Statistical Modeling with Spline Functions Methodology and Theory

Mark H. Hansen
University of California at Los Angeles

Jianhua Z. Huang
University of Pennsylvania

Charles Kooperberg
Fred Hutchinson Cancer Research Center

Charles J. Stone
University of California at Berkeley

Young K. Truong
University of North Carolina at Chapel Hill

January 5, 2006

7
Survival Analysis

7.1 An example

7.1.1 The bone marrow transplant data

Bone marrow transplantation is a procedure that is sometimes given to patients that have certain types of cancer, in particular, the so-called blood cancers such as leukemia, lymphoma, and multiple myeloma. Bone marrow transplantation can be either autologous, in which bone marrow from the patient is reinjected after treatment, or allogenic, in which bone marrow from another, healthy, person is inserted in the cancer patient. Our example uses data on allogenic bone marrow transplants. There are many complications with such bone marrow transplants. For example, the cancer patient may reject the transplanted material or, after some time, the cancer may reoccur. One particular well known form of rejection is Graft Versus Host Disease (GVHD). To reduce the chance of such rejection occurring, if the donor is not the patient, the bone marrow of the donor has to match that of the patient to a high degree. The chance that such a match occurs is considerably larger when the donor is a sibling than when the donor is an unrelated individual.

Bone marrow transplantation was first carried out in the late 1960s and was at that time highly experimental. Originally, virtually all transplants used siblings who matched at certain HLA loci, the so-called major histocompatibility or MHC loci. Since then bone marrow transplantation, while still a very complicated procedure, has become more standard.

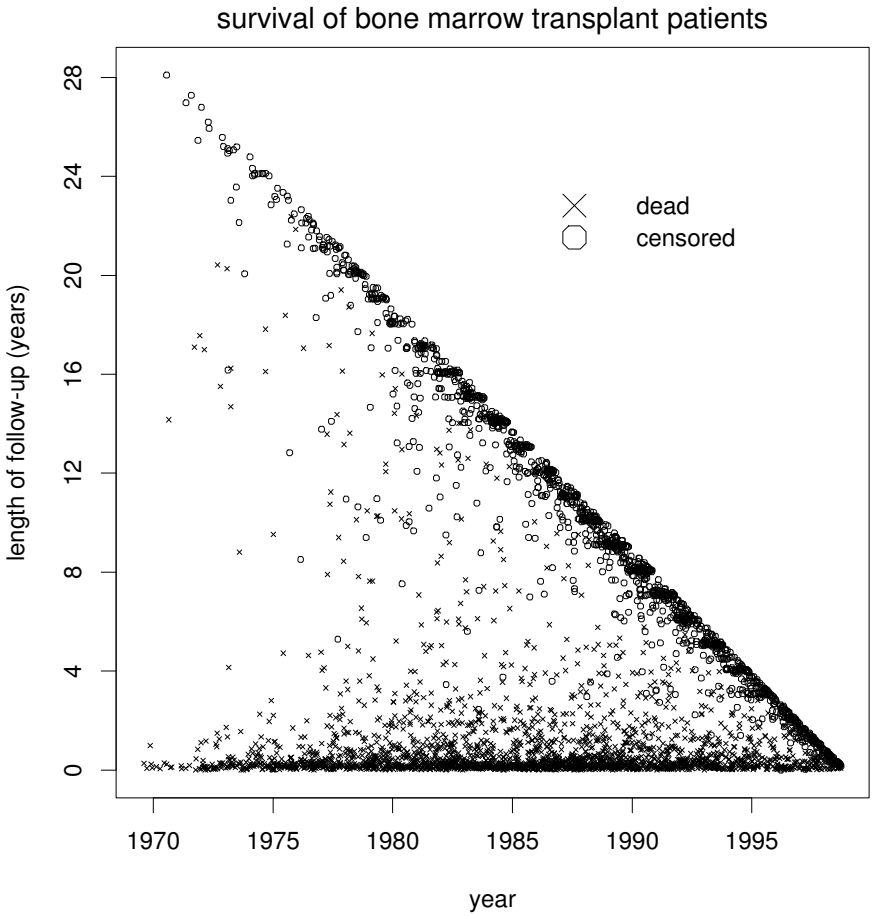

FIGURE 7.1. Survival of matched sibling bone marrow transplant patients at the Fred Hutchinson Cancer Research Center.

In Figure 7.1 we show survival data for 3887 bone marrow transplant patients, for whom the donor was a matched sibling and who received their transplant at the Fred Hutchinson Cancer Research Center (FHCRC) in Seattle between 1969, when bone marrow transplantation was first carried out, and September 1998. Our data set includes follow-up data through November 1998. At that time 2405 patients were known to be dead; the other patients were either known to be alive or lost to follow up. The diagonal in Figure 7.1 is clearly an artifact: it represents patients who were known to be alive as of a last contact, which typically was shortly before November 1998.

However, Figure 7.1 does not tell the complete story. In particular, not all patients have the same type of cancer. Our data set contains patients with a variety of leukemias, lymphomas, and myelomas. For our analysis we have

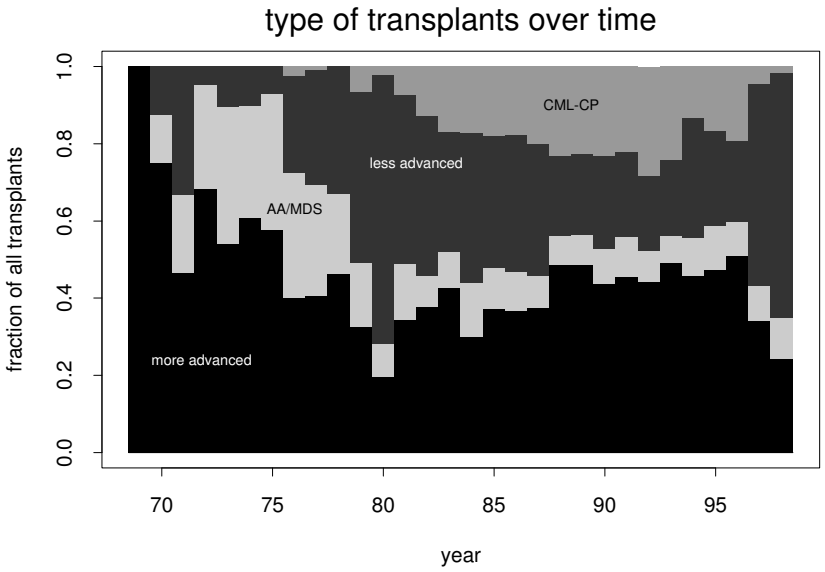

FIGURE 7.2. Disease types of the matched sibling bone marrow transplant patients at the Fred Hutchinson Cancer Research Center.

divided the patients into four groups: those with more advanced disease, those with aplastic anemia or myelodysplastic syndrome (AA/MDS, a less advanced disease category), those with chronic myeloid leukemia-chronic phase (CML-CP, a less advanced disease category), and those with other less advanced disease categories. As can be seen from Figure 7.2 the fractions of the patients in particular disease groups that were treated have changed considerably over time. This is relevant, as the survival probabilities are very different for the different disease categories: 21% of the patients with more advanced diseases are still alive, versus 56% of the patients with CML-CP, 66% of the patients with AA/MDS, and 41% of the patients with other less advanced diseases.

The number of matched sibling bone marrow transplants at the FHCRC has varied considerably over time, as can be seen from Figure 7.3. In the late 60s and 70s, a large fraction of transplants done in the world were done at the FHCRC. Later, other institutions started to carry out matched sibling bone marrow transplants. At the same time, the FHCRC started to carry out some of the more complicated transplants between unrelated individuals and reduced the number of more "standard" matched sibling transplants. Other things changed as well: patients became older (Figure 7.4), and the other medications/treatments that patients received in conjunction with their bone marrow transplant changed (Figure 7.5).

The medications and treatments that we use in our analysis are:

- An indicator for the use of total body irradiation (TBI). The conditioning regimen is a treatment that a patient undergoes prior to the

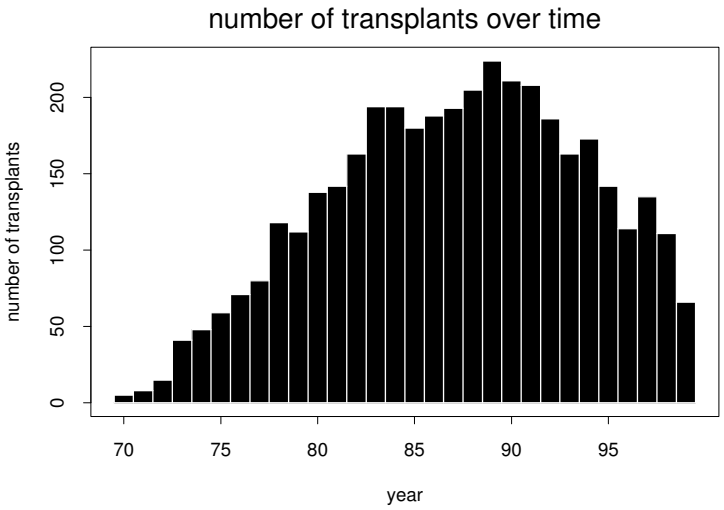

FIGURE 7.3. Number of matched sibling bone marrow transplant patients at the Fred Hutchinson Cancer Research Center.

bone marrow transplantation. It is used, essentially, to wipe out the immune system of the patient (and hopefully get rid of the disease as well). The various conditioning regimens consist of chemotherapy and/or total body irradiation (TBI). In the current analysis we use only TBI as a covariate.

- The use of acute GVHD prophylaxis. Almost all patients now receive methotrexate and cyclosporine, but this has not always been the case. The use of acute GVHD prophylaxis is generally regarded as being superior to known alternatives for prevention of acute GVHD.

- The use of fluconazole reflects a change in standard practice at the FHCRC. All patients now receive this drug for the purpose of prophylaxis for fungal infections. There was a randomized trial that compared fluconazole to placebo. Before this time the drug was not available to patients, but after the completion of the trial it was adopted as part of standard practice.

- The same also holds for the use of gancyclovir, although this drug is used only in patients who are cytomegalovirus (CMV) sero-positive. The drug is used for prophylaxis for CMV infections, and its use has drastically reduced the incidence of CMV disease (which can be fatal).

Some variables are not shown in these plots: we also have information about the gender of the patient, the age and the gender of the donor, and the dose of bone marrow that was received.

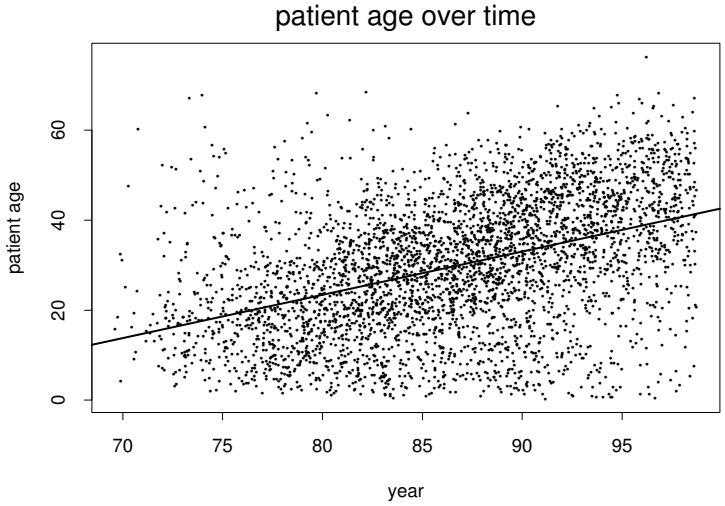

FIGURE 7.4. Age of matched sibling bone marrow transplant patients at the Fred Hutchinson Cancer Research Center (with regression line).

The goal in a survival analysis of a data set like this one is to find out which of the covariates the general survival depends on.[1] Ideally, we are able to identify the conditional distribution of the survival times given a set of covariates. In this chapter we will describe the hazard regression (Hare) methodology, which uses splines to model this conditional distribution.

7.1.2 Background

Consider data involving a positive response variable that may be (right-) censored and one or more covariates. We refer to the original, uncensored response variable as the *survival time* and think of it as having a conditional density function given the values of the covariates that is positive on $[0, \infty)$. The hazard function and its logarithm corresponding to this density function are referred to as the conditional hazard function and conditional log-hazard function, respectively.

A main goal in survival analysis is to determine how survival probabilities depend on the covariates. The most popular way to model this is dependence is by means of the proportional hazards model of Cox (1972), which assumes that the conditional log-hazard function is an additive function of time and the vector of covariates or, equivalently, that the conditional hazard function is a multiplicative function of time and the vector of co-

[1] It could also be of interest to look at other outcomes, such as disease free survival or time until GVHD. For these types of outcomes, death by other causes would be considered censoring. In our analysis we only consider general survival.

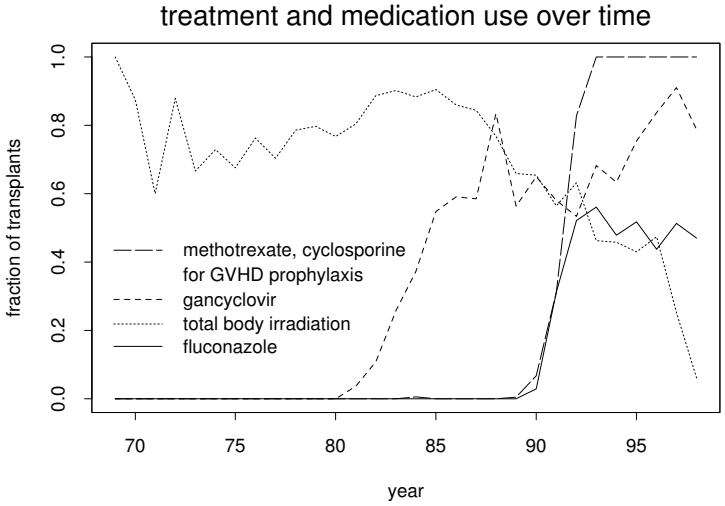

FIGURE 7.5. Additional treatments and medications received by matched sibling bone marrow transplant patients at the Fred Hutchinson Cancer Research Center.

variates. After the dependence on the covariates is estimated, the baseline distribution function (the conditional distribution function of the survival time for a particular fixed set of covariates) can be estimated using the Breslow estimator Breslow (1972), which is a generalization of the Kaplan–Meier estimator (Kaplan and Meier 1958).

The main complication in analyzing survival data is *censoring*. In November 1998, when the bone marrow transplant data set was put together, 1541 of the 3976 bone marrow recipients were still alive. Thus, for these patients the actual survival time is not known; rather it is known that they survived at least until November 1998 (or until they lost contact with the FHCRC).

In this chapter we focus primarily on the Hare (Hazard Regression) methodology for survival analysis, which was introduced in Kooperberg, Stone, and Truong (1995a). Kooperberg and Clarkson (1997) extended the Hare methodology to handle interval censored data and time-dependent covariates. These extensions are discussed in Section 7.4.

7.1.3 Linear models for the conditional log-hazard function

In this section we describe a linear model for the conditional log-hazard function that will be the basis for Hare. In the simplest examples (such as the bone marrow transplant data) all covariates are fixed throughout the study. The form of the log-likelihood does not change much when some of the covariates depend on time. Such *time-dependent* covariates often occur in medical studies, where covariates are measured at regular times during the study and may change between measurements. Initially, we assume that

none of the covariates are time-dependent. Time-dependent covariates will be discussed in detail in Section 7.4.2.

Let M be a nonnegative integer and let T be a positive random variable whose distribution may depend on a vector $\boldsymbol{x} = (x_1, \ldots, x_M)$ of M covariates. Note that T is commonly referred to as the survival time or the failure time. Suppose $\boldsymbol{x}$ lies in the subset $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_M$ of $\mathbb{R}^M$. The conditional hazard function $\lambda(\cdot|\boldsymbol{x})$ of T given $\boldsymbol{x}$, which is defined by

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

is assumed to exist and be positive on $(0, \infty)$. In survival analysis the (conditional) hazard function is often of interest, as $\lambda(t|\boldsymbol{x})dt$ can be interpreted as the "probability that someone dies in the next time interval of infinitesimal length dt, given that he is alive at time t". Let $f(\cdot|\boldsymbol{x})$ and $F(\cdot|\boldsymbol{x})$ denote the conditional density and distribution function, respectively. We refer to

$$\alpha(t|\boldsymbol{x}) = \log \lambda(t|\boldsymbol{x}) \tag{7.1.1}$$

and

$$\Lambda(t|\boldsymbol{x}) = \int_0^t \lambda(s|\boldsymbol{x})ds \tag{7.1.2}$$

as defining the conditional log-hazard and cumulative hazard function, respectively. Let $S(t|\boldsymbol{x}) = P(T > t|\boldsymbol{x})$ denote the conditional survival function.

Now

$$\lambda(t|\boldsymbol{x}) = \frac{f(t|\boldsymbol{x})}{1 - F(t|\boldsymbol{x})} = \frac{f(t|\boldsymbol{x})}{S(t|\boldsymbol{x})} = -\frac{d}{dt} \log S(t|\boldsymbol{x})$$

and $S(0|\boldsymbol{x}) = 1$, so

$$1 - F(t|\boldsymbol{x}) = S(t|\boldsymbol{x}) = \exp(-\Lambda(t|\boldsymbol{x})) \tag{7.1.3}$$

and hence

$$f(t|\boldsymbol{x}) = \lambda(t|\boldsymbol{x}) \exp(-\Lambda(t|\boldsymbol{x})). \tag{7.1.4}$$

In Hare, polynomial splines and selected tensor products are used to obtain a linear model for $\alpha(t|\boldsymbol{x})$. By modeling $\alpha(t|\boldsymbol{x})$, as opposed to $\lambda(t|\boldsymbol{x})$, $f(t|\boldsymbol{x})$ or $F(t|\boldsymbol{x})$, we do not have to worry about positivity constraints. Also, a model for the log-hazard function leads to a concave log-likelihood function, even in the context of right censoring[2]. This is not true, for example, for a linear model for the log-density function.

The most popular model for analyzing survival data is undoubtedly the Cox proportional hazards model (Cox 1972), which is given by

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

[2]The log-likelihood function is also concave when there is left truncation, as we will see in Section 7.4.3.

For the proportional hazards model, the regression function $\Psi(\boldsymbol{x})$ can be estimated using a partial likelihood function if it is given a parametric form that is independent of the (baseline) hazard function $\lambda_0(t)$. The most commonly used form for $\Psi(\boldsymbol{x})$ is

$$\Psi(\boldsymbol{x}) = \exp \sum_j \theta_j x_j,$$

while $\lambda_0(t)$ is left unspecified. The interpretation is that the hazard of someone with covariates $\boldsymbol{x}$ is proportional to the baseline hazard $\lambda_0(t)$ with hazard-ratio $\Psi(\boldsymbol{x})$. For this interpretation, the baseline hazard function is not important, although it is assumed that the hazards of all participants are indeed proportional to each other, and the integrated baseline hazard function

$$\Lambda_0(t) = \int_0^t \lambda_0(s)ds$$

can be estimated using the Breslow (1972) estimate. A proportional hazards model can be written as a linear model for the log-hazard function; thus by modeling the log-hazard function, Hare includes proportional hazard models as a special case.

Basis functions in Hare can depend on time, on one covariate, on two covariates, or on time and one covariate. If none of the basis functions depend on both time and a covariate, Hare effectively yields a proportional hazards model in which the baseline hazard function is modeled via the basis functions that only depend on time.

7.1.4 A Hare model for the bone marrow transplant data

The Hare (hazard regression) program was applied to a data set derived from the database of matched sibling bone marrow transplants at the FHCRC. In Figure 7.6 we show the conditional survival function $\widehat{S}(t|\boldsymbol{x})$ and the corresponding hazard function $\widehat{\lambda}(t|\boldsymbol{x}) = \widehat{f}(t|\boldsymbol{x})/(1 - \widehat{F}(t|\boldsymbol{x}))$ for the survival time of patients with several different types of cancer who received a bone marrow transplant in 1993 at age 30 and who got total body irradiation and also got methotrexate and cyclosporine as a prophylaxis against GVHD. Hare yielded a conditional distribution function that depends on the year of transplant, the age of the patient, whether a prophylaxis was applied, and on the type of cancer. In addition, Hare considered a number of other covariates, including patient and donor gender, age of the donor, dose of transplant, and whether various medications were used, which were not included in the final model. For the Hare methodology it is very easy to obtain graphical representations of the risk for any individual patient like those shown in Figure 7.6.

The model that was obtained by Hare is not a proportional hazards model, as is evident from the crossing hazard functions on the right side of

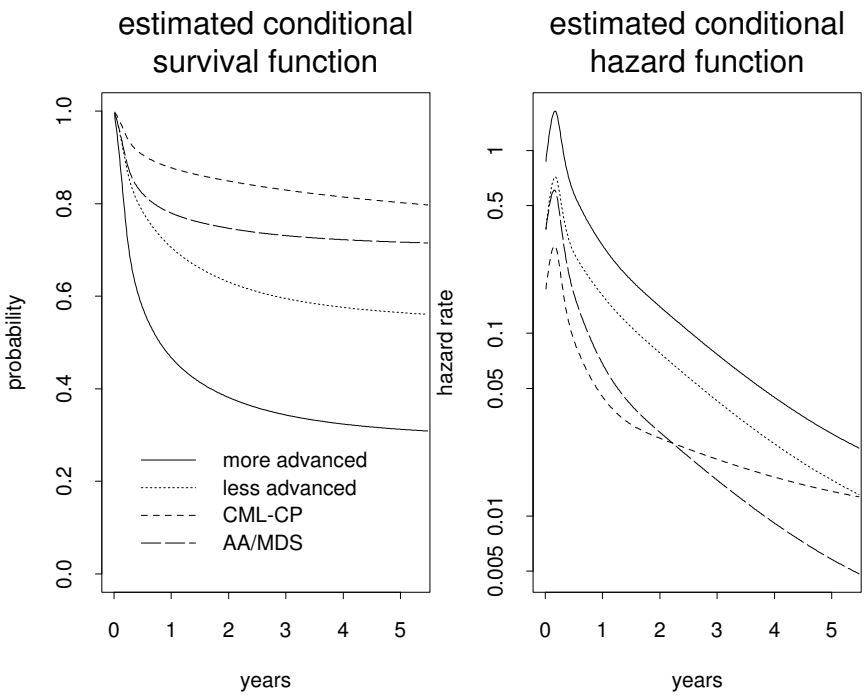

FIGURE 7.6. Estimated conditional survival and hazard functions for a 30 year old cancer patient who received a bone marrow transplant with prophylaxis in 1993.

Figure 7.6. The Hare fit suggests that there is a substantial risk of dying around 60 days after a patient is given a bone marrow transplant, but this risk decreases considerably over time and it depends significantly on the disease group. It does seem plausible that the peak around 60 days is associated with some increased risk related to the transplantation.

From the crossing hazard curves in Figure 7.6 it is clear that Hare did not yield a proportional hazards model. Actually, no two hazard functions in Figure 7.6 are proportional to each other. The fitted hazard functions also include an interaction between time and patient age: the hazard rate decreases faster for younger patients than for older patients. The effect of the use of prophylaxis interacts with the disease type. For the three less advanced disease groups, the use of prophylaxis reduces the hazard rate by 31%, but for the more advanced disease group the reduction is only 4%.

We will get back to this example in Section 7.3, where we will compare the results of Hare with a Hare model that does not include basis functions that depend on time and a covariate and is thus a proportional hazards model. There we will also assess the "significance" of the effects alluded to in this section.

7.2 The Hare methodology

7.2.1 The Hare model

Let $1 \leq p < \infty$, let $\mathbb{G}$ be a p-dimensional linear space of functions on $[0, \infty) \times \mathcal{X}$ such that $g(\cdot|\boldsymbol{x})$ is bounded on $[0, \infty)$ for $g \in \mathbb{G}$ and $\boldsymbol{x} \in \mathcal{X}$, and let $B_1, \ldots, B_p$ be a basis of this space. Consider the model

$$\alpha(t|\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{j=1}^{p} \theta_j B_j(t|\boldsymbol{x}), \qquad t \geq 0, \tag{7.2.1}$$

for the conditional log-hazard function, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$. Given $\boldsymbol{\theta} \in \mathbb{R}^p$ we define $\lambda(t|\boldsymbol{x}; \boldsymbol{\theta})$, $\Lambda(t|\boldsymbol{x}; \boldsymbol{\theta})$, $S(t|\boldsymbol{x}; \boldsymbol{\theta})$, $F(t|\boldsymbol{x}; \boldsymbol{\theta})$, and $f(t|\boldsymbol{x}; \boldsymbol{\theta})$ by imitating equations (7.1.1)–(7.1.4). We refer to model (7.2.1) as the "Hare model". Given such a model, we can obtain the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ in straightforward manner. Details are postponed until Section 7.2.4, after we have discussed the allowable spaces and model selection for Hare.

7.2.2 Allowable spaces

Piecewise linear Hare models

The family of allowable spaces for the version of Hare that employs linear splines and their tensor products is almost identical to the family of allowable spaces for PolyMARS as discussed in Chapter 3. Basis functions can depend on time and/or one or more covariates. In particular, the two differences between the allowable spaces for Hare and PolyMARS are that for Hare:

1. there is only one coefficient associated with each basis function;

2. we require all basis functions that depend on time to be constant when t goes to ∞, as this eliminates the need for imposing a positivity constraint on the coefficient of the right tail of $\alpha(t|\cdot)$.

We require a constant tail in time to guarantee that $\lim_{t\to\infty} \Lambda(t|\cdot) = \infty$ or, equivalently, that $\lim_{t\to\infty} S(t|\cdot) = 0$. To achieve the constant tail, we use basis functions of the form $B(t) = (t_k - t)_+$ in time, and we do not use a basis function $B(t) = t$. In the algorithm with stepwise addition and deletion of basis functions, basis functions that introduce a knot in time can thus be entered at any time into the model. Basis functions that can occur in a piecewise linear Hare model have the form 1, $(t_k - t)_+$, x_l, $(x_l - x_{lk})_+$, $(t_k - t)_+ x_l$, $(t_k - t)_+ (x_l - x_{lk})_+$, $x_l x_m$, $x_l (x_m - x_{mk})_+$, and $(x_l - x_{lk})_+ (x_m - x_{mk})_+$, where x_l and x_m are covariates, t_k is a knot in time, and x_{lk} and x_{mk} are knots in covariates l and m, respectively. A basis

function that depends on knots in a covariate can be in a Hare model only if the corresponding linear basis function in that covariate is in the model, and a tensor product basis function can be in a Hare model only if the corresponding basis functions in one variable are in the model.

The advantage of using linear splines for time in Hare is that all integrals that are required for the maximum likelihood estimation can be computed exactly without numerical integration. This is no longer true when either some of the covariates are time-dependent in a non-piecewise constant manner (see Section 7.4.2) or basis functions in time are not piecewise linear.

Piecewise cubic Hare models

The implementation of Hare by Insightful Corporation (see Section 7.6) also allows for the use of cubic splines and their tensor products for one or more of the covariates or time[3]. In this implementation, all tails of basis functions are required to be linear, except for the tails of basis functions in time when t goes to ∞, which are again required to be constant. For each covariate except for time this leads to basis functions that are identical to the Logspline basis (Chapter 6) for a covariate. A problem for Hare is that when knots are first entered in such a covariate, three knots have to be selected simultaneously[4]. The reason for this requirement is that a nonlinear twice continuously differentiable cubic spline with linear tails requires at least three jumps in the third derivative. In the Insightful implementation this problem is solved by considering only six order statistics as the first candidate knots; for the addition of further knots a search algorithm similar to the one for Polyclass (see Section 5.5.2 and the additional remark for cubic splines in Section 6.7.2) is employed. For time, again a similar basis to the Logspline basis is used, except that the basis function $B_2(t)$ is not considered.

7.2.3 Model selection

Model selection in Hare is carried out in a manner similar to that discussed in earlier chapters using a procedure that involves stepwise addition and stepwise deletion of basis functions and BIC to select the final model. Initially, we choose $\mathbb{G}$ as the minimum allowable space (which for Hare is the model $\alpha(t|\boldsymbol{x}) = \theta_0$). Then we proceed with stepwise addition. Here we successively replace the $(p-1)$-dimensional allowable space $\mathbb{G}_0$ by a p-dimensional allowable space $\mathbb{G}$ containing $\mathbb{G}_0$ as subspace, choosing among

[3]The Insightful implementation also allows for quadratic splines and step functions, which are not discussed here.

[4]This is not a problem in Logspline, since the corresponding initial model already contains knots, although the minimal allowable model for Logspline does contain three knots for exactly the same reason. A similar problem occurs in Lspec (Chapter 8).

the various candidates for a new basis function by a heuristic search that is designed approximately to maximize the corresponding Rao statistic.

During each step of the stepwise addition of basis functions, we consider all allowable spaces $\mathbb{G}$ of dimension $p+1$ that contain the current allowable space $\mathbb{G}_0$ of dimension p. Among the candidate basis functions we choose the one that corresponds to the largest Rao statistic. For the version of Hare that employs piecewise linear splines this means that we need to compute Rao statistics for the addition of basis functions that (i) are linear in a covariate, (ii) are tensor products of two existing basis functions involving different, single variables, (iii) introduce a new knot in time, or (iv) introduce a new knot in a covariate that is already in the model (we optimize the knot location using a procedure similar to the one described in Section 6.7.2).

For the version of Hare that employs cubic splines, the situation is more complicated. The first basis function in a covariate or time that depends on knots depends on three knots (for time) or two knots (for a covariate), as the basis function needs to be linear in one tail and constant in the other tail (for time) or linear in both tails (for a covariate). To keep the computations feasible, for the first knots in a covariate we only consider about 6 to 8 selected quantiles of the values of that covariate as knots. For the addition of subsequent knots we again use a procedure similar to the one described in Section 6.7.2). When we enter interactions between basis functions that depend on knots, we consider all allowable spaces that can be obtained by adding tensor products of two functions in $\mathbb{G}_0$ as a basis function. For example, suppose that a tensor product basis function will have knots in a covariate x_l. When cubic splines are used, this new basis function will have two knots in x_l. These two knots can be any two of the knots for x_l that are in the model; for example, if there currently are four knots in covariate x_l, there are $\binom{4}{2} = 6$ sets of knots that could be used for a bivariate basis function depending on x_l.

Upon stopping the stepwise addition process (for Hare the default value for the maximum dimension is given by $K_{\max} = \min(6n^{0.2}, n/4, 50)$; see Section 5.5.1 for a motivation of this type of default), we carry out stepwise deletion. Here we successively replace the p-dimensional allowable space $\mathbb{G}$ by a $(p-1)$-dimensional allowable subspace $\mathbb{G}_0$ until we arrive at the minimal allowable space, at each step choosing the candidate space $\mathbb{G}_0$ so that the Wald statistic for a basis function that is in $\mathbb{G}$ but not in $\mathbb{G}_0$ is smallest in magnitude. Again, this is more complicated using cubic splines than using linear splines, with complications for the cubic spline procedure similar to those during the addition of tensor product basis functions.

During the combination of stepwise addition and stepwise deletion, we get a sequence of models from which we select the final model by minimizing BIC, that is, AIC with the default penalty parameter $\log n$ (3.2.29).

7.2.4 Fitting Hare models

In survival analysis applications we typically try to model the distribution of the time when a particular event T happens. Sometimes we may not observe T, because the event to which T corresponds does not happen before the end of the study, or because another event (such as death from another cause) happens that prevents T from being observed. In such situations we still know that T is larger than the time of the end of the study or the time when the other event happens. Such a lower bound on T is known as a censoring time. Typically, we observe either T or the censoring time, and we know which of the two we observe.

Let the survival time T be a positive random variable whose distribution may depend on a vector $\boldsymbol{x}$ of covariates. Let the censoring time C be another positive random variable. Set $Y = \min(T, C)$ and $\delta = \text{ind}(T \leq C)$. It is assumed that T and C are conditionally independent and that T has conditional density function $f(\cdot|\boldsymbol{x})$ given $\boldsymbol{x} \in \mathcal{X}$. The random variable Y is said to be uncensored or censored according as $\delta = 1$ or $\delta = 0$.

Let $g(.|\boldsymbol{x})$ and $G(.|\boldsymbol{x})$ denote the conditional density (or probability) and distribution function of the censoring time C. For a random variable Y that is censored at y, we know that $T > y$ and $C = y$. Because of the conditional independence, the likelihood corresponding to $Y = y$, $\delta = 0$, $\boldsymbol{x}$, and $\boldsymbol{\theta}$ of such an observation is given by

$$g(y|\boldsymbol{x};\boldsymbol{\theta})P(T > y|\boldsymbol{x};\boldsymbol{\theta}) = g(y|\boldsymbol{x};\boldsymbol{\theta})[1 - F(y|\boldsymbol{x};\boldsymbol{\theta})].$$

For a random variable Y that is uncensored at y, we know that $T = y$ and $C > y$. Because of the conditional independence, the likelihood corresponding to $Y = y$, $\delta = 1$, $\boldsymbol{x}$ and $\boldsymbol{\theta}$ of such an observation is given by

$$P(C > y|\boldsymbol{x};\boldsymbol{\theta})f(y|\boldsymbol{x};\boldsymbol{\theta}) = [1 - G(y|\boldsymbol{x};\boldsymbol{\theta})]f(y|\boldsymbol{x};\boldsymbol{\theta}).$$

Since we are not interested in the conditional distribution of C, for the partial likelihood we ignore the parts depending on that distribution, so that the partial likelihood corresponding to $Y = y \geq 0$, $\delta \in \{0, 1\}$, $\boldsymbol{x}$, and $\boldsymbol{\theta}$ is given by $[f(y|\boldsymbol{x};\boldsymbol{\theta})]^{\delta}[1-F(y|\boldsymbol{x};\boldsymbol{\theta})]^{1-\delta}$; hence the (partial) log-likelihood is given by

$$\begin{aligned}
\phi(y, \delta|\boldsymbol{x};\boldsymbol{\theta}) &= \delta \log f(y|\boldsymbol{x};\boldsymbol{\theta}) + (1-\delta)\log(1 - F(y|\boldsymbol{x};\boldsymbol{\theta})) \\
&= \delta \log \lambda(y|\boldsymbol{x};\boldsymbol{\theta}) - \Lambda(y|\boldsymbol{x};\boldsymbol{\theta}) \\
&= \delta\alpha(y|\boldsymbol{x};\boldsymbol{\theta}) - \int_0^y \exp(\alpha(u|\boldsymbol{x};\boldsymbol{\theta}))du, \qquad (7.2.2)
\end{aligned}$$

where we have used (7.1.1)–(7.1.3). Consequently,

$$\frac{\partial}{\partial\theta_j}\phi(y, \delta|\boldsymbol{x};\boldsymbol{\theta}) = \delta B_j(y|\boldsymbol{x}) - \int_0^y B_j(u|\boldsymbol{x}) \exp(\alpha(u|\boldsymbol{x};\boldsymbol{\theta}))du$$

for $1 \leq j \leq p$ and

$$\frac{\partial^2}{\partial\theta_j \partial\theta_k}\phi(y, \delta|\boldsymbol{x}; \boldsymbol{\theta}) = -\int_0^y B_j(u|\boldsymbol{x}) B_k(u|\boldsymbol{x}) \exp(\alpha(u|\boldsymbol{x}; \boldsymbol{\theta}))du \qquad (7.2.3)$$

for $1 \leq j, k \leq p$.

The log-likelihood function corresponding to the observed data $(y_i, \delta_i, \boldsymbol{x}_i)$, $1 \leq i \leq n$, and the linear model for the conditional log-hazard function is given by

$$l(\boldsymbol{\theta}) = \sum_i \phi(y_i, \delta_i|\boldsymbol{x}_i; \boldsymbol{\theta}), \qquad \boldsymbol{\theta} \in \mathbb{R}^p.$$

It follows from (7.2.3) that $\phi(y, \delta|\boldsymbol{x}; \cdot)$ is a concave function on $\mathbb{R}^p$ for $y \geq 0$, $\delta \in \{0, 1\}$, and $\boldsymbol{x} \in \mathcal{X}$ and hence that $\ell(\boldsymbol{\theta})$ is a concave function on $\mathbb{R}^p$.

The maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ is given as usual by $l(\widehat{\boldsymbol{\theta}}) = \max_{\boldsymbol{\theta}} l(\boldsymbol{\theta})$, and the log-likelihood of the model is given by $\widehat{l} = l(\widehat{\boldsymbol{\theta}})$. The corresponding maximum likelihood estimates of the conditional log-hazard function, hazard function, density function, and distribution function are given by $\widehat{\alpha}(t|\boldsymbol{x}) = \alpha(t|\boldsymbol{x}; \widehat{\boldsymbol{\theta}})$, $\widehat{\lambda}(t|\boldsymbol{x}) = \lambda(t|\boldsymbol{x}; \widehat{\boldsymbol{\theta}})$, and so forth.

As the log-likelihood function for Hare is concave, the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ is easily obtained using a Newton–Raphson algorithm (see Section 2.5.3). The main complication in computing the maximum-likelihood estimate and Rao statistics is the need to evaluate the integrals in (7.2.2) and (7.2.3). Note that these integrals need to be carried out separately for *each* unique set of covariates. As such, it is essential that these integrals can be computed rapidly, thus the advantage of using a fast quadrature formula (e.g. Gaussian quadrature, see Abramowitz and Stegun 1965, Section 25) or to use linear splines for the basis functions in time so that the integrals can be computed exactly.

7.2.5 Inference

Since estimation for Hare models is carried out by the method of maximum likelihood, approximate standard errors for the coefficients in a Hare model that is specified a priori are readily available. However, as with any statistical procedure that involves formal or informal model selection, inference about parameters in a Hare model may be incorrect when stepwise model selection is carried out. For example, variables that end up in a Hare model are almost guaranteed to be "significant", since otherwise the stepwise model selection procedure would have dropped basis functions involving this variable. This may be misleading, for example, when another correlated variable was considered and did not end up in the model. Clearly, this is a problem not just with Hare, but with any statistical procedure that involves variable selection—even if this variable selection involves informal screening of variables by an expert before a parametric model is applied.

However, the problem may be more severe in Hare because of the large number of potential basis functions involved in the selection process.

The variability of Hare models can easily be assessed using simulation or bootstrap studies (see Section 7.3 below). However, these procedures cannot assess the possible bias of Hare estimates.

7.3 Further analysis of the bone marrow transplant data

7.3.1 Does a simpler model fit the data?

The Hare model for which survival and hazard functions were shown in Figure 7.6 uses cubic splines. The final model, as selected using BIC, has 18 basis functions, the log-likelihood is -4028.69, and the BIC value is 8206.16. The model is summarized in Table 7.1. Note that since we use cubic splines, 7 knots in time correspond to 5 basis functions: 4 basis functions for cubic polynomials plus 7 basis functions for knots minus 2 constraints to keep the left tail linear minus 3 constraints to keep the right tail constant minus 1 basis function because of the intercept.

We can also force the Hare algorithm to yield a proportional hazards model by not allowing any basis functions that depend on time and a covariate to enter the model. When we do this, we get a model with 15 basis functions, a log-likelihood of -4054.73, and a BIC value of 8233.45. The model has roughly the same basis functions as those listed in Table 7.1, except that, obviously, there are no interactions between time and other variables. The use of total body irradiation (TBI) and interactions between TBI and two of the disease categories also enter the model. If we force the Hare algorithm to fit an additive model (that is, without any basis functions depending on two variables), we get a model with 13 basis functions, a log-likelihood of -4075.43, and a BIC value of 8258.30. This model includes the same variables as the model in Table 7.1 and a nonlinear contribution from the variable year of transplant.

The difference in BIC values of $8233.45 - 8206.16 = 27.29$ between the full Hare model and the proportional hazards suggests that this data is not well modeled by a proportional hazards model. Clearly, BIC values are not the only basis for preferring one model over another. An advantage of proportional hazards models is, for example, that the exponents of the coefficients are interpretable as relative risk parameters (Kalbfleisch and Prentice 1980). Typically, when particular basis functions occur in a proportional and a nonproportional Hare model and the variables involved with these basis functions do not interact with time, the coefficients in the two models are quite similar. For example, we could compare the coefficients of patient age and prophylaxis use in Tables 7.1 and 7.4.

variables involved	type	number of basis fcts	(time) knots involved	coefficient	standard error
intercept	–	1	–	2.891	0.336
patient age	linear	1	–	0.02450	0.00255
prophylaxis use	linear	1	–	−0.3798	0.0731
year of transplant	linear	1	–	−0.03708	0.00405
more advanced	indicator	1	–	1.748	0.146
less advanced	indicator	1	–	0.9982	0.1288
disease type CML-CP	indicator	1	–	1.267	0.232
time	spline	5	.00, .19, .30, .59 1.97, 7.27, 8.59	misc	misc
more advanced × patient age	interaction	1	–	−0.01697	0.00300
more advanced × prophylaxis	interaction	1	–	0.3444	0.0916
patient age × time	interaction	1	.00, .19, .30	0.01803	0.00540
more advanced × time	interaction	1	.19, .30, 1.97	−0.7375	0.1853
less advanced × time	interaction	1	.30, .59, 1.97	−1.006	0.193
type CML-CP × time	interaction	1	1.97, 7.27, 8.59	−2.022	0.270

TABLE 7.1. Hare model for the bone marrow transplant data employing cubic splines.

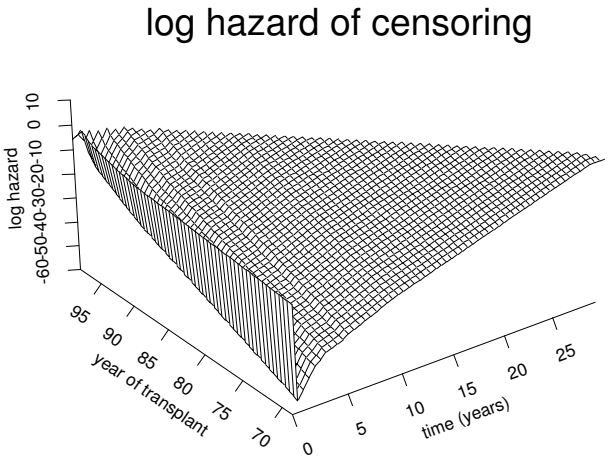

FIGURE 7.7. Estimated log-hazard of censoring as a function of time since transplant and year of transplant for the bone marrow transplant data.

To verify further that a nonproportional hazards model fits the data better than a proportional hazards model, we carried out a small simulation study. As the amount of cpu time involved in fitting Hare models with cubic splines is substantial, we used Hare with linear splines for this study. We proceeded as follows:

1. First we fitted three Hare models with linear splines to the bone marrow transplant data: an unrestricted model, a model that is forced to be a proportional hazards model, and a model that is forced to be an additive model. These three models are very similar to the models using cubic splines described above.
2. We used a Hare model with linear splines to estimate the censoring distribution, conditional on the covariates. This is easily done by applying Hare to the regular data, using $\delta^* = 1 - \delta$ as the censoring indicators. While all participants are censored at or before November 1998, the censoring distribution is still well modeled by a continuous distribution, as the time origin is date of transplantation, which is different for (virtually) every subject. Not surprisingly, this fitted model depends strongly on year of transplant, as patients who received their transplant in later years have to be censored sooner after their treatment. It does not depend on any other variables. In Figure 7.7 we show the conditional log-hazard rate of censoring as a function of time and year of transplant. It can be seen that initially the rate of censoring is fairly high, but that it drops considerably after about one year. The rate then increases again when it reaches the "diagonal", which corresponds to 1998, as there are no data on any patient beyond that time.

		fitted model	
		proportional	nonproportional
true model	additive	hazards	hazards
additive	205	24	21
proportional hazards	2	223	25
nonproportional hazards	1	10	239

TABLE 7.2. Number of times that data generated from a Hare model with a given structure resulted in a fitted Hare model with that or another structure, for a data structure similar to the bone marrow transplant data.

3. For each simulation we proceeded as follows: conditional on the actual observed variables, we generated a new set of survival times T_i^{*1}, $i = 1, \ldots, 3887$, from the unrestricted Hare model obtained in step (1) using the same covariates as in the original data. Similarly, we generated T_i^{*2}, $i = 1, \ldots, 3887$, from the Hare model restricted to be a proportional hazards model, and T_i^{*3}, $i = 1, \ldots, 3887$, from the Hare model restricted to be an additive model.
4. We generated a new set of censoring times C_i^*, $i = 1, \ldots, 3887$, from the Hare model obtained in step (2) using the same covariates as in the original data.
5. For $j = 1, \ldots, 3$, we then generated three new data sets $Y_i^{*j} = \min(T_i^{*j}, C_i^*), i = 1, \ldots, 3887$. To each of these data sets we applied Hare, and we counted how often Hare yielded an additive model, a proportional hazards model, and a nonproportional hazards model.

For each of the three models we generated 250 data sets. The results are summarized in Table 7.2. As can be seen from this table, when the true model is an additive model or a proportional hazards model, there is only about a 10% chance that Hare yields a nonproportional hazards model, while if the true model is nonproportional, there is a 95% chance that the fitted model is nonproportional. We conclude that there is fairly strong evidence that the "true" hazard function for the transplant data is nonproportional.

We can use both bootstrap and simulation techniques to assess the uncertainty in conditional hazard functions like those shown in Figure 7.6. In interpreting the results it is important to be aware of the sources of bias and variance in estimates of the (log-)hazard function:

- Estimation of the coefficients of a fixed set of basis functions (with fixed knots) is a parametric maximum likelihood problem. With these parameter estimates come standard errors that quantify uncertainty in the coefficient estimates.
- A second source of variability comes from the model selection. Clearly, there is considerable uncertainty about which basis functions end up

in the model. We discuss this topic in much more detail for Logspline and Triogram regression in Chapter 10.

- Finally, since the "true" model is not exactly a (linear, cubic) spline, we have bias as we cannot exactly fit the true model. Theoretical results (11) suggest that this bias is larger for linear than for cubic splines.

The parametric standard errors only address the first type of uncertainty. When we simulate from a fitted Hare model, as we did for the simulation study summarized in Table 7.2, we get an excellent handle on the first type of uncertainty and a reasonable handle on the second type (although the variation may be a bit different because now the "true model" is a spline function). When we use the bootstrap approach (sampling cases with replacement from our data set), we generate estimates that may not be centered around the fitted model, as the data are sampled from the true model (and not the fitted model).

The shaded area in the left side of Figure 7.8 shows the 2.5th and 97.5th percentiles of the fitted hazard functions for the 250 simulations from the fitted Hare model without any restrictions, which were used for the simulation described above; the dashed curves show the pointwise parametric 95% confidence bounds using the asymptotic covariance matrix of the estimated coefficients. The shaded area in the right side of this figure shows the pointwise 2.5th and 97.5th percentiles of the fitted hazard functions for the fits of 250 bootstrap samples from the original data; the dashed curve shows the mean of these 250 curves. To keep this figure readable, we only show two of the four groups of patients. All other covariates are as in Figure 7.6. The left side of Figure 7.8 shows that the parametric bands are indeed overly optimistic, in particular in the tail and close to the knot at about one year.

7.3.2 Partially linear Hare models

In many medical studies there may be one variable that represents a new treatment. When Hare is being used to analyze data from such a study it is rather unfortunate if this treatment does not end up in the model. In particular, when the goal is to make an inference about the effectiveness of the new treatment we may want to force one or more variables, for example, the treatment variable, to be linear in the model, without allowing for any interactions among these variables or between these variable and other variables. In practice we do this by starting with the variables that are forced to be linear in the initial model, by not considering any other basis functions involving these variables for entering in the model, and by not allowing the linear basis functions to be removed from the model.

For such a "partially linear" Hare model, inference about the coefficients of the variables that are forced to be linear is of particular interest. To

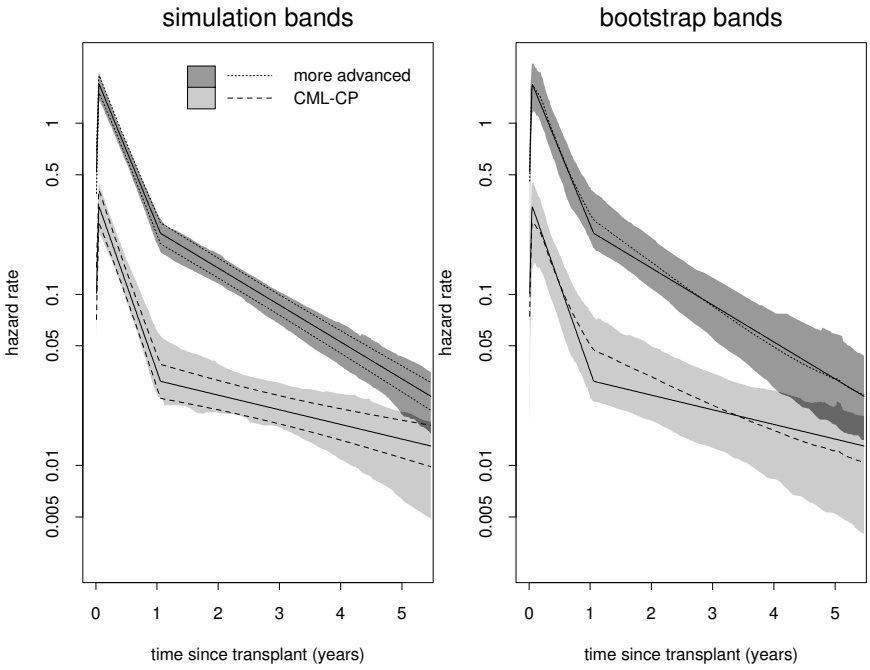

FIGURE 7.8. 95% simulation and bootstrap bands as shaded areas for two log-hazard functions. The left side includes the parametric pointwise "confidence" bands, and the right side includes the mean of 250 bootstrapped curves as dashed curves; the solid line is the fitted Hare model.

assess the accuracy of the standard errors of such coefficients, we carried out the following bootstrap study: for each of the ten covariates in the bone marrow transplant data other than the disease categories, we fitted a Hare model forcing a partially linear Hare model for that variable, allowing the other nine covariates to have interactions as usual. We then generated 250 bootstrap samples of the same size (drawing cases with replacement) from the bone marrow transplant data, and we fitted the same ten partially linear Hare models to the data. In Table 7.3, for each of the ten covariates, we compare its estimated coefficient and standard error based on the partially linear Hare model with the mean and standard deviation of the coefficient of that covariate over 250 partially linear Hare bootstrap models. We also provide the value of the absolute largest correlation between the particular variable and the other variables in the bone marrow transplant data. From this table we see that the parametric estimate of the standard error is always downwards biased, but that the magnitude of this bias seems to be strongly related to the correlation of the particular variable with the other variables. In Figure 7.9 we plot the ratio of the bootstrap standard error over the parametric standard error versus the absolute value of the largest correlation. The results of this bootstrap study suggest that for randomized

covariate	coefficient	bootstrap mean coef.	standard error	bootstrap std. dev.	largest correl.
patient age	0.01665	0.01527	0.00156	0.00319	0.916
donor age	0.002261	0.003700	0.003409	0.004670	0.916
use of					
– TBI	0.1518	0.1352	0.0590	0.0692	−0.434
– prophylaxis	−0.2534	−0.2098	0.0535	0.0700	0.555
– fluconazole	−0.1708	−0.1783	0.0720	0.1191	0.714
– gancyclovir	−0.1221	−0.0427	0.0765	0.0940	0.601
patient sex	−0.1159	−0.1204	0.0425	0.0455	0.098
donor sex	0.1067	0.1005	0.0413	0.0421	0.098
log(dose)	−0.09311	−0.08705	0.03184	0.03401	−0.258
year	−0.03765	−0.03440	0.00406	0.00677	0.715

TABLE 7.3. Estimate and standard error of coefficients in a partially linear Hare model together with the mean and standard deviation of the same coefficient in 250 bootstrapped models.

studies, where the variable of interest is uncorrelated with other variables, the parametric estimate of the standard error is likely to be fairly unbiased.

7.3.3 Proportional hazards regression

In Section 7.3.1 we concluded that a proportional hazards model did *not* fit the bone marrow transplant data well. In this section we compare two approaches to estimating the regression function $\Psi(\boldsymbol{x})$ using polynomial splines when we a priori decide to fit a proportional hazards model to the data, regardless of the validity of the proportional hazards assumption. The first approach uses a restricted version of Hare, as was done in Section 7.3.1, in which basis functions that depend both on time and a covariate are not allowed to enter in the Hare model. The basis functions that do not involve time yield an estimate of $\Psi(\boldsymbol{x})$, while those that do involve time yield an estimate of the baseline hazard function $\lambda_0(t)$. In the remainder of this section we refer to this restricted version of Hare as Hare_{ph}.

An alternative approach is directly to model the proportional hazards regression function $\Psi(\boldsymbol{x})$ and to use the method of maximum partial likelihood for estimation and model selection. Huang, Kooperberg, Stone, and Truong (2000) define the Proportional Hazards Regression (PHare) model as

$$\log \Psi(\boldsymbol{x}) = \sum_{j=1}^{p} \theta_j B_j(\boldsymbol{x}). \tag{7.3.1}$$

For fixed basis functions the parameters $\theta_1, \ldots, \theta_p$ in (7.3.1) are estimated using the method of maximum partial likelihood (see, for example, Kalbfleisch and Prentice 1980 for details about partial likelihood estimation). Model selection is carried out using the same stepwise algorithm as described in

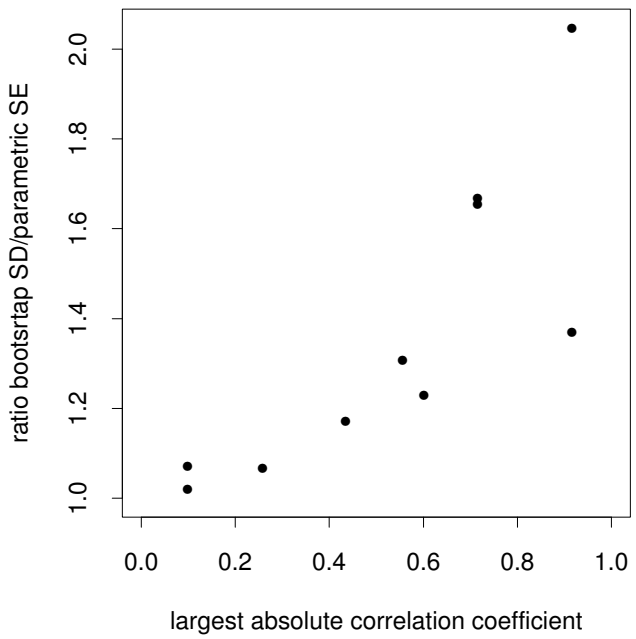

FIGURE 7.9. Largest absolute correlation coefficient and ratio between the bootstrap standard deviation and the parametric standard error for the partially linear bootstrap study.

Sections 7.2.2 and 7.2.3 with two differences: no basis function in time is considered; the initial model has $p = 0$ or, equivalently, $\Psi(\boldsymbol{x}) = 1$.

Implementation of PHare is considerably more straightforward than implementation of Hare, as most statistical packages have routines for estimating parameters in a Cox regression model, which can be used to estimate the parameters in (7.3.1) for fixed basis functions. In particular, estimation by maximum partial likelihood does not involve integration.

In Table 7.4 we compare the results of applying (piecewise linear versions of) Hare_{ph} and PHare. When examining these results, we should keep in mind that the two programs were developed independently and, as a result, many tiny programming details differ between them. In addition, the stepwise algorithm may enhance small earlier differences during later stages of the algorithm. To confirm that maximum likelihood and maximum partial likelihood estimation give very similar results, we fitted the Hare_{ph} model of Table 7.4 (ignoring the basis functions depending on time) using PHare and found that all coefficients that were fitted by maximum partial likelihood were within 1.4% of the value of those fitted by maximum likelihood, while the standard errors were all within 0.4%. This is consistent with the results reported in Huang, Kooperberg, Stone, and Truong (2000).

As a further comparison, we used maximum likelihood and maximum partial likelihood to estimate the parameters in Hare models that include

	Hare_{ph}		PHare	
basis function	coefficient	standard error	coefficient	standard error
intercept	−2.179	0.331	NA	NA
patient age	0.02611	0.00247	0.02640	0.00244
total body irradiation	NA	NA	0.6495	0.1233
prophylaxis use	−0.3911	0.0734	−0.3787	0.0726
dose	−0.09124	0.03150	NA	NA
year	−0.03730	0.00407	−0.03534	0.00425
more advanced	1.375	0.113	1.377	0.142
less advanced	0.4865	0.0795	0.6754	0.1177
disease type CML-CP	−0.3553	0.1054	−0.7145	0.1264
more advanced × patient age	−0.01671	0.00301	−0.01572	0.00301
more advanced × prophylaxis	0.3785	0.0920	0.3642	0.0923
more advanced × tbi	NA	NA	−0.5366	0.1478
less advanced × tbi	NA	NA	−0.7250	0.1582
$(0.0465 - t)_+$	−35.49	4.12	NA	NA
$(1.06 - t)_+$	1.543	0.079	NA	NA
$(6.73 - t)_+$	0.4585	0.0193	NA	NA

TABLE 7.4. Hare_{ph} and PHare model for the bone marrow transplant data employing linear splines. Basis functions for which the coefficient are listed as NA are not in the corresponding model.

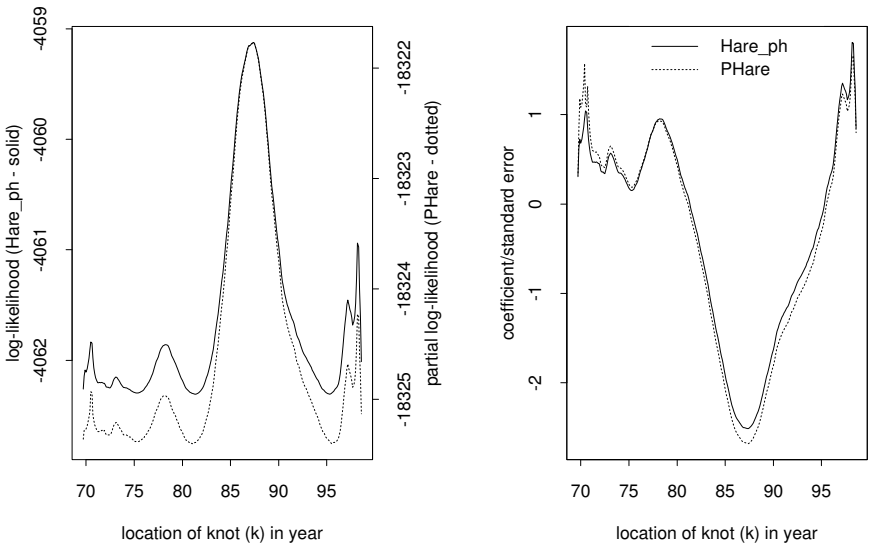

FIGURE 7.10. Log-likelihood for Hare_{ph} (solid) and partial log-likelihood for PHare (dotted) (left) and ratio coefficient/standard error (right) as a function of the location of the knot in year.

the eight basis functions that are both in the Hare_{ph} and the PHare model in Table 7.4 and also include a basis function $(\text{year} - k)_+$ for k in the range of year of transplant. For the model fitted by maximum likelihood we also included the intercept and the three basis functions that depend on time in Table 7.4. In Figure 7.10 we show the log-likelihood and the log-partial-likelihood for these two estimation methods and the ratio of the coefficient and the standard error for the basis function $(\text{year} - k)_+$ as a function of k. We note that there is little difference between estimation by maximum likelihood and estimation by maximum partial likelihood.

As the results for Hare_{ph} and PHare are virtually identical and the theoretical convergence rates are the same (Kooperberg, Stone, and Truong 1995b; Huang, Kooperberg, Stone, and Truong 2000), what are the advantages of one versus the other? Advantages of Hare:

- Hare also provides an estimate of the baseline hazard function;
- Hare can be used to assess whether the assumption of proportional hazards is appropriate.

Advantages of PHare:

- PHare requires less cpu time than Hare;
- developing a numerical implementation of PHare is easier than developing one for Hare;

- as Cox regression is the de facto standard for analyzing survival data, some scientists may be more easily convinced by a PHare analysis, which is closer in spirit to traditional Cox regression, than by a Hare analysis.

When a version of Hare is already available on one's system, the first two advantages of PHare may not be significant.

All in all, we do not see PHare as a serious alternative to Hare, but rather we see the similarity of the results of PHare and Hare_{ph} as an additional justification for the use of Hare, primarily addressed to those people used to analyzing data using Cox regression.

7.4 Extensions

7.4.1 The Colorado Plateau uranium miners data

The second example with survival data involves the Colorado Plateau Uranium Miners data. This data set comes from the U.S. Public Health Service database. The data set contains records on 4103 male uranium miners in the Colorado Plateau (located within the states of Colorado, Utah, New Mexico, and Arizona); only miners who worked in uranium mines at some time between 1950 and 1964 are included in the study. Vital status data on most of these (uranium) miners are available up to 1995, at which stage most miners were retired. See Hornung and Meinhardt (1987) and Hornung, Deddens, and Roscoe (1998) for details. The data set, as compiled by the National Institute for Occupational Safety and Health, contains information on radon exposure from uranium and possible hard rock mining and smoking patterns. Briefly, for each period in the life of one of these miners we know how many packs a day he smoked on the average and how large his radon exposure was (in units per time). For example, for one miner we may know that from age 0 to age 11 years and 8 months he did not smoke, from that age through age 23 years and 8 months he smoked half a pack per day, and from that age until the end of 1991, when this participant was 67 years old, he smoked one pack a day. The smoking information was obtained retrospectively using a questionnaire and is available through 1991. For miners who were still alive at that time we assumed that this miner continued smoking the same amount from then on. The information about radon exposure is organized similarly to the smoking information. For example, for a particular miner who had first (mining) radon exposure at age 22 years and 9 months, it could be that between age 22 years and 9 months and age 33 years and 9 months this miner had an exposure level of 0.303 working level months per month (WLM/m) and from that age until age 34 years and 1 month he had an exposure level of 3.5 WLM/m. (One WLM equals 3.5×10^{-3} joule-hours per cubic meter.) About one

quarter of the miners in the data set mined hard rock before they became uranium miners. For these miners we also have information for radon exposure from hard rock mining. See Luebeck, Heidenreich, Hazelton, Paretzke, and Moolgavkar (1999) for more details. In our analysis we excluded 109 miners with incomplete covariate or vital status data, as was done in the analysis of Luebeck et al. (1999), leaving 3,954 miners.

For this data set we would like to find out whether there is an effect of radon exposure on the hazard of dying from lung cancer among the uranium miners. Censoring combines loss to follow up, death by causes other than lung cancer, and being alive at the end of follow-up (1995). There are two complications in carrying out this analysis. The first is time-dependent covariates: the patterns of smoking and radon exposure may change over time. The second is left truncation: the study is restricted to men who were miners at some time during the period 1950–1964, but the natural time-origin when studying survival is time of birth. Thus males who died from lung cancer at a very young age were excluded from the study. The concept of left truncation, which may be new to readers unfamiliar with survival analysis, is discussed further in Section 7.4.3.

7.4.2 Time-dependent covariates

In our analysis of the uranium miners data we used two types of time-dependent covariates. For current amount of smoking and radon exposure we used a piecewise constant function; for example the radon exposure for the miner described above is 0.303 from year 22.75 through year 33.75 and is 3.5 from year 33.75 through year 34.083. The cumulative radon exposure for this miner is obtained by integrating the radon exposure over time. (Any exposure from sources other than mining is ignored; thus the cumulative radon exposure when the participant becomes a miner is zero.) Observe that the cumulative radon exposure is a piecewise linear function.

A third type of time-dependent covariates is more common in medical studies with periodic examinations: certain vital status measurements may be recorded periodically. Formally, we do not know how this covariate changes between measurements, but for analysis purposes we may either interpolate linearly or keep the previous value until the next measurement.

The proportional hazards model can easily deal with such covariates, as long as the covariate value is known at each event time (Kalbfleisch and Prentice 1980). For Hare the main change because of time-dependent covariates is that numerical integration may be needed for models that otherwise could be fit more easily.

We now discuss the complications of time-dependent covariates on the linear model for the log-hazard function and on maximum likelihood estimation of the coefficients, as discussed in Section 7.1.3 and 7.2. Here we assume that there is no left truncation.

We assume that the positive random variable T may depend on a vector $\boldsymbol{x}(t)$ of M (possibly time-dependent) covariates that lies in the subset $\mathcal{X}$ of $\mathbb{R}^M$ for each $0 \le t \le T$. Let $\lambda(t|\boldsymbol{x}(s), 0 \le s \le t)$ denote the conditional hazard function of T given $\boldsymbol{x}(s), 0 \le s \le t$. We assume that the conditional hazard function at time t depends only on the value of the covariates at that time; that is, we assume that

$$\lambda(t|\boldsymbol{x}(s), 0 \le s \le t) = \lambda(t|\boldsymbol{x}(t)) \tag{7.4.1}$$

and hence that the conditional log-hazard function $\alpha(t|\boldsymbol{x}(s), 0 \le s \le t) = \alpha(t|\boldsymbol{x}(t))$ has the same property. Let

$$\Lambda(t|\boldsymbol{x}(s), 0 \le s \le t) = \int_0^t \lambda(u|\boldsymbol{x}(u))du \tag{7.4.2}$$

denote the conditional cumulative hazard function.

From here we can get expressions for the partial likelihood $\phi(y, \delta|\boldsymbol{x}(s), 0 \le s \le y; \boldsymbol{\theta})$ and its derivatives for $y \ge 0$ and $\delta \in \{0, 1\}$ (compare with (7.2.2) and (7.2.3)) and see Andersen, Borgan, Gill, and Keiding (1993):

$$\begin{aligned} &\phi(y, \delta|\boldsymbol{x}(s), 0 \le s \le y; \boldsymbol{\theta}) \\ &\quad = \quad \delta\alpha(y|\boldsymbol{x}(y); \boldsymbol{\theta}) - \int_0^y \exp(\alpha(u|\boldsymbol{x}(u); \boldsymbol{\theta}))du, \end{aligned} \tag{7.4.3}$$

$$\begin{aligned} &\frac{\partial}{\partial\theta_j}\phi(y, \delta|\boldsymbol{x}(s), 0 \le s \le y; \boldsymbol{\theta}) \\ &\quad = \quad \delta B_j(y|\boldsymbol{x}(y)) - \int_0^y B_j(u|\boldsymbol{x}(u)) \exp(\alpha(u|\boldsymbol{x}(u); \boldsymbol{\theta}))du \end{aligned} \tag{7.4.4}$$

for $1 \le j \le p$, and

$$\begin{aligned} &\frac{\partial^2}{\partial\theta_j\partial\theta_k}\phi(y, \delta|\boldsymbol{x}(s), 0 \le s \le y; \boldsymbol{\theta}) \\ &\quad = \quad -\int_0^y B_j(u|\boldsymbol{x}(u))B_k(u|\boldsymbol{x}(u)) \exp(\alpha(u|\boldsymbol{x}(u); \boldsymbol{\theta}))du \end{aligned} \tag{7.4.5}$$

for $1 \le j, k \le p$. Thus, the log-likelihood function remains concave. The main complication occurs in computing the numerical integrals, as the exponent $\alpha(u|\boldsymbol{x}(u); \boldsymbol{\theta})$ may be a polynomial of order $2q - 1$ when splines of order q are being used, and some time-dependent covariates vary linearly with time. In particular, this means that even when linear splines are used numerical integration may be needed.

7.4.3 Left truncation

Left truncation occurs when failures before a particular time are excluded from the data, so nonfailures before that time cannot be used in estimating

the hazard rate. This happens frequently in epidemiological studies when the potential participants who have died before data collection started cannot be included in the study, because it is impossible to determine whether these potential participants actually would have been eligible for the study, because crucial covariates can no longer be determined, or because we simply are not aware of who these potential participants would have been. It also occurs in industrial settings when we are interested in the failure time of equipment, and the equipment that is being studied has already been in operation for a while. The Colorado Plateau uranium miners data described in Section 7.4.1 thus involves left truncation: the natural time-origin is birth, but only miners who worked in uranium mines at some time between 1950 and 1964 are included in the study.

Informally, it is the easiest to understand the problems associated with left truncated data by considering a simplified example. Suppose that all miners start mining at 25 years. By definition, none of these miners died of lung cancer before the age of 25, but this says nothing about the hazard rate for dying from lung cancer before that age. Keeping this in mind, it should now be clear that someone who become a miner at age r contributes to the information about $\lambda(t|\boldsymbol{x}(s), r \le s \le t)$ for $r \le t \le y$; here r is the left truncation time for this miner.

Consider a randomly selected individual for whom the survival time T is larger than the truncation time R. For this individual, let $C > R$ denote the censoring time and $\boldsymbol{x}$ the vector of covariates, which may depend on time, and set $Y = \min(T, C)$ and $\delta = \text{ind}(T \le C)$. The random variable Y is said to be uncensored or censored according as $\delta = 1$ or $\delta = 0$.

Compared to the treatment in Section 7.4.2, in (7.4.1) and (7.4.2), $0 \le s \le t$ gets replaced by $r \le s \le t$; in (7.4.2), $\int_0^t$ gets replaced by $\int_r^t$; and in (7.4.3)–(7.4.5) , $0 \le s \le y$ gets replaced by $r \le s \le y$ and $\int_0^y$ gets replaced by $\int_r^y$ (see Andersen, Borgan, Gill, and Keiding 1993). The log-likelihood function remains concave.

7.4.4 Analysis of the Colorado Plateau uranium miners data

We applied Hare using cubic splines to the Colorado Plateau uranium miners data as described in Section 7.4.1. The response T is the age of death from lung cancer. The left truncation time R is the time the miner joined the study, this typically is the maximum of the age in 1950 and the age of becoming a uranium miner. We used six covariates. Four of these covariates are time-dependent: current and cumulative radon exposure, and current and cumulative smoking. The other two are race, a binary variable indicating whether a miner was white (0) or not (1), and birth year, to account for a possible cohort effect.

The Hare model that was selected, using BIC, contains eight basis functions. Six of these basis functions, an intercept, two basis functions in-

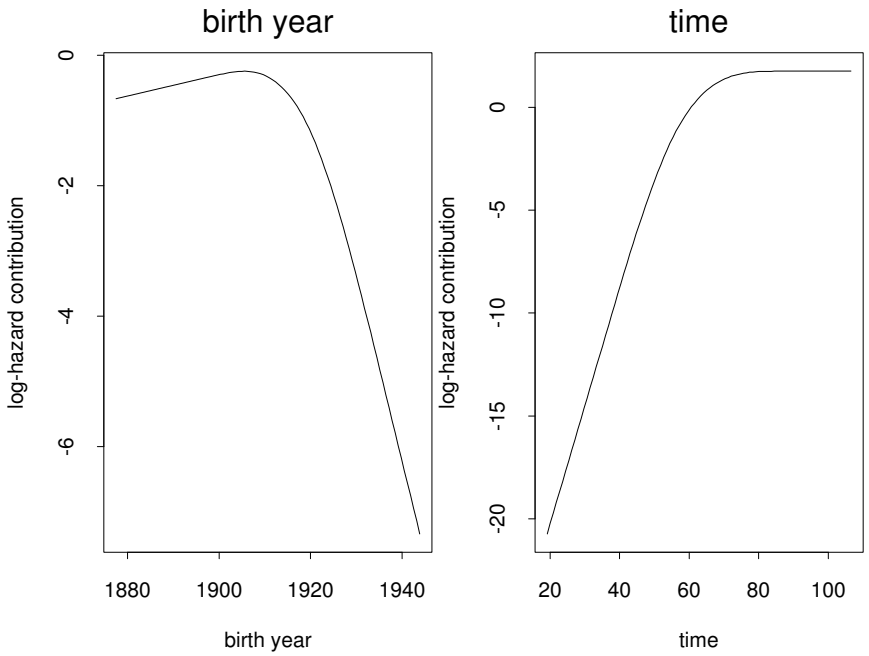

FIGURE 7.11. Contribution to the log-hazard in the Hare model of the univariate basis functions depending on birth year (left) and time (right). An interaction term with a much smaller effect, involving both of these variables, is not shown.

volving birth year, two basis functions involving time, and an interaction between time and birth year, do not involve smoking or radon status. The contributions to the log-hazard from the basis functions depending on time and birth year are shown in Figure 7.11. The interaction term is not shown in these figures; the contribution from it is small compared to the effects shown in Figure 7.11. The increasing hazard with time is not unexpected. The decreasing hazard with birth year presumably could be explained by the general improvement in health care during the 20th century.

Both cumulative smoking and cumulative radon exposure ended up linearly in the Hare model. Of all miners, 3096 (78%) smoked at some stage during their life. For those who did smoke, the median contribution to the log-hazard due to smoking at the time of censoring or death was 0.56, which corresponds to a relative risk of 1.75. For the 10th percentile among the smokers the contribution was 0.08, which corresponds to a relative risk of 1.08, and for the 90th percentile among the smokers the contribution was 1.09, which corresponds to a relative risk of 3.00. Sixty-four miners had fitted relative risks due to smoking of more than 5.00. All miners had some radon exposure, although for some this exposure was very low. The median contribution to the log-hazard due to radon exposure at the time of cen-

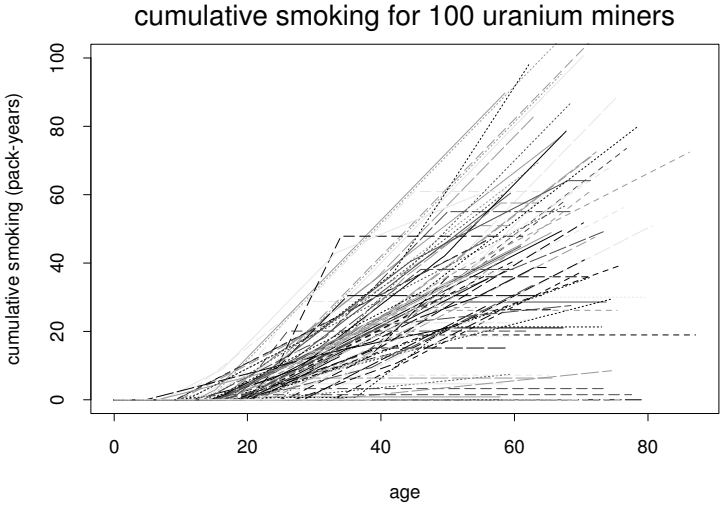

FIGURE 7.12. Relation between cumulative smoking and age for 100 randomly selected uranium miners.

soring or death was 0.14, which corresponds to a relative risk of 1.15. For the 10th percentile among the smokers the contribution was 0.02, which corresponds to a relative risk of 1.02, and for the 90th percentile among the smokers the contribution was 0.65, which corresponds to a relative risk of 1.91. Thirty-six miners had fitted relative risks due to radon exposure of more than 5.00.

Since *cumulative* smoking and, to a lesser extent, cumulative radon exposure are closely correlated with age (which is time in this example), it is clear that we have a problem of confounding. This relation is particularly evident from Figure 7.12: for those individuals who smoke, the relation between age and cumulative smoking is very close between age 20 and 60. Among the uranium miners, 79% have smoked and 65% have smoked at least 10 pack-years.

To attribute all of the variation in the log-hazard over time to smoking and radon exposure, we may want to fit a model without knots in time. By doing this, we essentially fit a proportional hazards model with a constant hazard rate. Both current and cumulative smoking and radon exposure variables end up in the fitted model, as well as an intercept and two basis functions involving birth year. The coefficient of cumulative smoking is considerably larger than in the earlier model, but this is partly offset by a negative coefficient for current smoking. If we combine these two terms, the maximum attained contribution to the log-hazard for the smokers increases by an average of about 0.18, corresponding to an additional relative risk of about 1.19. The results for radon exposure are harder to interpret. The coefficient of cumulative radon exposure is virtually unchanged, but there

is a substantial negative coefficient for current radon exposure, so that this model would suggest a protective effect of radon exposure for lung cancer until the effect of cumulative radon exposure is larger than the effect of current radon exposure. This is hard to believe.

The model that includes basis functions that depend on time and the model that does not include such basis functions both have eight basis functions. The model that does not involve time has a log-likelihood of -2334.60, while the model that does involve time has a log likelihood of -2280.89. The difference in log-likelihood of 53.71 and the inconsistent radon results together leads us to prefer the model which involves time.

7.4.5 Interval censored data

Many other types of censoring and truncation can occur in real data sets. Here, we briefly discuss interval censoring, probably the most common form of censoring other than right censoring. We say that an observation T is interval censored in the interval $C = [C_l, C_u]$ if it is known that $T \in C$, but the exact value of T is unknown. Hare with interval censored data is discussed extensively in Kooperberg and Clarkson (1997). We can also think of uncensored and right censored data in this manner. In particular, if T is uncensored, then $C = \{T\}$; if T is right censored at $C_l < T$, then $C = [C_l, \infty)$. For simplicity, we assume here that there are no time-dependent covariates. It is assumed that T is independent of the type of censoring given the vector $\boldsymbol{x}$ of covariates and that T and C are independent given $\boldsymbol{x}$. Let $\delta = 0$ if T is right censored, $\delta = 1$ if T is uncensored, and $\delta = 2$ if T is interval censored. When there is no left truncation, the partial likelihood corresponding to $C = [c_l, c_u]$, δ and $\boldsymbol{x}$ is given by

$$f(c_l|\boldsymbol{x})^{I(\delta=1)} \left[\int_{c_l}^{c_u} f(t|\boldsymbol{x})dt\right]^{I(\delta\neq 1)} = S(c_l|\boldsymbol{x})^{I(\delta=0)} f(c_l|\boldsymbol{x})^{I(\delta=1)} [S(c_u|\boldsymbol{x}) - S(c_l|\boldsymbol{x})]^{I(\delta=2)}.$$

Using this expression we can obtain formulas for the log-likelihood, the score function, and the Hessian. These formulas are rather tedious. See Kooperberg and Clarkson (1997) for details and for the analysis of a data set involving interval censoring.

As it turns out, the log-likelihood function is not necessarily concave when some observations are interval censored. In principle this may cause problems since the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ is no longer guaranteed to be unique and the Newton–Raphson algorithm is not guaranteed to find the MLE when it exists. In practice, however, we have not experienced problems in finding the MLE when some data are interval censored. See Section 6.5.2 for some discussion about maximization for nonconcave likelihoods.

7.4.6 Heft

When there are no covariates, the numerical problems involved with using cubic splines are reduced considerably. The Hazard Estimation with Flexible Tails (Heft) model uses cubic spline basis functions in time, and it has two additional basis functions that allow for a wider range of tail behavior; in particular, they allow Heft to fit Weibull and Pareto distributions exactly. The cubic spline basis functions for Heft are the same as those mentioned above for Hare. The two additional basis functions are defined as follows. Given a positive number c, set $B_{-1}(t) = \log(t/(t+c))$ and $B_0(t) = \log(t+c)$ for $t > 0$. Let $B_1(t), \ldots, B_p(t)$ be a cubic spline basis. Then $B_{-1}(t), B_0(t), B_1(t), \ldots, B_p(t)$ span an allowable space for Heft.

The two log terms in the model for the log-hazard function are easily motivated. Consider a positive density function f on $(0, \infty)$, and let F, h, and α denote, respectively, the associated distribution function, hazard function, and log-hazard function. Suppose first that $f(t) \approx at^\gamma$ for $t \approx 0$, where $a > 0$ and $\gamma > -1$. Then $\log f(t) \approx \gamma \log t$ for $t \approx 0$. Since $1-F(t) \approx 1$ for $t \approx 0$, we conclude that $\alpha(t) \approx \gamma \log t$ for $t \approx 0$. This motivates the inclusion of the term $\theta_{-1} B_{-1}(t)$ with $\theta_{-1} > -1$ in the model for the log-hazard function.

Suppose next that $f(t) \approx a \exp(-bt^\gamma)$ for $t \gg 1$, where $a > 0$, $b > 0$, and $\gamma > 0$. Then

$$1 - F(t) \approx \frac{a}{b\gamma t^{\gamma-1}} \exp(-bt^\gamma), \qquad t \gg 1,$$

so

$$\lambda(t) \approx b\gamma t^{\gamma-1}, \qquad t \gg 1,$$

and hence $\alpha(t) \approx (\gamma - 1) \log t$ for $t \gg 1$. This motivates the inclusion of the term $\theta_0 B_0(t)$ with $\theta_0 > -1$ in the model for the log-hazard function.

Suppose, instead, that $f(t) \approx at^{-b-1}$ for $t \gg 1$, where $a, b > 0$. Then $1 - F(t) \approx ab^{-1}t^{-b}$ for $t \gg 1$, so that $\lambda(t) \approx bt^{-1}$ for $t \gg 1$ and hence $\alpha(t) \approx (-1) \log t$ for $t \gg 1$. This motivates allowing the possibility that $\theta_0 = -1$ in the model for the log-hazard function.

Suppose now that $p = 1$ and $B_1 = 1$ and hence that

$$\alpha(t; \boldsymbol{\theta}) = \theta_{-1} \log \frac{t}{t+c} + \theta_0 \log(t+c) + \theta_1, \qquad t > 0.$$

This three-parameter model, which is the minimal allowable space for Heft, includes Weibull and Pareto distributions as special cases. As the default we chose the shift parameter c to be the upper quartile of the uncensored data.

Consider the Weibull density function f given by

$$f(t) = b\gamma t^{\gamma-1} \exp(-bt^\gamma), \qquad t > 0,$$

where $b > 0$ and $\gamma > 0$, whose distribution function is given by

$$F(t) = 1 - \exp(-bt^\gamma), \qquad t > 0. \tag{7.4.6}$$

The corresponding log-hazard function is given by $\alpha(t) = (\gamma - 1)\log t + \log b\gamma$ for $t > 0$. Thus $\alpha(\cdot) = \alpha(\cdot; \boldsymbol{\theta})$, where $\theta_{-1} = \theta_0 = \gamma - 1$ and $\theta_1 = \log b\gamma$. (Alternatively, we can get the Weibull model by setting $c = 0$, $\theta_{-1} = 0$, $\theta_0 = \gamma - 1$, and $\theta_1 = \log b\gamma$.)

Consider next the Pareto density function f given by

$$f(t) = \frac{bc^b}{(t+c)^{b+1}}, \qquad t > 0,$$

where $b > 0$ and $c > 0$, whose distribution function is given by

$$F(t) = 1 - \left(\frac{c}{t+c}\right)^b, \qquad t > 0. \tag{7.4.7}$$

The corresponding log-hazard function is given by $\alpha(t) = \log b - \log(t + c)$ for $t > 0$. Thus $\alpha(\cdot) = \alpha(\cdot; \boldsymbol{\theta})$, where $\theta_{-1} = 0$, $\theta_0 = -1$, and $\theta_1 = \log b$. (Here we have assumed that the parameter c of the three-parameter model coincides with the parameter c of the Pareto distribution; otherwise, the three-parameter model provides only an approximation to the Pareto distribution.)

Before applying Hare, we can use Heft to transform time so that the transformed unconditional hazard function will be approximately equal to one. To this end, let the Heft methodology be applied to (Y_i, δ_i), $1 \le i \le n$, to yield an estimate $\widehat{\lambda}_0$ of the unconditional hazard function. Let the Hare methodology then be applied to $(\widehat{q}_0(Y_i), \delta_i, \boldsymbol{x}_i)$, yielding an estimate $\widehat{\lambda}_1$ of the conditional hazard function for the transformed data and the estimate $\widehat{\lambda}(t|\boldsymbol{x}) = \widehat{\lambda}_0(t)\widehat{\lambda}_1(\widehat{q}_0(t)|\boldsymbol{x})$ of the conditional hazard function for the untransformed data; here $\widehat{q}_0 = -\log(1 - \widehat{F}_0)$ with $\widehat{F}_0$ being the distribution function corresponding to $\widehat{\lambda}_0$. (Note that if some of the covariates are time-dependent, the time argument of these covariates needs to be transformed as well.)

The unconditional hazard function of the transformation should be approximately constant on $[0, \infty)$. To see this, let T be a continuous random variable having distribution function F. Then $U = F(T)$ is uniformly distributed on $(0, \infty)$, so $-\log(1 - U) = -\log(1 - F(T))$ has the exponential distribution with mean 1, whose hazard function equals one on $[0, \infty)$.

There are two advantages of such a transformation. First, when applying Hare we can use either linear or cubic splines to model time. In either case we pay a price. When we use cubic splines the computational burden goes up considerably, since we need to resort to numerical integration. While we also need to use numerical computations for Heft, this is considerably cheaper since no covariates are present, which allows us to use some neat trickery (see Section 7.5.1). When linear splines are used the (baseline) hazard function may have big jumps in its first derivative at the various knots in time. However, the Hare model for the transformed data typically has

fewer knots in time, while the jumps in the first derivative of the baseline hazard function at these knots tend to be smaller. Secondly, because of the allowable spaces used for the Hare model, the fitted conditional hazard function beyond the last knot in time is necessarily constant. This is no longer true if the transformation based on Heft is made before applying Hare.

A disadvantage of using such a transformation is that after the data have been transformed, there is much less need for the addition of knots in time during the stepwise procedure in Hare, since the unconditional hazard function of the transformed data is approximately constant. (Some knots in time may still be added, since the largest model usually has many more basis functions than needed.) If there are far fewer knots in time available, there is less opportunity for Hare to fit nonproportional hazards models, since they require that first a knot in time be added; in addition, a basis function that is an interaction between time and a covariate has to generate an increase in the log-likelihood that is large enough to compensate for the addition of two basis functions in the BIC criterion.

Figure 7.13 shows the conditional hazard functions for the same four sets of covariates as in Figure 7.6 using three different modeling approaches: the left side of the figure shows the hazard functions using both the cubic spline version and the linear spline version of Hare (these can easily be distinguished by noting the corners in the hazard functions); the right side shows the hazard functions that are obtained after preprocessing the data using Heft. For this particular data set, the selected Hare model after preprocessing by Heft is proportional and does not contain any knots in time, although some knots in time and a nonproportional basis function in time entered during the stepwise model selection procedure. The effect on the hazard functions is particularly noticeable for the disease categories CML-CP and AA/MDS, where the tails of the hazard functions are switched.

7.4.7 Severe censoring and the penalty parameter

The value $\log n$ for the penalty parameter in BIC (3.2.29) may not be appropriate when numerous observations are censored. To see this, compare two data sets: one a regular survival analysis data set of size n; the other the same data set to which we have added $10n$ observations that are right censored at 0. Effectively, we do not add any information to the data; indeed, when we fit a fixed set of basis functions, both data sets will yield the same maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$. Rao and Wald statistics are also the same for both models. However, the penalty parameter is $\log n$ when the first model is fit and $\log 11n \approx \log n + 2.40$ when the second model is fit. This clearly seems to be inappropriate, and it suggests that $\log n$ may not be an appropriate penalty for Hare in the presence of heavy censoring.

There are other situations in survival analysis where censoring reduces the information content of various cases. For example, in power calculations

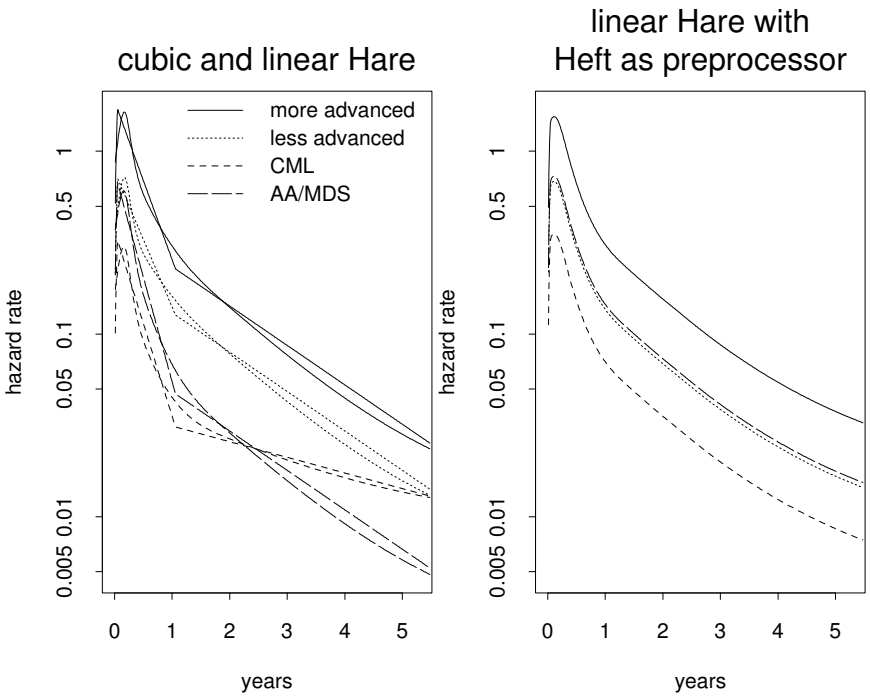

FIGURE 7.13. Estimated conditional hazard functions for the same sets of covariates as in Figure 7.6 using three different versions of Hare.

for clinical trials that are analyzed using the log-rank test, it is the number of events (uncensored observations) that is important, not the sample size (Fleming and Harrington 1991). For data sets with only right censored data, one could use $\log(\sum \delta_i)$ as the penalty parameter, but this seems inappropriate when the right censored observations are censored at a large quantile of their conditional survival distribution. In addition, counting the uncensored observations does not work when there is interval censoring. A reasonable, but admittedly ad hoc, modification is to use the penalty parameter $\log n'$ with

$$n' = n - \sum_i \left(\exp(-\widehat{\Gamma}(C_{il}|\boldsymbol{x})) - \exp(-\widehat{\Gamma}(C_{iu}|\boldsymbol{x})) \right).$$

This reduces to

$$n' = \sum_i \left(\delta_i + (1-\delta_i)(1 - \exp(-\widehat{\Gamma}(C_i|\boldsymbol{x}))) \right) \tag{7.4.8}$$

when there is no interval censoring and to

$$n' = \sum_i \left(\delta_i + (1-\delta_i)\widehat{F}(C_i|\boldsymbol{x}) \right)$$

when there are no time-dependent covariates. In practice any initial estimator of $\widehat{\Gamma}(\cdot|\boldsymbol{x})$ will do (for example, one which ignores the covariates and uses $\log n$ as the penalty parameter), since Hare is not very sensitive to the penalty parameter.

Example

For the uranium miners data $n = 3954$ but $\sum_i \delta_i = 386$. There is no interval censoring in this data set. Using as an initial estimator a Hare estimate that ignores all the covariates yields $n' = 760.3$. Thus, rather than using $\log 3954 \doteq 8.28$ as the penalty penalty, we could use $\log 760.3 \doteq 6.63$. As it turns out, for the Hare model that allows basis functions depending on time this change would result in five additional basis functions: an additional knot in time, a linear basis function for race, as well as interactions between race and cumulative radon exposure, race and birth year, and cumulative radon exposure and time. All coefficients of basis functions involving cumulative radon exposure are positive, so we could carry out an analysis for the effect of radon as was done in Section 7.4.4.

7.5 Technical details

7.5.1 Numerical integration for Heft

The main numerical task for the Heft algorithm is the computation of the log-likelihood $l(\boldsymbol{\theta})$, the score $\mathbf{S}(\boldsymbol{\theta})$, and the Hessian $\mathbf{H}(\boldsymbol{\theta})$ for various models and values of $\boldsymbol{\theta}$. The time-consuming aspect of this computation involves the numerical approximation of

$$\sum_i \int_0^{Y_i} \psi(u)du = \int_0^{\infty} N(u)\psi(u)du, \qquad N(u) = \#(\{i : Y_i \geq u\}),$$

for many functions ψ that are twice continuously differentiable on $(0, \infty)$ and infinitely differentiable on each of the intervals

$$(0, t_1],\ [t_1, t_2], \ldots,\ [t_{K-1}, t_K],\ [t_K, \infty).$$

Note that the function $N(\cdot)$ is piecewise constant, has jumps at the observations $Y_1, \ldots, Y_n$, and equals zero to the right of $Y_{(n)} = \max(Y_1, \ldots, Y_n)$.

Let $J_1, \ldots, J_M$ be a partition of $(0, Y_{(n)}]$ into disjoint intervals whose endpoints contain all of the initial knots. Then

$$\int_0^{\infty} N(u)\psi(u)du = \sum_{\nu} \int_{J_\nu} N(u)\psi(u)du.$$

Thus the time-consuming aspect of the computation involves the evaluation of $\int_J N(u)\psi(u)du$, where J is a bounded interval and ψ is an infinitely

differentiable function on a bounded interval J_0 containing J. Let b_1, b_2, b_3 and b_4 be distinct points in J_0, and let P be the cubic polynomial that interpolates the values of ψ at these points. We approximate $\int_J N(u)\psi(u)du$ by $\int_J N(u)P(u)du$. According to the Lagrange interpolation formula, $P(u) = \sum_l \psi(b_l)P_l(u)$, where $P_l(u) = \prod_{m\neq l}(u-b_m)/\prod_{m\neq l}(b_l-b_m)$. Observe that

$$\int_J N(u)P(u)du = \int_J N(u)\sum_l \psi(b_l)P_l(u) = \sum_l \psi(b_l)\int_J N(u)P_l(u)du,$$

where the quantities $\int_J N(u)P_l(u)du$ (which can be evaluated analytically) need to be obtained only once, right after the partition $J_1, \ldots, J_M$ and the four interpolation points corresponding to each of these intervals are determined.

7.6 Notes

Literature

There are numerous papers proposing methodologies based on splines for survival data. Most of the papers that appeared before 1990 either propose methods for estimating the unconditional hazard function or the baseline hazard function in the proportional hazards model using splines. The papers that appear around 1990 use splines to model a nonlinear regression function in the proportional hazards model. Many of the papers that appear after 1992 use a spline in time to create a time-varying coefficient for a covariate in a Cox model, which then becomes nonproportional.

Anderson and Senthilselvan (1990) and Senthilselvan (1987) used penalized likelihood to estimate the baseline hazard in a Cox model. The main difference between the two papers is the form of the penalty, which Anderson and Senthilselvan (1990) took to be $\int[\lambda(t)']^2dt$, while Senthilselvan (1987) used $\int[\sqrt{(}\lambda(t))']^2dt$. The later choice guarantees that the unconditional hazard function is a positive piecewise continuous function with jumps at the uncensored observations. In Senthilselvan (1987) the solution is called a hyperbolic spline. However, it is not a polynomial spline as defined in 2, but rather a function that involves exponentiations and polynomials. No algorithm for choosing the penalty parameter was provided.

Bloxom (1985) maximized an unpenalized spline to estimate the unconditional hazard function. He placed knots at the deciles of the data and added various constraints to the estimation of the parameters. His method did not deal with censored data. Klotz (1982) and Whittemore and Keller (1986) modeled the unconditional hazard function using (unpenalized) linear splines with knots at each uncensored data point. They also proposed using such splines for modeling the baseline hazard function in a proportional hazards model. Jarjoura (1988) used cubic splines to smooth the

unconditional hazard function. He first discretized the problem to regularize the solution and make the computations feasible. In addition, he added a smoothness penalty, which was optimized using a systematic search and cross-validation. Etezadi-Amoli and Ciampi (1987) used quadratic splines to model the baseline hazard function in proportional hazards models and accelerated failure time models. They used quadratic splines with very few knots. Interestingly, they numerically optimized the location of the knot.

O'Sullivan (1988a) was the first to model the log-hazard function. He used a penalty on the second derivative of this function. To circumvent the computational problem of having to put a knot at each data point, he used a smaller number of knots and B-spline basis functions. A cross-validation procedure yielded an automatic choice of the smoothing parameter.

Some density estimation procedures are also able to deal with censored data. Since there is a one-to-one correspondence between the density function and the hazard function, an estimate of the density function can be used to obtain an estimate of the hazard function. One example of a density estimation methodology that can deal with censored data is Logspline; another is the method of Abrahamowicz, Ciampi, and Ramsay (1992), which models the unconditional density function for possibly censored data using cubic splines. In this paper the number of knots, which were positioned at order statistics, were selected using AIC (with penalty parameter 2).

Herndon and Harrell (1990) modeled the unconditional hazard function using a cubic spline restricted to be linear in the tails. Initially they put four knots at restricted order statistics. Based on a visual inspection of the fit, they adjusted the number or locations of the knots. Herndon and Harrell (1995) extended the method of Herndon and Harrell (1990) for estimating the baseline hazard function in the proportional hazards model when some covariates are time-dependent. This allowed for nonproportionality for such covariates. Heinzl, Kaider, and Zlabinger (1996) proposed a model that differs only in details from that of Herndon and Harrell (1995). Rosenberg (1995) also used B-splines to model the baseline hazard function. He did not require the tails to be linear. The knots were positioned at order statistics. The number of knots was selected using AIC (with penalty parameter 2). His procedure allowed for interval censoring.

O'Sullivan (1988b) and Hastie and Tibshirani (1990) used cubic smoothing splines to estimate the effect of a covariate in a proportional hazards function. They put a penalty on the integrated second derivative of the relative risk function. Rather than putting a knot at every distinct value of the covariate, O'Sullivan (1988b) used a limited number of B-splines. Hastie and Tibshirani (1990) selected the smoothing parameter using an approximate "degrees of freedom" argument.

Sleeper and Harrington (1990) used cubic B-splines to estimate the (additive) effect of a covariate in a proportional hazards function. Parametric likelihood ratio tests were used to select the knots.

Hastie and Tibshirani (1993) discussed varying-coefficient models. In the context of survival analysis this allowed them to fit an additive model with time-varying coefficients. Gray (1992) extended the approach of Hastie and Tibshirani (1990) by also allowing for selected covariate interactions and time-varying coefficients. This relaxed the assumption of proportional hazards. Gray (1994) used models similar to those of Gray (1992) to test for (proportionality of the) covariate effects in a proportional hazards model. Hess (1994) used B-splines in time with three or four knots to fit time-varying coefficients for some covariates in the proportional hazards model. Abrahamowicz, MacKenzie, and Esdaile (1996) used a model similar to that of Hess (1994) to test for (proportionality of the) covariate effects in a proportional hazards model. Fahrmeier and Wagenpfeil (1996) and Fahrmeier and Klinger (1998) used penalized likelihood to develop varying coefficient models for discrete duration models and event history analysis models, respectively.

Within the context of the proportional hazards model, LeBlanc and Crowley (1999) used a MARS-like algorithm to model covariate effects. Their algorithm would likely yield results very similar to Hare when the underlying model satisfies the proportional hazards assumption or when Hare is used without allowing for basis functions that depend both on time and a covariate.

Gu (1994) proposed a smoothing spline procedure for estimating the conditional log-hazard function for censored survival data. In light of the computational issues involved, his procedure appears to be impractical for problems with many cases or covariates: in particular the selection of the smoothing parameter becomes intractable when the number of covariates increases. Gu (1996) developed the corresponding asymptotic theory.

Other than splines, the most popular adaptive nonparametric methods used in survival analysis extend the CART approach (Breiman, Friedman, Olshen, and Stone 1984) to survival data. There are many papers about "survival trees"; see, for example, Gordon and Olshen (1985), Segal (1988), Davis and Anderson (1989), and LeBlanc and Crowley (1992). Intrator and Kooperberg (1995) compared survival trees and Hare. Their paper contains many more references about survival tree methods.

There have been some applications to survival analysis of local polynomial methods (see, for example, Fan and Gijbels 1996) and kernel methods (Andersen, Borgan, Gill, and Keiding (1993) contains a number of applications), but the use of these approaches in survival analysis does not seem to have reached the level of splines and tree-based methods.

Software

The version of Hare and Heft that is described in Kooperberg, Stone, and Truong (1995a) is publically available from CRAN. This version of Hare employs linear splines and does not allow for interval censoring, time-

dependent covariates, or truncation. It is written in C and contains an interface to the statistical package R.

Several others have ported these S-Plus and C codes for easy installation on other platforms and under the R language. See Kooperberg's website

`http://bear.fhcrc.org/~clk/soft.html`

for current links.

A commercial version of Hare, implemented by Insightful Corporation, which includes all of the options described in this chapter (and many more), will be available in a future version of S-Plus.