BIOMETRICS 53, 1485–1494
December 1997

Hazard Regression with Interval-Censored Data

Charles Kooperberg

Department of Statistics, University of Washington,
Seattle, Washington 98195-4322, U.S.A.

and

Douglas B. Clarkson

Data Analysis Products Division of Mathsoft Inc.,
1700 Westlake N #500, Seattle, Washington 98109, U.S.A.

SUMMARY

In a recent paper, Kooperberg, Stone, and Truong (1995a) introduced hazard regression (HARE), in which linear splines and their tensor products are used to estimate the conditional log-hazard function based on possibly censored, positive response data and one or more covariates. Model selection is carried out in an adaptive fashion using maximum likelihood estimation of the unknown coefficients, Rao and Wald statistics to carry out stepwise addition and deletion of basis functions, and the Bayesian Information Criterion (BIC) to select the final model. In the present paper, the HARE methodology is extended to accommodate interval-censored data, time-dependent covariates, and cubic splines. The presence of interval-censored data means that the log-likelihood function may no longer be concave, presenting additional numerical challenges. The extended methodology is applied to a data set containing both interval-censoring and time-dependent covariates. The new software will be available in a future release of S-Plus.

1. Introduction

Consider (survival) data involving a possibly censored, positive response variable and one or more covariates. Assume that the uncensored response variable has a conditional density function, given the values of the covariates, that is positive on $[0, \infty)$.

A basic assumption of the proportional hazards model (Cox, 1972) is 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 covariates. In the hazard regression (HARE) methodology (Kooperberg, Stone, and Truong, 1995a; hereafter referred to as KST), a practical approach to modeling the conditional hazard function that does not depend on the validity of this assumption is developed. In particular, a general framework for modeling the logarithm of the conditional hazard function with linear models is described. Maximum likelihood is used to estimate the unknown parameters of the model, and a fully automatic method involving stepwise addition, stepwise deletion, and the Bayesian Information Criterion (BIC) is used to select the final model.

In HARE, splines and selected tensor products are used to estimate the logarithm of the conditional hazard function. The method is similar in spirit to Multivariate Adaptive Regression Splines (MARS; Friedman, 1991). One advantage of HARE models is that they include proportional hazards models as a subclass. The presence or absence of interaction terms between time and one or more covariates in the final model can be regarded as a check on the proportional hazards assumption.

Under suitable conditions, Kooperberg, Stone, and Truong (1995b) obtained the L_2 rate of convergence for a nonadaptive version of the methodology treated in KST. This result lends theoretical

Key words: Cubic splines; HARE; MARS; Model selection; Survival analysis; Time-dependent covariates.

support to HARE and, in particular, to the use of polynomial splines and their tensor products in defining the allowable spaces used in these procedures.

There are several other nonparametric approaches to the modeling of conditional hazard functions. The KST approach is one of the few that does not start with a proportional hazards model, but includes it as a special case. [For further discussion of the literature, see KST and Abrahamowicz, Ciampi, and Ramsay (1992).]

Two limitations of the HARE methodology, as described in KST, are that HARE cannot deal with time-dependent covariates and that it is not applicable to interval-censored data. Such data may occur, for example, if subjects are periodically monitored for the presence of a symptom-free disease. In this situation, the event (start of the disease) is interval censored between the first examination at which the patient has the disease and the examination immediately preceding it (or the start of study). Proportional hazards models have been applied to both interval-censored and right-censored data as well as to data with related censoring schemes (see, e.g., Finkelstein, 1986, and Huang, 1996, and the references therein). However, we know of no other nonparametric methodology that has been applied to interval-censored data. The parametric survival models in SURVREG (Preston and Clarkson, 1983), for example, can deal with interval-censored data.

Time-dependent covariates occur especially in situations where covariates are measured at each examination. 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). However, for the HARE methodology, in which the complete conditional hazard function is modeled, time-dependent covariates pose additional numerical challenges.

In this paper, we extend the hazard regression methodology to interval-censored data and time-dependent covariates. In the next section, we describe the HARE model, including the extensions for interval-censored data, and we also summarize the model-selection techniques developed in KST. An example of the use of the HARE methodology is described in Section 3. This example comes from an ongoing study of the natural history of anal dysplasia in gay men. The data involves interval-censoring and time-dependent covariates. The software for the new version of HARE, which employs cubic splines, will be available in a future version of S-PLUS. The version described in KST is publicly available from STATLIB. An extension of this program, which employs linear splines and can deal with interval-censored data but not with time-dependent covariates, is available from the first author.

2. The HARE Model

2.1 *Linear Models for the Conditional Log-Hazard Function*

Let M be a positive integer, and let T be a positive random variable whose distribution may depend on a vector of M (possibly time-dependent) covariates $\mathbf{x}(t) = (x_1(t), \ldots, x_M(t))$, $t \geq 0$. Suppose $\mathbf{x}(t)$ lies in the subset $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_M$ of $\mathbf{R}^M$ for each $t \geq 0$. Let $\lambda(\cdot \mid \mathbf{x}(s), s \geq 0)$ denote the conditional hazard function of T given $\mathbf{x}(s), s \geq 0$, which is assumed to exist and to be positive on $(0, \infty)$, and let $\alpha(\cdot \mid \mathbf{x}(s), s \geq 0)$ denote the conditional log-hazard function. 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 $\lambda(t \mid \mathbf{x}(s), s \geq 0) = \lambda(t \mid \mathbf{x}(t))$ and hence that $\alpha(t \mid \mathbf{x}(s), s \geq 0) = \alpha(t \mid \mathbf{x}(t))$. Standard algebra now yields that the values at time t of the corresponding conditional density function $f(t \mid \mathbf{x}(s), s \geq 0)$, conditional survival function $S(t \mid \mathbf{x}(s), s \geq 0)$, and conditional cumulative hazard function $\Lambda(t \mid \mathbf{x}(s), s \geq 0)$ depend on the values of the covariates only up to time t. Thus, we can write these values as $f(t \mid \mathbf{x}(s), s \leq t)$, $S(t \mid \mathbf{x}(s), s \leq t)$, and $\Lambda(t \mid \mathbf{x}(s), s \leq t)$.

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

$$\alpha(t \mid \mathbf{x}(t); \boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_j B_j(t \mid \mathbf{x}(t)), \qquad t \geq 0, \tag{1}$$

for the conditional log-hazard function, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$. Observe that $\lambda(t \mid \mathbf{x}(t); \boldsymbol{\beta}) = \exp \alpha(t \mid \mathbf{x}(t); \boldsymbol{\beta})$, $\Lambda(t \mid \mathbf{x}(s), s \leq t; \boldsymbol{\beta}) = \int_0^t \lambda(s \mid \mathbf{x}(s); \boldsymbol{\beta}) ds$, $f(t \mid \mathbf{x}(s), s \leq t; \boldsymbol{\beta}) = \lambda(t \mid \mathbf{x}(t); \boldsymbol{\beta}) \exp(-\Lambda(t \mid \mathbf{x}(s), s \leq t; \boldsymbol{\beta}))$, and $S(t \mid \mathbf{x}(s), s \leq t; \boldsymbol{\beta}) = \exp(-\Lambda(t \mid \mathbf{x}(s), s \leq t; \boldsymbol{\beta}))$ for $\boldsymbol{\beta} \in \mathbf{R}^p$ and $t \geq 0$.

2.2 *Censoring and Likelihood*

Given possibly censored survival data and a set $B_1, \ldots, B_p$ of basis functions, we will estimate the coefficients $\boldsymbol{\beta}$ in (1) by maximum likelihood.

Let T be the survival time, let $\mathbf{x}$ be the vector of covariates for a randomly selected individual, and let $C = [C_l, C_u]$ be a (random) subinterval of $[0, \infty)$ so that it is known only that $T \in C$. If T is uncensored, then $C = \{T\}$; if T is right-censored at $C_l < T$, then $C = [C_l, \infty)$; if T is interval-censored, then $0 \le C_l < T \le C_u < \infty$. It is assumed that T is independent of the type of censoring given $\mathbf{x}$, that when T is censored it is independent of C given $\mathbf{x}$, and that T has conditional hazard function $\lambda(\cdot \mid \mathbf{x}(s), s \ge 0)$ given $\mathbf{x}$. Let $\delta = 0$ if T is right-censored, $\delta = 1$ if T is uncensored, and $\delta = 2$ if T is interval-censored. Note that the partial likelihood corresponding to $C = (c_l, c_u)$, δ, and $\mathbf{x}$ is given by

$$\begin{aligned}&[f(c_l \mid \mathbf{x}(s), s \le c_l)]^{I(\delta=1)} \left[\int_{c_l}^{c_u} f(t \mid \mathbf{x}(s), s \le t)dt\right]^{I(\delta\ne 1)} \\ &\quad = [S(c_l \mid \mathbf{x}(s), s \le c_l)]^{I(\delta=0)}[f(c_l \mid \mathbf{x}(s), s \le c_l)]^{I(\delta=1)} \\ &\qquad \cdot [S(c_l \mid \mathbf{x}(s), s \le c_l) - S(c_u \mid \mathbf{x}(s), s \le c_u)]^{I(\delta=2)}.\end{aligned}$$

Set

$$\phi(c_l, t, \delta \mid \mathbf{x}(s), s \le t; \boldsymbol{\beta}) = \delta\alpha(c_l \mid \mathbf{x}(c_l); \boldsymbol{\beta}) - \Lambda(c_l \mid \mathbf{x}(s), s \le c_l), \qquad t \ge c_l \ge 0 \text{ and } \delta \in \{0, 1\},$$

and

$$\phi(c_l, c_u, 2 \mid \mathbf{x}(s), s \le c_u) = \log[S(c_l \mid \mathbf{x}(s), s \le c_l) - S(c_u \mid \mathbf{x}(s), s \le c_u)], \qquad c_u > c_l \ge 0.$$

These are the contributions to the log-likelihood corresponding to $\delta \in \{0, 1\}$ and $\delta = 2$, respectively.

To maximize the likelihood function and to examine its concavity, we need expressions for the partial derivatives of $\phi(\cdot \mid \cdot)$. For notational convenience, set $S(t) = S(t \mid \mathbf{x}(s), s \le t; \boldsymbol{\beta})$,

$$D_j(t) = -\frac{\partial\Lambda(t \mid \mathbf{x}(s), s \le t; \boldsymbol{\beta})}{\partial\beta_j} = -\int_0^t B_j(u \mid \mathbf{x}(u)) \exp\alpha(u \mid \mathbf{x}(u); \boldsymbol{\beta})du, \qquad 1 \le j \le p,\ t \ge 0,$$

and

$$\begin{aligned}E_{jk}(t) &= -\frac{\partial^2\Lambda(t \mid \mathbf{x}(s), s \le t; \boldsymbol{\beta})}{\partial\beta_j\partial\beta_k} \\ &= -\int_0^t B_j(u \mid \mathbf{x}(u))B_k(u \mid \mathbf{x}(u)) \exp\alpha(u \mid \mathbf{x}; \boldsymbol{\beta})du, \qquad 1 \le j, k \le p,\ t \ge 0. \qquad (2)\end{aligned}$$

Then

$$\begin{aligned}\frac{\partial}{\partial\beta_j}\phi(c_l, t, \delta \mid \mathbf{x}(s), s \le t; \boldsymbol{\beta}) &= \delta B_j(c_l \mid \mathbf{x}(c_l)) + D_j(c_l), \qquad t \ge c_l \ge 0 \text{ and } \delta \in \{0, 1\}, \\ \frac{\partial}{\partial\beta_j}\phi(c_l, c_u, 2 \mid \mathbf{x}(s), s \le c_u; \boldsymbol{\beta}) &= \frac{D_j(c_l)S(c_l) - D_j(c_u)S(c_u)}{S(c_l) - S(c_u)}, \qquad c_u > c_l \ge 0, \\ \frac{\partial^2}{\partial\beta_j\partial\beta_k}\phi(c_l, t, \delta \mid \mathbf{x}(s), s \le t; \boldsymbol{\beta}) &= E_{jk}(c_l), \qquad t \ge c_l \ge 0 \text{ and } h\delta \in \{0, 1\}, \qquad (3)\end{aligned}$$

and

$$\begin{aligned}&\frac{\partial^2}{\partial\beta_j\partial\beta_k}\phi(c_l, c_u, 2 \mid \mathbf{x}(s), s \le c_u; \boldsymbol{\beta}) \\ &\quad = \frac{(E_{jk}(c_l) + D_j(c_l)D_k(c_l))S(c_l) - (E_{jk}(c_u) + D_j(c_u)D_k(c_u))S(c_u)}{S(c_l) - S(c_u)} \\ &\qquad - \frac{(D_j(c_l)S(c_l) - D_j(c_u)S(c_u))(D_k(c_l)S(c_l) - D_k(c_u)S(c_u))}{(S(c_l) - S(c_u))^2}, \qquad c_u > c_l \ge 0.\end{aligned}$$

It follows from (2) and (3) that $\phi(t, \delta \mid \mathbf{x}; \cdot)$ is a concave function on $\mathbf{R}^p$ if $\delta \in \{0, 1\}$.

2.3 *Maximum Likelihood Estimation*

Consider n randomly selected individuals. For $1 \le i \le n$, let T_i be the survival time, $C_i = [C_{il}, C_{iu}]$ the censoring interval, and $\mathbf{x}_i(t)$, $t \le C_{iu}$, the covariates for the ith such individual. The log-likeli-

hood function corresponding to the observed data $(C_i, \delta_i, \mathbf{x}_i(t), t \le C_{iu})$, $1 \le i \le n$, and the linear model for the conditional log-hazard function discussed in the previous section is given by

$$\ell(\boldsymbol{\beta}) = \sum_i \phi(C_{il}, C_{iu}, \delta_i \mid \mathbf{x}_i(t), t \le C_{iu}; \boldsymbol{\beta}), \qquad \boldsymbol{\beta} \in \mathbf{R}^p. \tag{4}$$

If $\delta_i \in \{0, 1\}$ for all i, this is a concave function on $\mathbf{R}^p$ and the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ can be found using the Newton–Raphson method (see KST). However, if $\delta_i = 2$ for some i, then $\ell(\boldsymbol{\beta})$ may not be concave.

One numerical procedure for calculating $\hat{\boldsymbol{\beta}}$ when the log-likelihood function is not necessarily concave is to modify the Newton–Raphson method slightly by subtracting a small positive constant from the diagonal of the Hessian so that the modified Hessian is negative definite. In particular, let $\mathbf{S}(\boldsymbol{\beta})$ denote the score at $\boldsymbol{\beta}$ [that is, the p-dimensional column vector with entries $\partial l(\boldsymbol{\beta})/\partial\beta_j$], and let $\mathbf{H}(\boldsymbol{\beta})$ denote the Hessian at $\boldsymbol{\beta}$ [that is, the $p \times p$ matrix with entries $\partial^2 l(\boldsymbol{\beta})/\partial\beta_j\partial\beta_k$]. The modified Newton–Raphson method for computing $\hat{\boldsymbol{\beta}}$ is to start with an initial guess $\hat{\boldsymbol{\beta}}^{(0)}$ and iteratively determine $\hat{\boldsymbol{\beta}}^{(m+1)}$ from $\hat{\boldsymbol{\beta}}^{(m)}$ according to the formula

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \hat{\boldsymbol{\beta}}^{(m)} - 2^{-\mu}[\mathbf{G}(\hat{\boldsymbol{\beta}}^{(m)})]^{-1}\mathbf{S}(\hat{\boldsymbol{\beta}}^{(m)}).$$

Here, $\mathbf{G}(\hat{\boldsymbol{\beta}}^{(m)}) = \mathbf{H}(\hat{\boldsymbol{\beta}}^{(m)}) - a\mathbf{I}$, with a slightly larger than the largest eigenvalue of $\mathbf{H}(\hat{\boldsymbol{\beta}}^{(m)})$ if this eigenvalue is positive and $a = 0$ otherwise. This ensures that $\mathbf{G}$ is negative definite (see Kennedy and Gentle, 1980, Section 10.2.2). Furthermore, μ is the smallest nonnegative integer such that

$$l(\hat{\boldsymbol{\beta}}^{(m)} - 2^{-\mu}[\mathbf{G}(\hat{\boldsymbol{\beta}}^{(m)})]^{-1}\mathbf{S}(\hat{\boldsymbol{\beta}}^{(m)})) \ge l(\hat{\boldsymbol{\beta}}^{(m)} - 2^{-\mu-1}[\mathbf{G}(\hat{\boldsymbol{\beta}}^{(m)})]^{-1}\mathbf{S}(\hat{\boldsymbol{\beta}}^{(m)})).$$

The routine we used, the S-Plus function nlminb (FORTRAN version developed by Gay, 1984), incorporated these and other features expected in a modern optimization algorithm. We experienced no difficulties in any of the interval-censored examples we have tried so far. This experience is consistent with that of Kooperberg and Stone (1992) in the context of logspline density estimation.

2.4 *Model Selection*

When modeling the log-hazard function with a linear model (1), the remaining issue to be resolved is the choice of G. Initially, we let G be the space of constant functions. Thus, $\alpha(t \mid \mathbf{x}(t); \boldsymbol{\beta}) = \beta_1$ for $t \ge 0$, so that α does not depend on t or the vector $\mathbf{x}$ of covariates. Then we proceed with stepwise addition, successively replacing a $(p-1)$-dimensional space G_0 by a p-dimensional space G containing G_0 as a subspace. The candidates for the new basis function (a function that together with a basis of G_0 spans G) depend on which functions are already in G. For the model selection described here, it does not matter whether a covariate is time dependent.

Here we describe model selection in HARE when linear splines are used. The new implementation of HARE also allows the use of cubic splines, complicating the variable selection strategy, though the basic ideas remain essentially unchanged. [See Clarkson and Kooperberg (1996) for details.] Some remarks about other issues that arise when using higher-order splines are discussed in the next section.

Functions that are always allowed as basis functions of G are (i) piecewise linear splines in time that are of the form $B(t \mid \mathbf{x}(t)) = (t_k - t)_+$, where $(t)_+ = \max(t, 0)$ and t_k is a fixed positive number, called a knot, and (ii) linear functions in any of the covariates $x_m(t)$ for $1 \le m \le M$. Supposing that a linear function $B(t \mid \mathbf{x}(t)) = x_m(t)$ is in G, piecewise linear splines in that covariate of the form $B(t \mid \mathbf{x}(t)) = (x_m(t) - x_{mk})_+$ can also be used as basis functions (but not otherwise); here the knot x_{mk} is a fixed number in the range $\mathcal{X}_m$ of $x_m(t)$. Tensor products of any two basis functions in the model that depend on different (single) variables are also allowed. Thus, for example, if $(t_k - t)_+$ and $x_m(t)$ are basis functions of G, then $B(t \mid \mathbf{x}(t)) = (t_k - t)_+ x_m(t)$ is also allowed in the model. There is one other rule: tensor products of a basis function $B(t \mid \mathbf{x}(t)) = (x_m(t) - x_{mk})_+$ and any other basis function can only be a basis function when the tensor product between x_m and that basis function is in G.

To decide which basis function to add to a space G_0, we compute Rao (score) statistics:

1. for all spaces that can be obtained from G_0 by adding a basis function $B(t \mid \mathbf{x}(t)) = x_m(t)$ to the model;
2. for all allowable spaces that can be obtained from G_0 by adding a basis function to G_0 that is a tensor product of two basis functions that depend on different variables that are in G_0;

3. for a space that can be obtained from G_0 by adding a basis function corresponding to a potential new knot in time, using a heuristic search algorithm to find the best location for this new knot, where every observed or censored time in the data set is a candidate;
4. for a space that can be obtained from G_0 by adding a basis function corresponding to a potential new knot in a covariate that is already in the model, using a heuristic search algorithm to find the best location for this new knot, where every value of the covariate in its range in the data is a candidate.

As the new space G, we choose the one corresponding to the largest Rao statistic. The use of Rao statistics can be motivated by examining a quadratic approximation to the log-likelihood function near the maximum likelihood estimate $\hat{\boldsymbol{\beta}}_0$ of $\boldsymbol{\beta}_0$ in G_0. In particular, it is easily shown that if the log-likelihood function were exactly quadratic, then the Rao statistic would be twice the increase of the log-likelihood for adding a basis function. [See Kooperberg, Bose, and Stone (1997) for details.]

Upon stopping the stepwise addition stage, we proceed to stepwise deletion. Here we successively replace the p-dimensional space G by a $(p-1)$-dimensional subspace G_0. The basis functions of the subspace G_0 must satisfy the same rules listed above that applied during stepwise addition. In particular, at each step we choose the candidate space G_0 such that the Wald statistic for a basis function that is in G but not in G_0 is smallest in magnitude. The use of Wald statistics during stepwise deletion can be motivated in the same way as the use of Rao statistics during stepwise addition.

During the combination of stepwise addition and stepwise deletion, we get a sequence of models indexed by ν, with the νth model having p_ν parameters. Let $\hat{l}_\nu$ denote the log-likelihood of the νth model, and let

$$\mathrm{AIC}_{a,\nu} = -2\hat{l}_\nu + ap_\nu \tag{5}$$

be the Akaike Information Criterion with penalty parameter a for this model. We select the model corresponding to the value $\hat{\nu}$ of ν that minimizes $\mathrm{AIC}_{a,\nu}$. In light of KST and our experience in the present investigation, we recommend choosing $a = \log n$ as in BIC due to Schwarz (1978). Section 3 contains a further discussion of the choice of the penalty parameter a in the presence of censoring.

Since the log-likelihood function may not be concave, good starting values for the (modified) Newton–Raphson iterations are even more crucial than usual. In the context of stepwise addition, we use the maximum likelihood estimate from the previous model, which is possible since the new linear space contains the previous one as a proper subspace. In the context of stepwise deletion, the starting value $\hat{\boldsymbol{\beta}}^{(0)}$ can be obtained by considering a quadratic approximation to the log-likelihood $\ell(\cdot)$.

2.5 *Innovations*

The new implementation of HARE has a number of additional features, some of which have already been mentioned.

Cubic splines. A one-dimensional function that is continuous and piecewise linear is called a linear spline. The basis functions used in the model-selection procedure of the previous section are linear splines. A twice continuously differentiable, piecewise cubic function is called a cubic spline. The new implementation of HARE can use either linear or cubic splines for its basis functions. In either case, we require that the conditional log-hazard function be linear in its tails. Additionally, time-dependent covariates must be constant in the upper tail.

The model-selection strategy with cubic splines is generally the same as the one with piecewise linear splines outlined above. A complication is the fact that cubic polynomial basis functions depend on more than one knot. The tail conditions also complicate the basis functions and thus knot-selection procedures. Even so, it is still possible to develop model-selection algorithms in which addition and deletion of basis functions is, essentially, addition and deletion of knots. [See Clarkson and Kooperberg (1996) for details.]

Categorical covariates. The new implementation of HARE can also accommodate categorical covariates. Let V_i denote a subset of the possible values of the categorical covariate. HARE allows as basis functions indicator variables that are set to 1 if the value of the covariate is in V_i and 0 otherwise. HARE considers all possible subsets V_i not currently in the model and chooses the subset yielding maximum Rao statistic when building a model. Wald statistics are used for removing basis functions from the model. Basis functions for categorical covariates can also enter the model in tensor products with other basis functions.

Time-dependent covariates. So far, we have acted as if time-dependent covariates are known at every time t. In practice, time-dependent covariates are often measured at discrete time points, usually when patients come in for examinations. For continuous covariates, the new implementation of HARE allows for piecewise linear (constant in the tails) interpolation of the time-dependent covariates. For categorical covariates, the new implementation allows piecewise constant time-dependent covariates.

While the implementation of time-dependent covariates is conceptually straightforward, their presence considerably complicates the coding. For example, since the number of times at which the covariates are measured may be different for each subject, it is no longer convenient to use a matrix to store the data. Rather, a list of vectors (where each vector may have a different length) is used. Also, even for piecewise linear splines, numerical integration methods may be required to compute the integrals in the log-likelihood when continuous time-dependent covariates are in the model. Numerical integration is not as complicated as it might first appear, however, since all integrals are sums of integrals of the form $\int_a^b p_i(t) \exp p_2(t) dt$, where p_1 and p_2 are polynomials. Integrals such as these are easily computed using Gaussian quadrature (Abramowitz and Stegun, 1964), but complicated bookkeeping is required to keep track of the limits a and b of integration.

3. An Example

The data for our example come from an ongoing study of the natural history of anal dysplasia in gay HIV-positive and HIV-negative men who are enrolled in the AIDS Prevention Project of the Seattle–King County Department of Public Health. At enrollment, subjects undergo an interview, anal examination with collection of specimens, in particular for cytology (pap smear for detection of cancer or precancer), and collection of blood for CD4 count and other serologic tests. Subjects return every 3 to 6 months for follow-up interviews, collections of specimens for cytology and CD4 count, and anal exams. Evidence of precancer is our event of interest. The data is interval-censored since we do not know the precise time between two interviews when the precancerous conditions developed. In all, there were 897 subjects, but since in our analysis we include only patients who had no precancerous conditions when they entered the study, 626 subjects remain. As the event, we consider the first detection of precancer (high- or low-grade squamous intraepithelial lesions, SIL). Among the 626 subjects, there were 250 events. Of the corresponding 250 subjects, 99 had developed SIL at their first return visit. The observation time for these subjects was thus interval censored between 0 and the time of their first visit, the median of which was 21 days. The other 151 events were interval-censored, with median times to the lower and upper end of the censoring interval being 241 and 434 days, respectively, and the median length of the interval being 161 days. The remaining 399 subjects were right-censored with a median follow-up time of 441 days.

We used 14 covariates. Six time-dependent covariates were measured at every visit, including three types of HPV (human papillomavirus, which causes genital warts): HPV 16/18/45 status, HPV 31/33/35 status, HPV 6/11 status; HIV status; CD4 count; and whether the subject had a new partner since his previous visit. The other eight covariates were all fixed at entry of the study: the age of the person, the age at first receptive anal intercourse, a variable indicating the cumulative number of male partners (coded on a scale from 0 to 6), smoking, IV drug use, a history of warts, syphilis serology, and race. While in practice the assumption that the conditional hazard function only depends on the current value of the covariates may be questionable, it seems reasonable to make that assumption in the current situation.

Prior to our analysis, it was hypothesized by Professor Critchlow (private communication) that HPV, and in particular HPV 16/18/45, is the primary cause of SIL; that HIV and/or low CD4 count increases risk of SIL; and that there is an interaction between low-risk HPV (HPV 6/11) and HIV in that these HPV infections are more likely to result in SIL among HIV-positive than among HIV-negative subjects. This interaction had not been established by other analyses.

The model that HARE selected (see the second and third columns of Table 1) included a constant term, a basis function in time, linear basis functions for three of the HPV variables, HIV status, age, age at first anal intercourse, and interactions between HPV 6/11 and time and between HPV 6/11 and HIV status. The basis function in time is a cubic spline that decreases smoothly and is constant beyond 40 days. Because of the HPV 6/11 $\times$ time interaction, the model is not a proportional hazards model.

From the coefficients in the second column of Table 1, we note that positive HPV 16/18/45, HPV 31/33/35, and HIV status increase the risk of SIL. Decreasing age, a low age of first receptive anal intercourse, and a large number of partners also increase the risk of SIL. Since the coefficient of the HPV 6/11 $\times$ HIV interaction is negative, we note that HPV 6/11 is a larger risk factor for

Table 1
HARE analysis of the cytology data

	Default value for a			$8.71 < a < 31.56$		
Basis function	Coefficient	Std error	Coef/Std err	Coefficient	Std error	Coef/Std err
Constant	−8.9801	0.4752	−18.8959	−9.6415	0.2436	− 39.5747
Time	2.7670	0.4074	6.7924	3.8101	0.1871	20.3689
HPV 16/18/45	0.4484	0.1108	4.0483	0.5999	0.1025	5.8541
HPV 6/11	0.8399	0.1856	4.5233	1.1883	0.1688	7.0389
HIV	1.6623	0.2974	5.6486	1.9322	0.2780	6.9510
HIV × HPV 6/11	−0.9760	0.1995	−4.8936	−0.9558	0.1996	−4.7892
HPV 31/33/35	0.4010	0.1225	3.2740	NA	NA	NA
CD4	−0.0007	0.0002	−2.9186	NA	NA	NA
Age	−0.0387	0.0105	−3.6694	NA	NA	NA
Age at first anal intercourse	0.0521	0.0128	4.0880	NA	NA	NA
Number of partners	0.1466	0.0495	2.9621	NA	NA	NA
HPV 6/11 × time	0.8129	0.2737	2.9702	NA	NA	NA

HIV-negative men, which contradicts the hypothesized effect. Further exploratory data analysis suggested that in this data set the interaction should have a negative sign.

An advantage of HARE over traditional proportional hazards modeling is that HARE not only estimates the effect of covariates, but it also provides a (smooth) estimate of the underlying hazard function. In Figure 1, we show the log-hazard function for the cytology data for an HIV-negative subject of age 30, age at first anal intercourse of 18, with HPV 16/18/45 and HPV 31/33/35 equal to 0 throughout, 5 partners, with a CD4 count of 1000, and with HPV 6/11 equal to 0 (negative, dashed curve) and 1 (positive, solid curve). The nonproportionality of the HARE model is evident from this figure since the spacing between the log-hazard functions changes with time. Since these conditional hazard functions are constant beyond 45 days, we truncated the plot at 80 days.

In practice, the values of the time-dependent covariates change during the study. In Figure 2, we show the estimated conditional log-hazard function for three patients in the data set. The patient whose conditional hazard function is the solid line is an HIV-negative male who first had SIL diagnosed after 1149 days. His HPV 6/11 status increased from 0 to 2 between 904 days and 1145

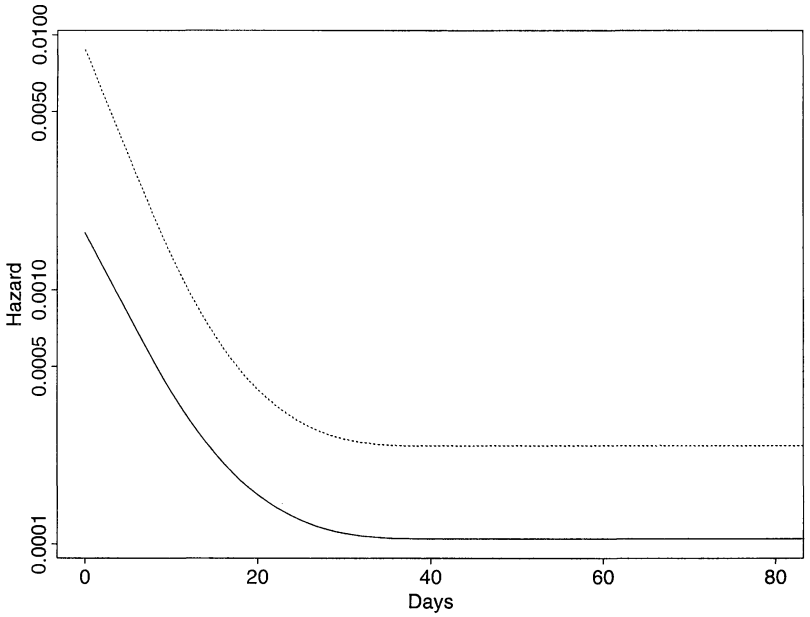

Figure 1. Log-hazard functions for two hypothetical subjects for whom the values of the time-dependent covariates do not change. Solid: HPV 6/11 is 1; dashed HPV 6/11 is 0. The distance between the curves at 0 days is almost twice as large as at 80 days.

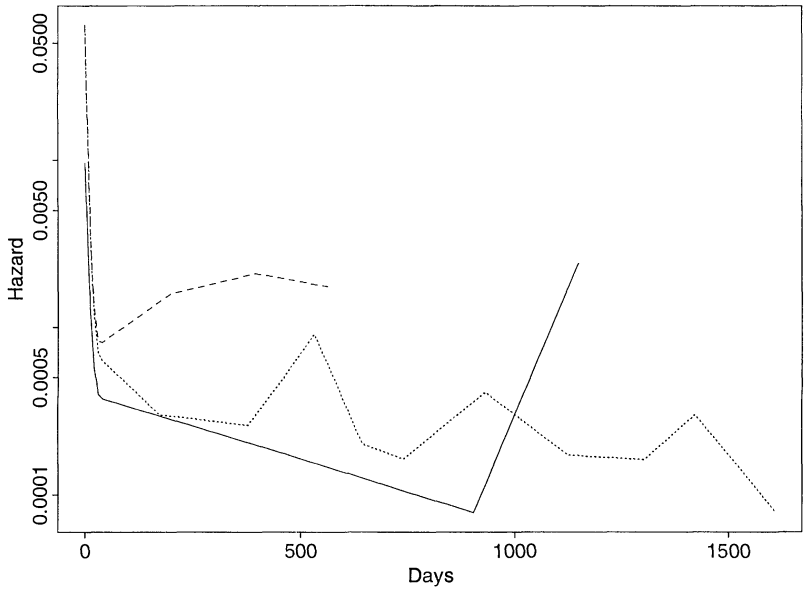

Figure 2. Log-hazard functions for three subjects for whom the value of HPV 6/11 varied over time.

days, explaining the sharp rise in the hazard function. The dotted line is for an HIV-negative male whose HPV 6/11 and HPV 16/18/45 status changed a number of times during the study, which explains why his hazard function goes up and down so much. The dashed line is for an HIV-positive subject; his hazard function is clearly much higher than those for the other two subjects.

Since the time origin is arbitrary, it appeared strange that the hazard functions in Figure 1, for subjects whose time-dependent covariates do not change, were not constant. However, personal communication with Professor Cathy Critchlow revealed that some of the subjects do not have their SIL status measured at their first visit, due to the extra amount of administration needed at that time, and have absence of SIL recorded in their record. Some of these subjects presumably already had SIL at that time. Since these patients get diagnosed at their second visit, we have a number of extra subjects who are interval censored between the first and the second visit that are erroneously added to the data set, yielding a peak in the hazard rate close to 0.

It is of interest to look at smaller models selected by HARE. As it turned out, the model summarized in the second and third column of Table 1 was optimal for a between 5.54 and 7.08, while the default value for a was $\log 626 = 6.44$. Were a chosen between 8.71 and 31.56, a much smaller model with only six basis functions would be selected. Interesting enough, this model still contains the three effects that were hypothesized by Professor Critchlow prior to our analysis as being the most influential: an effect of HPV 16/18/45, an effect of HIV status, and an interaction between HPV 6/11 and HIV status. (If a were chosen between 7.08 and 8.71, a nonproportional model with eight basis functions would be selected.)

Since this smaller model does not contain any interactions with time, it is a proportional hazards model. It is of interest that HARE automatically compares proportional and nonproportional hazards models. Actually, the interface to HARE makes it possible to force HARE to fit a proportional hazards model. When we did this, we again obtained the model summarized in columns four and five of Table 1.

In general, it is reasonable to examine models for a variety of choices of the penalty parameter a in (5). In particular, it is plausible that, if a substantial number of the observations are censored, then a should be smaller than $\log n$. For example, if we add a subject that is right-censored at 0 or interval-censored between 0 and ∞, this would increase n by 1, and hence it would increase $a = \log n$, but it clearly would not increase the amount of signal in the data.

4. Concluding Remarks

The HARE methodology, as described in KST and extended in this paper, should be a useful addition to the survival analysis toolkit. Its features make it easy to try a variety of models. In particular, linear proportional hazard models, additive proportional hazards models, proportional

hazards models with time-varying coefficients, and nonproportional hazards models can conveniently be fitted and compared.

There are many studies that yield interval-censored data. However, most survival analysis methods cannot deal with such data. The extension of HARE to deal with time-dependent covariates and interval censoring should increase its applicability.

Acknowledgements

We would like to thank Charles J. Stone and Cathy W. Critchlow for many suggestions, which improved this paper considerably, and Glenn Satten for spotting a mistake in an earlier version of this manuscript. This work was supported in part by National Science Foundation grant DMS-9403371 and National Institutes of Health grants CA 61937 and CA 65156. The data for the example were kindly provided to us by Professor Cathy W. Critchlow of the Department of Epidemiology of the University of Washington. Data collection was supported by National Institutes of Health grant CA 55488.

Résumé

Dans un article récent Kooperberg, Stone et Truong (1995a) ont défini une régression de force de mortalité ('hazard regression' (HARE)) dans laquelle des fonctions splines linéaires et leur produits tensoriels sont utilisés pour estimer la fonction du logarithme de la force de mortalité conditionnelle basée sur des réponses positives pouvant être censurées et sur une ou plusieurs covariables. Un modèle de sélection adaptatif est mis au point utilisant, l'estimation du maximum de vraisemblance des paramètres inconnus, les statistiques de Rao et Wald pour l'entrée ou la sortie des fonctions de base, et le critère de l'information bayésienne (BIC) pour la sélection du modèle final. Dans le présent article la méthodologie HARE est généralisée pour prendre en compte des données censurée par intervalle, des covariables temps-dépendantes et des fonctions splines cubiques. La présence de données censurées par intervalle signifie que la fonction de log-vraisemblance ne peut plus être concave, introduisant des problèmes numériques supplémentaires. La méthode généralisée est appliquée à un jeu de données ayant à la fois des censures par intervalle et des covariables temps-dépendantes. Le nouveau logiciel sera disponible dans une version prochaine de S-Plus.

References

Abrahamowicz, M., Ciampi, A., and Ramsay, J. O. (1992). Nonparametric density estimation for censored survival data: Regression-spline approach. *The Canadian Journal of Statistics* **20,** 171–185.

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions.* Washington, D.C.: National Bureau of Standards.

Clarkson, D. B. and Kooperberg, C. (1996). *Hazard regression user's guide.* Technical Report, MathSoft, Seattle.

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

Finkelstein, D. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* **42,** 845–854.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *The Annals of Statistics* **19,** 1–141.

Gay, D. M. (1984). A thrust region approach to linearly constrained optimization. In *Numerical Analysis,* F. A. Lootsma (ed), 171–189. Berlin: Springer.

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *The Annals of Statistics* **24,** 540–568.

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

Kennedy, W. J. and Gentle J. E. (1980). *Statistical Computing.* New York: Marcel Dekker.

Kooperberg, C. and Stone, C. J. (1992). Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics* **1,** 301–328.

Kooperberg, C., Stone, C. J., and Truong, Y. K. (1995a). Hazard regression. *Journal of the American Statistical Association* **90,** 78–94.

Kooperberg, C., Stone, C. J., and Truong, Y. K. (1995b). The L_2 rate of convergence for hazard regression. *Scandinavian Journal of Statistics* **22,** 143–157.

Kooperberg, C., Bose, S., and Stone, C. J. (1997). Polychotomous regression. *Journal of the American Statistical Association* **92,** 117–127.

Preston, D. L. and Clarkson, D. B. (1983). SURVREG—An interactive program for the analysis of survival regression models. *The American Statistician* **37,** 174.

Schwarz (1978). Estimating the dimension of a model. *The Annals of Statistics* **6,** 461–464.

Received August 1995; *revised October* 1996 *and April* 1997; *accepted April* 1997.