"Smooth" Semiparametric Regression Analysis for Arbitrarily Censored Time-to-Event Data

Min Zhang* **and Marie Davidian**

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203, U.S.A.

**email:* mzhang4@stat.ncsu.edu

SUMMARY. A general framework for regression analysis of time-to-event data subject to arbitrary patterns of censoring is proposed. The approach is relevant when the analyst is willing to assume that distributions governing model components that are ordinarily left unspecified in popular semiparametric regression models, such as the baseline hazard function in the proportional hazards model, have densities satisfying mild "smoothness" conditions. Densities are approximated by a truncated series expansion that, for fixed degree of truncation, results in a "parametric" representation, which makes likelihood-based inference coupled with adaptive choice of the degree of truncation, and hence flexibility of the model, computationally and conceptually straightforward with data subject to any pattern of censoring. The formulation allows popular models, such as the proportional hazards, proportional odds, and accelerated failure time models, to be placed in a common framework; provides a principled basis for choosing among them; and renders useful extensions of the models straightforward. The utility and performance of the methods are demonstrated via simulations and by application to data from time-to-event studies.

KEY WORDS: Accelerated failure time model; Heteroscedasticity; Information criteria; Interval censoring; Proportional hazards model; Proportional odds model; Seminonparametric (SNP) density; Time-dependent covariates.

1. Introduction

Regression analysis of censored time-to-event data is of central interest in health sciences research, and the most widely used approaches are based on semiparametric models. While representing some feature of the relationship between time-to-event and covariates by a parametric form, these models leave other aspects of their distribution unspecified.

Among such models, Cox's proportional hazards model (PH) (Cox, 1972) is unquestionably the most popular and is used almost by default in practice when data are right-censored, owing to straightforward, widely-available implementation. The hazard given covariates is represented as a parametric form modifying multiplicatively an unspecified baseline hazard function. This proportional hazards assumption is often not checked; however, effects of prognostic covariates often do not exhibit proportional hazards (e.g., Gray, 2000). Accordingly, there is considerable interest in alternative semiparametric regression models.

The accelerated failure time model (AFT) (Kalbfleisch and Prentice, 2002, sec. 2.2.3), in contrast to the PH model, where survival time and covariate effects are modeled indirectly through the hazard, represents the logarithm of event time directly by a parametric function of covariates plus a deviation with unspecified distribution, lending it practical appeal. However, this and similar models are used infrequently, likely due to computational challenges that undoubtedly dictate lack of commercially-available software. Although the iterative fitting method of Buckley and James (1979) (see also, e.g., Lin and Wei, 1992) for right-censored data is simple to program, it can exhibit problematic behavior, such as oscillation between two "solutions" (Jin, Lin, and Ying, 2006). Competing approaches based on rank tests (e.g., Tsiatis, 1990; Wei, Ying, and Lin, 1990; Jin et al., 2003) may also admit multiple solutions (or have no solutions at all), can be computationally intensive (Lin and Geyer, 1992), and/or can involve rather complicated estimation of sampling variance.

The proportional odds (PO) model (Murphy, Rossini, and van der Vaart, 1997; Yang and

Prentice, 1999) instead represents the logarithm of the ratio of the odds of survival given covariates to the baseline odds as a parametric function of covariates, where the associated baseline survival function is left unspecified. Despite its pleasing interpretation, the PO model is rarely used, again likely due to difficulty of implementation.

Hence, although the regression parameters in all of these models have intuitive interpretations, and although one model may be more suitable for representing the data than another, only the PH model is widely used. The PH and PO models are special cases of the linear transformation model (Cheng, Wei, and Ying, 1995; Chen, Jin, and Ying, 2002); Cheng, Wei, and Ying (1997) and Scharfstein, Gilbert, and Tsiatis (1998) also discuss a general class of models that includes both. The AFT and PH models are cases of the "extended" hazards model of Chen and Jewell (2001), including the "accelerated hazards" model of Chen and Wang (2000). However, there is no accessible framework that includes all three models and, indeed, further competitors, in which selection among them may be conveniently placed.

Moreover, the majority of developments for semiparametric time-to-event regression have been for right-censored (independently given covariates) event times. Fitting the PH model is straightforward under these conditions, but with interval censoring, specialized methods are required (Finkelstein, 1986; Satten, Datta, and Williamson, 1998; Goetghebeur and Ryan, 2000; Pan, 2000; Betensky et al., 2002), as they are for alternative models (e.g., Betensky, Rabinowitz, and Tsiatis, 2001; Sun, 2006). This requires the analyst to seek out specialized, distinct techniques for different censoring patterns, even for the familiar PH model.

In this paper, we propose a general framework for semiparametric regression analysis of censored time-to-event data that (i) provides a foundation for selection among competing models; (ii) unifies handling of different patterns of censoring, obviating the need for specialized techniques; and (iii) is computationally tractable regardless of model or censoring pattern. To achieve simultaneously (i)–(iii), we sacrifice a bit of generality relative to tradi-

tional semiparametric methods by making the unrestrictive assumption that the distribution associated with unspecified model components has a density satisfying mild "smoothness" assumptions. Indeed, large sample theory for traditional methods requires similar assumptions (e.g., Ritov, 1990; Tsiatis, 1990; Jin et al., 2006). We assume that densities lie in a broad class whose elements may be approximated by the "SemiNonParametric" (SNP) density estimator of Gallant and Nychka (1987), tailored to provide an excellent approximation to virtually any plausible survival density. Many authors have used smoothing techniques in time-to-event regression (e.g., Kooperberg and Clarkson, 1997; Joly, Commenges, and Letenneur, 1998; Cai and Betensky, 2003; Komárek, Lesaffre, and Hilton, 2005). Our SNP approach endows likelihood-based inference for any of these models with "parametric-like" features and a virtually closed-form objective function under arbitrary censoring patterns that admits tractable implementation with standard optimization software. Competing models may be placed in a unified likelihood-based framework, providing a convenient, defensible basis for choosing among them via standard model selection techniques.

In Section 2, we review the SNP representation and describe its use in approximating any plausible survival density. We discuss SNP-based semiparametric time-to-event regression analysis with arbitrary censoring in Section 3. Simulation studies in Section 4 demonstrate performance. In Section 5, we apply the methods to two well-known data sets.

2. SNP Representation of a Survival Density

Gallant and Nychka (1987) gave a mathematical description of a class $\mathcal{H}$ of k-dimensional "smooth" densities that are sufficiently differentiable to rule out "unusual" features such as jumps or oscillations but that may be skewed, multi-modal, or fat- or thin-tailed. When $k = 1$, $\mathcal{H}$ includes almost any density that is a realistic model for a (possibly transformed), continuous time-to-event random variable and excludes implausible candidates. For $k = 1$, densities $h \in \mathcal{H}$ may be expressed as an infinite Hermite series $h(z) = P_\infty^2(z)\psi(z)$ plus a

lower bound on the tails, where $P_\infty(z) = a_0 + a_1 z + a_2 z^2 + \cdots$ is an infinite-dimensional polynomial; $\psi(z)$ is the "standardized" form of a known density with a moment generating function, the "base density;" and $h(z)$ has the same support as $\psi(z)$. The base density is almost always taken as $\mathcal{N}(0,1)$, but need not be (see below). For practical use, the lower bound is ignored and the polynomial is truncated, yielding the so-called SNP representation

$$h_K(z) = P_K^2(z)\psi(z), \quad P_K(z) = a_0 + a_1 z + a_2 z^2 + \cdots + a_K z^K, \quad \boldsymbol{a} = (a_0, a_1, \ldots, a_K)^T, \quad (1)$$

With $\boldsymbol{a}$ such that $\int h_K(z)\, dz = 1$ and K suitably chosen, $h_K(z)$ provides a basis for estimation of $h(z)$. The SNP representation has been widely used, particularly in econometric applications. Web Appendix A gives more detail on the SNP and its properties.

Zhang and Davidian (2001) noted that requiring $\int h_K(z)\, dz = \int P_K^2(z)\psi(z)\, dz = 1$ is equivalent to requiring $E\{P_K^2(U)\} = \boldsymbol{a}^T \boldsymbol{A} \boldsymbol{a} = 1$, where U has density ψ, and $\boldsymbol{A}$ is a known positive definite matrix easily calculated for given ψ. Thus, $\boldsymbol{a}^T \boldsymbol{A} \boldsymbol{a} = \boldsymbol{c}^T \boldsymbol{c} = 1$, suggesting the spherical transformation $c_1 = \sin(\phi_1)$, $c_2 = \cos(\phi_1)\sin(\phi_2)$, $\ldots$, $c_K = \cos(\phi_1)\cos(\phi_2)\cdots\cos(\phi_{K-1})\sin(\phi_K)$, $c_{K+1} = \cos(\phi_1)\cos(\phi_2)\cdots\cos(\phi_{K-1})\cos(\phi_K)$ for $-\pi/2 < \phi_j \leq \pi/2$, $j = 1, \ldots, K$. Web Appendix B presents examples of this formulation. Thus, for fixed K, (1) is "parameterized" in terms of $\boldsymbol{\phi}$ $(K \times 1)$ and we write $h_K(z; \boldsymbol{\phi}) = P_K^2(z; \boldsymbol{\phi})\psi(z)$; estimation of the finite-dimensional "parameter" $\boldsymbol{\phi}$ leads to an estimator for $h(z)$.

With $K = 0$ in (1), $P_K^2(z) \equiv 1$, and $h_K(z)$ reduces to the base density; i.e., $h_K(z) = \psi(z)$. Values $K > 1$ control the extent of departure from ψ and hence flexibility for approximating the true $h(z)$ (K is not the number of components in a mixture). Several authors (e.g., Fenton and Gallant, 1996; Zhang and Davidian, 2001) have shown that $h_K(z; \boldsymbol{\phi})$ with $K \leq 4$ can well-approximate a diverse range of true densities.

We now describe how we use (1) to approximate the assumed "smooth" density $f_0(t)$ of a continuous, positive, time-to-event random variable T_0 with survival function $S_0(t) =$

$P(T_0 > t)$, $t > 0$. As T_0 is positive, we assume that we may write

$$\log(T_0) = \mu + \sigma Z, \quad \sigma > 0, \tag{2}$$

where Z takes values in $(-\infty, \infty)$. We consider two formulations that together are sufficiently rich to support an excellent approximation to virtually any $f_0(t)$. In (2), it is natural to assume that Z has density $h \in \mathcal{H}$ that may be approximated by (1) with $\mathcal{N}(0,1)$ base density $\psi(z) = \varphi(z)$ for suitably chosen K, so that T_0 is lognormally distributed when $K = 0$. Although this can approximate very skewed or close-to-exponential f_0 with large enough K, an alternative formulation better suited to this case is to assume that $Z^* = e^Z$ has density $h \in \mathcal{H}$ that may be approximated by (1) with standard exponential base density $\psi(z) = \mathcal{E}(z) = e^{-z}$, so that Z has an extreme value distribution (Kalbfleisch and Prentice, 2002, sec. 2.2.1) when $K = 0$. As discussed in Section 3, we propose choosing the representation (normal or exponential) and associated K best supported by the data.

In both cases, approximations for $f_0(t)$ and $S_0(t)$ follow straightforwardly. Under the normal base density representation, for fixed K and $\boldsymbol{\theta} = (\mu, \sigma, \boldsymbol{\phi}^T)^T$, we have for $t > 0$

$$\begin{aligned} f_{0,K}(t;\boldsymbol{\theta}) &= (t\sigma)^{-1} P_K^2\{(\log t - \mu)/\sigma;\, \boldsymbol{\phi}\}\, \varphi\{(\log t - \mu)/\sigma\}, \\ S_{0,K}(t;\boldsymbol{\theta}) &= \int_{(\log t - \mu)/\sigma}^{\infty} P_K^2(z;\boldsymbol{\phi})\, \varphi(z)\, dz. \end{aligned} \tag{3}$$

Because $P_K^2(z;\boldsymbol{\phi})$ may be written as $\sum_{k=0}^{2K} d_k z^k$, where the d_k are functions of the elements of $\boldsymbol{\phi}$, $S_{0,K}(t;\boldsymbol{\theta})$ in (3) may be written as a linear combination of integrals of the form $I(k,c) = \int_c^\infty z^k \varphi(z) dz$ that satisfy $I(k,c) = c^{k-1}\varphi(c) + (k-1)I(k-2,c)$ for $k \geq 2$, where $I(0,c) = 1 - \Phi(c)$, $I(1,c) = \varphi(c)$, and $\Phi(\cdot)$ is the $\mathcal{N}(0,1)$ cumulative distribution function (cdf). For the exponential base density representation, we have approximations

$$\begin{aligned} f_{0,K}(t;\boldsymbol{\theta}) &= (\sigma e^{\mu/\sigma})^{-1}\, t^{(1/\sigma - 1)}\, P_K^2\{(t/e^\mu)^{1/\sigma};\boldsymbol{\phi}\}\, \mathcal{E}\{(t/e^\mu)^{1/\sigma}\} \\ S_{0,K}(t;\boldsymbol{\theta}) &= \int_{(t/e^\mu)^{1/\sigma}}^{\infty} P_K^2(z;\boldsymbol{\phi})\, \mathcal{E}(z)\, dz, \end{aligned} \tag{4}$$

where, similar to the normal base case, the integral in (4) may be calculated using the recursion $I(k,c) = c^k\mathcal{E}(c) + kI(k-1,c)$, $k > 0$, with $I(0,c) = e^{-c}$. Note, then, that for fixed K, except for the need for a routine to calculate the normal cdf, the approximations of $f_0(t)$ and $S_0(t)$ using either base density representation are in a closed form depending on the "parameter" $\boldsymbol{\theta}$, whose finite dimension depends on K. This offers computational advantages and makes handling of arbitrary censoring patterns straightforward, as we demonstrate next.

3. Censored Data Regression Analysis Based on SNP

3.1 *Popular Regression Models*

Let $\boldsymbol{X}_i$ be a vector of time-independent covariates and T_i be the event time, with $(T_i, \boldsymbol{X}_i)$ independent and identically distributed (iid) for $i = 1, \ldots, n$. The usual PH model is

$$\lambda(t|\boldsymbol{X};\boldsymbol{\beta}) = \lim_{\delta\to 0^+} \delta^{-1}P(t \le T < t+\delta \,|\, T \ge t, \boldsymbol{X}) = \lambda_0(t)\exp(\boldsymbol{X}^T\boldsymbol{\beta}), \quad t > 0, \quad (5)$$

where $\lambda_0(t)$ is the baseline hazard function corresponding to $\boldsymbol{X} = \boldsymbol{0}$. Letting $S(t|\boldsymbol{X};\boldsymbol{\beta}) = P(T > t|\boldsymbol{X})$ be the conditional survival function for T given $\boldsymbol{X}$, it is straightforward (Kalbfleisch and Prentice, 2002, sec. 4.1) to show that $S(t|\boldsymbol{X};\boldsymbol{\beta}) = S_0(t)^{\exp(\mathbf{X}^T\boldsymbol{\beta})}$, where $S_0(t) = \exp\{-\int_0^t \lambda_0(u)\,du\}$ is the baseline survival function associated with $\lambda_0(t)$. Usually, $\lambda_0(t)$ is left completely unspecified, whereupon (5) is a semiparametric model, and $\boldsymbol{\beta}$ characterizing the hazard relationship is estimated via partial likelihood (PL; Kalbfleisch and Prentice, 2002, sec. 4.2). We instead impose the mild restriction that $S_0(t)$ is the survival function of a random variable T_0 satisfying (2) with density $f_0(t)$, and that $f_0(t)$ and $S_0(t)$ may be approximated by either (3) or (4). Letting the conditional density of $T|\boldsymbol{X}$ be $f(t\,|\,\boldsymbol{X};\boldsymbol{\beta})$, we obtain approximations to $S(t\,|\,\boldsymbol{X};\boldsymbol{\beta})$ and $f(t\,|\,\boldsymbol{X};\boldsymbol{\beta})$ for fixed K given by

$$S_K(t|\boldsymbol{X};\,\boldsymbol{\beta},\boldsymbol{\theta}) = S_{0,K}(t;\boldsymbol{\theta})^{\exp(\mathbf{X}^T\boldsymbol{\beta})}, \quad f_K(t|\boldsymbol{X};\,\boldsymbol{\beta},\boldsymbol{\theta}) = e^{\mathbf{X}^T\boldsymbol{\beta}}\lambda_{0,K}(t;\boldsymbol{\theta})S_K(t|\boldsymbol{X};\,\boldsymbol{\beta},\boldsymbol{\theta}), \quad (6)$$

where $\lambda_{0,K}(t;\boldsymbol{\theta}) = f_{0,K}(t;\boldsymbol{\theta})/S_{0,K}(t;\boldsymbol{\theta})$. As we demonstrate shortly, the approximations in (6) may be substituted into a likelihood function appropriate for the censoring pattern of

interest, upon which estimation of $(\boldsymbol{\beta}, \boldsymbol{\theta})$ and choice of K and the base density may be based.

We propose a similar formulation for the usual AFT model

$$\log(T_i) = \boldsymbol{X}_i^T \boldsymbol{\beta} + e_i, \quad e_i \text{ iid.} \tag{7}$$

Rather than taking the distribution of the "errors" e_i to be completely unspecified, we assume that $e_i = \log(T_{0i})$, where T_0 has survival function $S_0(t)$ and "smooth" density $f_0(t)$ that may be approximated by (3) or (4). For fixed K, this leads to approximations to the conditional survival and density functions of $T|\boldsymbol{X}$, $S(t \,|\, \boldsymbol{X}; \boldsymbol{\beta})$ and $f(t|\boldsymbol{X}; \boldsymbol{\beta})$, given by

$$S_K(t \,|\, \boldsymbol{X};\, \boldsymbol{\beta}, \boldsymbol{\theta}) = S_{0,K}(te^{-\mathbf{X}^T\boldsymbol{\beta}}; \boldsymbol{\theta}), \quad f_K(t \,|\, \boldsymbol{X};\, \boldsymbol{\beta}, \boldsymbol{\theta}) = e^{-\mathbf{X}^T\boldsymbol{\beta}} f_{0,K}(te^{-\mathbf{X}^T\boldsymbol{\beta}}; \boldsymbol{\theta}). \tag{8}$$

The same principle may be applied to the PO model, which in its usual form assumes

$$\frac{S(t \,|\boldsymbol{X};\, \boldsymbol{\beta})}{1 - S(t \,|\boldsymbol{X};\, \boldsymbol{\beta})} = \left\{ \frac{S_0(t)}{1 - S_0(t)} \right\} \exp(-\boldsymbol{X}^T\boldsymbol{\beta}), \tag{9}$$

where $S_0(t)$ is the baseline survival function, assumed to have density $f_0(t)$, and $S(t \,|\, \boldsymbol{X};\, \boldsymbol{\beta})$ is the conditional survival function given $\boldsymbol{X}$ with density $f(t \,|\, \boldsymbol{X};\, \boldsymbol{\beta})$. Model (9) implies $S(t \,|\, \boldsymbol{X};\, \boldsymbol{\beta}) = S_0(t)/\{e^{\mathbf{X}^T\boldsymbol{\beta}} + S_0(t)(1 - e^{\mathbf{X}^T\boldsymbol{\beta}})\}$; thus, assuming $S_0(t)$ and $f_0(t)$ may be approximated by (3) or (4), $S(t \,|\, \boldsymbol{X};\, \boldsymbol{\beta})$ and $f(t \,|\, \boldsymbol{X};\, \boldsymbol{\beta})$ may be approximated by

$$S_K(t \,|\, \boldsymbol{X};\, \boldsymbol{\beta}, \boldsymbol{\theta}) = S_{0,K}(t; \boldsymbol{\theta}) a_{0,K}^{-1}(t, \boldsymbol{X}; \boldsymbol{\beta}, \boldsymbol{\theta}), \quad f_K(t \,|\, \boldsymbol{X};\, \boldsymbol{\beta}, \boldsymbol{\theta}) = f_{0,K}(t; \boldsymbol{\theta}) e^{\mathbf{X}^T\boldsymbol{\beta}} a_{0,K}^{-2}(t, \boldsymbol{X}; \boldsymbol{\beta}, \boldsymbol{\theta}) \tag{10}$$

for fixed K, where $a_{0,K}(t, \boldsymbol{X}; \boldsymbol{\beta}, \boldsymbol{\theta}) = e^{\mathbf{X}^T\boldsymbol{\beta}} + S_{0,K}(t; \boldsymbol{\theta})(1 - e^{\mathbf{X}^T\boldsymbol{\beta}})$.

We may now exploit these developments. Assuming as usual that the censoring mechanism is independent of T given $\boldsymbol{X}$, we demonstrate when T may be (i) interval-censored, known only to lie in an interval $[L, R]$; (ii) right-censored at L (set $R = \infty$); or (iii) observed (set $T = L = R$). For (i) and (ii), $\Delta = 0$; else, $\Delta = 1$ (iii). With iid data $(L_i, R_i, \Delta_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$, assuming that $f(t \,|\, \boldsymbol{X})$ and $S(t \,|\, \boldsymbol{X})$ may be represented as in (6), (8), or (10),

for fixed K, the loglikelihood for $(\boldsymbol{\beta}, \boldsymbol{\theta})$, conditional on the $\boldsymbol{X}_i$, is

$$\ell_K(\boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \Big[\Delta_i \log\{f_K(L_i \mid \boldsymbol{X}_i;\, \boldsymbol{\beta}, \boldsymbol{\theta})\} + (1-\Delta_i) \log\{S_K(L_i \mid \boldsymbol{X}_i;\, \boldsymbol{\beta}, \boldsymbol{\theta}) - S_K(R_i \mid \boldsymbol{X}_i;\, \boldsymbol{\beta}, \boldsymbol{\theta})\} \Big].$$

For fixed K, base density, and model, $\ell_K(\boldsymbol{\beta}, \boldsymbol{\theta})$ may be maximized in $(\boldsymbol{\beta}, \boldsymbol{\theta})$ using standard optimization routines; we use the SAS IML optimizer `nlpqn` (SAS Institute, 2006). Choice of starting values is critical for ensuring that the global maximum is reached. In Web Appendix C, we recommend an approach where $\ell_K(\boldsymbol{\beta}, \boldsymbol{\theta})$ is maximized for each of several starting values found by fixing $\boldsymbol{\phi}$ over a grid and using "automatic" rules to obtain corresponding starting values for $(\mu, \sigma, \boldsymbol{\beta})$. Although elements of $\boldsymbol{\phi}$ are restricted to certain ranges, unconstrained optimization virtually always yields a valid transformation so that $\int h_K(z; \boldsymbol{\phi}) dz = 1$. The declared estimates correspond to the solution(s) yielding the largest $\ell_K(\boldsymbol{\beta}, \boldsymbol{\theta})$.

Following other authors (e.g., Gallant and Tauchen, 1990; Zhang and Davidian, 2001), for a given model (PH, AFT, PO), we propose selecting adaptively the K-base density combination by inspection of an information criterion over all combinations of base density (normal, exponential) and $K = 0, 1, \ldots, K_{\max}$. Our extensive studies show $K_{\max} = 2$ is generally sufficient to achieve an excellent fit. With $q = \dim(\boldsymbol{\beta}, \boldsymbol{\theta})$, criteria of the form $-\ell_K(\boldsymbol{\beta}, \boldsymbol{\theta}) + qc$ have been advocated, with small values preferred. Ordinarily, the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Hannan-Quinn (HQ) criteria take $c = 1$, $\log(n)/2$, and $\log\{\log(n)\}$, respectively; AIC tends to select "larger" models and BIC "smaller" models, with HQ intermediate. As noted by Kooperberg and Clarkson (1997, Section 3), with censored data, dependence of c on n may be suspect; for right-censored data, replacing n by d = number of failures has been proposed (e.g., Volinsky and Raftery, 2000), although a similar adjustment under interval censoring is not obvious. It is nonetheless common practice to base c on n. We have found in the current context that replacing n by d has little effect on the K-base density choice. We use HQ with $c = \log\{\log(n)\}$ in the sequel.

The SNP approach is an alternative to traditional semiparametric methods such as PL

when one is willing to adopt the assumption of a "smooth" density $f_0(t)$. The formulation also supports selection among competing models (e.g., PH, AFT, PO): that for which the chosen K-base density combination yields the most favorable value of the information criterion may be viewed as "best supported" by the data. This may be used objectively or in conjunction with other evidence, e.g., the outcome of a formal test of the proportional hazards assumption (e.g., Gray, 2000; Lin, Zhang, and Davidian, 2006), in adopting a final model.

To obtain standard errors and confidence intervals for the estimator for $\boldsymbol{\beta}$, the parameter ordinarily of central interest, as well as for any other functional of the conditional distribution of $T|\boldsymbol{X}$ based on a final selected representation, we follow other authors and treat the chosen K, base density, and model as predetermined. That is, we approximate the sampling variance of the resulting estimator $\widehat{\boldsymbol{\beta}}$ or of any functional $d(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})$ via the inverse "information matrix" acting as if the chosen $\ell_K(\boldsymbol{\beta}, \boldsymbol{\theta})$ were the loglikelihood under a predetermined parametric model. This matrix is readily obtained from optimization software. For $\widehat{\boldsymbol{\beta}}$, the square root of the relevant diagonal element of this matrix yields immediately our proposed standard error; for general functionals, we use the delta method. Assuming that these quantities have approximately normal sampling distributions, $100(1-\alpha)\%$ Wald confidence intervals may be constructed as the estimate $\pm$ the normal critical value $\times$ the estimated standard error. Although the choice of K and base density is made adaptively, which would seem to invalidate this practice, results cited in Web Appendix A support it, and simulations in Section 4 demonstrate that this approach yields reliable inferences in realistic sample sizes.

Several useful byproducts follow from the SNP approach. Selection of a model with $K = 0$ suggests evidence favoring the parametric model implied by the chosen base density; e.g., the AFT model with $K = 0$ and normal base density corresponds to assuming T given $\boldsymbol{X}$ is lognormally distributed. Because "smooth" estimates of baseline densities and survival functions are immediate, predictors of survival probabilities and calculation of associated

confidence intervals as in Cheng et al. (1997) are easily handled.

3.2 *Model Extensions*

A "parametric" representation makes otherwise difficult-to-implement extensions of standard time-to-event regression models straightforward. In Web Appendices D and E, we exhibit two possibilities: extension of the AFT model to incorporate so-called "heteroscedastic errors" and extension of this model to accommodate time-dependent covariates.

4. Simulation Studies

We report on simulations to evaluate performance of the SNP approach. For all SNP fits, we considered $K_{\max} = 2$ and both the normal and exponential base densities.

In the first set of scenarios, data were generated under the PH model (5) with continuous covariate X uniformly distributed on $(0, 1)$ and 25% independent uniform right censoring, with true baseline hazard $\lambda_0(t)$ corresponding to a lognormal with mean 2.9 and scale 0.66; a Weibull with shape 0.9 and scale 25.0; a gamma with shape and scale 2.0; and a log-mixture of normals found by exponentiating draws from the bimodal normal mixture $0.3\mathcal{N}(0.2, , 0.36) + 0.7\mathcal{N}(1.8, 0.36)$. In all cases, the true value of $\beta = 2.0$, $n = 200$, and 1000 Monte Carlo (MC) data sets were generated. For each, the PH model was fitted by PL via SAS `proc phreg` and by the SNP approach, with comparable results, as shown in Table 1(a). The SNP-based AFT and PO models (7) and (9) were also fitted to each data set, and Table 1(a) summarizes how often HQ selected each model. Percentages do not necessarily add to 100% across the three models; because fits with $K = 0$ and exponential base density lead to the same value of HQ for the AFT and PH models, HQ supports more than one model when this configuration is selected, so the percentages reflect the proportions of times this occurred. Under the Weibull, selection of the AFT or PH model corresponds to choosing a PH model (Kalbfleisch and Prentice, 2002, p. 44). "Correct" indicates the percentage of data sets for which HQ supported selection of the (true) PH model under these conditions

and indicates that inspection of HQ across SNP fits of competing models can be useful for deducing the appropriate model, a capability that grows with sample size. The true PH model was identified over 90% of the time for all distributions when $n = 500$.

Informally, the information on $\lambda_0(t)$ and β is roughly "orthogonal," so it is not unexpected that imposing smoothness assumptions on $\lambda_0(t)$ and fitting the SNP-based PH model does not yield increased precision for estimating β relative to PL. For the PH model, the real advantage of the SNP approach is the ease with which it handles interval- and other arbitrarily-censored data. Under the gamma and log-mixture-normal scenarios, we generated interval-censored data for each subject by drawing five random examination times, where the times between each were independently lognormally distributed, and then generated independently an event time from the PH model. This led to the percentages of right- and interval-censored data in Table 1(a). Results of fitting the PH model by SNP for 1000 MC data sets with $n = 200$ show that the approach leads to reliable inferences.

The second set of scenarios involved a true PO model (9) with X and either independent 25% uniform right censoring or interval censoring as above, $\beta = 2.0$ or -2.0, $n = 200$, and 1000 data sets generated with $f_0(t)$ lognormal with mean and scale 13.8 and 0.53, log-mixture normal from $0.3\mathcal{N}(1.2, 0.36) + 0.7\mathcal{N}(-1.8, 0.36)$, and Weibull with shape and scale 1.0 and 5.0. From Table 1(b), when the true PO model is fitted via SNP, reliable inferences on β obtain, and HQ is able to identify the true PO model well except for the Weibull; for this case, performance improves with increasing sample size.

In the third set of scenarios, data were generated from the AFT model (7) with X as above; $\beta = 2.0$; and $f_0(t)$ lognormal with mean 0.5 and scale 1.31, Weibull with shape 2.0 and scale 16.0, gamma with shape and scale 2.0, and the log-mixture of normals $0.3\mathcal{N}(1.2, , 0.36)+$ $0.7\mathcal{N}(-1.8, 0.36)$ (bimodal). For each of 1000 data sets with independent uniform right censoring, the AFT model was fitted via SNP; the Buckley-James method; and the rank-

based method of Jin et al. (2003), with both logrank and Gehan-type weight functions, using the `R` function `aft.fun` (Zhou, 2006). Table 2 shows that the SNP method yields reliable inferences and compares favorably to competing semiparametric methods, achieving marked relative gains in efficiency in some cases. With $n = 200$ and 50% censoring, the SNP procedure continues to perform well. Undercoverage of SNP Wald confidence intervals in the log-mixture-normal case is resolved for $n = 500$. The PH and PO models were also fitted. In all but the gamma scenario, HQ strongly supports the AFT model; increasing to $n = 500$ in the gamma case vastly improves identification of the correct model. The similarity of the gamma distribution to a Weibull may be responsible for the difficulty the criterion has distinguishing the AFT and PH models for the smaller sample size.

Byproducts of the SNP approach for any model are estimates of the corresponding density $f_0(t)$ and survival function $S_0(t)$. Figure 1 shows the 1000 estimates of $S_0(t)$ under two of the AFT scenarios, demonstrating that its true form can be recovered with impressive accuracy.

We also carried out simulations for the true AFT model for gamma and log-mixture-normal scenarios under interval censoring, each with 1000 data sets generated as for the PH model to yield the censoring patterns in Table 2. The AFT model was fitted to each using the SNP approach; as for PH, the results demonstrate the reliable performance of the method, with undercoverage of confidence intervals for $n = 200$ under the log-mixture-normal.

For all three model scenarios, Table 3 presents for selected configurations the number of times each K-base density combination was chosen by HQ when fitting the true model. Not surprisingly, the normal base density is chosen most often when $f_0(t)$ is lognormal and log-mixture normal, and the exponential base density is preferred for the Weibull and gamma.

Undercoverage of Wald intervals in some instances with $n = 200$ in Table 2 suggests that the delta method approximation may be less reliable for the AFT model than for PH and PO. We thus investigated replacing delta method standard errors by those from a nonparametric

bootstrap, where the K-base density for the AFT model is chosen by HQ for each bootstrap data set. E.g., for the interval censored log-mixture scenario, for the first 300 of the 1000 MC data sets, the MC mean, standard deviation, average of delta method standard errors, and associated coverage are 2.05, 0.26, 0.23, and 90.3, respectively. The MC average of bootstrap standard errors using 50 bootstrap samples and coverage of the associated interval are 0.28 and 94.3, suggesting that this approach can correct underestimation of sampling variation.

The foregoing results involved simple models with a single covariate in order to allow reporting for a number of scenarios with straightforward interpretation. Given failure to achieve nominal coverage for some settings for the AFT model, further evaluation of the SNP-based approach in this case is warranted. Moreover, demonstration of computational stability and feasibility of the proposed methods for all three models under more complex conditions is required. Accordingly, we carried out additional simulations. We report on representative scenarios, each involving 1000 MC data sets with $n = 200$ or 500 and 25% independent uniform right censoring and generated from the AFT model (7), where $\boldsymbol{X} = (X_1, X_2, X_3)^T$ with X_1 distributed as uniform on (0,2), X_2 Bernoulli with $P(X_2 = 1) = 0.5$, and $X_3 \sim \mathcal{N}(0.5, 1)$ and the true value of $\boldsymbol{\beta} = (2.5, 0.5, -0.8)^T$. Table 4 shows results for fitting the AFT model when the true $f_0(t)$ was lognormal with mean 54.6 and scale 7.3, gamma with with shape 2.0 and scale 6.0, and log-mixture of normals $0.3\mathcal{N}(1.2, , 0.36) + 0.7\mathcal{N}(-1.8, 0.36)$ (bimodal). In all scenarios, no computational issues were encountered for any data sets, and performance is similar to that for the simpler models above, with analogous undercoverage of Wald intervals for components of $\boldsymbol{\beta}$ in some cases. HQ chose K-base density combinations in proportions similar to those in Table 3 in all cases. For the gamma and log-mixture scenarios with $n = 200$, we used a nonparametric bootstrap with 50 bootstrap replicates as described above to obtain alternative standard errors for the first 300 MC data sets; results are indicated by an asterisk in Table 4 and suggest that, as above, use of bootstrap standard

errors to form Wald intervals yields reasonable performance.

We also carried out analogous simulations under the PH and PO models, representative results of which are in Web Appendix F. Again, computation was stable and straightforward in every situation we tried, and, as in the single covariate case reported above, coverage of delta method intervals achieved the nominal level for both models.

Overall, the simulations demonstrate that the SNP approach is computationally straightforward and yields reliable performance under the "smoothness" assumption and provides a tool for practical model selection. Simulations showing performance of the SNP approach for the model extensions described in Section 3.2 are given in Web Appendices D and E.

5. Applications

5.1 *Cancer and Leukemia Group B Protocol 8541*

Lin et al. (2006) discuss Cancer and Leukemia Group B (CALGB) protocol 8541, a randomized clinical trial comparing survival for high, moderate, and low dose regimens of cyclophosphamide, adriamycin, and 5-flourouacil (CAF) in women with early stage, node-positive breast cancer. Following the primary analysis, interest focused on the prognostic value of baseline characteristics. We consider estrogen receptor (ER) status; ER-positive tumors are more likely to respond to anti-estrogen therapies than those that are ER-negative. ER status is available for 1437 of the 1479 subjects, of whom 64% were ER-positive, with 64% right-censored survival times. Figure 1 of Lin et al. (2006) suggests that the relationship of survival to ER status does not exhibit proportional hazards, a finding corroborated by their spline-based test for departures from proportional hazards (p-value=< 0.001).

We fit the AFT, PH, and PO models with binary covariate X_i (=1 if ER-positive) using SNP; here and in Section 5.2 we considered the normal and exponential base densities and $K_{\max} = 2$. The HQ criterion was 10177, 10197, and 10192 for the preferred K-base density combinations for AFT, PH, and PO, respectively. Both the PL and SNP fits of the PH

model yielded an estimated hazard ratio of 0.77. The AFT model is best supported by the data, consistent with the evidence discrediting the PH model. The preferred AFT fit takes $K = 1$ with normal base density, with an estimate of β of 0.45 (SE 0.08). Here, the effect of a covariate is multiplicative on time itself rather than on the hazard, leading to the interpretation that failure times for ER-positive women are "decelerated" relative to those for ER-negative women: the probability that an ER-positive woman survives to time t is the same as the probability that an ER-negative woman survives to time $0.64t$, so that, roughly, being ER-negative curtails survival times by 64% relative to being ER-positive. A possible explanation is that ER-positive women may have received anti-estrogen therapy during follow-up, enhancing their survival.

Further analyses of the CALBG 8541 data are in Web Appendix F.

5.2 *Breast Cosmesis Study*

The famous breast cosmesis data (Finkelstein and Wolfe, 1985) involve time to cosmetic deterioration of the breast in early breast cancer patients who received radiation alone ($X = 0$, 46 patients) or radiation+adjuvant chemotherapy ($X = 1$, 48 patients). Deterioration times were right-censored for 38 women. Times for the 56 women experiencing deterioration were interval-censored due to its evaluation only at intermittent clinic visits. Numerous authors have used these data to demonstrate methods for interval censored data.

We fitted the AFT, PH, and PO models using the SNP approach, obtaining HQ values of 309, 309, 317, respectively, for the chosen K-base density combination, supporting AFT and PH. The preferred fit for each uses $K = 0$ and the exponential base density; this configuration is equivalent to a Weibull regression model, for which the PH and AFT models are the same. This is consistent with the adoption of the PH (e.g., Goetghebeur and Ryan, 2000; Betensky et al., 2002) or AFT (e.g., Tian and Cai, 2006) models by many authors. The SNP estimate of $\beta = 0.95$ (SE 0.280) in (5) is consistent with the results from several methods for fitting

the PH model with interval censored data reported by Goetghebeur and Ryan (2000) and Betensky et al. (2002). The corresponding Wald statistic for testing $\beta = 0$ is 3.35, in line with the score statistic of Finkelstein (1986) of 2.86 and Wald statistics implied in Table 2 of Goetghebeur and Ryan (2000). Figure 2 shows the SNP estimates of $S(t \mid X = 0)$ and $S(t \mid X = 1)$ based on the PH fit; compare to Figure 1 of Goetghebeur and Ryan (2000).

6. Discussion

We have proposed a general framework for regression analysis of arbitrarily censored time-to-event data under the mild assumption of a "smooth" density for model components ordinarily left unspecified under a semiparametric perspective. The methods are straightforward to implement using standard optimization software, and computation is stable across a range of conditions. A SAS macro is available from the first author. Although we focused on the PH, AFT, and PO models, the approach allows any competing models, such as generalizations of (7), models with nonlinear covariate effects, and linear transformation models to be placed in a common framework, providing a basis for model selection. Standard errors and Wald confidence intervals may be obtained using standard parametric asymptotic theory in most cases; however, this approximation is less reliable for the AFT model, so we recommend using a nonparametric bootstrap with small samples/numbers of failures in this case. A rigorous proof of consistency and asymptotic normality of the estimators for $\boldsymbol{\beta}$ and functionals of $f_0(t)$ in the general censored-data regression formulation here is an open problem.

It should be possible to adapt the approach to problems involving both censoring and truncation (Joly et al., 1998; Pan and Chappell, 2002). Because with the SNP representation $f(t|\boldsymbol{X};\boldsymbol{\beta})$ and $S(t|\boldsymbol{X};\boldsymbol{\beta})$ are in "parametric" form, the likelihood function is straightforward under the usual assumption that censoring and truncation are independent of event time.

A further advantage, not illustrated here, is that an efficient rejection sampling algorithm for simulation from a fitted SNP density is available (Gallant and Tauchen, 1990). This may

be used to simulate draws from the fit of $f_0(t)$ under the preferred model and hence draws of T_i from $f(t \mid \boldsymbol{X})$ for any $\boldsymbol{X}$, allowing any functional of this distribution to be approximated.

Supplementary Materials

Web Appendices A–F, referenced in Sections 2, 3.1, 3.2, 4, and 5.1, are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

This work was supported by NIH grants R01-CA085848 and R37-AI031789. The authors gratefully acknowledge the comments of the co-editor, associate editor, and two reviewers, which led to improvements to the paper and Supplementary Materials.

References

Betensky, R. A., Lindsey, J. C., Ryan, L. M., and Wand, M. P. (2002). A local likelihood proportional hazards model for interval censored data. *Statistics in Medicine* **21,** 263–275.

Betensky, R.A., Rabinowitz, D., and Tsiatis, A.A. (2001). Computationally simple accelerated failure time regression for interval censored data. *Biometrika* **88,** 703–711.

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

Cai, T. and Betensky, R. A. (2003). Hazard regression for interval-censored data with penalized spline. *Biometrics* **59,** 570–579.

Chen, K., Jin, Z., and Ying, Z. (2002). Semiparametric analysis of transformation models with censored data. *Biometrika* **89,** 659–668.

Chen, Y. Q. and Jewell, N. P. (2001). On a general class of semiparametric hazards regression model. *Biometrika* **88,** 677–702.

Chen, Y. Q. and Wang, M. C. (2000). Analysis of accelerated hazards model. *Journal of American Statistical Association* **95,** 608–618.

Cheng, S. C., Wei, L. J., and Ying, Z. (1995). Analysis of transformed models with censored data. *Biometrika* **82,** 835–842.

Cheng, S. C., Wei, L. J., and Ying, Z. (1997). Predicting survival probabilities with semiparametric transformation models. *Journal of American Statistical Association* **92,** 227–235.

Cox, D. R. (1972). Regression models and life tables (with Discussion). *Journal of the Royal Statistical Society, Series B* **34,** 187–200.

Fenton, V. M. and Gallant, A. R. (1996). Qualitative and asymptotic performance of SNP density estimators. *Journal of Econometrics* **7**4, 77–118.

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

Finkelstein, D. and Wolfe, R. M. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics* **41,** 933–945.

Gallant, A. R. and Nychka, D. W. (1987). Seminonparametric maximum likelihood estimation. *Econometrica* **55,** 363–390.

Gallant, A. R. and Tauchen, G. E. (1990). A nonparametric approach to nonlinear time series analysis: Estimation and simulation. In *New Directions in Time Series Analysis, Part II,* D. Brillinger, P. Caines, J. Geweke, E. Parzen, M. Rosenblatt, and M.S. Taqqu (eds.) New York: Springer, pp. 71–92.

Goetghebeur, E. and Ryan, L. (2000). Semiparametric regression analysis of interval-censored data. *Biometrics* **56,** 1139–1144.

Gray, R. J. (2000). Estimation of regression parameters and the hazard function in transformed linear survival models. *Biometrics* **56,** 571–576.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data, Second Edition.* New York: John Wiley and Sons.

Komárek, A., Lesaffre, E., and Hilton, J.F. (2005). Accelerated failure time model for ar-

bitrarily censored data with smoothed error distribution. *Journal of Computational and Graphical Statistics* **14,** 726–745.

Jin, Z., Lin, D. Y., Wei, L. J., and Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika* **90,** 341–353.

Jin, Z., Lin, D. Y., and Ying, Z. (2006). On least-squares regression with censored data. *Biometrika* **93,** 147–162.

Joly, P., Commenges, D., and Letenneur, L. (1998). A penalized likelihood approach for arbitrarily censored and truncated data: Application to age-specific incidence of dementia. *Biometrics* **54,** 185–194.

Kooperberg, C. and Clarkson, D. B. (1997). Hazard regression with interval-censored data. *Biometrics* **53,** 1485–1494.

Lin, D.Y. and Geyer, C. J. (1992). Computational methods for semiparametric linear regression with censored data. *Journal of Computational and Graphical Statistics* **1,** 77–90.

Lin, J., Zhang, D., and Davidian, M. (2006). Smoothing spline-based score tests for proportional hazards models. *Biometrics* **62,** 803–812.

Lin, J. S. and Wei, L. J. (1992). Linear regression analysis based on Buckley-James estimating equation. *Biometrics* **48,** 679–681.

Murphy, S. A., Rossini, A. J., and van der Vaart, A. W. (1997). Maximum likelihood estimation in proportional odds model. *Journal of the American Statistical Association* **92,** 968–976.

Pan, W. (2000). A multiple imputation approach to Cox regression with interval-censored data. *Biometrics* **56,** 199–203.

Pan, W. and Chappell, R. (2002). Estimation in the Cox proportional hazards model with left-truncated and interval-censored data. *Biometrics* **58,** 64–70.

Ritov, Y. (1990). Estimation in a linear regression model with censored data. *Annals of*

Statistics **18,** 303–328.

SAS Institute, Inc. (2006). *SAS Online Doc 9.1.3*. Cary, NC: SAS Institute, Inc.

Satten, G. A., Datta, S., and Williamson, J. M. (1998). Inference based on imputed failure times for the proportional hazards model with interval-censored data. *Journal of the American Statistical Association* **93,** 318–327.

Scharfstein, D. O., Tsiatis, A. A., Gilbert, P. B. (1998). Semiparametric efficient estimation in the generalized odds-rate class of regression models for right-censored time-to-event data. *Lifetime Data Analysis* **4,** 355–391.

Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*. New York: Springer.

Tian, L. and Cai, T. (2006). On the accelerated failure time model for current status and interval censored data. *Biometrika* **93,** 329–342.

Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *Annals of Statistics* **18,** 354–372.

Volinsky, C.T. and Raftery, A.E. (2000). Bayesian information criterion for censored survival models. *Biometrics* **56,** 256–262.

Wei, L. J., Ying, Z., and Lin, D. Y. (1990). Linear regression analysis of censored survival data based on rank tests. *Biometrika* **77,** 845–851.

Yang, S. and Prentice, R. L. (1999). Semiparametric inference in the proportional odds regression model. *Journal of the American Statistical Association* **94,** 125–136.

Zhang, D. and Davidian, M. (2001) Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics* **57,** 795–802.

Zhou, M. (2006). The `rankreg` package.

`http://cran.r-project.org/src/contrib/Descriptions/rankreg.html`

Table 1

Simulation results based on 1000 Monte Carlo data sets when the true model is the PH or PO model with baseline density $f_0(t)$. Mean is Monte Carlo mean of the 1000 estimates of β when the true model is fitted, SD is their Monte Carlo standard deviation, SE is the average of the 1000 estimated delta method standard errors, and CP is Monte Carlo coverage probability, expressed as a percent, of 95% Wald confidence intervals. For right-censored data, SNP and PL indicate fitting using the SNP approach with K and the base density chosen via HQ and partial likelihood, respectively. All of the AFT, PH, and PO models were fitted to each data set under right-censoring; the columns AFT, PH, and PO indicate the percentage of 1000 data sets for which that model was chosen based on HQ, and Correct indicates the percentage of data sets supporting the PH model; see the text. For interval-censored data, only SNP was used. (a) PH model: true value of $\beta = 2.0$ in all cases. (b) PO model: true value of $\beta = 2.0$ (lognormal, Weibull) or $\beta = -2.0$ (log-mixture).

$f_0(t)$	n	Cens. rate	Method	Mean	SD	SE	CP	AFT	PH	PO	Correct
(a) True PH model											
Right-censored data											
lognormal	200	25%	SNP	2.02	0.32	0.31	95.4	9.4	86.5	5.2	86.5
			PL	2.00	0.32	0.32	96.3				
Weibull	200	25%	SNP	2.02	0.31	0.31	95.5	86.7	88.0	2.8	97.2
			PL	2.01	0.32	0.32	96.3				
gamma	200	25%	SNP	2.06	0.32	0.31	94.3	66.8	81.6	6.3	81.6
			PL	2.02	0.32	0.32	95.2				
log-mixture	200	25%	SNP	2.04	0.34	0.33	94.9	2.4	73.6	24.0	73.6
			PL	2.02	0.33	0.33	95.5				
Interval-censored data											
gamma	200	18% right, 82% interval	SNP	2.04	0.30	0.29	93.9				
log-mixture	200	20% right, 80% interval	SNP	2.04	0.30	0.30	94.8				
(b) True PO model											
Right-censored data											
lognormal	200	25%	SNP	2.01	0.18	0.18	94.6	0.7	2.2	97.1	97.1
Weibull	200	25%	SNP	2.00	0.46	0.45	94.6	29.4	19.7	65.1	65.1
log-mixture	200	25%	SNP	−1.99	0.46	0.45	95.3	1.4	15.4	83.2	83.2
Interval-censored data											
log-mixture	200	20% right, 80% interval	SNP	−2.01	0.48	0.47	95.7				

Table 2

Simulation results based on 1000 Monte Carlo data sets when the true model is the AFT model with baseline density $f_0(t)$. For right-censored data, SNP, BJ, Gehan, and LR indicate fitting using the SNP approach with K and the base density chosen via HQ, the Buckley-James method, and the rank-based method of Jin et al. (2003) using Gehan-type and log-rank weight functions, respectively. All other entries are as in Table 1. For interval-censored data, only SNP was used. True value of $\beta = 2.0$ in all cases.

$f_0(t)$	n	Cens. rate	Method	Mean	SD	SE	CP	AFT	PH	PO	Correct
				Right-censored data							
lognormal	200	25%	SNP	2.02	0.27	0.26	94.4	80.2	8.0	12.3	80.2
			BJ	2.02	0.27	0.26	94.2				
			Gehan	2.02	0.27	0.28	95.3				
			logrank	2.01	0.29	0.29	94.7				
	200	50%	SNP	2.02	0.30	0.30	93.3				
Weibull	200	25%	SNP	2.00	0.14	0.15	95.9	71.0	88.7	1.1	98.9
			BJ	2.00	0.18	0.19	96.3				
			Gehan	2.00	0.17	0.17	95.7				
			logrank	2.00	0.14	0.15	96.1				
	200	50%	SNP	2.01	0.20	0.20	94.8				
gamma	200	25%	SNP	2.00	0.20	0.19	94.0	65.0	66.4	11.4	65.0
			BJ	2.00	0.22	0.23	96.0				
			Gehan	2.00	0.21	0.22	95.4				
			logrank	2.00	0.20	0.21	95.4				
	200	50%	SNP	2.00	0.26	0.24	92.9				
	500	25%	SNP	2.00	0.06	0.06	94.0	98.8	1.2	0.0	98.8
log-mixture	200	25%	SNP	1.99	0.19	0.18	91.9	100.0	0.0	0.0	100.0
			BJ	1.99	0.42	0.28	80.5				
			Gehan	1.99	0.29	0.29	95.6				
			logrank	2.00	0.41	0.43	96.5				
	200	50%	SNP	1.98	0.23	0.22	91.5				
	500	25%	SNP	2.00	0.05	0.05	94.8				
	500	50%	SNP	2.00	0.06	0.06	93.5				
				Interval-censored data							
gamma	200	20% right, 80% interval	SNP	2.01	0.22	0.21	92.2				
gamma	500	16% right, 84% interval	SNP	2.00	0.06	0.06	94.7				
log-mixture	200	17% right, 83% interval	SNP	2.05	0.27	0.23	90.3				
log-mixture	500	17% right, 83% interval	SNP	2.00	0.07	0.06	94.0				

Table 3

Numbers of times each K-base density combination was chosen by the HQ criterion when fitting the true model (PH, AFT, PO) for selected configurations in Tables 1 and 2.

Base Density			Standard Normal			Standard Exponential		
			K			K		
$f_0(t)$	n	Cens. rate	0	1	2	0	1	2
			True PH Model					
lognormal	200	25%	873	43	19	11	33	21
Weibull	200	25%	9	0	35	854	74	28
gamma	200	25%	140	12	33	644	143	28
log-mixture	200	25%	239	26	721	0	0	14
gamma	200	18% right, 82% interval	302	28	8	645	12	5
log-mixture	200	20% right, 80% interval	505	62	241	9	6	177
			True AFT Model					
lognormal	200	25%	873	49	18	10	30	20
Weibull	200	25%	3	0	24	890	54	29
gamma	200	25%	65	16	49	624	212	34
log-mixture	200	25%	0	153	847	0	0	0
gamma	200	20% right, 80% interval	223	32	17	640	68	20
log-mixture	200	18% right, 82% interval	0	567	432	0	0	1
			True PO Model					
lognormal	200	25%	878	81	19	0	6	16
Weibull	200	25%	31	3	40	830	82	14
log-mixture	200	25%	0	225	774	0	0	1
log-mixture	200	18% right, 82% interval	0	620	350	0	6	24

Table 4

Simulation results for the SNP approach based on 1000 Monte Carlo data sets when the true model is the AFT model with baseline density $f_0(t)$ and multiple covariates under right censoring. Entries are as in Table 1. The True $\boldsymbol{\beta}$ column gives the true values of the elements of $\boldsymbol{\beta}$. Entries with an asterisk (*) at the sample size indicate results for the first 300 Monte Carlo data sets, for which both delta method and nonparametric bootstrap standard errors were used, where SE_{boot} and CP_{boot} denote the average of bootstrap standard error and Monte Carlo coverage probability expressed as a percent, of 95% Wald confidence intervals using the bootstrap standard errors, respectively. For each scenario, N_K and E_K, $K = 0, 1, 2$, indicate the number of times the configuration of normal (N) or exponential (E) base density with the indicated K was chosen by HQ.

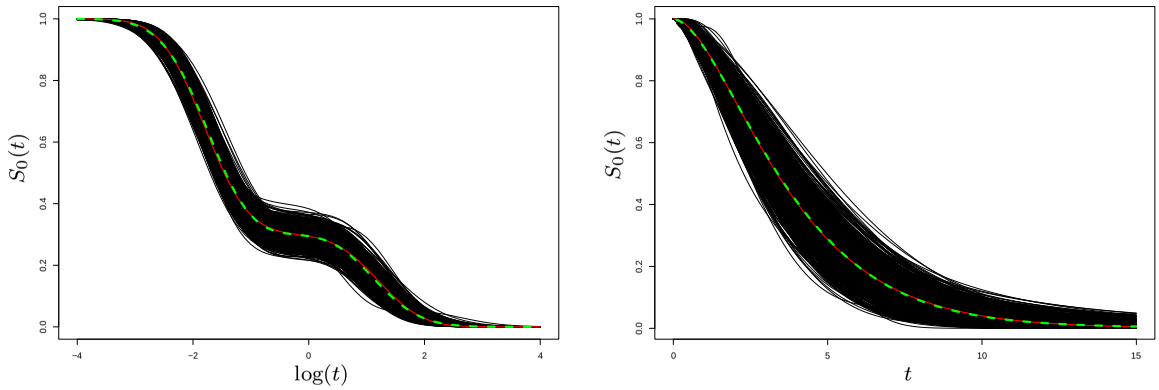
$f_0(t)$	n	Cens. rate	True $\boldsymbol{\beta}$	Mean	SD	SE	CP	SE_{boot}	CP_{boot}
lognormal	200	25%	2.5	2.51	0.27	0.26	94.8		
			0.5	0.49	0.15	0.15	93.5		
			-0.8	-0.81	0.30	0.29	94.6		
	($N_0 = 853$, $N_1 = 64$, $N_2 = 19$, $E_0 = 8$, $E_1 = 38$, $E_2 = 18$)								
gamma	200	25%	2.5	2.50	0.16	0.11	93.3		
			0.5	0.50	0.06	0.06	92.3		
			-0.8	-0.80	0.12	0.11	92.1		
	($N_0 = 50$, $N_1 = 16$, $N_2 = 60$, $E_0 = 594$, $E_1 = 237$, $E_2 = 43$)								
	200*	25%	2.5	2.50	0.11	0.11	93.3	0.13	96.0
			0.5	0.51	0.07	0.06	92.7	0.07	95.0
			-0.8	-0.79	0.12	0.11	91.3	0.13	96.3
	500	25%	2.5	2.50	0.07	0.07	93.7		
			0.5	0.50	0.04	0.04	93.1		
			-0.8	-0.80	0.07	0.07	93.8		
	($N_0 = 0$, $N_1 = 3$, $N_2 = 68$, $E_0 = 418$, $E_1 = 464$, $E_2 = 47$)								
log-mixture	200	25%	2.5	2.49	0.10	0.09	94.1		
			0.5	0.50	0.06	0.05	92.8		
			-0.8	-0.80	0.11	0.10	91.8		
	($N_0 = 0$, $N_1 = 197$, $N_2 = 803$, $E_0 = 0$, $E_1 = 0$, $E_2 = 0$)								
	200*	25%	2.5	2.49	0.09	0.09	95.3	0.10	96.7
			0.5	0.50	0.06	0.05	93.0	0.06	94.0
			-0.8	-0.80	0.11	0.10	91.7	0.11	93.0
	500	25%	2.5	2.49	0.06	0.06	94.7		
			0.5	0.50	0.03	0.03	94.7		
			-0.8	-0.80	0.07	0.06	94.5		
	($N_0 = 0$, $N_1 = 16$, $N_2 = 984$, $E_0 = 0$, $E_1 = 0$, $E_2 = 0$)								

FIGURE CAPTIONS

(figures follow, one per page, in order)

Figure 1. SNP estimates of $S_0(t)$ for the AFT model based on 1000 Monte Carlo data sets, with the true $S_0(t)$ (white solid line) and average of 1000 estimates (dashed line) superimposed. (a) log-normal mixture scenario with $n = 500$, 50% right censoring. (b) gamma scenario with $n = 200$, 25% right censoring.

Figure 2. Estimated survival functions for time to cosmetic deterioration for the radiation only group (solid line) and radiation+chemotherapy group (dashed line) based on the SNP fit of the PH (AFT) model to the breast cosmesis study data.



(a) (b)

Figure 1.

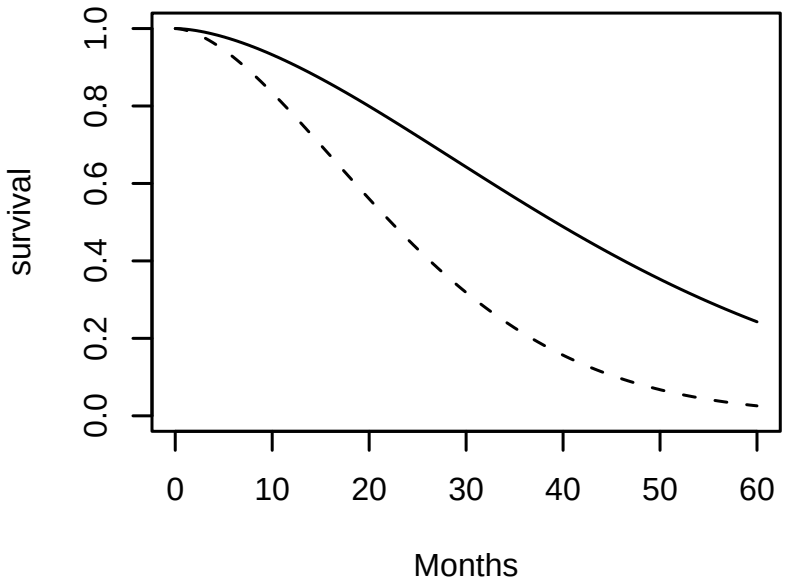


Figure 2.