and the log of the rate of the occurrence of some adverse health outcome.

Such parametric functional forms may be problematic. The continuity of the underlying biological mechanisms implies that effects of discretized birth weight indicators provide biased estimates of effects at specific birth weights. If medical protocols postulate interventions that condition on birth weight then the benefits of the intervention depend on the accuracy with which the relation between birth weight and outcome is es-

[1] There has been an increasing interest in efficient estimation of duration models, see for example Bearse, Canals-Cerdá and Rilstone (2007) and Rezat and Rilstone (2015) for discrete duration data. Hausman and Woutersen (2008) provide an overview. Ridder and Woutersen (2003) examine models with unobserved heterogeneity and provide conditions under which the information matrix is nonsingular.

timated. Note also that birth weight effects on mortality are plausibly non-monotonous. For example, Ahlgren et al. (2007) demonstrate positive associations between birth weight and the rates of almost any type of cancer at higher ages.

This calls for a semiparametric approach in which the long-run effect of birth weight on mortality is not restricted by a parametric functional form. Our method is particularly well-suited for this because of the consensus about the functional form for the dependence of the mortality rate on the current age, for ages up to 90. Specifically, we may adopt the Gompertz functional form for this. It is well known that the parameters of this functional form vary by gender and socio-economic class (see references above). Our method can deal with this, as well as with variation of the birth weight effect by these personal characteristics.

Clearly, the application requires data of individuals born many decades ago, for whom birth weight and age at death are recorded with high accuracy. We use the Uppsala Birth Cohort Study, UBCoS, which is a lifelong follow-up study of a representative sample of individuals born in 1915–1929. Upon birth, the birth weight was recorded in grams by qualified nurses. The data set contains additional information registered at birth, notably the socio-economic characteristics of the parental household. We conjecture that this data set provides the best data in the world to relate birth weight and high-age mortality.

The paper is organized as follows. Section 2 presents our semiparametric model and explains how it contains models in the literature as special cases. In Section 3 we introduce the counting process formulation of our model. In Section 4 we define the estimators for the parameter θ and the nonparametric function g. In Section 5 we introduce the asymptotic distribution theory. In Section 6 we derive the local linear version of our estimator g and show the simulation results for the local constant and the local linear estimator to assess their performance under different bandwidth selection techniques. Section 7 contains the empirical application. Section 8 concludes.

2 The semiparametric model

This section presents the semiparametric model and explains how it contains models in the literature as special cases. Our model has the stochastic hazard rate

$$\lambda(t) = \alpha\{X(t);\theta\}g\{Z(t)\}. \tag{3}$$

Here, $\alpha(\cdot;\theta)_{\theta\in\Theta}$ is a parametric class of functions whereas $g(\cdot)$ is unspecified apart from smoothness assumptions to be discussed below. Obviously, α and g must be nonnegative. The vectors $X(t)$ and $Z(t)$ are covariate or marker processes with dimensions d_x and d_z, respectively. For sake of exposition, we take $d_x \geq 1$ and $d_z \geq 1$. Note that $d_z = 0$ leads to a fully parametric model while $d_x = 0$ leads to a fully nonparametric model. The elements of $X(t)$ and $Z(t)$ may include the elapsed duration or time t. The elements of the vector $X(t)$ can be discretely or continuously distributed. Concerning the elements of $Z(t)$, for obvious reasons, we restrict attention to continuously distributed variables. We discuss exogeneity requirements on the covariate processes below.

Example 1: The Cox model with a time-varying covariate process is obtained as a special case by taking $Z(t) := t$ and $\alpha\{X(t);\theta\} := \exp\{\theta' X(t)\}$. In this setting, g is the baseline hazard capturing duration dependence of the hazard while α is the so-called systematic part of the hazard.

Example 2: The Stratified Cox model (Kalbfleisch and Prentice, 1980) extends the Cox model by allowing strata to have different baseline hazards. This can be captured in our model by specifying $Z(t) := (W, t)$ with W being discrete and finite, and $\alpha\{X(t);\theta\} := \exp\{\theta' X(t)\}$. Here, different values of W capture different strata.

Example 3: Nielsen, Linton and Bickel (1998) consider a model with $X(t) := t$ and in which $Z(t)$ has only one element,

$$\lambda(t) = \alpha(t;\theta)g\{Z(t)\}, \tag{4}$$

Like (3), this model does not impose a functional form on the covariate effect. However, it is more restrictive in that it does not allow the time effect $\alpha(t;\theta)$ to depend on individual characteristics, and it only deals with one covariate Z. In general, in duration analysis, it is advisable to include all relevant observed covariates in the model, to prevent bias due to omitted unobserved heterogeneity; see the overview in van den Berg (2001).

Example 4: Dabrowska (1997) considers a model that can be expressed as

$$\lambda(t) = \exp\{\theta' X(t)\}g\{Z(t)\} \tag{5}$$

in the same notation as above. This is a special case of (3) because it assumes a specific functional form for the function α.[2]

Our general model lends itself to other interesting specifications, for example,

$$\lambda(t) = \alpha(t;\theta)g_\beta\{Z(t)\} \tag{6}$$

where g_β is a parametric function that does not necessarily satisfy $g_\beta\{Z(t)\} = \exp\{\beta' Z(t)\}$. One could for example imagine instead that $g_\beta\{Z(t\}) = \beta' Z(t)$.

[2] A number of other models have been proposed in the literature. For example, Linton, Nielsen and van de Geer (2003) study a model that is a hybrid between a semiparametric and nonparametric model; it assumes that the stochastic hazard is a multiplicative or additive function of unspecified functions of single elements of $Z(t)$. In this paper we do not consider that model. Neither do we consider semiparametric transformation models for duration data, since these are difficult to interpret in terms of hazard rate properties. See Dabrowska (2006) for an example of an estimator for such a model. Towards the end of the paper we briefly discuss semiparametric models with single-index structures for the dependence of the hazard rate on $Z(t)$.

In general, the inclusion of t as an element of $X(t)$ and/or $Z(t)$ allows for non-proportional hazard specifications, that is, specifications where the hazard effects of t on the one hand and the covariates on the other are not multiplicative. Allowing for non-proportionality is useful, as proportionality is often hard to justify. For example, in the study of mortality, where it is natural to model the parametric effect of age t on the hazard by way of $\exp(\theta t)$, the coefficient θ may vary with individual characteristics. In the study of unemployment durations, the hazard rate of interest is the transition rate out of unemployment into employment. Economic-theoretical models predict that the decrease of this rate with the elapsed unemployment duration is stronger if aggregate labor market conditions are unfavorable (Blanchard and Diamond, 1994) or if the difference between the unemployment insurance level and the welfare level is large (van den Berg, 1990, 2001). We should emphasize that our model does not rule out that $X(t)$ and $Z(t)$ have common elements. This has particular statistical implications to which we turn in subsequent sections.

Semiparametric models are typically developed in conjunction with estimation methods tailored to the model. The Cox model and the corresponding partial likelihood estimation method are a case in point. It is useful to discuss some key properties of the estimators developed for the semiparametric models of Nielsen et al. (1998) and Dabrowska (1997) and other models and contrast them with properties of the estimator developed in our paper. Nielsen et al. (1998) show that their estimator of θ in (4) is efficient. This estimator has two stages. In the first stage, they estimate g best possible under the assumption that θ is actually known. In the second stage, they use this estimator $\widehat{g}_\theta$ of g in a profile likelihood, recognizing that the stochastic hazard $\widehat{\lambda}(t) = \alpha_\theta(t)\widehat{g}_\theta\{Z(t)\}$ has a parametric specification family of hazards, enabling the application of standard maximum likelihood methodology. Our estimator generalizes this. Dabrowska (1997) proves asymptotic square-root-n consistency and asymptotic normality of her estimator of θ. However, she does not achieve efficiency as we do with our approach.

We end this section by mentioning fully nonparametric approaches to statistical inference as an alternative approach to semiparametric inference. Our estimator for the function g will be inspired by the nonparametric estimators developed in Nielsen and Linton (1995) and Nielsen (1998). These studies develop local constant and local linear kernel hazard estimators, respectively, for a model framework where the stochastic hazard is fully unspecified as a function of a vector $Z(t)$ which may include t. As methods for statistical inference on hazard rates, such estimators have the disadvantage that they suffer from the curse of dimensionality. Of course, this also applies to other estimators for nonparametric duration models, such as the estimators of Dabrowska (1987) and Spierdijk (2008).

3 Counting process formulation of the model

We follow the counting process formulations of e.g. Mammen and Nielsen (2007) and restrict ourselves to an independent identically distributed sampling and one-jump counting process case. Let $N(t) = (N_1(t), ..., N_n(t))$ be an n-dimensional collection of n one-jump counting processes with respect to an increasing, right-continuous, complete filtration $(\mathcal{F}_t : t \in [0, T])$. Specifically, N is adapted to the filtration and has components N_i taking values in $\{0, 1\}$, indicating, by the value 1, whether or not an observed jump has been registered for the i th individual. The N_i's are right-continuous step functions, zero at time zero. The variable $N_i(t)$ is defined over the whole period $[0, T]$, where T is finite. Suppose that N_i has predictable intensity, see Andersen et al. (1993),

$$\lambda_i(t)dt = E\{dN_i(t)|\mathcal{F}_{t-}\} = \alpha\{X_i(t); \theta_0\}g\{Z_i(t)\}Y_i(t)dt \qquad (7)$$

where Y_i is a predictable process taking values in $\{0, 1\}$ indicating, by the value 1, when the *ith* individual is at risk, whereas X_i is a d_x dimensional and Z_i a d_z dimensional predictable covariate process with support in some compact set $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ and $\mathcal{Z} \subseteq \mathbb{R}^{d_z}$, respectively.

We assume that the stochastic processes $(N_1, X_1, Z_1, Y_1), ..., (N_n, X_n, Z_n, Y_n)$ are independent and identically distributed for the n individuals. Let

$$\mathcal{F}_{t,i} = \sigma\{N_i(u), X_i(u), Y_i(u), Z_i(u); u \leqslant t\} \quad \text{and} \quad \mathcal{F}_t = \vee_{i=1}^n \mathcal{F}_{t,i}.$$

It follows that λ_i is predictable with respect to $\mathcal{F}_{t,i}$ and hence $\mathcal{F}_t$, and the processes $M_i(t) = N_i(t) - \Lambda_i(t)$, $i = 1, ..., n$, with compensators $\Lambda_i(t) = \int_0^t \lambda_i(u)du$, are square integrable martingales with respect to $\mathcal{F}_{t,i}$ on the time interval $[0, T]$. Hence, $\Lambda_i(t)$ is the compensator of $N_i(t)$ with respect to both the filtration $\mathcal{F}_{t,i}$ and the filtration $\mathcal{F}_t$.

4 Definition of the estimators of θ and g

4.1 Three-step approach

We use a semiparametric profile likelihood estimation method in three steps.

Step (i). The nonparametric function g is estimated via a Nadaraya-Watson type estimator (to be explained in Subsection 4.2) under the assumption that the true parameter θ is known. This estimator of g depends on θ and on a smoothing parameter b. We make use of a leave-one-out version denoted by $\widehat{g}_{b,\theta,-i}(z)$ if the i-th observation is left out.

Step (ii). We derive the likelihood function for the observable data assuming that the true g is known. The parameter θ is now estimated from the pseudo-likelihood that arises when g is replaced by $\widehat{g}_\theta(z)$. This estimator depends on a bandwidth b and we therefore denote the estimator by $\widehat{\theta}_b$. The leave-one-out version of the estimator is denoted by $\widehat{\theta}_{b,-i}$.

Step (iii). The final estimator of g is now calculated by assuming that $\widehat{\theta}_b$ is the true parameter and by using kernel smoothing using a bandwidth b^*. Therefore, the final estimator of g is of the form $\widehat{g}_{b^*,\widehat{\theta}_b}(z)$. The leave-one-out version is denoted by $\widehat{g}_{b^*,\widehat{\theta}_{b,-i},-i}(z)$.

The two bandwidth vectors b and b^* should not be chosen to be identical. In order to obtain an asymptotically unbiased estimator of θ we need an undersmoothing bandwidth b. Thus b should be of smaller order than b^*. In our empirical application, we will choose the tuple (b, b^*) jointly data-adaptively such that the following overall cross-validation criterion is minimized:

$$Q_{CV}(b, b^*) = n^{-1} \left[\sum_{i=1}^{n} \int \widehat{a}^2_{-i}\{X_i(s), Z_i(s)\} Y_i(s) ds - 2 \sum_{i=1}^{n} \int \widehat{a}_{-i}\{X_i(s), Z_i(s)\} dN_i(s) \right], \tag{8}$$

where

$$\widehat{a}_{-i}(x, z) = \alpha(x, \widehat{\theta}_{b,-i}) \widehat{g}_{b^*,\widehat{\theta}_{b,-i},-i}(z),$$

see also Linton and Nielsen (1995) for a similar criterion for the choice of one bandwidth vector. Our introduction of double cross-validation is very flexible in the sense that it provides the smoothing or the parametric part that is best from the point of view of a global goodness of fit. And getting the best global fit is not necessarily the same as getting the best possible parametric estimator. We see the full advantage of this flexibility in our misspecification study illustrated in Figure 3. The double-cross-validation approach provides us with that parametric value that is best for the following nonparametric minimization.

4.2 Definition of $\widehat{g}_\theta$

As the Nadaraya-Watson type estimator of g in Step (i) we may take a local constant estimator. The approach immediately generalizes to the notationally slightly more burdensome local linear approach. The finite sample analyses in our paper illustrates that the local linear methodology performs on average better in practice than the local constant approach.

In this subsection we present the local constant estimator of the nonparametric function g. For any value of θ, we use the following leave-one-out procedure:

$$\widehat{g}_{b,\theta,-i}(z)=\frac{\sum_{j\neq i}\int K_b\{z-Z_j(u)\}dN_j(u)}{\sum_{j\neq i}\int K_b\{z-Z_j(u)\}\alpha\{X_j(u);\theta\}Y_j(u)du}, \tag{9}$$

where K is a multivariate kernel function with $K_b(\cdot)=b_{prod}^{-1}K(B^{-1}\cdot)$ for any multivariate $b=(b_1^0,...,b_{d_z}^0)^T$. Here, B is the diagonal matrix with diagonal entries $b_1^0,...,b_{d_z}^0$ and $b_{prod}=b_1^0\cdot...\cdot b_{d_z}^0$. We will not always indicate dependence on the bandwidth b and write $\widehat{g}_{\theta,-i}(z)$ instead of $\widehat{g}_{b,\theta,-i}(z)$. Under our regularity conditions, we have that $\widehat{g}_{\theta,-i}(z)-\widehat{g}_{\theta}(z)=o_P(1)$, uniformly in θ, i and z, where

$$\widehat{g}_{\theta}(z)=\widehat{g}_{b,\theta}(z)=\frac{\sum_{j=1}^n\int K_b\{z-Z_j(u)\}dN_j(u)}{\sum_{j=1}^n\int K_b\{z-Z_j(u)\}\alpha\{X_j(u);\theta\}Y_j(u)du}. \tag{10}$$

For z that lie in a neighbourhood of the boundary we replace the kernel K_b by a boundary kernel $K_{z,b}(u)$, see Assumption (A3). This is not indicated in the notation.

Furthermore, $\widehat{g}_{\theta_0}$ consistently estimates $g(z)$ (see Nielsen and Linton, 1995), and, away from the true parameter value,

$$\widehat{g}_{\theta}(z)\rightarrow_p g_{\theta}(z)\equiv\frac{g(z)e_{\theta_0}(z)}{e_{\theta}(z)}, \tag{11}$$

where $e_{\theta}(z)=\int\alpha(x;\theta)f_u(x,z)y(u)du\;dx$ with $y(u)=\text{pr}(Y_i(u)=1)$. Let

$$g^*_{\theta,-i}(z)=\frac{\sum_{j\neq i}\int K_b\{z-Z_j(u)\}\lambda_j(u)du}{\sum_{j\neq i}\int K_b\{z-Z_j(u)\}\alpha\{X_j(u);\theta\}Y_j(u)du} \tag{12}$$

and note that

$$\widehat{g}_{\theta,-i}(z) - g^*_{\theta,-i}(z) = \frac{\sum_{j\neq i}\int K_b\{z - Z_j(u)\}dM_j(u)}{\sum_{j\neq i}\int K_b\{z - Z_j(u)\}\alpha\{X_j(u);\theta\}Y_j(u)du}. \tag{13}$$

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

4.3 Definition of $\widehat{\theta}$

In this subsection we present the expression for the estimator $\widehat{\theta}$ of the parameter θ. Conditional on Y, X and Z, the standard 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 $g(z)$ were known, we would maximize the following likelihood function over θ

$$\ell(\theta) = \sum_{i=1}^n \int \mu_\theta\{X_i(u), Z_i(u)\}dN_i(u) - \sum_{i=1}^n \int \exp[\mu_\theta\{X_i(u), Z_i(u)\}]Y_i(u)du \tag{14}$$

where $\mu_\theta(x,z) = \ln\{\alpha(x;\theta)g(z)\}$ is the logarithmic hazard. Consequently, the maximum likelihood estimator $\widehat{\theta}_g$ for θ given known g is given by $\arg \underset{\theta}{\max} \ell(\theta)$.

Since $g(z)$ is not known, we substitute $\hat{\mu}_{\theta,-i}(x,z)$ for $\mu_\theta(x,z)$ where $\hat{\mu}_{\theta,-i}(x,z) = \ln\{\alpha(x;\theta)\widehat{g}_{\theta,-i}(z)\}$:

$$\hat{\ell}(\theta) = \sum_{i=1}^n \int \hat{\mu}_{\theta,-i}\{X_i(u), Z_i(u)\}dN_i(u) - \sum_{i=1}^n \int \exp[\hat{\mu}_{\theta,-i}\{X_i(u), Z_i(u)\}]Y_i(u)du. \tag{15}$$

The pseudo-maximum likelihood estimator $\widehat{\theta}$ is defined as

$$\widehat{\theta} = \arg \max_{\theta\in\mathcal{N}_0} \hat{\ell}(\theta). \tag{16}$$

Here, $\mathcal{N}_0$ is a fixed compact subset of Θ having θ_0 as an interior point.

4.4 Model identification

As shown in the next section, under regularity assumptions, a consistent nonparametric estimator of the hazard function $a(x,z) = \alpha(x,\theta)g(z)$ can be constructed. We now discuss the question if this implies that the function g and the parameter θ are identified. We argue that, in general, this is indeed the case if $X_i(t)$ and $Z_i(t)$ have no common elements. Suppose that the support of $X_i(t)$ and $Z_i(t)$ does not depend on t and that the joint support is equal to the product of the marginal supports $\mathcal{X}$ and $\mathcal{Z}$. Then $a(x,z)$ is identified for $x \in \mathcal{X}$ and $z \in \mathcal{Z}$. Thus $\int_{\mathcal{Z}} a(x,z)\mathrm{d}z$ identifies $\alpha(x,\theta)$ up to a multiplicative factor. Suppose now additionally that the parametrization of α is chosen such that the ratio $\alpha(x,\theta_1)/\alpha(x,\theta_2)$ is non-constant for all parameters $\theta_1 \neq \theta_2$. Then we have that the function $\alpha(x,\theta)$ and, in particular, if the map $\theta \to \alpha(\cdot,\theta)$ is invertible, the parameter θ is identified. We also get that $g(z) = a(x,z)/\alpha(x,\theta)$ is identified. This discussion also applies if one of the two covariate vectors, $X_i(t)$ or $Z_i(t)$, contains time t as an element.

The situation changes if $X_i(t)$ and $Z_i(t)$ have common elements. Let us consider the case that both covariate vectors have t as a common factor and, in abuse of notation, let us write the model as $a(x,z,t) = \alpha(x,t,\theta)g(z,t)$ where $x \in \mathcal{X}$, $z \in \mathcal{Z}$ and $0 \leq t \leq T$. Here the function g and the parameter θ is identified if for each pair of parameters $\theta_1 \neq \theta_2$ there exists a value of t such that $x \to \alpha(x,t,\theta_1)/\alpha(x,t,\theta_2)$ is non-constant in x. But, identification relies here strongly on the chosen parametric model for α which down-weighs the importance of this fact. An illustrative simple example where g and θ are not identified is given by models of the form $a(t) = \alpha(t,\theta)g(t)$. Trivially, in this model θ and g are not identified. But, one could search for the value of θ that minimizes a global error criterion for the estimation of the product $a(t) = \alpha(t,\theta)g(t)$. We will come back to the discussion of non-identified models in our simulations where we will consider the model $a(t,z) = t^{\theta-1}(1-t)z(1-z)$. It turns out that our estimation procedure outlined in this section with using the adaptive bandwidth selector (8) leads to a much improved estimation compared to estimators that choose a fixed value of θ. Here, our approach is related to proposals where first a parametric model is fitted and then in a second step the parametric fit is improved by a nonparametric estimator, see e.g. Hjort and Glad (1995) and Hjort and Jones (1996). But there is an essential difference to our approach. We are searching for the parametric fit that leads to the best two-step procedure of a. For this purpose the parametric fit has to be adapted to the chosen nonparametric procedure of the second step. We conjecture that this is achieved by our data adaptive bandwidth selector (8).

5 Asymptotic results

In this section we derive asymptotic results for our estimator for the identified case, i.e. for the case where $X_i(t)$ and $Z_i(t)$ have no common elements. The results will be for deterministic bandwidths.

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

$$\overline{\ell}(\theta) = \sum_{i=1}^{n} \int \overline{\mu}_\theta\{X_i(u), Z_i(u)\} dN_i(u) - \sum_{i=1}^{n} \int \exp[\overline{\mu}_\theta\{X_i(u), Z_i(u)\}] Y_i(u) du \qquad (17)$$

with $\overline{\mu}_\theta(x,z) = \ln\{\alpha(x,\theta)g_\theta(z)\}$. We show in the appendix that $\overline{Q}_n(\theta)$ approaches

$$Q(\theta) = \int\int \left[\ln\left\{\frac{\alpha(x;\theta)e_{\theta_0}(z)}{\alpha(x;\theta_0)e_\theta(z)}\right\} - \frac{\alpha(x,\theta)e_{\theta_0}(z)}{\alpha(x;\theta_0)e_\theta(z)} + 1\right] \alpha(x;\theta) f_u(x,z) y(u) du\ dz, \quad (18)$$

in probability, uniformly over any compact neighborhood of θ_0. This implies consistency of $\widehat{\theta}$. $f_u(x,z)$ is the conditional multivariate density of $\{X_i(u), Z_i(u)\}$ given that $Y_i(u)$ is equal to one. We allow $f_u(x,z)$ to have one possible Dirac element such that up to one of the d_x elements of the covariate $X_i(u)$ is identically equal to time u or vice versa for up to one of the elements of $Z_i(u)$.

In a next step we show asymptotic normality of $\widehat{\theta}$. The score vector $\hat{s}_\theta$ and the Hessian matrix $\hat{H}_{\theta\theta}$ are defined as the first and second derivatives of the pseudo-likelihood $\hat{\ell}$ standardized by sample size:

$$\begin{aligned} \hat{s}_\theta(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\int \frac{\partial \hat{\mu}_{\theta,-i}}{\partial\theta}\{X_i(u), Z_i(u)\}dN_i(u) \qquad (19)\\ &\quad -\frac{1}{n}\sum_{i=1}^{n}\int \frac{\partial \hat{\mu}_{\theta,-i}}{\partial\theta}\{X_i(u), Z_i(u)\}\alpha\{X_i(u);\theta\}\widehat{g}_{\theta,-i}\{Z_i(u)\}Y_i(u)du,\\ \hat{H}_{\theta\theta}(\theta) &= n^{-1}\sum_{i=1}^{n}\int \frac{\partial^2 \hat{\mu}_{\theta,-i}}{\partial\theta\partial\theta^T}\{X_i(u), Z_i(u)\}dN_i(u)\\ &\quad -n^{-1}\sum_{i=1}^{n}\int \left(\frac{\partial^2 \hat{\mu}_{\theta,-i}}{\partial\theta\partial\theta^T} + \frac{\partial \hat{\mu}_{\theta,-i}}{\partial\theta}\frac{\partial \hat{\mu}_{\theta,-i}}{\partial\theta^T}\right)\{X_i(u), Z_i(u)\}\alpha\{X_i(u);\theta\}\widehat{g}_{\theta,-i}\{Z_i(u)\}Y_i(u)du. \end{aligned}$$

By the mean value theorem

$$0 = n^{1/2}\hat{s}_\theta(\theta_0) + \hat{H}_{\theta\theta}(\breve{\theta})n^{1/2}(\widehat{\theta} - \theta_0), \qquad (20)$$

where $\breve{\theta}$ lies between θ_0 and $\widehat{\theta}$. We first analyze the pseudoscore vector evaluated at the true θ_0, using (19) with $\theta = \theta_0$:

$$\begin{aligned} \hat{s}_\theta(\theta_0) &= \frac{1}{n}\sum_{i=1}^{n}\int \frac{\partial \hat{\mu}_{\theta_0,-i}}{\partial\theta}\{X_i(u), Z_i(u)\}dM_i(u) + \frac{1}{n}\sum_{i=1}^{n}\int \frac{\partial \hat{\mu}_{\theta_0,-i}}{\partial\theta}\{X_i(u), Z_i(u)\}d\Lambda_i(u)\\ &\quad -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \hat{\mu}_{\theta_0,-i}}{\partial\theta}\{X_i(u), Z_i(u)\}\alpha\{X_i(u);\theta_0\}g\{Z_i(u)\}Y_i(u)du \qquad (21)\\ &\quad -\frac{1}{n}\sum_{i=1}^{n}\int \frac{\partial \hat{\mu}_{\theta_0,-i}}{\partial\theta}\{X_i(u), Z_i(u)\}\alpha\{X_i(u);\theta_0\}\left[\widehat{g}_{\theta_0,-i}\{Z_i(u)\} - g\{Z_i(u)\}\right]Y_i(u)du. \end{aligned}$$

Here we have substituted N by $M+\Lambda$ and $\widehat{g}_{\theta_0,-i}$ by $g+\widehat{g}_{\theta_0,-i}-g$. By the definition of Λ_i, we find that the second and third term on the right hand side of (21) cancel. We then break $\widehat{g}_{\theta_0,-i}-g$ into stable and variable terms. Using the decomposition (13), we find, after interchanging the order of summation and integration, that

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

where

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

Now substitute $\partial\overline{\mu}_{\theta_0}/\partial\theta+\partial\ln\widehat{g}_{\theta_0,-i}/\partial\theta-\partial\ln g_{\theta_0,-i}/\partial\theta$ for $\partial\hat{\mu}_{\theta_0,-i}/\partial\theta$ in the first term on the right hand side of (21). Collecting everything together we obtain that

$$\begin{aligned}
\hat{s}_\theta(\theta_0) \ = \ & n^{-1}\sum_{i=1}^{n}\int\frac{\partial\overline{\mu}_{\theta_0}}{\partial\theta}\{X_i(u),Z_i(u)\}dM_i(u) \qquad (22)\\
&-n^{-1}\sum_{i=1}^{n}\int\frac{\partial\hat{\mu}^*_{\theta_0,-i}}{\partial\theta}\{Z_i(u)\}dM_i(u)\\
&+n^{-1}\sum_{i=1}^{n}\int\left\{\frac{\partial\ln\widehat{g}_{\theta_0,-i}}{\partial\theta}-\frac{\partial\ln g_{\theta_0}}{\partial\theta}\right\}\{X_i(u)\}dM_i(u)\\
&-n^{-1}\sum_{i=1}^{n}\int\frac{\partial\hat{\mu}_{\theta_0,-i}}{\partial\theta}\{X_i(u),Z_i(u)\}\alpha\{X_i(u);\theta_0\}\{g^*_{\theta_0,-i}-g\}\{Z_i(u)\}Y_i(u)du.
\end{aligned}$$

We have written $\hat{s}_\theta$ as a sum of four terms: the last term is a stochastic average of $g^*_{\theta_0,-i}-g$ that arises from the bias obtained in the estimation of g: it is asymptotically negligible if a sufficiently small bandwidth is chosen. Undersmoothing is necessary in many semiparametric estimation problems; see Bickel et al. (1993) for a discussion. In the appendix we show that the second and third term on the right hand side of (22) are also $o_p(n^{-1/2})$. Because the integrands converge to zero in probability, this would immediately follow if the integrands are predictable. But the latter is not the case, and therefore the formal proof is more complicated, see the appendix. The proof makes use of the approach to the predictability issue developed in Mammen and Nielsen (2007). We have that

$$\begin{aligned} n^{1/2}\hat{s}_\theta(\theta_0) &= n^{1/2}s^e_\theta(\theta_0) + o_p(1), \text{ where} \\ s^e_\theta(\theta_0) &= n^{-1}\sum_{i=1}^n \int \frac{\partial \overline{\mu}_{\theta_0}}{\partial\theta}\{X_i(u), Z_i(u)\}dM_i(u). \end{aligned} \tag{23}$$

since $\partial \ln \overline{\mu}_\theta\{X_i(u), Z_i(u)\}/\partial\theta$ is a predictable process, we can apply Rebolledo's martingale central limit theorem to $s^e_\theta(\theta_0)$ and we get that

$$n^{1/2}s^e_\theta(\theta_0) \to N(0, \mathcal{I}_0), \text{ in distribution,} \tag{24}$$

where

$$\mathcal{I}_0 = \int\int \frac{\partial\overline{\mu}_{\theta_0}}{\partial\theta}\frac{\partial\overline{\mu}_{\theta_0}}{\partial\theta^T}(x,z)\alpha(x,\theta_0)g(z)f_u(x,z)y(u)du\ dz$$

with

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

In the appendix, we also show that the Hessian matrix $\hat{H}_{\theta\theta}(\theta)$ satisfies

$$\sup_{\theta\in\mathcal{N}_n} |\hat{H}_{\theta\theta}(\theta) - \mathcal{I}_0| \to_p 0, \tag{25}$$

for $\mathcal{N}_n = \{\theta : |\theta - \theta_0| \le \delta_n\}\delta_n \to 0$ is a shrinking neighborhood of θ_0. In conclusion, we get from (20), (23), (24) and (25) that $n^{1/2}(\widehat{\theta} - \theta_0) \to N(0, \mathcal{I}_0^{-1})$, in distribution.

Theorem 1 summarizes our discussion. Its proof is in the appendix. It makes use of the following assumptions:

(A1) For $0 \leq t \leq 1$ it holds that $\mathrm{pr}\{Z_i(t) \in \mathcal{Z}\} = 1$ and $\mathrm{pr}\{X_i(t) \in \mathcal{X}_1 \times \mathcal{X}_2\} = 1$ for compact subsets $\mathcal{Z}$, $\mathcal{X}_1$ of $\mathbb{R}^{d_z}$ or $\mathbb{R}^{d_x^1}$, respectively, and for a finite set $\mathcal{X}_2 \subset \mathbb{R}^{d_x^2}$ with $d_z \geq 1$, $d_x^1, d_x^2 \geq 0$ and $d_x := d_x^1 + d_x^2 \geq 1$. The sets $\mathcal{Z}$, $\mathcal{X}_1$ and $\mathcal{X}_2$ do not depend on t. The covariate vector $\{X_i(t), Z_i(t)\}$ has a density $f_t(x, z)$ with respect to $\nu = \nu_x \times \nu_z$ where ν_z is the Lebesgue measure on $\mathbb{R}^{d_z}$ and ν_x is a product of a d_x^1-dimensional Lebesgue measure and the counting measure on $\mathcal{X}_2$. For a neighborhood $\mathcal{N}_0$ of θ_0 we assume that for fixed x_2 the functions $g(z)$, $\alpha(x_1, x_2; \theta)$ and $f_t(x_1, x_2, z)$ are strictly positive and continuous on $\mathcal{Z}$, $\mathcal{X}_1 \times \mathcal{N}_0$, and $[0, T] \times \mathcal{X}_1 \times \mathcal{Z}$, respectively. Furthermore, for $\theta \in \mathcal{N}_0$ and $z \in \mathcal{Z}$ the function $g_\theta(z)$ has 2κ derivatives that are continuous in θ and z. For the definition of e_θ see equation (11).

(A2) For $\theta \in \mathcal{N}_0$ the function $\alpha(x; \theta)$ is twice differentiable w.r.t. θ. The derivatives are bounded for all $\theta \in \mathcal{N}_0$ and $x \in \mathcal{X}_1 \times \mathcal{X}_2$. Furthermore, there exists a constant $C > 0$ such that $\|\frac{\partial^2}{\partial\theta\partial\theta^T}\alpha(x; \theta_1) - \frac{\partial^2}{\partial\theta\partial\theta^T}\alpha(x; \theta_2)\| \leq C\|\theta_1 - \theta_2\|$ for all $\theta_1, \theta_2 \in \mathcal{N}_0$ and $x \in \mathcal{X}_1 \times \mathcal{X}_2$.

(A3) The kernel K is a multivariate kernel function $K(y) = k(y_1) \cdot ... \cdot k(y_{d_z})$ where k is a symmetric with compact support, say $[-1, 1]$. It is a kernel of order 2κ, i.e. $\int u^l k(u)\mathrm{d}u = 0$ for $l = 1, ..., 2k-1$, $\int k(u)\mathrm{d}u = 1$, $\int u^{2\kappa}k(u)\mathrm{d}u \neq 0$. If the support of the kernel $K_b(z - \cdot)$ is not contained in $\mathcal{Z}$ we replace K_b by a boundary kernel $K_{z,b}$ that fulfills $\int_{\mathcal{Z}} K_{z,b}(z-u)\mathrm{d}u = 1$, $\int_{\mathcal{Z}}(z_1-u_1)^{l_1} \cdot ... \cdot (z_{d_z}-u_{d_z})^{l_{d_z}} K_{z,b}(z-u)\mathrm{d}u = 0$ for $0 \leq l_1 + ... + l_{d_z} \leq 2\kappa - 1$, $|K_{z,b}| \leq C\frac{1}{b_{prod}}$ and that has a subset of $[-b_1^0, b_1^0] \times ... \times [-b_{d_z}^0, b_{d_z}^0]$ as support.It holds that $b_{max} := \max\{b_1^0, ..., b_{d_z}^0\} \to 0$ and that $nb_{prod}^2 \to \infty$.

(A4) For all $\theta \in \mathcal{N}_0$ it holds that $\alpha(x_1, x_2; \theta)/e_\theta(z) \neq \alpha(x_1, x_2; \theta_0)/e_{\theta_0}(z)$ with positive ν-measure.

(A5) It holds that $b_{max} = o(n^{-1/(4\kappa)})$.

(A6) The semiparametric information matrix $\mathcal{I}_0$ is finite and nonsingular.

(A7) θ_0 is an interior point of Θ.

Note that (A1)–(A7) are standard assumptions. (A1) and (A2) state standard smoothness assumptions. In (A3) we assume that the kernel K is a kernel of order 2κ and that appropriate modifications of the kernel are used at the boundary. The assumption that $nb_{prod}^2 \to \infty$ is used in the proof of our main result to verify claim (40). In this claim the integrand of a martingale integral is replaced by a leave-one-out expression. Here we use brute force bounds that require $nb_{prod}^2 \to \infty$. At all other places of the proof we only need the weaker assumption $(nb_{prod})^{-1}(\log n) \to 0$. We conjecture that for covariates $Z_i(t)$ that do not depend on time it suffices to require only that $nb_{prod}^{\gamma} \to \infty$ for some $1 < \gamma < 2$. Assumption (A4) is needed to get identifiability of the parameter θ. Assumptions (A5) guarantees that the bias $g_\theta^* - g_\theta$ is of order $o_P(n^{-1/2})$.

Theorem 1. *Make the assumptions (A1)–(A4).*
(i) With probability tending to one, there exists a maximizer $\widehat{\theta}$ *in (16). All (measurable) choices of the maximizer result in a consistent estimator:* $\widehat{\theta} \xrightarrow{p} \theta_0$.
(ii) Make the additional assumptions (A5)–(A7). Then

$$n^{1/2}(\widehat{\theta} - \theta_0) \xrightarrow{d} N(0, \mathcal{I}_0^{-1}). \tag{26}$$

(iii) The asymptotic covariance matrix $\mathcal{I}_0^{-1}$ *is consistently estimated by* $\hat{H}_{\theta\theta}^{-1}(\widehat{\theta})$.

We now argue that our estimator of θ achieves the semiparametric efficiency bound.

For this purpose consider the following parametric specification of the hazard function:

$$\lambda_i(t;\theta) = \alpha\{X_i(t);\theta\}g_\theta\{Z_i(t)\}Y_i(t). \tag{27}$$

The pseudo-maximum likelihood estimator in the model is the maximizer $\overline{\theta}$ of the likelihood function $\overline{\ell}(\theta)$. By classical theory one gets that

$$n^{1/2}(\overline{\theta} - \theta_0) = \mathcal{I}_0^{-1} n^{-1/2} \sum_{i=1}^{n} \int \frac{\partial \overline{\mu}_{\theta_0}}{\partial \theta}\{X_i(u), Z_i(u)\}dM_i(u) + o_P(1).$$

Thus, $\overline{\theta}$ has the same asymptotic limit distribution as $\widehat{\theta}$ and the specification (27) is the hardest parametric sub-model of our semiparametric model. In particular, we get that $\mathcal{I}_0$ is the semiparametric information matrix.

In our simulations and in our empirical application we also use a local linear estimator of the functions g_θ. It can be shown that this also leads to efficient estimation of θ.

In the final estimation step an estimator of g is calculated. This can be done by $\widehat{g}_{b^*,\widehat{\theta}}(z)$ where $\widehat{\theta}$ is plugged in for the parameter θ. As discussed above, the bandwidth vector b^* should differ from b. We also consider a local linear estimator $\widehat{g}^{LL}_{b^*,\widehat{\theta}}(z)$. For a definition of $\widehat{g}^{LL}_{b^*,\widehat{\theta}}(z)$ see Appendix B.1. In corollary 1 we only discuss the case $d_z = 1$, $\kappa = 1$.

Corollary 1. *Suppose that assumptions (A1)-(A7) hold with $d_z = 1$, $\kappa = 1$ hold and that z is an interior point of $\mathcal{Z}$. Then*

$$\begin{aligned}\sqrt{nb^*}\{\widehat{g}_{b^*,\widehat{\theta}}(z) - g(z) - b^{*2}\beta(z)\} &\rightarrow N(0,\nu(z)),\\ \sqrt{nb^*}\{\widehat{g}^{LL}_{b^*,\widehat{\theta}}(z) - g(z) - b^{*2}\beta^{LL}(z)\} &\rightarrow N(0,\nu(z)),\end{aligned}$$

where

$$\begin{aligned}
\beta(z) &= \frac{\gamma^2}{2}\mu_2(K)\left\{2\frac{\partial g}{\partial z}(z)\frac{\partial \ln e_{\theta_0}}{\partial z}(z) + \frac{\partial^2 g}{\partial z^2}(z)\right\}, \\
\beta^{LL}(z) &= \frac{\gamma^2}{2}\mu_2(K)\frac{\partial^2 g}{\partial z^2}(z), \\
\nu(z) &= \gamma^{-1}||K||^2\frac{g(z)}{e_{\theta_0}(z)}
\end{aligned}$$

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

$$\hat{\nu}(z) = \frac{nb^* \sum_{i=1}^n \int K_{b^*}\{z - Z_i(u)\}^2 dN_i(u)}{\sum_{i=1}^n [K_{b^*}\{z - Z_i(u)\}\alpha\{X_i(u); \widehat{\theta}\}Y_i(u)du]^2}$$

is a consistent estimator of $\nu(z)$, *i.e.*

$$\hat{\nu}(z) \rightarrow_p \nu(z).$$

6 Simulation study

In this section we present the core results from our simulation study. We present additional simulation evidence regarding the bias and variance, both for the overall variance and for the local bias and variance, as well as empirical coverage in the supplementary Section 9.[3]

To study the performance of our estimator, we simulate data from the following models:

$$\textbf{Model 1:} \qquad \lambda(t) = \exp\{\theta t\}\gamma \times z(1-z),$$

[3] In the simulations we did not use boundary kernels because for $d_z = 1$ because the b neighborhood of z is of size b and the bias is also of order b, which results in an error of size b^2 which is negligible under our assumptions.

$$\textbf{Model 2:} \qquad \lambda(t) = \gamma\theta t^{\theta-1} \exp\left\{-\frac{1}{2}\cos(2\pi z) - \frac{3}{2}\right\},$$

$$\textbf{Model 3:} \qquad \lambda(t) = \exp\{\theta t\} \exp\left\{-\frac{1}{2}\cos(2\pi z) - \frac{3}{2}\right\},$$

$$\textbf{Model 4:} \qquad \lambda(t) = t^{\theta-1}(1-t)z(1-z). \tag{28}$$

with $\theta = 1.5$ and $\gamma = 1$. The two-dimensional hazards as functions of t and z are shown in Figure 1.

6.1 Results: model performance

We report the estimation results from 100 simulated samples using a discretized version of the local constant estimator and the local linear estimator (see Subsection B.2 in the appendix). We simulate on a grid $R \times R'$ with size 100 × 100 (i.e., $R = 100, R' = 100$) which seems sufficient for our purposes. The sample size is either $n = 10000$ or $n = 5000$ observations. Our estimator is evaluated along three dimensions: (1) **bandwidth selection:** We evaluate whether feasible bandwidth selection methods work to choose the two bandwidths b and b^*, (2) **parameter estimate:** we compare the true parameter θ with its estimate, and (3) **Integrated Squared Error (ISE):** we evaluate the integrated squared error of our estimator of the function g. Table 1 reports the results for the cross-validated bandwidths, the ISE bandwidths and the resulting parameter estimates in terms of the average absolute deviation from the true parameter. In general, the estimator performs well regardless of the true form of the hazard and independent of whether we use the ISE bandwidths or the bandwidths selected by minimizing the cross-validation criterion. The parameter is estimated with precision, regardless of the method, and the parameter estimates are in general not sensitive to bandwidth choice. It seems that the local constant is as good or even better as the local linear estimator for estimating the parameter, although the differences are small. Overall, in terms of the distribution of the ISE, the local linear estimator performs better than the local constant estimator, in some models even in smaller sample sizes, which suggests that the local linear is better suited to capture the nonparametric function, which is not a surprising result, considering the well-known shortcomings of the local constant estimator in boundary regions.

In almost all cases, the standard errors on b are rather large, at least compared with the standard errors on b^*. This result reflects how little the parameter estimate depends on the bandwidth choice. This suggests that applied researchers might find it practical to fix b to be very small and only consider different bandwidths for b^*.

Figure 2 visualizes the empirical distribution of the integrated squared error for all 100 samples, for both sample sizes and the two different estimators and for all four models. In general, the local linear estimator performs better than the local constant estimator. Increasing the sample size leads on average to a reduction of the ISE and a reduction in the variance of the distribution of the ISE. However, while we can retrieve the parameter with relative precision, cross-validation tends towards undersmoothing in many of the cases that were considered. While this is not surprising, better feasible bandwidth selection methods, such as “do-validation” (Gámiz Pérez et al., 2013) might improve performance.

We also compare the ISE for the nonparametric local constant and local linear estimators to the ISE for our semiparametric estimator. In our simulation setting, the former estimators target a function that has a dimensionality that exceeds the dimensionality of the function g with one. We find that if the semiparametric model is true, then, unsurprisingly, it improves estimation accuracy enormously to impose this semiparametric structure from the outset rather than using a fully nonparametric approach. In all cases, local linear estimation performs significantly better than local constant estimation, irrespectively of whether a semiparametric or a nonparametric is considered. These results are unsurprising and the results are not listed here; however, they do provide us with a helpful sanity check of the estimation and modeling approach of this paper.

6.2 The "unidentified" case

To demonstrate what happens when we include the same covariate both in the nonparametric function and the parametric function, we perform an additional simulation study using only Model 4 from the previous section. Disregarding z, we simulate the data with $\lambda(t) = t^{\theta-1}(1-t)$, but in the likelihood function estimating θ specify $\alpha = t^{\theta-1}$ and include t in the estimation of $\hat{g}$. The estimated hazard is then calculated as $\hat{\lambda} = \alpha_{\hat{\theta}}(t)\hat{g}_{\hat{\theta}}(t)$.

The results for the whole model are depicted in Figure 3. We compare the performance of our estimator, in 100 simulated samples, with the fully nonparametric local linear estimator. For the sake of comparability we compare the estimators in the simulated samples that correspond to the 25th, 50th and 75th percentile of the mean integrated squared error. The estimated parameter is relatively close to the true parameter, although the estimated parameter varies more with the choice of b than in the identified case (in that case the choice of b does not matter much). The overall model estimate (the blue dashed line) tracks the true model very well (see Figure 4), "compensating" for the missing part of the likelihood function. The different components of the model estimate are depicted in the right panel of Figure 4. Additionally to the overall model estimate $\hat{\lambda}$, we depict here the estimates for the misspecified parametric function (red solid line) and the estimated nonparametric function of our model (black, small-dashed line). The relative difference between mean integrated squared errors of the fully nonparametric estimator and our semiparametric estimator is depicted on the left panel of Figure 4 as a function of sample size ranging from $N = 500, ..., 5000$ observations.

7 Empirical application: the effect of birth weight on later-life mortality

7.1 The Uppsala Birth Cohort Study data

The Uppsala Birth Cohort Study is a lifelong follow-up study of birth cohorts of individuals born in Uppsala in 1915–1929. Rajaleid et al. (2008) demonstrate that it is representative of birth cohorts in Sweden in the years 1915–1929. Information on early-life characteristics of these newborns and social characteristics of their parents was retrieved from the neonatal register of the hospital in Uppsala. Mortality is observed from parish records and national death registers. Loss of follow-up due to emigration is observed from censuses, starting with the 1960 census, routine administrative registers, starting in 1961 or later, and archives. In the data at our disposal, individuals are followed over time up to the end of 2002, so that the highest observed death age is 87. Leon et al. (1998) and Rajaleid et al. (2008) provide detailed descriptions of the data.

The birth and death dates and the resulting individual lifetime durations are observed in days. Not all variables are observed for all of individuals, but birth date, lifetime duration or time until loss of follow-up, and birth weight are observed for virtually every individual. We omit all individuals who were stillborn or died within one day. This leads to a sample size of 13668 individuals.

Birth weight was recorded in grams. We trim the data by discarding 2 observations with birth weight below 1000 g and 27 observations with birth weight above 5000 g. For 13 of the remaining individuals, birth weight is not observed. This leads to the final sample size of $n = 13639$ individuals. The socio-economic status or social class at birth is a grouped hierarchically ordered version of the Swedish SEI code which in turn is based on the occupation of the main breadwinner in the household. The values run from 1 (highest class) to 7.

In the sample, 50% are observed to die before 2002 and 50% have right-censored lifetime durations, almost all of the latter are still alive at the end of 2002. Table 2 gives some sample statistics of the main variables that were made accessible for our study.

To interpret the results it is useful to emphasize that living conditions in Sweden in the birth years 1915–1929 were relatively good in comparison to most other countries at the time and in comparison to many developing countries today, see references below. Life expectancy was among the highest in the world, and infant mortality among the lowest (around 5%). The public health care system was modern, with institutionalized maternal and child health care in urban areas. At the time of birth, most individuals in our data resided in or around the city of Uppsala. In the years 1915–1929, the population of the city of Uppsala was stable at the level of around 30,000 inhabitants. The two largest sectors in the city's labor market were manufacturing and trade, occupying 45% and 25% of the workforce, respectively. Electricity was available everywhere. Lobell et al. (2007) provide details of the Swedish economy in these years and the surrounding decades. National Central Bureau of Statistics (1969) provide detailed descriptions of demographic developments. Sundin and Willner (2007) contains a detailed history of public health in Sweden. Modin (2002) describes local conditions in Uppsala around the 1920s. Notice that contemporary birth weight values are in the same ball park as those in the data.

The data have been used by a number of studies on long-run effects of birth weight. All of these estimate Cox Proportional Hazard models with partial likelihood. Leon et al. (1998) and Rajaleid et al. (2008) use discrete birth weight indicators based on a small number of weight intervals. van den Berg and Modin (2013) assume that the log cardiovascular mortality rate is a linear function of the log birth weight.

7.2 Model specification and results

For the parametric function $\alpha(.;\theta)$ in the hazard rate we adopt a Gompertz functional form, that is, $\alpha(.;\theta) = \exp\{\theta t\}$. In words, with age-invariant covariates this means that the log mortality rate is linear in age. This has been shown to accurately capture the age dependence of mortality for the ages covered by our observation window, in cohorts born in the first half of the 20th century. Indeed, as mentioned in Section 1, it is common in the study of adult mortality to model the effect of age t on the mortality rate by way of this specification, especially when conditioning on covariates or when considering relatively homogeneous sub-populations, and provided that extreme ages are not taken into consideration. See e.g. Wetterstrand (1981) and Gavrilov and Gavrilova (1991) for overviews of the evidence. As a starting point, we thus take the mortality rate to equal

$$\lambda(t) = \exp\{\theta t\} g(z) \tag{29}$$

with z being the birth weight. In model extensions we allow θ to vary with other covariates x (see below).

We discretize the time dimension in 150 intervals and the covariate in 100 intervals ($R = 150$, $R' = 100$), where the covariate is rescaled to lie on the unit interval $[0, 1]$, according to the formula $z_u = (z - z^{min})/(z^{max} - z^{min})$. We use the Epanechnikov Kernel given by $K(u) = \frac{3}{4}(1 - u^2)\mathbb{I}_{|u \leq 1|}$. As a robustness check we also use the kernel used in Nielsen and Tanggaard (2001), but the choice of kernel does not alter our results in any substantial way. The confidence intervals are calculated using the bootstrap procedure for kernel hazard estimators introduced by Fledelius et al. (2004).

The estimate of the shape parameter θ is practically unaffected by the bandwidth choice. The first line of Table 3 gives estimates for the local linear estimator.

Figure 5 shows the estimates of the nonparametric function g, using the local linear estimator. The x-axis depicts birth weight and the y-axis shows the estimated values of the nonparametric part of the hazard function. The estimated function varies over z in an inverted J- or a U-shape, indicating that mortality risk decreases as birth weight increases and then increases again at very high birth weights. The results can be interpreted in the following way: compared to an infant born in the optimal birth weight range of about 3000-3500g, the relative risk is about 2.5 as high as for an infant born with 1000g and 1.5 times as high as for an infant born weighting 5000g.

The local constant estimator does not perform satisfactorily. Specifically, it loses structure very quickly and becomes flat when the bandwidth is increased. This did not occur in the simulations and may be due to the scarcity of observations in the boundary regions in the application. Using a local linear framework is therefore strictly preferred.

The association between birth weight and mortality at low ages may be strongly affected by medical interventions in the first years of life. In contrast, at higher ages, biological mechanisms may drive the association. At the same time, survival up to late adulthood means survival into the 1960s and beyond, allowing the individual to benefit from medical innovations in the mid 20th century. A single function g is not necessarily able to fit such widely differing explanations. It is therefore interesting to see whether the estimated g changes if we truncate longevity from below at, say, age 40. Figure 6 plots the shape of g for that case. The results do not fundamentally differ from those in Figure 5. The mortality rate at very low birth weights is now point-estimated to be lower. This may be due to improvements in medical technology in the mid 20th century. Alternatively, dynamic selection may cause the frailest individuals among those with low birth weight to have died before age 40, causing an attenuation of the association beyond age 40. The “dynamic selection” explanation is at odds with the model that does not allow for systematic ex ante unobserved heterogeneity. However, note that the truncation of low longevities does not in fact entail the kind of simple attenuation of birth weight effects that one may expect to observe in case of dynamic selection. Specifically, the point estimate for the "optimal birth weight" shifts slightly to the right and lies now at about 4000g. In any case, whatever the cause for the small differences between Figures 5 and 6, one should keep in mind that the confidence bands in Figure 6 are wider than in Figure 5, especially at extreme birth weight values.

7.3 Comparison to a parametric specification for g

To compare the performance of the estimator to a parametric specification, we replace the nonparametric function g with a quadratic polynomial,

$$g(z;\beta) = \exp\{\beta_0 + \beta_1 z + \beta_2 z^2\} \tag{30}$$

The parameters β_0, β_1 and β_2 are estimated with maximum likelihood. The estimates (standard errors) for β_1 and β_2 are: $-4.64(0.65)$ and $3.98(0.44)$, respectively. Further, $\widehat{\theta} = 9.56e^{-5}(2.67e^{-12})$, which is very close to what we find in the semiparametric analysis. The results are shown in Figure 7. While the differences are not large, the parametric analysis overestimates the mortality risk at high birth weights. The larger point is, of course, that it is not possible, ex-ante, to know the exact parametric form of the hazard.

One may argue that the inclusion of z^2 as a covariate in the parametric $g(z;\beta)$ in (30) is likely to lead to a bad fit at very high values of z. As an alternative, we replace z and z^2 in $g(z;\beta)$ by $\log z$ and $(\log z)^2$. However, it turns out that the estimation results do not add new insights to those above.

7.4 Including additional covariates

Our approach allows us to extend the vector $X(t)$ to include more covariates than just time t. As mentioned above, it is important to avoid omitted covariates in order to prevent unobserved heterogeneity bias. To proceed, we parameterize our parametric function as $\alpha_\theta(X(t)) = \exp\{\theta_1 X^d + \theta_2 t\}$, where X^d denotes parental social class at the birth of the individual. The relevant estimation results are depicted in Figure 8. The shape of the estimated risk is not materially different from the estimate ignoring social class. The parameter estimates are reported in the first line in Table 3. Belonging to a lower social class increases mortality hazard. For a fully parametric model the results are in row 2 in Table 3. The parameter estimates are very similar to those for our semiparametric model.

Gender is known to have a large effect on mortality. We stratify our empirical analysis by gender and estimate the impact of birth weight, age and social class separately for men and women. The parameter estimates are shown in Table 3. The estimates for the impact of social class (θ_1) do not differ substantially between men and women, whereas the age dependence estimate (θ_2) is larger for men than for women. For birth weight, the effects differ by gender; see Figure 9. The left panel depicts the effect for men and the right panel for women. The increased risk at high birth weight is much more pronounced for men than for women. Apparently, birth weights above 4000g present a risk factor for men but not for women.

8 Conclusion

In the paper we specify a general class of semiparametric duration models and we develop an estimation technique for these models. The class of models includes models in which the hazard rate is a nonparametric function of covariates. We argue that our paper serves a need for estimation methods for such models, since they cannot be recast in the Cox model. Indeed, our class of models is more general than other semiparametric model classes studied in the literature. We prove that our estimator is consistent *and* efficient. In simulations we show that our estimator performs well with sample sizes that are common in epidemiology and econometrics. In the estimation procedure, we recommend to use local linear kernel estimation for the nonparametric function of the covariates.

We apply the estimator to study the association between birth weight and late-life mortality, which is seen as an issue of great interest due to its relevance for the "developmental origins" theory of late-life health. This application allows us to assess the performance of the estimator under realistic empirical conditions, with a sample size of about 13,000 individuals of which about half have right-censored lifetimes. We find a non-monotonic relationship. This is preserved if we control for social class at birth. The relationship cannot be captured with a simple parametric polynomial, confirming the usefulness of our approach. Separate analyses by gender show that the non-monotonicity is mostly due to an increased later-life mortality risk for men with high birth weight.

The application very much focuses on the flexible estimation of covariate effects. We should point out that our approach is also useful if one aims to estimate a parametric part of the hazard rate in the presence of some covariates whose effects cannot be captured parametrically because there is insufficient prior knowledge on their functional form. The effects of such covariates are then nuisance functions, but they nevertheless need to be taken into account when estimating the parameters of interest. Our approach deals with that.

This is potentially important because a model specification with many covariates leads to a curse of dimensionality, while at the same time the omission of covariates without controlling for unobserved heterogeneity may lead to biased inference. A different but related topic for further research may be to reduce the dimensionality of the model by assuming a single-index structure for the parametric part of the hazard rate as a function of covariates and markers.

As an obvious topic for further research one may consider the inclusion of unobserved heterogeneity or frailty terms in the individual hazard rates. Indeed, recent advances in the econometric literature about duration models have emphasized the need for a flexible estimation structure in the context of mixed proportional hazard estimation, see for example Bijwaard et al. (2013), Wolter (2016) and Hausman and Woutersen (2014). While these papers make important contributions in the way that they relax certain parametric assumptions, they share the inflexibility of the parametric form of the covariate function, so there lies potential for further research.

Acknowledgements

We thank the Associate Editor, three anonymous Referees, Nikolay Angelov and participants at conferences in Mannheim, Cambridge and Cologne and seminars in Bonn, Frankfurt, Bergen and Fribourg for their useful comments. We are grateful to CHESS (Stockholm University) and its UBCoS board members Ilona Koupil and Denny Vågerö for permission to use the UBCoS Multigen data. Gerard van den Berg, Enno Mammen and Lena Janys thank the German Science Foundation (DFG) for financial support through the FOR916 program. We thank Axel Munk for the coordination of this program. Enno Mammen's research was supported by the DFG through the Research Training Group RTG 1953 and it was prepared at the National Research University Higher School of Economics, Moscow, Russian Federation within the framework of a subsidy granted to the HSE by the Government of the Russian Federation for the implementation of the Global Competitiveness Program. Lena Janys acknowledges funding through the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - GZ 2047/1, Projekt-ID 390685813.

Appendix A Technical Appendix

A.1 Proof of Theorem 1

Proof of (i). We will show that

$$\sup_{\theta\in\mathcal{N}_0} |Q_n(\theta) - Q(\theta)| = o_P(1). \tag{31}$$

We now argue that this implies the claim of (i). Put

$$d_\theta(x,z) = \{\alpha(x_1,x_2;\theta)e_{\theta_0}(z)\}/\{\alpha(x_1,x_2;\theta_0)e_\theta(z)\}.$$

From (A1) and (A4) we get that $\ln d_\theta(x,z) - d_\theta(x,z) + 1 \neq 0$ with positive ν-measure for $\theta \in \mathcal{N}_0$ with $\theta \neq \theta_0$. Note that $\ln(x) - x + 1 < 0$ for $x \neq 1$. Thus we have that $Q(\theta) < Q(\theta_0)$ for $\theta \in \mathcal{N}_0$ with $\theta \neq \theta_0$. Since $Q(\theta)$ is continuous in θ we get the statement of (i), see e.g. Theorem 5.7 in van der Vaart (2000). It remains to show (31). We will show that

$$\sup_{\theta\in\mathcal{N}_0} \left|Q_n(\theta) - \overline{Q}_n(\theta)\right| = o_P(1), \tag{32}$$

$$\sup_{\theta\in\mathcal{N}_0} \left|\overline{Q}_n(\theta) - Q(\theta)\right| = o_P(1). \tag{33}$$

Claim (33) follows by a uniform law of large numbers. Note that $\overline{Q}_n(\theta)$ is an average of i.i.d. summands that are continuous in θ and uniformly bounded. For the proof of (32) it suffices to show that

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

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

These two claims follow from

$$\sup_{1\le i\le n,\theta\in\mathcal{N}_0,z\in\mathcal{Z}} |\widehat{g}_{\theta,-i}(z)-g_\theta(z)| = O_P\{(nb_{prod})^{-1/2}(\log n)^{1/2}+b_{max}\} = o_P(1),$$

see Condition (A3). The result on the uniform convergence of $\widehat{g}_{\theta,-i}$ follows by standard kernel smoothing theory. One uses that $|\widehat{g}_{\theta,-i}(z)-\widehat{g}_\theta(z)| = O_P((nb_{prod})^{-1})$, uniformly for $1\le i\le n, \theta\in\mathcal{N}_0, z\in\mathcal{Z}$. Then one argues that it suffices to prove uniform convergence over a grid of points θ and z values where the number of grid points increases polynomially. At this point one uses Lipschitz continuity of the kernel K and α, see (A1) and (A2). Then one shows uniform convergence over this grid by application of an exponential inequality for $\widehat{g}_{\theta,-i}(z)-g_\theta(z)$. □

Proof of (ii). As outlined in Section (5) we have to show (23) and (25). For the proof of (23) it suffices to show the following claims, see also (22).

$$n^{-1/2}\sum_{i=1}^n \int \left\{\frac{\partial\hat{\mu}_{\theta_0,-i}}{\partial\theta}-\frac{\partial\overline{\mu}_{\theta_0}}{\partial\theta}\right\}\{X_i(u),Z_i(u)\}dM_i(u)\to_p 0, \tag{34}$$

$$n^{-1/2}\sum_{i=1}^n \int \frac{\partial\hat{\mu}^*_{\theta_0,-i}}{\partial\theta}\{Z_i(u)\}dM_i(u)\to_p 0, \tag{35}$$

$$n^{-1/2}\sum_{i=1}^n \int \frac{\partial\hat{\mu}_{\theta_0,-i}}{\partial\theta}\{X_i(u),Z_i(u)\}\alpha\{X_i(u);\theta_0\} \tag{36}$$
$$\times(g^*_{\theta_0,-i}-g)\{Z_i(u)\}Y_i(u)du\to_p 0.$$

Claim (36) follows from $\sup_{z\in\mathcal{Z},1\le i\le n}|(g^*_{\theta_0,-i}-g)(z)| = O_P(b^{2\kappa}) = o_P(n^{-1/2})$, see (A5). For the proof of (34) we apply the results in Mammen and Nielsen (2007). For the determination of $\frac{\partial \hat{\mu}_{\theta_0,-i}}{\partial\theta}(x,z)$ one has to calculate

$$\begin{aligned}
\hat{v}_{\theta,-i}(z) &= n^{-1}\sum_{j\neq i}\int K_b\{z-Z_j(u)\}dN_j(u),\\
\hat{w}^0_{\theta,-i}(z) &= n^{-1}\sum_{j\neq i}\int K_b\{z-Z_j(u)\}\alpha\{X_j(u);\theta\}Y_j(u)du,\\
\hat{w}^1_{\theta,-i}(z) &= n^{-1}\sum_{j\neq i}\int K_b\{z-Z_j(u)\}\frac{\partial\alpha}{\partial\theta}\{X_j(u);\theta\}Y_j(u)du.
\end{aligned}$$

Define $\frac{\partial \hat{\mu}^c_{\theta_0,-i}}{\partial\theta}(x,z)$ as $\frac{\partial \hat{\mu}_{\theta_0,-i}}{\partial\theta}(x,z)$ but with $\hat{v}_{\theta,-i}(z), \hat{w}^0_{\theta,-i}(z), \hat{w}^1_{\theta,-i}(z)$ replaced by

$$\begin{aligned}
\hat{v}^c_{\theta,-i}(z) &= \min\left(c^{-1}, \max\left[c, n^{-1}\sum_{j\neq i}\int K_b\{z-Z_j(u)\}dN_j(u)\right]\right),\\
\hat{w}^{0,c}_{\theta,-i}(z) &= \min\left(c^{-1}, \max\left[c, n^{-1}\sum_{j\neq i}\int K_b\{z-Z_j(u)\}\alpha\{X_j(u);\theta\}Y_j(u)du\right]\right),\\
\hat{w}^{1,c}_{\theta,-i}(z) &= \min\left(c^{-1}, \max\left[c, n^{-1}\sum_{j\neq i}\int K_b\{z-Z_j(u)\}\frac{\partial\alpha}{\partial\theta}\{X_j(u);\theta\}Y_j(u)du\right]\right).
\end{aligned}$$

If $c > 0$ is chosen small enough one can check that $\hat{v}^c_{\theta,-i}(z) = \hat{v}_{\theta,-i}(z)$, $\hat{w}^{0,c}_{\theta,-i}(z) = \hat{w}^0_{\theta,-i}(z)$, $\hat{w}^{1,c}_{\theta,-i}(z) = \hat{w}^1_{\theta,-i}(z)$ for all $1 \le i \le n$, $\theta \in \mathcal{N}_0$ and $z \in \mathcal{Z}$, with probability tending to one. Thus, $\frac{\partial \hat{\mu}^c_{\theta_0,-i}}{\partial\theta}(x,z) = \frac{\partial \hat{\mu}_{\theta_0,-i}}{\partial\theta}(x,z)$ for all $z \in \mathcal{Z}$ and $x \in \mathcal{X}_1 \times \mathcal{X}_2$, with probability tending to one. We now apply Corollary 2 in Mammen and Nielsen (2007) with $h^{(n)}_i\{X_i(u), Z_i(u)\}$ equal to the leave-one-out version $n^{-1/2}\left(\frac{\partial \hat{\mu}^c_{\theta_0,-i}}{\partial\theta} - \frac{\partial \overline{\mu}_{\theta_0}}{\partial\theta}\right)\{X_i(u), Z_i(u)\}$ and with $h^{(n)}_{i,j}$ in this corollary equal to two-leave-out analogues. Then Corollary 2 implies (34) if one verifies that

$$\sum_{i=1}^n \rho_i^2 + n\sum_{i=1}^n \delta_i^2 \to 0, \tag{37}$$

where $\rho_i^2 = E[\int h^{(n)}_i\{X_i(u), Z_i(u)\}^2\alpha\{X_i(u);\theta\}g\{Z_i(u)\}\,du]$ and $\delta_i^2 = \max_{1\le j\le n} E[\int\{h^{(n)}_i - h^{(n)}_{i,j}\}\{X_i(u), Z_i(u)\}^2\alpha\{X_i(u);\theta\}g\{Z_i(u)\}\,du]$. Now, (37) can be easily verified because of $\max_{1\le i\le n}\rho_i^2 = O(n^{-2}b_{prod}^{-1})$ and $\max_{1\le i\le n}\delta_i^2 = O(n^{-3}b_{prod}^{-1})$. Thus, we get (34).

For the proof of (23) it remains to check (35). For the proof of this claim it suffices to show that

$$\begin{aligned} n^{-1/2}\sum_{i=1}^{n}\int \frac{\partial \mu^*_{\theta_0}}{\partial\theta}\{Z_i(u)\}dM_i(u) &= 0, && (38)\\ n^{-1/2}\sum_{i=1}^{n}\int \left\{\frac{\partial \hat{\mu}^{**}_{\theta_0,-i}}{\partial\theta} - \frac{\partial \mu^*_{\theta_0}}{\partial\theta}\right\}\{Z_i(u)\}dM_i(u) &\rightarrow_p 0, && (39)\\ n^{-1/2}\sum_{i=1}^{n}\int \left\{\frac{\partial \hat{\mu}^{**}_{\theta_0,-i}}{\partial\theta} - \frac{\partial \hat{\mu}^*_{\theta_0,-i}}{\partial\theta}\right\}\{Z_i(u)\}dM_i(u) &\rightarrow_p 0, && (40) \end{aligned}$$

where

$$\frac{\partial \mu^*_{\theta_0}}{\partial\theta}\{Z_i(u)\} = e^{-1}_{\theta_0}\{Z_i(u)\}\int \frac{\partial \overline{\mu}_{\theta_0}}{\partial\theta}\{x, Z_i(u)\}\alpha\{x;\theta_0)\}f_t\{Z_i(u), x\}y(t)dt\ dx, \qquad (41)$$

and where $\frac{\partial \hat{\mu}^{**}_{\theta_0,-i}}{\partial\theta}\{Z_i(u)\}$ is the following leave one out version of $\frac{\partial \hat{\mu}^{*}_{\theta_0,-i}}{\partial\theta}\{Z_i(u)\}$.

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

with

$$\frac{\partial}{\partial\theta}\hat{\mu}_{\theta_0,-j,-i}\{X_j(t), Z_j(t)\} = \frac{\partial}{\partial\theta}\ln\{\alpha[X_j(t),\theta_0]\} + \frac{\frac{1}{n}\sum_{k\neq j,i}\int K_b\{Z_j(t) - Z_k(v)\}\frac{\partial\alpha}{\partial\theta}\{X_k(v),\theta_0\}Y_k(v)dv}{\frac{1}{n}\sum_{k\neq j,i}\int K_b\{Z_j(t) - Z_k(v)\}\alpha\{X_k(v),\theta_0\}Y_k(v)dv}.$$

For the proof of (38) we now argue that that $\partial\mu^*_{\theta_0}\{Z_i(u)\}/\partial\theta = 0$. This follows by plugging the following terms into the definition (41) of $\partial\mu^*_{\theta_0}\{Z_i(u)\}/\partial\theta$:

$$\frac{\partial\overline{\mu}_{\theta_0}}{\partial\theta}(x,z) = \frac{\partial \ln\alpha}{\partial\theta}(x;\theta) - \frac{\partial \ln e_\theta}{\partial\theta}(z),\ \frac{\partial e_\theta}{\partial\theta}(z) = \int \frac{\partial\alpha}{\partial\theta}(x;\theta) f(z,x)y(u)du.$$

For the proof of (39) we proceed similarly as in the proof of (34) above. Again, we apply Corollary 2 in Mammen and Nielsen (2007), now with

$$h_i^{(n)}\{Z_i(u)\} = n^{-1/2}\left\{\frac{\partial\hat{\mu}^{**,c}_{\theta_0,-i}}{\partial\theta}\right\}\{Z_i(u)\} = n^{-1/2}\left\{\frac{\partial\hat{\mu}^{**,c}_{\theta_0,-i}}{\partial\theta} - \frac{\partial\mu^*_{\theta_0}}{\partial\theta}\right\}\{Z_i(u)\},$$

where $\left\{\frac{\partial\hat{\mu}^{**,c}_{\theta_0,-i}}{\partial\theta}\right\}\{Z_i(u)\}$ is a truncation modification of $\left\{\frac{\partial\hat{\mu}^{**,c}_{\theta_0,-i}}{\partial\theta}\right\}\{Z_i(u)\}$, similarly constructed as the modification $\left(\frac{\partial\hat{\mu}^{c}_{\theta_0,-i}}{\partial\theta}\right)\{X_i(u), Z_i(u)\}$ of $\left(\frac{\partial\hat{\mu}_{\theta_0,-i}}{\partial\theta}\right)\{X_i(u), Z_i(u)\}$ in the proof of (34). Again, $h_{i,j}^{(n)}$ are chosen as the two-leave-out analogues of $h_i^{(n)}$. Then Corollary 2 in Mammen and Nielsen (2007) implies (39) if we verify (37) with the updated definitions of ρ_i and δ_i. By lengthy but straight forward calculations one can show that $\max_{1\le i\le n}\rho_i^2 = O(n^{-2}b_{prod}^{-1})$ and $\max_{1\le i\le n}\delta_i^2 = O(n^{-3}b_{prod}^{-1})$. This implies (37) because of assumption (A3) and concludes the proof of (39). For the proof of (40) note that

$$\begin{aligned}\Delta &= n^{-1/2}\sum_{i=1}^n \int \left(\frac{\partial\hat{\mu}^{**}_{\theta_0,-i}}{\partial\theta} - \frac{\partial\hat{\mu}^{*}_{\theta_0,-i}}{\partial\theta}\right)\{Z_i(u)\}dM_i(u)\\ &= n^{-1}\sum_{i=1}^n [\Delta_{1,i}(u) + \Delta_{2,i}(u)]dM_i(u),\end{aligned}$$

where

$$\begin{aligned}
\Delta_{1,i}(u) &= n^{-1/2}\sum_{j\neq i}\int\left(\frac{\frac{1}{n}\sum_{k\neq j,i}\int K_b\{Z_j(t)-Z_k(v)\}\frac{\partial\alpha}{\partial\theta}\{X_k(v),\theta_0\}Y_k(v)dv}{\left[\frac{1}{n}\sum_{k\neq j,i}\int K_b\{Z_j(t)-Z_k(v)\}\alpha\{X_k(v),\theta_0\}Y_k(v)dv\right]^2}\right.\\
&\qquad\left.-\frac{\frac{1}{n}\sum_{k\neq j}\int K_b\{Z_j(t)-Z_k(v)\}\frac{\partial\alpha}{\partial\theta}\{X_k(v),\theta_0\}Y_k(v)dv}{\left[\frac{1}{n}\sum_{k\neq j}\int K_b\{Z_j(t)-Z_k(v)\}\alpha\{X_k(v),\theta_0\}Y_k(v)dv\right]^2}\right)\\
&\qquad\times\alpha\{X_j(t);\theta_0\}Y_j(t)K_b\{Z_j(t)-Z_i(u)\}dt,
\end{aligned}$$

$$\begin{aligned}
\Delta_{2,i}(u) &= n^{-1/2}\sum_{j\neq i}\int\left(\frac{1}{\left[\frac{1}{n}\sum_{k\neq j,i}\int K_b\{Z_j(t)-Z_k(v)\}\alpha\{X_k(v),\theta_0\}Y_k(v)dv\right]^2}\right.\\
&\qquad\left.-\frac{1}{\left[\frac{1}{n}\sum_{k\neq j}\int K_b\{Z_j(t)-Z_k(v)\}\alpha\{X_k(v),\theta_0\}Y_k(v)dv\right]^2}\right)\\
&\qquad\times\frac{\partial\alpha}{\partial\theta}\alpha\{X_j(t);\theta_0\}Y_j(t)K_b\{Z_j(t)-Z_i(u)\}dt.
\end{aligned}$$

Using brute force bounds one gets with a random variable $R=O_P(1)$ that for $j\in\{1,2\}, i\in\{1,...,n\}, 0\leq u\leq T$

$$\begin{aligned}
|\Delta_{j,i}(u)| &\leq n^{-1/2}R\int\frac{1}{b_{prod}}1\left(\max_{1\leq i\leq d_z}\left|\frac{Z_i(u)-Z_i(v)}{b_i^0}\right|\leq 2\right)\mathrm{d}v\\
&\leq n^{-1/2}R\frac{T}{b_{prod}}\\
&= o_P(1),
\end{aligned}$$

because we have assumed that $nb_{prod}^2\to\infty$. This implies that $\Delta=o_P(1)$ and concludes the proof of (40).

For the proof of statement (ii) of the theorem it remains to check (25). We will show the following expansions for sequences δ_n with $\delta_n\to 0$. These expansions immediately imply (25).

$$\sup_{|\theta-\theta_0|\le\delta_n}\left|\hat{H}_1(\theta)+\int\int\frac{\partial\overline{\mu}_\theta}{\partial\theta}\frac{\partial\overline{\mu}_\theta}{\partial\theta^T}(x,z)\alpha(x,\theta)g_\theta(z)f_u(x,z)y(u)dz\ dx\ du\right| = o_p(1), \tag{42}$$

$$\sup_{|\theta-\theta_0|\le\delta_n}\left|\hat{H}_j(\theta)\right| = o_p(1),\ \text{ for } j=2,...,5 \tag{43}$$

where

$$\begin{aligned}
\hat{H}_1(\theta) &= -n^{-1}\sum_{i=1}^n\int\frac{\partial\hat{\mu}_{\theta,-i}}{\partial\theta}\frac{\partial\hat{\mu}_{\theta,-i}}{\partial\theta^T}\{X_i(u),Z_i(u)\}\alpha\{X_i(u);\theta\}g_\theta\{Z_i(u)\}Y_i(u)du,\\
\hat{H}_2(\theta) &= n^{-1}\sum_{i=1}^n\int\frac{\partial^2\hat{\mu}_{\theta,-i}}{\partial\theta\partial\theta^T}\{X_i(u),Z_i(u)\}\\
&\quad\times[\alpha\{X_i(u);\theta_0\}g\{Z_i(u)\}-\alpha\{X_i(u);\theta\}g_\theta\{Z_i(u)\}]Y_i(u)du,\\
\hat{H}_3(\theta) &= n^{-1}\sum_{i=1}^n\int\left(\frac{\partial^2\hat{\mu}_{\theta,-i}}{\partial\theta\partial\theta^T}-\frac{\partial^2\hat{\mu}_{\theta_0,-i}}{\partial\theta\partial\theta^T}\right)\{X_i(u),Z_i(u)\}dM_i(u),\\
\hat{H}_4(\theta) &= n^{-1}\sum_{i=1}^n\int\frac{\partial^2\hat{\mu}_{\theta_0,-i}}{\partial\theta\partial\theta^T}\{X_i(u),Z_i(u)\}dM_i(u),\\
\hat{H}_5(\theta) &= -n^{-1}\sum_{i=1}^n\int\left\{\frac{\partial^2\hat{\mu}_{\theta,-i}}{\partial\theta\partial\theta^T}+\frac{\partial\hat{\mu}_{\theta,-i}}{\partial\theta}\frac{\partial\hat{\mu}_{\theta,-i}}{\partial\theta^T}\right\}\{X_i(u),Z_i(u)\}\\
&\quad\times\alpha\{X_i(u);\theta\}\{\widehat{g}_{\theta,-i}-g_\theta\}\{Z_i(u)\}Y_i(u)du.
\end{aligned}$$

Note that $\hat{H}_{\theta\theta}(\theta)=\sum_{j=1}^5\hat{H}_j(\theta)$. For the proof of (42)–(43) one uses results on the uniform convergence of $\widehat{g}_\theta$ and its first two partial derivatives w.r.t. θ and uniform laws of large numbers. Compare also the proof of part (i) of the theorem for the proof of (43) for $j=4$.

□

Proof of (iii). This follows immediately from (25) and the consistency of $\widehat{\theta}$. □

A.2 Proof of Corollary 1

The asymptotic distribution of $\widehat{g}$ follows directly from $\sqrt{n}$-consistency of $\widehat{\theta}$, see also Nielsen, Linton and Bickel (1998).

Appendix B The local linear estimator and the discretized estimator

B.1 The local linear estimator

In this subsection we give a definition of the local linear estimator $\widehat{g}^{LL}_{b,\widehat{\theta}}(z)$. This estimator is defined as γ_0 where (γ_0, γ_1) solves

$$
\begin{aligned}
0 &\overset{!}{=} \sum_{i=1}^{n} \int_0^T \begin{pmatrix} 1 \\ z - Z_i(s) \end{pmatrix} \alpha_\theta\{X_i(s)\} K_b\{z - Z_i(s)\} \alpha_\theta\{X_i(s)\}^{-1} dN_i(s) \\
&- \sum_{i=1}^{n} \int_0^T \begin{pmatrix} 1 & z - Z_i(s) \\ z - Z_i(s) & (z - Z_i(s))^2 \end{pmatrix} \begin{pmatrix} \gamma_0 \\ \gamma_1 \end{pmatrix} \alpha_\theta\{X_i(s)\}^2 K_b\{z - Z_i(s)\} \alpha_\theta\{X_i(s)\}^{-1} Y_i(s) ds.
\end{aligned}
$$

Thus we have that

$$
\widehat{g}^{LL}_{b,\widehat{\theta}}(z) = \frac{a_{22}(\theta) y_1 - a_{12}(\theta) y_2}{a_{11}(\theta) a_{22}(\theta) - a_{12}(\theta)^2} \tag{44}
$$

with

$$
\begin{aligned}
y_1 &= \sum_{i=1}^{n} \int_0^T K_b\{z - Z_i(s)\} dN_i(s) ds, \\
y_2 &= \sum_{i=1}^{n} \int_0^T \{z - Z_i(s)\} K_b\{z - Z_i(s)\} dN_i(s) ds, \\
a_{11}(\theta) &= \sum_{i=1}^{n} \int_0^T K_b\{z - Z_i(s)\} \alpha_\theta\{X_i(s)\} Y_i(s) ds, \\
a_{12}(\theta) &= \sum_{i=1}^{n} \int_0^T \{z - Z_i(s)\} K_b\{z - Z_i(s)\} \alpha_\theta\{X_i(s)\} Y_i(s) ds, \\
a_{22}(\theta) &= \sum_{i=1}^{n} \int_0^T \{z - Z_i(s)\}^2 K_b\{z - Z_i(s)\} \alpha_\theta\{X(s)\} Y_i(s) ds.
\end{aligned}
$$

B.2 Discretized estimators

We use a discrete version of the pseudolikelihood equation (15). Let $E_{rr'}$ be the number of exposures at the point rr' in the two-dimensional grid (with $R \times R'$ gridpoints) and $O_{rr'}$ the number of occurrences (or failures). We can calculate the occurrences and the exposures as follows:

$$O_{r,r'} = \sum_{j=1}^{n_{r'}} \int_{t_{r-1}}^{t_r} dN_{r',j}(s),$$

and

$$E_{r,r'} = \sum_{j=1}^{n_{r'}} \int_{t_{r-1}}^{t_r} Y_{r',j}(s) ds,$$

for $r = 1, \ldots, R$ and $r' = 1, \ldots, R'$. Note that $O_{r,r'}$ represents the observed occurrences of the counting processes $\{N_{r',1}, \ldots, N_{r',n_{r'}}\}$, and $E_{r,r'}$ represents the observed exposures from the counting processes $\{Y_{r',1}, \ldots, Y_{r',n_{r'}}\}$ in the interval $[t_{r-1}, t_r)$ (for $r = 1, \ldots, R$ and $r' = 1, \ldots, R'$). In case of local constant estimation of g, the discrete estimator for g is:

$$\widehat{g}_\theta(z) = \frac{\sum_{r'=1}^{R'} \sum_{r=1}^{R} K_b\{z - Z_{r'}(r)\} O_{rr'}}{\sum_{r'=1}^{R'} \sum_{r=1}^{R} K_b\{z - Z_{r'}(r)\} \alpha\{X(r);\theta\} E_{rr'}} \tag{45}$$

The discrete estimator for θ follows from the discrete version of the likelihood function:

$$\hat{\ell}(\theta) = \sum_{r'=1}^{R'} \sum_{r=1}^{R} \ln[\alpha(X(r);\theta\}\widehat{g}_\theta(z)] O_{rr'} - \sum_{r'=1}^{R'} \sum_{r=1}^{R} \alpha\{X(r);\theta\}\widehat{g}_\theta(z) E_{rr'} \tag{46}$$

in which (45) can be inserted. This can in turn be straightforwardly modified into a discretized leave-one-out estimator. Bandwidth selection is accordingly modified, along the lines of Subsection B.2.

$$\begin{aligned}\hat{\ell}(\theta) \;\; = \;\; & \sum_{r'=1}^{R'} \sum_{r=1}^{R} \ln\left[\alpha\{X(r);\theta\} \frac{\sum_{r'=1}^{R'} \sum_{r=1}^{R} K_b\{z - Z_{r'}(r)\} O_{rr'}}{\sum_{r'=1}^{R'} \sum_{r=1}^{R} K_b\{z - Z_{r'}(r)\} \alpha\{X(r);\theta\} E_{rr'}}\right] O_{rr'} \\ & - \sum_{r'=1}^{R'} \sum_{r=1}^{R} \alpha\{X(r);\theta\} \frac{\sum_{r'=1}^{R'} \sum_{r=1}^{R} K_b\{z - Z_{r'}(r)\} O_{rr'}}{\sum_{r'=1}^{R'} \sum_{r=1}^{R} K_b(t-r) \alpha\{X(r);\theta\} E_{rr'}} E_{rr'}. \end{aligned} \tag{47}$$

The discrete leave-one-out estimator for $\boldsymbol{\theta}$ makes use of the discretized version of the likelihood function that is defined as follows:

$$\hat{\ell}^{loo}(\theta) = \sum_{r'=1}^{R'} \sum_{r=1}^{R} \ln[\alpha\{X(r);\theta\}\widehat{g}_\theta^{loo}(z)] O_{rr'}^{loo} - \sum_{r'=1}^{R'} \sum_{r=1}^{R} \alpha\{X(r);\theta\}\widehat{g}_\theta^{loo}(z) E_{rr'}. \tag{48}$$

$\widehat{g}^{loo}$ is the leave on out version of the estimator where $O_{rr'}$ is replaced with the leave-one-out version $O_{rr'}^{loo}$.

References

Aalen, O. (1978), Nonparametric inference for a family of counting processes, The Annals of Statistics 6(4), 701–726.

Ahlgren, M., J. Wohlfahrt, L.W. Olsen, T.I.A. Sørensen and M. Melbye (2007), Birth weight and risk of cancer, Cancer 110, 412–419.

Almond, D. and Currie, J. (2011), Killing me softly: The fetal origins hypothesis, *Journal of Economic Perspectives* 25, 153–172.

Andersen, P.K., Borgan,Ø., Gill, R.D. and Keiding, N. (1993), Statistical models based on counting processes, *Springer*.

Bearse, P., Canals-Cerdá, J. and Rilstone, P. (2007), Efficient Semiparametric Estimation of Duration Models with Unobserved Heterogeneity, Econometric Theory 23(2), 281–308.

van den Berg, G.J. (1990), Nonstationarity in job search theory, Review of Economic Studies 57, 255–277.

van den Berg, G.J. (2001), Duration models: specification, identification, and multiple durations, in: J.J. Heckman and E. Leamer (eds.), Handbook of Econometrics, Volume V, North-Holland, Amsterdam.

van den Berg, G.J. and B. Modin (2013), Economic conditions at birth, birth weight, ability, and the causal path to cardiovascular mortality, Working paper, IZA Bonn.

Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J.A. (1993), Efficient and adaptive estimation for semiparametric models, Baltimore: Johns Hopkins University Press.

Bijwaard, G., Ridder, G., Woutersen, T. (2013), A Simple GMM Estimator for the Semiparametric Mixed Proportional Hazard Model, Journal of Econometric Methods 2(1), 1–23.

Blanchard, O.J. and Diamond, P. (1994), Ranking, unemployment duration, and wages, Review of Economic Studies 61, 417–434.

Borgan, Ø. (1984), Maximum likelihood estimation in parametric counting process models, with applications to censored failure time data, Scandinavian Journal of

Statistics 11 (1) 1–16.

Calonico, S., Cattaneo, M.D. and Farrell, M.H. (2018), On the effect of bias estimation on coverage accuracy in nonparametric inference, Journal of the American Statistical Association, 113 (522), 767–779.

Cattaneo, M.D, Crump, R. and Jansson, M. (2013), Generalized Jackknife Estimators of Weighted Average Derivatives (with comments and rejoinder), Journal of the American Statistical Association 108 (504), 1243–1268.

Cox, D.R. (1972), Regression models and life tables, Journal of the Royal Statistical Society. Series B 34 (2), 187–220.

Dabrowska, D.M. (1987), Non-parametric regression with censored survival time data, Scandinavian Journal of Statistics 14, 181–197.

Dabrowska, D.M. (1997), Smoothed Cox regression, The Annals of Statistics 25(4), 1510–1540.

Dabrowska, D.M. (2006), Estimation in a class of semiparametric transformation models, IMS Lecture Notes – Monograph Series 2nd Lehmann Symposium – Optimality 49, 131–169.

Davey Smith, G. (2005), Epidemiological Freudianism, International Journal of Epidemiology 34, 1–2.

Fledelius, P., Guillen, M., Nielsen, J.P. and Vogelius, M. (2004), Two-dimensional hazard estimation for longevity analysis, *Scandinavian Actuarial Journal* 2, 133–156.

Gámiz Pérez, M.L., Janys, L., Martínez Miranda, M.D. and Nielsen, J.P. (2013), Bandwidth selection in marker dependent kernel hazard estimation, Computational Statistics and Data Analysis 68, 155–169.

Gavrilov, L. and N. Gavrilova (1991), The Biology of Life Span: A Quantitative Approach, Harwood, Chur.

Hausman, J.A, Woutersen, T. (2014), Estimating a semi-parametric duration model without specifying heterogeneity, Journal of Econometrics 178, 114–131

Hjort, N.L. and Glad, I.K. (1995), Nonparametric density estimation with a parametric start, The Annals of Statistics 23(3), 882–904.

Hjort, N.L. and Jones, M.C. (1996), Locally parametric nonparametric density estimation, The Annals of Statistics 24(4), 1619–1647.

Huxley, R., C.G. Owen, P.H. Wincup, D.G. Cook, J. Rich-Edwards, G. Davey Smith and R. Collins (2007), Is birth weight a risk factor for ischemic heart disease in later life?, American Journal of Clinical Nutrition 85, 1244–1250.

Jovanovic, B. (1984), Wages and turnover: a parametrization of the job-matching model, Studies in Labor Market Dynamics, edited by G.R. Neumann and N. Westergård-Nielsen, Springer-Verlag.

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

Kuh, D. and Ben-Shlomo, Y. (2004), A Life Course Approach to Chronic Disease Epidemiology, Oxford University Press, Oxford.

Leon, D.A., Lithell, H.O., Vågerö, D., Koupilová, I., Mohsen, R., Berglund, L., Lithell, U.B. and McKeigue P.M. (1998), Reduced fetal growth rate and increased risk of death from ischaemic heart disease: cohort study of 15000 Swedish men and women born 1915–29, British Medical Journal 317, 241–245.

Linton, O.B., Nielsen, J.P. and van de Geer, S. (2003), Estimating multiplicative and additive hazard functions by kernel methods, Annals of Statistics 31, 464–492.

Lobell, H., Schön, L. and Krantz, O. (2007), Observations from the new Swedish Historical National Accounts, Working paper, Lund University.

Mammen, E., Nielsen, J.P. (2007), A general approach to the predictability issue in survival analysis with applications, *Biometrika* 94(4), 873–892.

Modin, B. (2002), Setting the Scene for Life: Longitudinal Studies of Early Social Disadvantage and Later Life Chances, Centre for Health Equity Studies, Stockholm.

National Central Bureau of Statistics (1969), Historical Statistics of Sweden, Part 1. Population, Second edition, 1720–1967, National Central Bureau of Statistics, Stockholm.

Nielsen, J.P. and Linton, O. (1995), Kernel estimation in a nonparametric marker dependent hazard model, The Annals of Statistics, 1735–1748.

Nielsen, J.P., Linton, O. and Bickel, P.J. (1998), On a semiparametric survival model with flexible covariate effect, The Annals of Statistics, 215–241.

Nielsen, J.P. (1998), Marker dependent kernel hazard estimation from local linear estimation, Scandinavian actuarial journal 2, 113–124.

Nielsen, J.P. and Tanggaard, C. (2001), Boundary and bias correction in kernel hazard estimation, Scandinavian Journal of Statistics 28(4), 675–698.

Osler, M., Andersen, A.M.N., Due, P., Lund, R., Damsgaard, M.T. and Holstein, B.E. (2003), Socioeconomic position in early life, birth weight, childhood cognitive function, and adult mortality. A longitudinal study of Danish men born in 1953, Journal of Epidemiology and Community Health 57, 681–686.

Poulter, N.R., Chang, C.L., MacGregor, A.J., Snieder, H. and Spector, T.D. (1999), Association between birth weight and adult blood pressure in twins: historical cohort study, British Medical Journal 319, 1330–1333.

Rajaleid, K., Manor, O. and Koupil, I. (2008), Does the strength of the association between foetal growth rate and ischaemic heart disease mortality differ by social circumstances in early or later life?, Journal of Epidemiology and Community Health 62(5), e6.

Rasmussen, K.M. (2001), The "fetal origins" hypothesis: challenges and opportunities for maternal and child nutrition, Annual Review of Nutrition 21, 73–95.

Rezat, S., Rilstone P. (2015), Semiparametric efficiency bounds and efficient estimation of discrete duration models with unspecified hazard rate, Econometric Reviews 35(5), 693–726.

Ridder, G., Woutersen, T. (2003), The Singularity of the Information Matrix of the Mixed Proportional Hazard Model, Econometrica 71, 1579–1589.

Spierdijk, L. (2008), Nonparametric conditional hazard rate estimation: a local linear approach, Computational Statistics and Data Analysis 52, 2419–2434.

Sundin, J. and S. Willner (2007), Social Change and Health in Sweden: 250 Years of Politics and Practice, Swedish National Institute of Public Health, Östersund.

van der Vaart, A.W. (2000), Asymptotic Statistics, Cambridge University Press.

Wetterstrand, W. (1981), Parametric models for life insurance mortality data: Gompertz's law over time, *Transactions of the Society of Actuaries* 33, 159–175.

Wolter, J.L. (2016), Kernel estimation of hazard functions when observations have dependent and common covariates, Journal of Econometrics 193, 1-16.

Tables

Integrated Squared Error

Model	n	Bandwidths				Parameter, $\overline{e}$	
		LC		LL		LC	LL
		b	b^*	b	b^*		
1	10000	0.3084 (0.1709)	0.0780 (0.0126)	0.2244 (0.2269)	0.1476 (0.0179)	0.011	0.031
	5000	0.2872 (0.1858)	0.0872 (0.0126)	0.1688 (0.1985)	0.1674 (0.0237)	0.020	0.035
2	10000	0.1044 (0.1701)	0.1346 (0.0249)	0.1024 (0.1761)	0.1398 (0.0244)	0.010	0.010
	5000	0.0418 (0.1082)	0.1466 (0.0246)	0.0456 (0.1214)	0.1448 (0.0254)	0.027	0.027
3	10000	0.4049 (0.0998)	0.1235 (0.016)	0.4571 (0.0974)	0.1305 (0.0147)	0.014	0.022
	5000	0.3250 (0.1888)	0.1314 (0.0266)	0.3474 (0.1965)	0.1432 (0.0264)	0.015	0.023
4	10000	0.1642 (0.2643)	0.1628 (0.0258)	0.1500 (0.2373)	0.2468 (0.0357)	0.030	0.030
	5000	0.1360 (0.1839)	0.1684 (0.038)	0.1302 (0.1703)	0.2354 (0.0495)	0.054	0.054

Cross-Validation

Model	n	Bandwidths				Parameter, $\overline{e}$	
		LC		LL		LC	LL
		b	b^*	b	b^*		
1	10000	0.177 (0.1554)	0.026 (0.0117)	0.276 (0.2250)	0.029 (0.0164)	0.028	0.030
	5000	0.104 (0.1159)	0.021(0.0034)	0.190 (0.2028)	0.021 (0.0052)	0.038	0.039
2	10000	0.035 (0.0643)	0.022 (0.0095)	0.036 (0.0727)	0.022 (0.0095)	0.011	0.011
	5000	0.193 (0.2427)	0.021 (0.0001)	0.691 (0.0192)	0.626 (0.0090)	0.031	0.034
3	10000	0.182 (0.1505)	0.038 (0.0154)	0.213 (0.1792)	0.039 (0.0154)	0.036	0.036
	5000	0.171 (0.1577)	0.032 (0.0061)	0.202 (0.1946)	0.032 (0.0064)	0.033	0.033
4	10000	0.092 (0.1809)	0.032 (0.0060)	0.082 (0.1692)	0.032 (0.0058)	0.031	0.031
	5000	0.101 (0.1508)	0.031 (0.0044)	0.098 (0.1469)	0.031 (0.0039)	0.055	0.055

Table 1: Simulation results for the models in equation (28), with $\theta = 1.5$ as the true parameter, for two different sample sizes (5000, 10000). The numbers are averages over 100 simulated samples. The upper panel shows the results for bandwidths chosen by the infeasible strategy of minimizing the ISE. b and b^* refer to the two associated bandwidths. Standard errors are in parentheses. The lower panel shows the results for bandwidths chosen by the feasible bandwidth selection criterion of minimizing the cross-validation score (CV). LC and LL refer to the use of the local constant and the local linear estimator, respectively. In the last column, the parameter estimate is reported in terms of the average of the estimation error $\overline{e} = abs(\widehat{\theta}_b - \theta_0)$ over 100 samples.

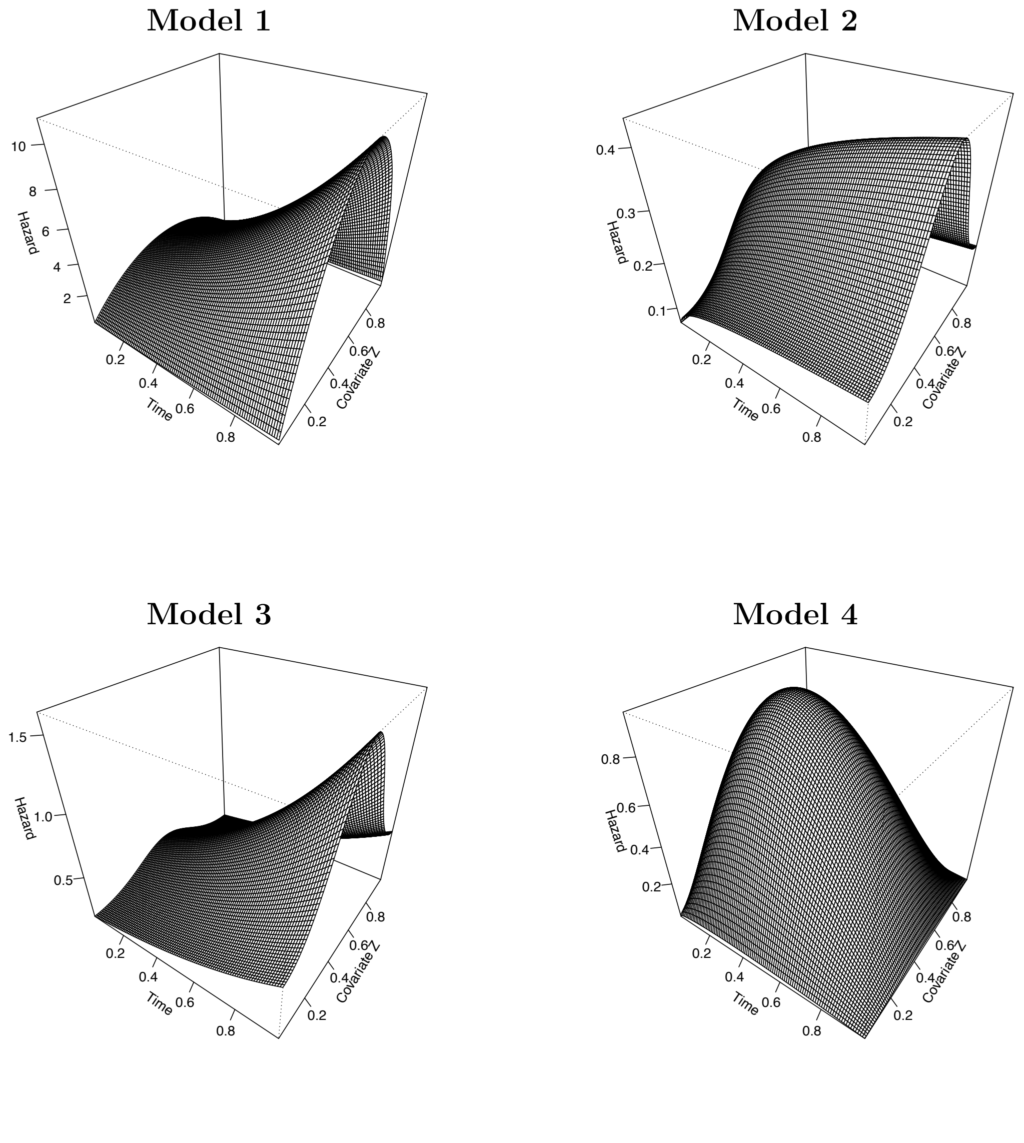
	10^{th}	mean	90^{th}
variable	percentile		perc.
right-censored durations		0.50	
duration (years) if uncensored	0.9	54.6	80.0
duration (years) if censored	73.7	77.6	84.6
birth weight (grams)	2750	3416	4080
birth year	1916	1922.6	1928
social class at birth (1 to 7: high to low)	2	4.2	6
male		0.52	
male birth weight	2810	3478	4140
female birth weight	2700	3349	4000

Table 2: Summary statistics of the sample

	$\widehat{\theta_1}$		$\widehat{\theta_2}$	
Semiparametric Model (only θ_2)	–		9.6×10^{-5}	(3×10^{-12})
Semiparametric Model	0.041	(4.2e-05)	0.00095	(3×10^{-9})
Parametric Model	0.042	(4.8e-06)	0.00095	(3×10^{-10})
Men	0.04	(7e-06)	0.0001	(5×10^{-12})
Women	0.036	(0.0001)	0.001	(8×10^{-11})

Table 3: Parameter estimates for θ_1 and θ_2 in a model with parental social class at birth, standard errors are shown in parentheses.

Figures



Model 1

Model 2

Model 3

Model 4

Figure 1: The two-dimensional hazard functions of Models 1–4, see (28).

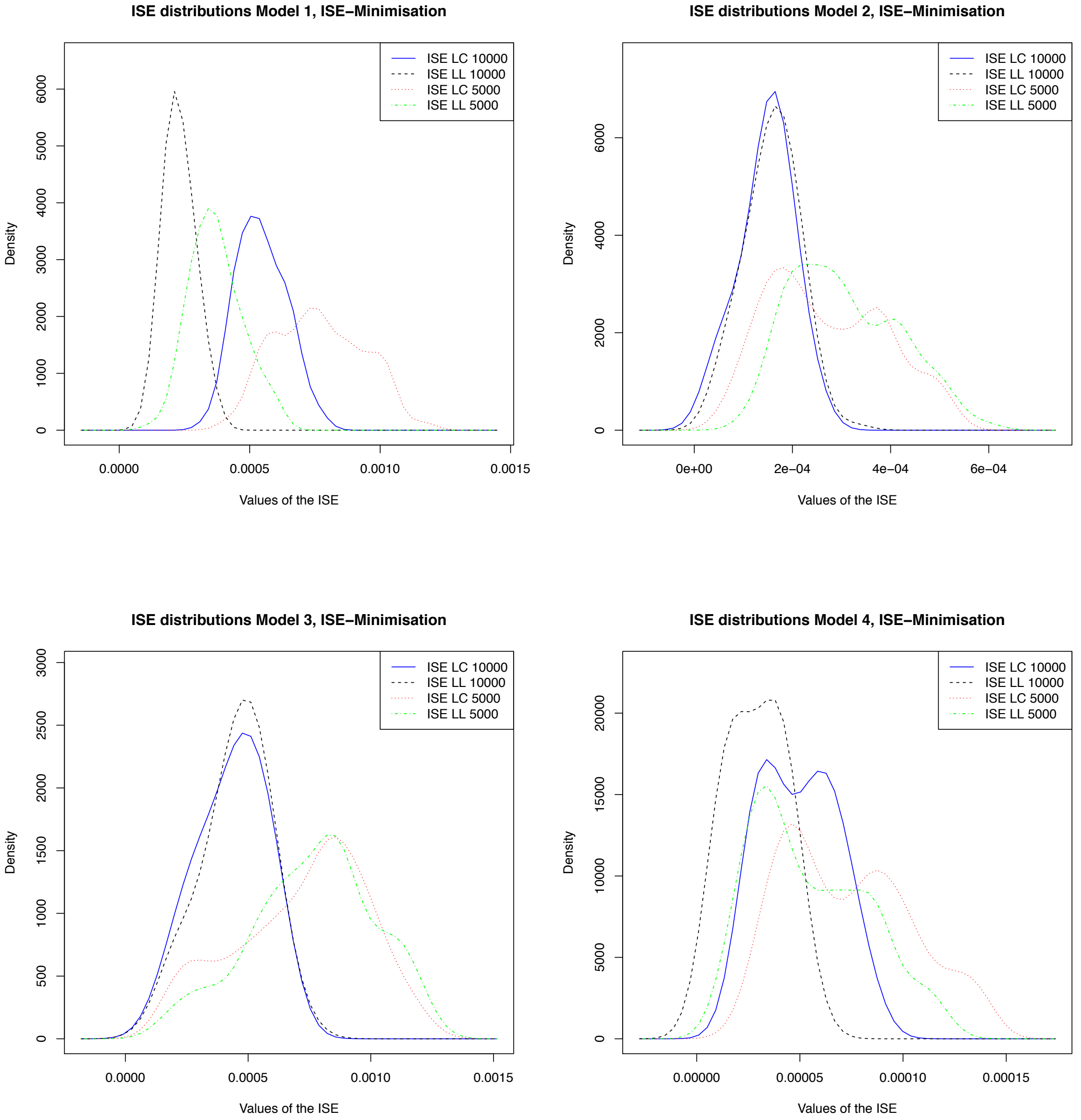
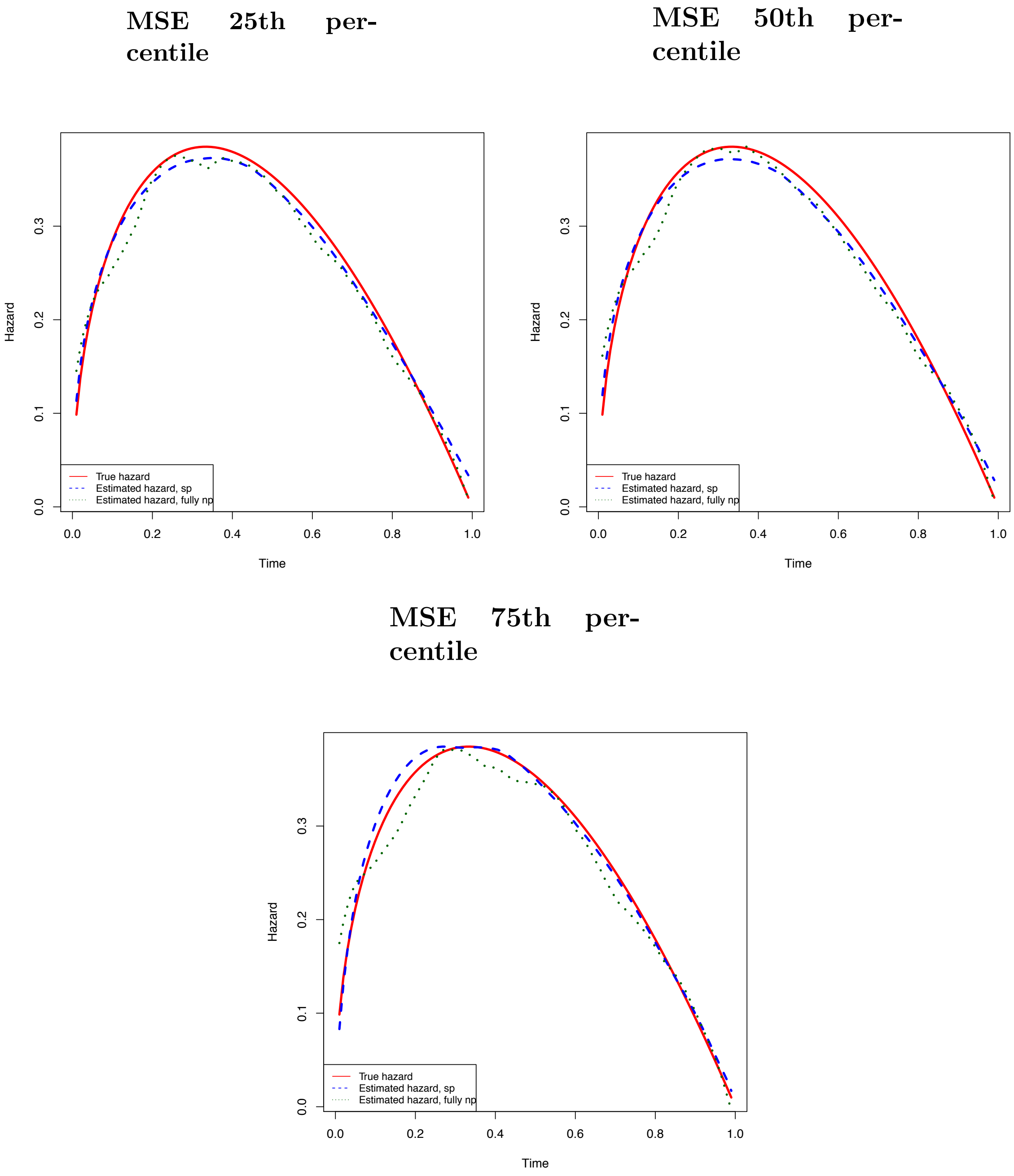


Figure 2: Kernel density estimates of the integrated squared error over 100 samples. The solid line represents the local constant estimator with $n = 10000$, the dashed line represents the local linear estimator with $n = 10000$. The dotted line represents the local constant estimator with $n = 5000$ and the dot-dash line represents the local linear estimator with $n = 5000$.



MSE 25th percentile

MSE 50th percentile

MSE 75th percentile

Figure 3: The true hazard function, the estimated semiparametric hazard function and the fully nonparametric estimate for the local linear estimator. Bandwidths were chosen by cross-validation (semiparametric estimator) and minimizing the infeasible ISE-Criterion (for the fully nonparametric estimator) for 5000 observations, comparing the simulations corresponding to the 25th percentile, 50th percentile and 75th percentile of the mean integrated squared error.

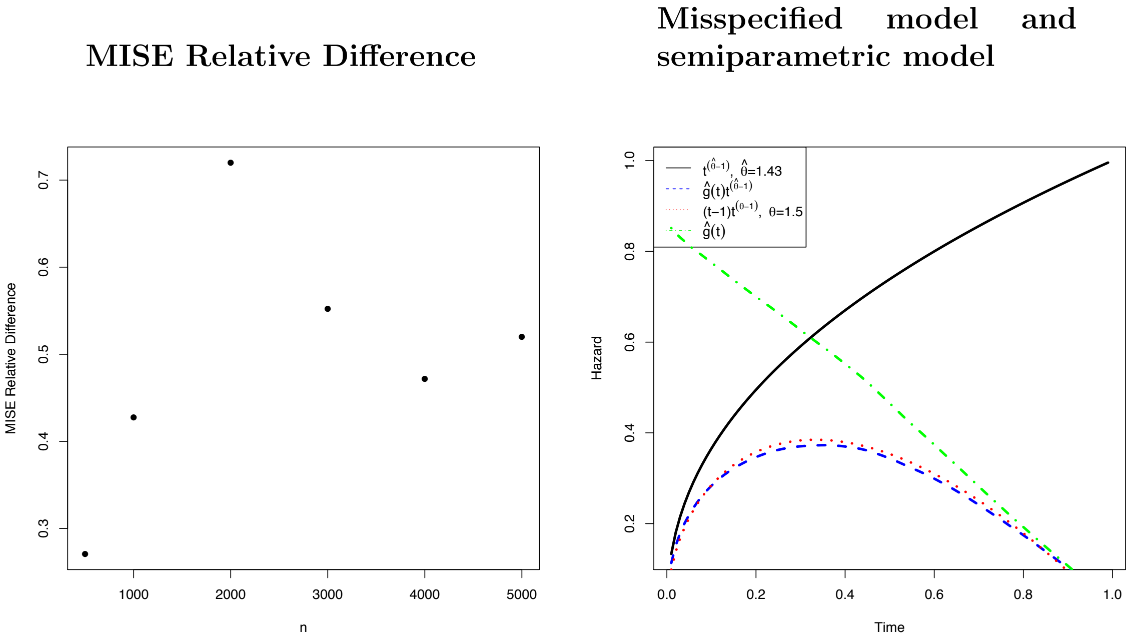


Figure 4: The left panel shows the evolution of the relative difference in mean integrated squared error depending on sample size for 100 simulation runs between the fully nonparametric estimator and the semiparametric estimator with misspecified parametric function. The right panel shows the estimated hazard when the misspecified likelihood function it used, the true hazard and our semiparametric estimator.

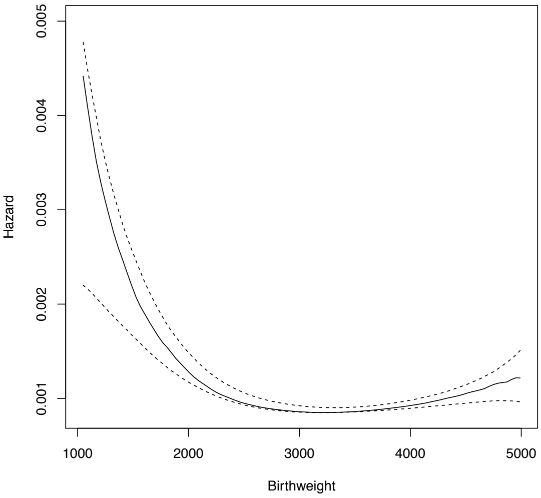


Figure 5: Estimation results for g using the local linear estimator. The y-axis reports the estimated hazard, the x-axis depicts birth weights.

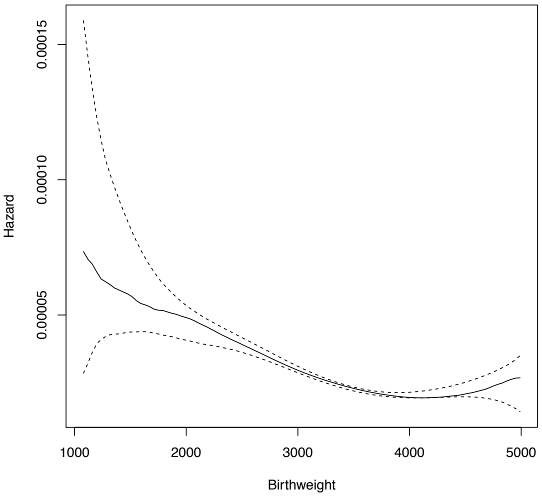


Figure 6: The estimated g when using a truncated sample of individuals with longevity exceeding age 40.

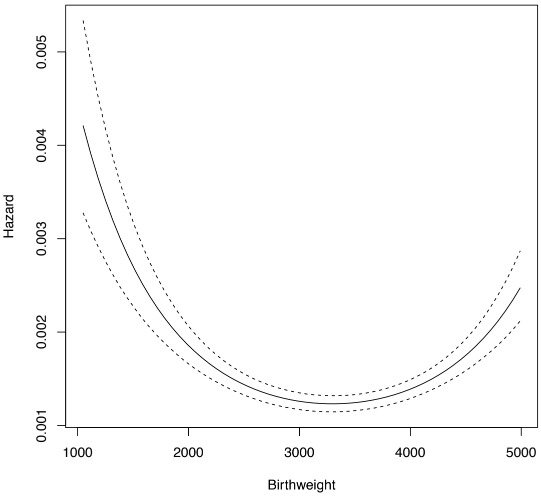


Figure 7: Estimated parametric covariate hazard function.

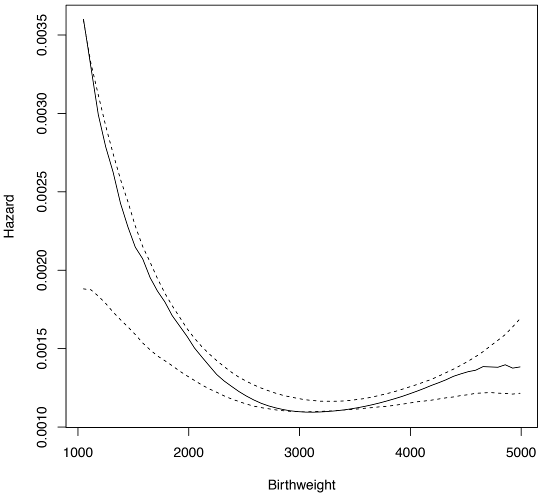
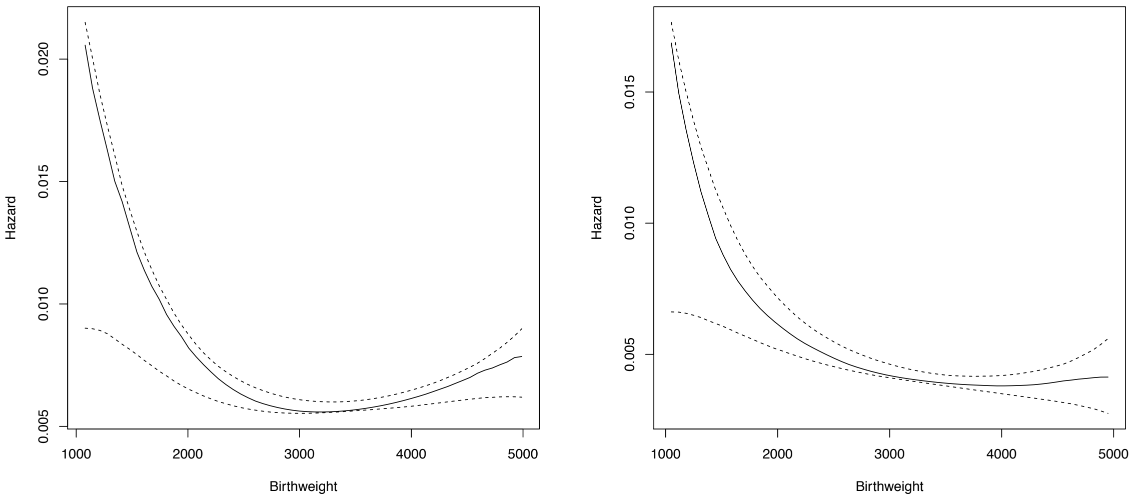


Figure 8: The estimated nonparametric function with social class contained in the covariate vector, included in the baseline hazard, full range of birth weights.



Males. *Females.*

Figure 9: Semiparametric estimation results for g using the local linear estimator, controlling for social class at birth and stratified by gender.

9 Supplementary Materials

In the following we present some additional simulation evidence on the performance of our estimator in terms of variance and bias; that is, both the model (overall) bias and the local bias with optimally chosen bandwidths.

9.1 Empirical coverage and interval length

Table 4 shows the empirical coverage of the bootstrapped, nominal 95% confidence intervals for all four models. We show the coverage for different levels of under- and oversmoothing, both for the "parametric bandwidth" b and the nonparametric bandwidth b^*. Even with undersmoothing, confidence bands do not achieve nominal coverage, although we can get very close, especially for Models 1 and 4 if the degree of undersmoothing is appropriately chosen. Calonico, Cattaneo and Farrell (2018) propose a method for bias correction in density estimation that could potentially alleviate the problems of undercoverage. In Figure 10 we show average interval length for all models depending on the degree of undersmoothing and on the number of observations. Clearly, average interval length decreases with both smaller degrees of undersmoothing and with an increase in the number of observations.

9.2 Results: Variance, bias and empirical coverage

Since in the local linear estimation clearly outperforms the local constant estimator in the simulations above, the results below are calculated using the local linear specification for $\hat{g}$ and $N = 5000$ observations.

Let $m = 1, ..., M$ with M the number of simulations. For each $t = r, z = r'$ calculate an estimate for the model $\hat{\lambda}_m(t, z)$.

- For the **model variance** we calculate the average hazard for each cell r, r' over all simulations
$$\bar{\lambda}\{x_r, z_{r'}\} = \frac{1}{M}\sum_{m=1}^{M} \hat{\lambda}_m\{x_r, z_{r'}\}$$

Model1

	10th perc.	25th perc.	50th perc.	75th perc.	90th perc.
$b\ b^*$	0.57	0.63	0.79	0.67	0.56
$b \cdot 1.2\ b^*$	0.69	0.74	0.82	0.80	0.56
$b\ b^* \cdot 1.2$	0.73	0.78	0.93	0.89	0.62
$b\ b^* \cdot 0.8$	0.66	0.71	0.78	0.75	0.64
$b \cdot 0.8\ b^*$	0.69	0.74	0.86	0.80	0.58
$b\ b^* \cdot 0.2$	0.84	0.91	0.83	0.85	0.84

Model2

	10th perc.	25th perc.	50th perc.	75th perc.	90th perc.
$b\ b^*$	0.57	0.63	0.79	0.67	0.56
$b \cdot 1.2\ b^*$	0.69	0.74	0.82	0.80	0.56
$b\ b^* \cdot 1.2$	0.73	0.78	0.93	0.89	0.62
$b\ b^* \cdot 0.8$	0.66	0.71	0.78	0.75	0.64
$b \cdot 0.8\ b^*$	0.69	0.74	0.86	0.80	0.58
$b\ b^* \cdot 0.2$	0.84	0.91	0.83	0.85	0.84

Model3

	10th perc.	25th perc.	50th perc.	75th perc.	90th perc.
$b\ b^*$	0.57	0.63	0.79	0.67	0.56
$b \cdot 1.2\ b^*$	0.69	0.74	0.82	0.80	0.56
$b\ b^* \cdot 1.2$	0.73	0.78	0.93	0.89	0.62
$b\ b^* \cdot 0.8$	0.66	0.71	0.78	0.75	0.64
$b \cdot 0.8\ b^*$	0.69	0.74	0.86	0.80	0.58
$b\ b^* \cdot 0.2$	0.84	0.91	0.83	0.85	0.84

Model4

	10th perc.	25th perc.	50th perc.	75th perc.	90th perc.
$b\ b^*$	0.57	0.63	0.79	0.67	0.56
$b \cdot 1.2\ b^*$	0.69	0.74	0.82	0.80	0.56
$b\ b^* \cdot 1.2$	0.73	0.78	0.93	0.89	0.62
$b\ b^* \cdot 0.8$	0.66	0.71	0.78	0.75	0.64
$b \cdot 0.8\ b^*$	0.69	0.74	0.86	0.80	0.58
$b\ b^* \cdot 0.2$	0.84	0.91	0.83	0.85	0.84

Table 4: Pointwise Coverage Probabilities: Model 1–4, for the 10th, 25th, 50th, 75th, 90th percentile of the covariate z. b and b^* chosen optimally and multiplied for different degrees of over- undersmoothing.

- and then calculate

$$Var_{\lambda_{r,r'}} = \frac{1}{M}\sum_{m=1}^{M}\left[\hat{\lambda}_m\{x_r, z_{r'}\} - \bar{\lambda}\{x_r, z_{r'}\}\right]^2$$

- For the **model bias** we calculate the difference between the estimated- and the true hazard for each cell.

$$B_{\lambda_{r,r'}} = \frac{1}{M}\sum_{M=1}^{M}|\hat{\lambda}_m\{x_r, z_{r'}\} - \lambda\{x_r, z_{r'}\}|$$

We then calculate the total variance/bias in each model as the sum over all cells, i.e.

$$Var_{model} = \sum_{r=1}^{R}\sum_{r'=1}^{R'} Var_{\lambda_{r,r'}} \text{ and } B_{model} = \sum_{r=1}^{R}\sum_{r'=1}^{R'} B_{\lambda_{r,r'}}. \tag{49}$$

In order to examine the local properties of bias and variance, we calculate the variance and bias for each combination of the covariate and time, i.e. we perform the above calculations but refrain from summing over all cells and examine each cell individually.

We examine the evolution of the model variance for all four models in Tables 5–8. The table in the top panel generally shows the evolution of the variance, the table in the bottom panel the evolution of the bias.

Summarizing the results for all four models, we generally observe that a smaller b means a larger variance for a given b^* and that with an increasing bandwidth b^*, the variance decreases, which is what we would expect. In order of magnitude, increasing b^* matters much more for decreasing the model variance than increasing b, which is also in line with our expectations, as the bandwidth generally only affects the value of the parameter as second order. As for the bias, generally a smaller b^* is associated with a lower bias and the choice of b does not seem to influence the size of the bias by much

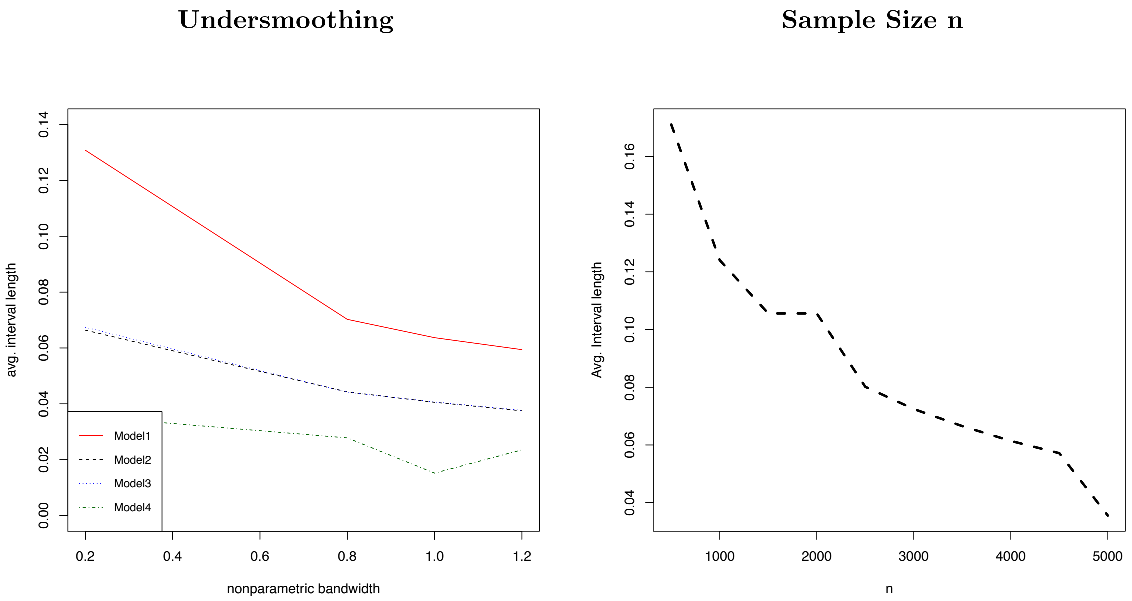


Figure 10: *The left panel shows average interval length over 100 simulations for the intervals at $z = 0.10, 0.25, 0.5, 0.75, 0.9$ for under-/oversmoothing factors $(0.2, 0.8, 1.0, 1.2)$ of the nonparametric bandwidth (b^*) while keeping b at the optimal level. The right panel shows the average interval length over 100 observations for z for an increasing number of observations from $n = 500 - 5000$ for Model 1.*

or even not at all.

The local variance for the four different models is graphically depicted in Figure 11, using one particular bandwidth combination. The local variance differs from model to model, so it is not easy to make generalizations about our estimator's performance based on these simulations.

The local bias is depicted in Figure 12. As with the local variance, the shape of the local bias depends on the model, although in all cases, the bias increases at the boundary values of the covariate z.

b^*/b	0.07	0.09	0.11	0.13	0.15	0.17	0.19	0.21	0.23	0.25	0.27
0.07	0.1725	0.1725	0.1724	0.1725	0.1725	0.1724	0.1724	0.1724	0.1724	0.1723	0.1724
0.09	0.1713	0.1713	0.1712	0.1712	0.1712	0.1712	0.1712	0.1711	0.1711	0.1711	0.1711
0.11	0.1698	0.1698	0.1698	0.1698	0.1698	0.1698	0.1698	0.1697	0.1697	0.1697	0.1697
0.13	0.1682	0.1682	0.1681	0.1681	0.1681	0.1681	0.1681	0.1681	0.1681	0.1680	0.1681
0.15	0.1663	0.1663	0.1663	0.1663	0.1663	0.1663	0.1663	0.1662	0.1662	0.1662	0.1662
0.17	0.1642	0.1642	0.1642	0.1642	0.1642	0.1642	0.1642	0.1641	0.1641	0.1641	0.1641
0.19	0.1620	0.1620	0.1619	0.1619	0.1619	0.1619	0.1619	0.1618	0.1619	0.1618	0.1619
0.21	0.1595	0.1595	0.1595	0.1595	0.1595	0.1595	0.1595	0.1594	0.1594	0.1594	0.1594
0.23	0.1570	0.1570	0.1569	0.1569	0.1569	0.1569	0.1569	0.1568	0.1568	0.1568	0.1568
0.25	0.1542	0.1542	0.1542	0.1542	0.1542	0.1542	0.1542	0.1541	0.1541	0.1541	0.1541
0.27	0.1514	0.1514	0.1514	0.1514	0.1514	0.1513	0.1514	0.1513	0.1513	0.1513	0.1513

b^*/b	0.07	0.09	0.11	0.13	0.15	0.17	0.19	0.21	0.23	0.25	0.27
0.07	0.0454	0.0454	0.0454	0.0454	0.0454	0.0455	0.0454	0.0454	0.0454	0.0454	0.0454
0.09	0.0413	0.0413	0.0413	0.0413	0.0413	0.0413	0.0413	0.0413	0.0412	0.0412	0.0412
0.11	0.0380	0.0380	0.0380	0.0380	0.0380	0.0380	0.0380	0.0380	0.0380	0.0379	0.0380
0.13	0.0354	0.0354	0.0353	0.0354	0.0353	0.0354	0.0353	0.0354	0.0353	0.0353	0.0353
0.15	0.0337	0.0337	0.0337	0.0337	0.0337	0.0337	0.0337	0.0337	0.0337	0.0337	0.0337
0.17	0.0335	0.0335	0.0335	0.0335	0.0335	0.0335	0.0335	0.0335	0.0335	0.0335	0.0335
0.19	0.0345	0.0345	0.0345	0.0346	0.0345	0.0346	0.0345	0.0346	0.0346	0.0346	0.0345
0.21	0.0368	0.0368	0.0368	0.0368	0.0368	0.0368	0.0368	0.0368	0.0368	0.0368	0.0368
0.23	0.0404	0.0404	0.0404	0.0404	0.0404	0.0404	0.0404	0.0404	0.0404	0.0404	0.0404
0.25	0.0454	0.0455	0.0455	0.0455	0.0455	0.0455	0.0455	0.0455	0.0455	0.0455	0.0455
0.27	0.0515	0.0515	0.0515	0.0515	0.0515	0.0515	0.0515	0.0516	0.0516	0.0516	0.0516

Table 5: The top panel shows the evolution of the model variance for Model 1 for b^* (depicted over the rows) and b (depicted over the columns). The table is centered around the optimal, infeasible b^*. The bottom panel shows the model bias.

b^*/b	0.05	0.07	0.09	0.11	0.13	0.15	0.17	0.19	0.21	0.23	0.25
0.05	0.0470	0.0470	0.0470	0.0470	0.0470	0.0470	0.0470	0.0470	0.0470	0.0469	0.0469
0.07	0.0467	0.0467	0.0467	0.0467	0.0467	0.0467	0.0467	0.0467	0.0467	0.0467	0.0467
0.09	0.0465	0.0465	0.0465	0.0465	0.0465	0.0465	0.0465	0.0464	0.0464	0.0464	0.0464
0.11	0.0462	0.0462	0.0462	0.0462	0.0462	0.0462	0.0462	0.0462	0.0462	0.0462	0.0462
0.13	0.0459	0.0459	0.0459	0.0459	0.0459	0.0459	0.0459	0.0459	0.0459	0.0459	0.0459
0.15	0.0455	0.0455	0.0455	0.0455	0.0455	0.0455	0.0455	0.0455	0.0455	0.0455	0.0455
0.17	0.0452	0.0452	0.0451	0.0451	0.0451	0.0451	0.0451	0.0451	0.0451	0.0451	0.0451
0.19	0.0447	0.0447	0.0447	0.0447	0.0447	0.0447	0.0447	0.0447	0.0447	0.0447	0.0447
0.21	0.0442	0.0442	0.0442	0.0442	0.0442	0.0442	0.0442	0.0442	0.0442	0.0442	0.0442
0.23	0.0437	0.0437	0.0437	0.0437	0.0437	0.0437	0.0437	0.0437	0.0437	0.0437	0.0437
0.25	0.0431	0.0431	0.0431	0.0431	0.0431	0.0431	0.0431	0.0431	0.0431	0.0431	0.0431

b^*/b	0.05	0.07	0.09	0.11	0.13	0.15	0.17	0.19	0.21	0.23	0.25
0.05	0.0430	0.0430	0.0430	0.0430	0.0430	0.0430	0.0430	0.0430	0.0430	0.0430	0.0431
0.07	0.0369	0.0369	0.0369	0.0369	0.0369	0.0369	0.0369	0.0369	0.0369	0.0369	0.0369
0.09	0.0328	0.0328	0.0328	0.0328	0.0328	0.0328	0.0328	0.0328	0.0328	0.0329	0.0329
0.11	0.0301	0.0301	0.0301	0.0302	0.0302	0.0302	0.0302	0.0302	0.0302	0.0302	0.0302
0.13	0.0288	0.0288	0.0288	0.0289	0.0289	0.0289	0.0289	0.0289	0.0289	0.0289	0.0289
0.15	0.0287	0.0287	0.0287	0.0287	0.0287	0.0287	0.0287	0.0288	0.0288	0.0288	0.0288
0.17	0.0294	0.0294	0.0294	0.0295	0.0295	0.0295	0.0295	0.0295	0.0295	0.0295	0.0295
0.19	0.0308	0.0308	0.0308	0.0308	0.0308	0.0308	0.0308	0.0308	0.0309	0.0309	0.0309
0.21	0.0325	0.0325	0.0325	0.0325	0.0326	0.0326	0.0326	0.0326	0.0326	0.0326	0.0326
0.23	0.0345	0.0345	0.0346	0.0346	0.0346	0.0346	0.0346	0.0346	0.0346	0.0346	0.0346
0.25	0.0368	0.0368	0.0368	0.0368	0.0368	0.0369	0.0369	0.0369	0.0369	0.0369	0.0369

Table 6: The top panel shows the evolution of the model variance for Model 2 for b^* (depicted over the rows) and b (depicted over the columns). The table is centered around the optimal, infeasible b^*. The bottom panel shows the model bias.

b^*/b	0.05	0.07	0.09	0.11	0.13	0.15	0.17	0.19	0.21	0.23	0.25
0.05	0.3648	0.3646	0.3646	0.3645	0.3644	0.3643	0.3642	0.3641	0.3640	0.3639	0.3638
0.07	0.3632	0.3630	0.3629	0.3628	0.3628	0.3627	0.3626	0.3625	0.3624	0.3623	0.3621
0.09	0.3614	0.3612	0.3611	0.3611	0.3610	0.3609	0.3608	0.3607	0.3606	0.3605	0.3604
0.11	0.3593	0.3591	0.3591	0.3590	0.3589	0.3588	0.3587	0.3586	0.3585	0.3584	0.3583
0.13	0.3569	0.3567	0.3566	0.3565	0.3565	0.3564	0.3563	0.3562	0.3561	0.3560	0.3559
0.15	0.3540	0.3539	0.3538	0.3537	0.3536	0.3535	0.3535	0.3534	0.3533	0.3531	0.3530
0.17	0.3508	0.3507	0.3506	0.3505	0.3504	0.3503	0.3503	0.3502	0.3501	0.3500	0.3498
0.19	0.3472	0.3471	0.3470	0.3469	0.3469	0.3468	0.3467	0.3466	0.3465	0.3464	0.3463
0.21	0.3433	0.3431	0.3430	0.3430	0.3429	0.3428	0.3428	0.3427	0.3426	0.3425	0.3423
0.23	0.3390	0.3388	0.3387	0.3387	0.3386	0.3385	0.3384	0.3384	0.3383	0.3382	0.3380
0.25	0.3343	0.3342	0.3341	0.3340	0.3340	0.3339	0.3338	0.3337	0.3336	0.3335	0.3334

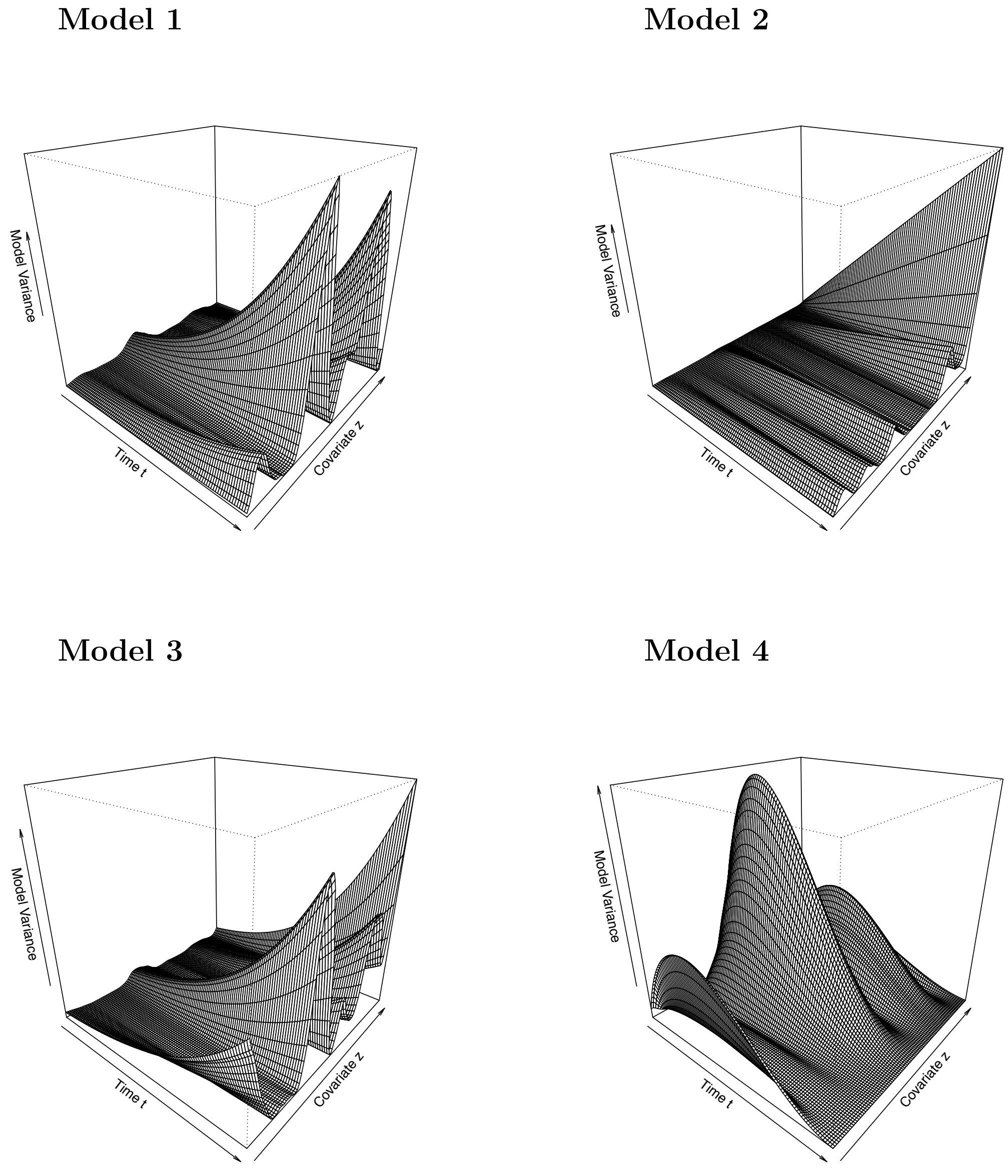
b^*/b	0.05	0.07	0.09	0.11	0.13	0.15	0.17	0.19	0.21	0.23	0.25
0.05	0.0706	0.0705	0.0705	0.0705	0.0705	0.0705	0.0704	0.0704	0.0704	0.0704	0.0704
0.07	0.0626	0.0625	0.0625	0.0625	0.0625	0.0624	0.0624	0.0624	0.0624	0.0623	0.0623
0.09	0.0578	0.0578	0.0578	0.0577	0.0577	0.0577	0.0577	0.0577	0.0576	0.0576	0.0576
0.11	0.0543	0.0543	0.0543	0.0543	0.0543	0.0542	0.0542	0.0542	0.0542	0.0541	0.0541
0.13	0.0522	0.0521	0.0521	0.0521	0.0521	0.0521	0.0521	0.0520	0.0520	0.0520	0.0520
0.15	0.0519	0.0518	0.0518	0.0518	0.0518	0.0518	0.0518	0.0517	0.0517	0.0517	0.0517
0.17	0.0534	0.0534	0.0534	0.0534	0.0534	0.0533	0.0533	0.0533	0.0533	0.0533	0.0532
0.19	0.0567	0.0567	0.0567	0.0567	0.0567	0.0567	0.0566	0.0566	0.0566	0.0566	0.0566
0.21	0.0615	0.0615	0.0615	0.0614	0.0614	0.0614	0.0614	0.0614	0.0614	0.0614	0.0614
0.23	0.0672	0.0672	0.0672	0.0672	0.0672	0.0672	0.0672	0.0672	0.0672	0.0672	0.0672
0.25	0.0736	0.0736	0.0736	0.0735	0.0735	0.0735	0.0735	0.0735	0.0735	0.0735	0.0735

Table 7: The top panel shows the evolution of the model variance for Model 3 for b^* (depicted over the rows) and b (depicted over the columns). The table is centered around the optimal, infeasible b^*. The bottom panel shows the model bias.

b^*/b	0.13	0.15	0.17	0.19	0.21	0.23	0.25	0.27	0.29	0.31	0.33
0.13	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014
0.15	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014
0.17	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014
0.19	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014
0.21	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013
0.23	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013
0.25	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013
0.27	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013
0.29	0.0012	0.0012	0.0012	0.0012	0.0012	0.0012	0.0012	0.0012	0.0012	0.0012	0.0012
0.31	0.0012	0.0012	0.0012	0.0012	0.0012	0.0012	0.0012	0.0012	0.0012	0.0012	0.0012
0.33	0.0012	0.0012	0.0012	0.0012	0.0012	0.0012	0.0012	0.0012	0.0012	0.0012	0.0012

b^*/b	0.13	0.15	0.17	0.19	0.21	0.23	0.25	0.27	0.29	0.31	0.33
0.13	0.0128	0.0128	0.0128	0.0128	0.0128	0.0128	0.0128	0.0128	0.0128	0.0128	0.0128
0.15	0.0123	0.0123	0.0123	0.0123	0.0123	0.0123	0.0123	0.0123	0.0123	0.0123	0.0123
0.17	0.0120	0.0120	0.0120	0.0120	0.0120	0.0120	0.0120	0.0120	0.0120	0.0120	0.0120
0.19	0.0118	0.0118	0.0118	0.0118	0.0118	0.0118	0.0118	0.0118	0.0118	0.0118	0.0118
0.21	0.0117	0.0117	0.0117	0.0117	0.0117	0.0117	0.0117	0.0117	0.0117	0.0117	0.0117
0.23	0.0116	0.0116	0.0116	0.0116	0.0116	0.0116	0.0116	0.0116	0.0116	0.0116	0.0116
0.25	0.0117	0.0117	0.0117	0.0117	0.0117	0.0117	0.0117	0.0117	0.0117	0.0117	0.0117
0.27	0.0118	0.0118	0.0118	0.0118	0.0118	0.0118	0.0118	0.0118	0.0118	0.0118	0.0118
0.29	0.0121	0.0121	0.0121	0.0121	0.0121	0.0121	0.0121	0.0121	0.0121	0.0121	0.0121
0.31	0.0125	0.0125	0.0125	0.0125	0.0125	0.0125	0.0125	0.0125	0.0125	0.0125	0.0125
0.33	0.0130	0.0130	0.0130	0.0130	0.0130	0.0130	0.0130	0.0130	0.0130	0.0130	0.0130

Table 8: The top panel shows the evolution of the model variance for Model 4 for b^* (depicted over the rows) and b (depicted over the columns). The table is centered around the optimal, infeasible b^*. The bottom panel shows the model bias.



Model 1

Model 2

Model 3

Model 4

Figure 11: *The evolution of local variance for each combination of the covariate z and time t for a fixed pair of (optimally chosen) bandwidths b, b^* for all four models.*

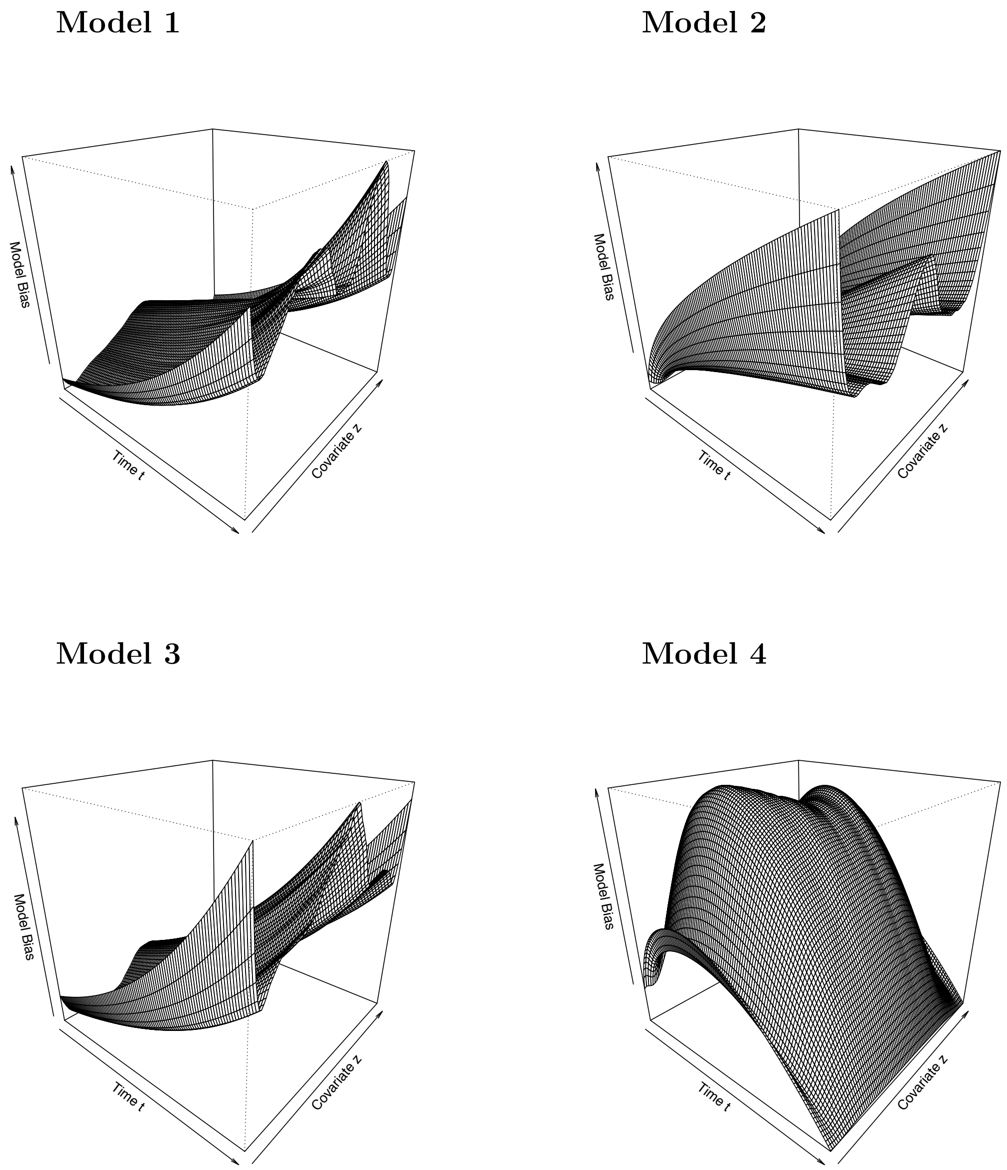


Figure 12: *The evolution of local bias for each combination of the covariate z and time t for a fixed pair of (optimally chosen) bandwidths b, b^* for all four models.*