Statistica Sinica **20** (2010), 1709-1723

REGRESSION ANALYSIS OF CASE II INTERVAL-CENSORED FAILURE TIME DATA WITH THE ADDITIVE HAZARDS MODEL

Lianming Wang, Jianguo Sun and Xingwei Tong

Unversity of South Carolina,
University of Missouri and Beijing Normal University

Abstract: Interval-censored failure time data often arise in clinical trials and medical follow-up studies, and a few methods have been proposed for their regression analysis using various regression models (Finkelstein (1986); Huang (1996); Lin, Oakes, and Ying (1998); Sun (2006)). This paper proposes an estimating equation-based approach for regression analysis of interval-censored failure time data with the additive hazards model. The proposed approach is robust and applies to both noninformative and informative censoring cases. A major advantage of the proposed method is that it does not involve estimation of any baseline hazard function. The implementation of the propsoed approach is easy and fast. Asymptotic properties of the proposed estimates are established and some simulation results and an application are provided.

Key words and phrases: Additive hazards model, counting processes; estimating equation, informative censoring, interval-censored data, semiparametric regression analysis.

1. Introduction

Interval-censored failure time data usually refer to the data in which the failure time of interest is observed only to belong to an interval instead of being known exactly (Kalbfleisch and Prentice (2002); Sun (2006)). Such data arise naturally in medical follow-up studies where the event of interest (failure) is not observed directly but only detected by some laboratory tests; then the failure time is known only to lie between the two monitoring times that are the last monitoring time at which the event has not occurred and the first monitoring time at which the event has already occurred. A well-known example of interval-censored data is discussed in Finkelstein (1986), the event of interest being the occurrence of breast retraction among early breast cancer patients. Another example is the HIV data studied by Zeng, Cai, and Shen (2006), where the failure is the first active Cytomegalovirus (CMV) infection.

A special case of interval-censored failure time data occurs if each study subject is observed only once as in cross-sectional studies. In this case, the

failure time of interest is known only to be either smaller or larger than the observation time, giving either left- or right-censored observation, respectively. This type of data is commonly referred to as case I interval-censored data or current status data (Huang (1996); Lin, Oakes, and Ying (1998); Martinussen and Scheike (2002)). In this paper, we study general or case II interval-censored data that are a mixture of left-, interval-, and right-censored observations.

For regression analysis of interval-censored data, a few methods have been proposed. For example, Finkelstein (1986) considered fitting the proportional hazards model to general interval-censored data and Hunag (1996) studied the efficient estimation problem for current status data using the same model. Lin, Oakes, and Ying (1998) and Martinussen and Scheike (2002) considered regression analysis of current status data, and Zeng, Cai, and Shen (2006) discussed regression analysis of case II interval-censored data, all using the additive hazards model. In particular, Zeng, Cai, and Shen (2006) studied the efficient estimation for the regression parameters and proposed to apply the full likelihood approach. However, its implementation can be quite complicated because of the need for estimation of the baseline cumulative hazard function, which is time consuming especially when the monitoring variables are continuous. Betensky, Rabinowitz, and Tsiatis (2001) developed a relatively easy estimation method for general interval-censored data using the accelerated failure time model.

In this paper, we develop an approach that is easy to implement for case II interval-censored data. Three situations are considered with respect to the observed intervals or the minitoring process: in the first, we assume that there exist only two monitoring times independent of the failure time of interest given the covariate process; in the second, it is assumed that there exist a sequence of monitoring times that are independent of the failure time of interest given the covariate process; in the third, there are also two monitoring times but they may be dependent of the failure time of interest given the covariate process. In all these situations, we allow that the monitoring times are random and continuous, and assume that the failure time of interest follows the additive hazards model, that the monitoring times follow Cox-type models (Cox (1972)). A major advantage of the proposed approach is that it does not require estimation of any nuisance baseline hazard functions.

The remainder of the paper is organized as follows. We present the proposed approach for the first situation in Sections 2 and 3. In particular, Section 2 introduces some notation and the assumed models, and some estimating equations for regression parameters are presented in Section 3. The asymptotic properties of the proposed estimates are given in Section 3. In Section 4, we generalize the proposed method to the second and third situations described above. Section 5 presents some results of a simulation study, and Section 6 illustrates the proposed

methodology in the breast cancer example. Section 7 contains some concluding remarks.

2. Notation and Models

Let T denote the failure time of interest. Here we focus on the situation in which there are only two monitoring variables U and V characterizing the monitoring process, and they are observable. Let Z denote a possibly time-dependent p-dimensional covariate vector that is assumed to be completely observed. Unless mentioned otherwise, we assume that failure time T is independent of monitoring times U and V given covariate Z.

Denote by $(T_i, U_i, V_i, Z_i(\cdot))$ the n i.i.d. replicates of $(T, U, V, Z(\cdot))$ and define indicators $\delta_{1\,i} = I(T_i < U_i)$, $\delta_{2\,i} = I(U_i \le T_i < V_i)$, and $\delta_{3\,i} = 1 - \delta_{1i} - \delta_{2i}$. These indicators determine whether the failure for subject i has occurred before U_i, during the examination interval $[U_i, V_i)$, or after V_i, respectively. The observed data are $(U_i, V_i, \delta_{i1}, \delta_{i2}, \delta_{i3}, Z_i(\cdot))$.

Throughout, we model the failure time with the additive hazards model. Specifically, we assume that T_i has the hazard function

$$\lambda_i(t \,|\, Z_i) = \lambda_0(t) + \beta_0' \, Z_i(t) \tag{2.1}$$

given the covariate process up to t, where λ_0 is an unknown baseline hazard function and β_0 denotes the p-dimensional vector of regression parameters. Our primary interest is in the estimation of β_0.

Due to the strict order restriction between the monitoring variables U and V, it is natural to regard them as recurrent events and model them with the Cox-type hazard functions (Cox (1972))

$$\lambda_i^U(t \,|\, Z_i) = \lambda_1(t)\, e^{\gamma_0' \, Z_i(t)}, \tag{2.2}$$

$$\lambda_i^V(t \,|U_i,\, Z_i) = I(t > U_i)\lambda_2(t)\, e^{\gamma_0' \, Z_i(t)}\,. \tag{2.3}$$

This $\lambda_1(t)$ and $\lambda_2(t)$ denote unspecified baseline hazard functions and γ_0 is a p-dimensional vector of unknown regression parameters.

Model (2.3) essentially assumes that the gap time between the two monitoring times U and V follows a Cox type model conditional on U. Similar models have been discussed for regression analysis of recurrent event data and multivariate data in Kelly and Lim (2000) and Prentice, Williams and Peterson (1981), respectively. There are several motivations for considering the models described. First, the Cox model is the most widely used model due to its modeling flexibility and it is well studied and easy to implement. Second, under the current model setup, there is an easy procedure, as shown below, to estimate regression coefficients without the need to estimate the baseline hazard functions. Third,

the model assumptions can be easily checked since we have complete data for the monitoring times.

For each i, define a 0-1 counting process $N_i^{(1)}(t) = (1-\delta_{1i})\, I(U_i \le t)$ and, conditional on U_i, define $N_i^{(2)}(t) = \delta_{3i}\, I(V_i \le t)$ if $t \ge U_i$ and 0 if $t < U_i$. The definition of $N_i^{(2)}$ is naturally based on the order restriction between U_i and V_i, and indicates that V_i is considered only after U_i has been observed. Following the same arguments as those in Lin, Oakes, and Ying (1998) and under models (2.1)$\sim$(2.3), we obtain intensity functions for $N_i^{(1)}(t)$ and $N_i^{(2)}(t)$ as

$$\lambda_i^{(1)}(t \,|\, Z_i) = \lambda_1(t)\, e^{-\Lambda_0(t)}\, e^{-\beta_0' Z_i^*(t) + \gamma_0' Z_i(t)} \tag{2.4}$$

$$\lambda_i^{(2)}(t \,|\, U_i,\, Z_i) = I(t > U_i)\lambda_2(t) e^{-\Lambda_0(t)} e^{-\beta_0' Z_i^*(t) + \gamma_0' Z_i(t)}, \tag{2.5}$$

where $Z_i^*(t) = \int_0^t Z_i(s)\, ds$ and $\Lambda_0(t) = \int_0^t \lambda_0(s)\, ds$.

Clearly models (2.4) and (2.5) are Cox type models similar to models (2.2) and (2.3). Note that (2.5) is a conditional model since the starting time point is the observed monitoring time U_i. In the next section, we use (2.2)$-$(2.5) to construct estimating equations for regression coefficients β_0 and γ_0.

3. Estimation of Regression Parameters

To estimate β_0 and γ_0, for $j = 0, 1$, let

$$S_{1,\beta}^{(j)}(t,\beta,\gamma) = n^{-1} \sum_{i=1}^{n} I(t \le U_i)\, e^{-\beta' Z_i^*(t) + \gamma' Z_i(t)}\, Z_i^{*\,(j)}(t),$$

$$S_{2,\beta}^{(j)}(t,\beta,\gamma) = n^{-1} \sum_{i=1}^{n} I(U_i < t \le V_i)\, e^{-\beta' Z_i^*(t) + \gamma' Z_i(t)}\, Z_i^{*\,(j)}(t)\ ,$$

where $Z_i^{*\,(0)}(t) = 1$ and $Z_i^{*\,(1)}(t) = Z_i^*(t)$. Motivated by Lin, Oakes, and Ying (1998) who considered regression analysis of current status data using models (2.1) and (2.2), we propose the estimating function $U_\beta(\beta,\gamma)$ as

$$\sum_{i=1}^{n} \left[\int_0^\infty \Big\{ Z_i^*(t) - \frac{S_{1,\beta}^{(1)}(t,\beta,\gamma)}{S_{1,\beta}^{(0)}(t,\beta,\gamma)} \Big\} dN_i^{(1)}(t) + \int_0^\infty \Big\{ Z_i^*(t) - \frac{S_{2,\beta}^{(1)}(t,\beta,\gamma)}{S_{2,\beta}^{(0)}(t,\beta,\gamma)} \Big\} dN_i^{(2)}(t) \right]$$

$$= \sum_{i=1}^{n} (1-\delta_{1i}) \Big\{ Z_i^*(U_i) - \frac{S_{1,\beta}^{(1)}(U_i,\beta,\gamma)}{S_{1,\beta}^{(0)}(U_i,\beta,\gamma)} \Big\} + \sum_{i=1}^{n} \delta_{3i} \Big\{ Z_i^*(V_i) - \frac{S_{2,\beta}^{(1)}(V_i,\beta,\gamma)}{S_{2,\beta}^{(0)}(V_i,\beta,\gamma)} \Big\},$$

for estimation of β_0 given γ.

In this expression for $U_\beta(\beta,\gamma)$, the first term is the partial likelihood score function under (2.4) if one has only current status data and thus is unbiased.

The second term is the partial likelihood score function obtained under model (2.5) if one considers only current status data given by the V_i's, and thus also has mean 0 at γ_0 and β_0 due to the fact that each integral is a martingale given the observed U_i; thus, $U_\beta(\beta,\gamma)$ is unbiased. The key idea here is to reduce general interval-censored data to current status data, and similar ideas have been used by Betensky, Rabinowitz, and Tsiatis (2001), among others.

For estimation of γ_0, one can easily develop an estimating function that is similar to $U_\beta(\beta,\gamma)$ utilizing (2.4) and (2.5). On the other hand, note that for U_i's and V_i's, or models (2.2) and (2.3), complete data are available and thus it is more efficient to directly estimate γ_0 from them. To this end, let $\tilde{N}_i^{(1)}(t) = I(U_i \le t)$ and $\tilde{N}_i^{(2)}(t) = I(V_i \le t)$ if $t \ge U_i$ and 0 if $t < U_i$ given the observed U_i, $i = 1,\ldots,n$. Also take

$$S_{1,\gamma}^{(j)}(t,\gamma) = n^{-1}\sum_{i=1}^{n} I(t \le U_i)\, e^{\gamma' Z_i(t)}\, Z_i^{(j)}(t),$$

$$S_{2,\gamma}^{(j)}(t,\gamma) = n^{-1}\sum_{i=1}^{n} I(U_i < t \le V_i)\, e^{\gamma' Z_i(t)}\, Z_i^{(j)}(t)\ ,$$

for $j = 0, 1$, where $Z_i^{(0)}(t) = 1$ and $Z_i^{(1)}(t) = Z_i(t)$.

The similarity between models (2.2)$\sim$(2.3) and models (2.4)$\sim$(2.5) suggests an estimating function $U_\gamma(\gamma)$ for γ_0 as

$$\sum_{i=1}^{n}\left[\int_0^\infty \left\{Z_i(t) - \frac{S_{1,\gamma}^{(1)}(t,\gamma)}{S_{1,\gamma}^{(0)}(t,\gamma)}\right\} d\tilde{N}_i^{(1)}(t) + \int_0^\infty \left\{Z_i(t) - \frac{S_{2,\gamma}^{(1)}(t,\gamma)}{S_{2,\gamma}^{(0)}(t,\gamma)}\right\} d\tilde{N}_i^{(2)}(t)\right]$$
$$= \sum_{i=1}^{n}\left\{Z_i(U_i) - \frac{S_{1,\gamma}^{(1)}(U_i,\gamma)}{S_{1,\gamma}^{(0)}(U_i,\gamma)}\right\} + \sum_{i=1}^{n}\left\{Z_i(V_i) - \frac{S_{2,\gamma}^{(1)}(V_i,\gamma)}{S_{2,\gamma}^{(0)}(V_i,\gamma)}\right\}.$$

This estimating function is exactly the same as the partial likelihood score function under models (2.2) and (2.3) for complete data that is found in Lin (1994).

Let $\hat{\gamma}$ be the solution to $U_\gamma(\gamma) = 0$. Then we can estimate β_0 by $\hat{\beta}$ defined as the root of $U_\beta(\beta,\hat{\gamma}) = 0$. Let $\hat{A}_\beta(\beta,\gamma) = -n^{-1}\,\partial U_\beta(\beta,\gamma)/\partial\beta$ and A_β denote the limit of $\hat{A}_\beta(\beta,\gamma)$ at $\beta = \beta_0$ and $\gamma = \gamma_0$. It can be easily shown that $\hat{\gamma}$ is consistent and has an asymptotic normal distribution (Lin (1994); Wei, Lin and Weissfeld (1989)). The consistency of $\hat{\beta}$ can be similarly proved by noting that $\hat{A}_\beta(\beta,\hat{\gamma})$ is positive semidefinite and that its limit is assumed to be positive definite at β_0.

For the asymptotic distribution of $\hat{\beta}$, we note that $n^{-1/2}U_\beta(\beta_0,\hat{\gamma})$ converges in distribution to a normal distribution with mean zero and a covariance matrix that can be consistently estimated see in the Appendix. Then a Taylor series expansion of $U_\beta(\hat{\beta},\hat{\gamma})$ around β_0 shows that the distribution of $n^{1/2}\,(\,\hat{\beta}-\beta_0\,)$ can

be asymptotically approximated by the normal distribution with mean zero and a covariance matrix Σ that can be consistently estimated.

For determination of $\hat{\beta}$ and $\hat{\gamma}$, note that both estimating functions $U_\beta(\beta,\gamma)$ and $U_\gamma(\gamma)$ are similar to the partial likelihood score functions arising from right-censored failure time data under a stratified proportional hazards models, or for multivariate right-censored failure time data under marginal proportional hazards models. Thus $\hat{\beta}$ and $\hat{\gamma}$ can easily obtained using any statistical software designed for these situations.

4. Two Generalizations

In the previous sections we assumed that there were only two monitoring times for each subject and, in practice, there may be more than two. Also we assumed that monitoring times are independent of the survival time of interest given covariates, which may not be true. In this section, we generalize the approach in Section 3 to situations where there exist k monitoring time points for each subject, and where the monitoring times may depend on the failure time.

4.1. Inference with k monitoring variables

In this subsection, we consider the situation where the monitoring process is characterized by $k\,(\geq 2)$ monitoring variables. Let $(V_1,\cdots,V_k)$ be the k monitoring variables and assume that they are independent of the failure time T given the covariate process Z. Let $(V_{i1},\cdots,V_{ik})$ be the realizations of the k monitoring times for subject i, and $V_{i0}=0$ for the sake of notation convenience. Generalizing models (2.2) and (2.3), we assume that given V_{il-1}, the hazard function of V_{il} is

$$\lambda_i^{V_l}\Big(t|\,(V_{i0},\cdots,V_{i\,l-1}),Z(t)\Big)=I(t>V_{i\,l-1})\lambda_l(t)e^{\gamma_0'Z_i(t)},\qquad l=1,\cdots,k,$$

where λ_l is the baseline hazard functions for V_l, $l=1,\ldots,,k$. These models deal naturally with the order restriction between the monitoring variables, and were studied by Prentice, Williams and Peterson (1981) for consecutive failure times.

Let $\delta_{il}=I(V_{i\,l-1}\leq t<V_{il})$ and $N_i^{(l)}(t)=(1-\sum_{h=1}^{l}\delta_{ih})I(V_{i\,l}\leq t)$ if $t>V_{i\,l-1}$ and 0 if $t\leq V_{i\,l-1}$, for $l=1,\cdots,k$ and $i=1,\ldots,,n$. Similarly to models (2.4) and (2.5), we have the intensity function

$$I(V_{i\,l-1}\leq t<V_{il})\lambda_l(t)\,e^{-\Lambda_0(t)}\,e^{-\beta_0'Z_i^*(t)+\gamma_0'Z_i(t)}$$

for $N_i^{(l)}(t)$, $l=1,\cdots,k$ and $i=1,\ldots,,n$. Similarly to those in Section 3, estimation equations can be taken as

$$U_\beta(\beta,\gamma)=\sum_{l=1}^{k}\sum_{i=1}^{n}(1-\sum_{h=1}^{l}\delta_{ih})\Big\{\,Z_i^*(V_{il})-\frac{S_{l,\beta}^{(1)}(V_{il},\beta,\gamma)}{S_{l,\beta}^{(0)}(V_{il},\beta,\gamma)}\,\Big\}=0,$$

$$U_\gamma(\gamma) = \sum_{l=1}^{k} \sum_{i=1}^{n} \left\{ Z_i(V_{il}) - \frac{S^{(1)}_{l,\gamma}(V_{il},\gamma)}{S^{(0)}_{l,\gamma}(V_{il},\gamma)} \right\} = 0,$$

where

$$S^{(j)}_{l,\beta}(t,\beta,\gamma) = n^{-1} \sum_{i=1}^{n} I(V_{il-1} < t \le V_{il})\, e^{-\beta' Z_i^*(t) + \gamma' Z_i(t)}\, Z_i^{*(j)}(t)\,,$$

$$S^{(j)}_{l,\gamma}(t,\gamma) = n^{-1} \sum_{i=1}^{n} I(V_{il-1} < t \le V_{il})\, e^{\gamma' Z_i(t)}\, Z_i^{(j)}(t)\,,$$

for $j = 0, 1$ and $l = 1, \ldots, , k$.

As in Section 3, we can obtain $\hat\gamma$ by solving $U_\gamma(\gamma) = 0$, and then solving $U_\beta(\beta, \hat\gamma) = 0$ for $\hat\beta$ as an estimate of β. The asymptotic theory for $\hat\beta$ can be established as it was for the case of two monitoring variables. The essence of the proposed method is to utilize the complete information about the monitoring variables, and this idea was also used in Betensky, Rabinowitz, and Tsiatis (2001).

4.2. Inference with informative censoring

In this subsection, we consider the situation where the monitoring times may depend on the failure time of interest. This can be the case if the observed interval is given or is formed by the two closest monitoring times containing the failure time. Thus we have informative interval censoring. In the following, we consider a situation where the dependence between the monitoring times and the failure time of interest is induced by sharing a common latent random process in their hazard functions. We show that the approach proposed in Section 3 can apply directly to this situation without any change.

Assume that there exists an unobservable random process b that characterizes the dependency between the monitoring times and the failure time and that, given the covariate process and process b, the monitoring times U and V and the failure time T are independent. The same idea has been used by, for example, Zhang, Sun, and Sun (2005) for current status data. We further assume that

$$\lambda_i^T(t\,|Z_i(s), b_i(s), s \le t) = \lambda_0(t) + \beta_0' Z_i(t) + b_i(t), \tag{4.1}$$

$$\lambda_i^U(t\,|Z_i(s), b_i(s), s \le t) = \lambda_1(t)\, e^{\gamma_0' Z_i(t) + b_i(t)}, \tag{4.2}$$

$$\lambda_i^V(t\,|U_i = u_i, Z_i(s), b_i(s), s \le t) = \begin{cases} \lambda_2(t) e^{\gamma_0' Z_i(t) + b_i(t)} & \text{if} \quad t \ge u_i \\ 0 & \text{if} \quad t < u_i\ , \end{cases} \tag{4.3}$$

where b_i's are i.i.d. realizations of an unobservable stochastic process b, which is assumed to have mean 0. We remark that the model setups (4.1)−(4.3) are quite general since the law of random process b is totally unspecified.

It is easy to show that (4.1) is essentially an additive hazard model since the survival function can be derived as

$$\Pr(T > t|Z(s), s \le t) = \mathbf{E}_b(Pr(T > t|Z(s), b)) = \mathbf{E}_b(B_i(t)) \exp(-\Lambda(t) - \beta' Z_i^*(t)),$$

where $B_i(t) = \int_0^t b_i(s)ds$ and $Z_i^*(t)$ is defined as above. The $\mathbf{E}_b$ term denotes the expectation with respect to b and is not subject specific in the above expression. Let $N_i^{(1)}$, $N_i^{(2)}$, $\tilde{N}_i^{(1)}$, and $\tilde{N}_i^{(2)}$ be defined as in Section 3 and under models (4.1)−(4.3), one can derive their intensity functions as

$$I(U_i \ge t)\, \mathbf{E}_b\{e^{-\int_0^t b_i(s)ds}\, e^{b_i(t)}\}\, e^{-\Lambda_0(t)}\, \lambda_1(t)\, e^{-\beta_0' Z_i^*(t) + \gamma_0' Z_i(t)},$$
$$I(u_i < t \le V_i)\, \mathbf{E}_b\{e^{-\int_0^t b_i(s)ds}\, e^{b_i(t)}\}\, e^{-\Lambda_0(t)}\, \lambda_2(t)\, e^{-\beta_0' Z_i^*(t) + \gamma_0' Z_i(t)},$$
$$I(U_i \ge t)\, \mathbf{E}_b\{b_i(t)\}\, \lambda_1(t)\, e^{\gamma_0' Z_i(t)} \quad \text{and} \quad I(u_i < t \le V_i)\, \mathbf{E}_b\{b_i(t)\}\, \lambda_2(t)\, e^{\gamma_0' Z_i(t)},$$

respectively. It can be seen that these intensity functions are the same as those in (2.2)−(2.5) except for an extra $\mathbf{E}_b$ term.

Note that none of the $\mathbf{E}_b$ terms is subject specific. Since none of the baselines nor the law of $b(t)$ is specified, these nonparametric parts can be put together as one function in each intensity function. Therefore, using the strategy of Sections 2 and 3, we can construct $U_\gamma(\gamma)$ and $U_\beta(\beta, \gamma)$ exactly as in Section 3 and can follow the estimation procedure there. It is easily shown that the asymptotic properties of the obtained estimates given before still hold. In other words, the proposed estimation procedure of Section 3 is robust and applies to the informative censoring case here.

5. A Simulation Study

An extensive simulation study was carried out to assess the finite sample performance of the estimation approach proposed in the previous sections, with the focus on the case of two monitoring variables. In the study, we considered non-informative censoring and informative censoring cases. In the non-informative censoring case, the failure times T_i's were generated from (2.1) and the censoring times U_i's and V_i's were generated from (2.2) and (2.3), respectively, and in the informative censoring case they were generated from (4.1)−(4.3), respectively. In both cases we considered a two-sample problem: the one covariate was bernoulli with success probability 0.5, and we took the baseline hazard functions $\lambda_0(t)$, $\lambda_1(t)$, and $\lambda_2(t)$ to be constants as 2, 4, and 2, respectively, so the proportions of left-, interval- and right-censored observations were 1/3 when $\beta_0 = \gamma_0 = 0$. The true regression parameter β_0 was taken to be 0.5, 0, or −0.5, and γ_0 to be 0.5, 0, or −0.5, resulting in nine setups in each simulation case. Under informative censoring, we took $b_i(t) \equiv b_i$ for simplicity, where b_i's were i.i.d. random

Table 1. Simulation results for estimation of β_0 and γ_0.

TRUE	EST	Non-informative censoring				Informative censoring			
		BIAS	SSD	SEE	CP	BIAS	SSD	SEE	CP
γ_0 = 0.0	$\hat{\gamma}$	-0.0015	0.1460	0.1430	0.939	-0.0051	0.1498	0.1436	0.941
β_0 = 0.0	$\hat{\beta}$	-0.0117	0.6393	0.5773	0.940	0.0121	0.5802	0.5615	0.948
γ_0 = 0.0	$\hat{\gamma}$	0.0090	0.1492	0.1430	0.943	-0.0062	0.1461	0.1438	0.952
β_0 = 0.5	$\hat{\beta}$	0.0423	0.6829	0.6450	0.954	0.0329	0.6709	0.6384	0.947
γ_0 = 0.0	$\hat{\gamma}$	-0.0076	0.1432	0.1431	0.954	-0.0025	0.1467	0.1436	0.948
β_0 = -0.5	$\hat{\beta}$	-0.0201	0.5569	0.5171	0.945	-0.0411	0.5361	0.5073	0.945
γ_0 = 0.5	$\hat{\gamma}$	0.0035	0.1487	0.1482	0.955	-0.0340	0.1460	0.1483	0.946
β_0 = 0.0	$\hat{\beta}$	0.0076	0.6417	0.6263	0.960	-0.0079	0.6498	0.6093	0.946
γ_0 = 0.5	$\hat{\gamma}$	0.0118	0.1514	0.1484	0.946	-0.0266	0.1546	0.1486	0.939
β_0 = 0.5	$\hat{\beta}$	0.0595	0.7389	0.6955	0.940	-0.0177	0.7275	0.6826	0.951
γ_0 = 0.5	$\hat{\gamma}$	0.0143	0.1517	0.1479	0.942	-0.0227	0.1528	0.1492	0.943
β_0 = -0.5	$\hat{\beta}$	0.0025	0.5973	0.5670	0.948	-0.0064	0.7245	0.6733	0.937
γ_0 = -0.5	$\hat{\gamma}$	-0.0165	0.1536	0.1482	0.945	0.0276	0.1484	0.1487	0.951
β_0 = 0.0	$\hat{\beta}$	-0.0023	0.6121	0.5812	0.940	0.0397	0.5943	0.5687	0.951
γ_0 = -0.5	$\hat{\gamma}$	-0.0096	0.1539	0.1484	0.944	0.0179	0.1548	0.1486	0.942
β_0 = 0.5	$\hat{\beta}$	0.0292	0.7071	0.6520	0.948	0.0486	0.6094	0.5696	0.942
γ_0 = -0.5	$\hat{\gamma}$	0.0001	0.1502	0.1477	0.940	0.0252	0.1548	0.1485	0.943
β_0 = -0.5	$\hat{\beta}$	0.0420	0.5644	0.5173	0.950	-0.0126	0.5518	0.5084	0.944

effects generated from $N(0,1)/4$. The small variance of b_i was taken to ensure positive hazards when generating T_i's in all setups.

Table 1 presents the simulation results based on 1,000 replicates for each setup with sample size $n = 100$. For each setup, the results include the bias (BIAS) given by the average of 1,000 point estimates minus the true value, the sample standard deviation (SSD) of the 1,000 point estimates, the average of 1,000 estimated standard errors (SEE), and the 95% empirical coverage probability. It can be seen from Table 1 that the proposed approach worked very well in both non-informative and informative censoring cases: the biases of the proposed estimates were small, the sample standard deviation and the estimated standard error were quite close, and the empirical coverage probabilities seemed quite close to 95% in all setups.

It is interesting to note that the variance estimates of the regression parameters under informative censoring were not larger than those under non-informative censoring even though there is more variability of data in the former setting. This is not surprising, and is due to the unspecified baseline hazard functions and the unspecified law of b as seen in Subsection 4.2. We observed better performance of the proposed approach when sample size was increased to $n = 200$ (results not shown). The program coded in Matlab is very fast: just

Table 2. Analysis results for breast cancer data.

Parameter	Point estimate	Standard error	Z-score	P-value
γ_0	-0.4261	0.1811	-2.3532	0.0186
β_0	-0.0164	0.0067	-2.4368	0.0148
γ_1	-0.0422	0.2891	-0.1459	0.8840
γ_2	-0.6362	0.2301	-2.7647	0.0057
β_0	-0.0149	0.0061	-2.4568	0.0140

over a minute for each setup with $n = 100$ on Dell laptop Latitude D830 with Intel(R) core 2 Duo CPU and 3.5 GB of RAM.

6. An Illustration

This section applies the proposed approach to the breast cancer data discussed in Finkelstein (1986), among others. The study consisted of 94 early breast cancer patients who were given either radiation therapy alone (46), or radiation therapy plus adjuvant chemotherapy (48). During the study, patients were supposed to be seen at clinic visits every 4 to 6 months. Actual visit times differed from patient to patient, and times varied between visits. At the visits, physicians evaluated features such as breast retraction, a response that has a negative impact on overall cosmetic appearance. The goal of the study was to compare the two treatments with respect to the time to breast retraction, and only interval-censored data were available.

To apply the proposed method, we assume that the time to breast retraction and the monitoring times can be described by models (4.1)−(4.3). The breast cancer data are given in the form $[L_i, R_i)$: we have a mixture of left-, interval-, and right-censored observations. We made an adjustment: for subject i with $[L_i, R_i)$, if $L_i = 0$, we took $U_i = R_i$ and V_i to be the largest observation in the study; if $R_i = \infty$, we took $V_i = L_i$ and U_i to be the smallest observation time in the study; when $L_i \neq 0$ and $R_i \neq \infty$, we took $U_i = L_i$ and $V_i = R_i$. Corresponding to this adjustment, we adjusted $U_\gamma(\gamma)$ to

$$U_\gamma(\gamma) = \sum_{i=1}^{n}(1-\delta_{3i})\left\{Z_i(U_i) - \frac{S^{(1)}_{1,\gamma}(U_i,\gamma)}{S^{(0)}_{1,\gamma}(U_i,\gamma)}\right\} + \sum_{i=1}^{n}(1-\delta_{1i})\left\{Z_i(V_i) - \frac{S^{(1)}_{2,\gamma}(V_i,\gamma)}{S^{(0)}_{2,\gamma}(V_i,\gamma)}\right\},$$

and kept $U_\beta(\beta,\gamma)$ unchanged. This essentially treats U_i as missing in the right-censored case, and V_i as missing in the left-censoring case.

We let $Z_i = 1$ if the ith patient was given radiation therapy alone and 0 otherwise. Table 2 shows the analysis results using the proposed method when U and V share the same covariate effect γ_0, and when they have different covariate effects γ_1 and γ_2, respectively. The difference between the estimates of γ_1 and

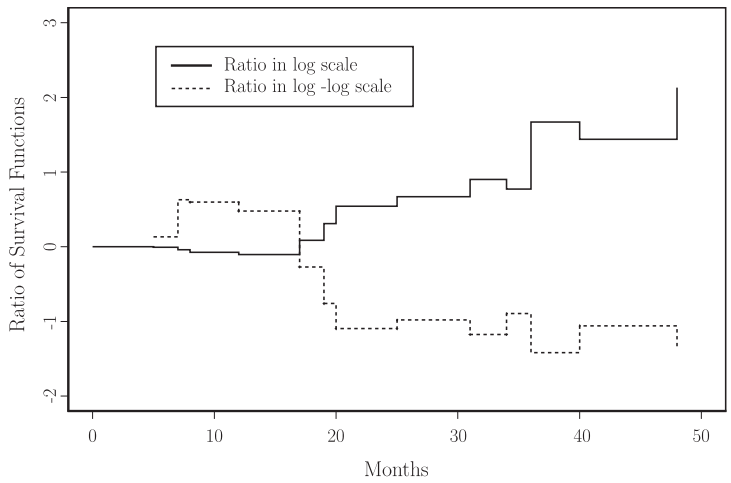

Figure 1. Ratios of estimated survival functions in log and log-log scales.

γ_2 suggests that the models with different covariate effects on U and V are more realistic than the models with the common covariate effect. In that analysis we obtained $\hat{\beta} = -0.0149$, with estimated standard error 0.0061 and the p-value 0.0140 for testing $\beta_0 = 0$. These results suggest that the patients given radiation therapy alone had a significantly lower risk to develop breast retraction than those given radiation therapy plus adjuvant chemotherapy. This conclusion agrees with that given by Finkelstein (1986) using the proportional hazards model.

Although the result given here is similar to that obtained using the proportional hazards model, it is of interest to assess which of the two models is more appropriate for the given data set. To this end, we determined the nonparametric maximum likelihood estimators of the survival functions for the two treatment groups and plotted the log ratio of the survival function estimators and the log ratio of the log survival function estimators in Figure 1. Note that the former should give a straight line passing the origin under the additive hazards model and the latter should give a line parallel to x-axis under the proportional hazards model. Although nothing is clear cut, Figure 1 suggests that the additive hazards model is more reasonable in terms of the global effect.

7. Concluding Remarks

We have discussed regression analysis of case II interval-censored failure time data using the additive hazards model. Some estimating equation-based approaches were developed for estimation of regression parameters and the asymptotic properties of the proposed estimates were established. The proposed approaches apply to both noninformative censoring and informative censoring cases. One major

advantage of the presented method is that it does not involve estimation of any baseline hazard function.

The proposed methodology involves modeling gap times between the adjacent monitoring times using the Cox model. An alternative to this is to directly model all the monitoring times with the Cox model marginally. However, such modeling requires stricter conditions due to the order relationship, as shown in Yang and Ying (2001). In contrast, the gap time modeling approach is very flexible.

The proposed approach provides an alternative to the full likelihood method given in Zeng, Cai, and Shen (2006), who explored efficient estimation for regression analysis of case II interval-censored data. One merit of the full likelihood method is that it does not impose distribution assumptions on the monitoring times U and V, but it can be time-consuming and sometimes infeasible since it involves estimation of the infinite-dimensional cumulative baseline hazard function Λ_0. The advantage of the approaches proposed here lies in easy and fast implementation. Moreover, the proposed methods can deal with the informative censoring that is common for interval-censored data.

Acknowledgement

The authors thank the Editor, an associate editor, and two reviewers for their critical comments and suggestions that have greatly improved the original presentation. The third author's work is partly supported by NSFC 10971015.

Appendix

Asymptotic normality of $n^{-1/2}\, U_\beta(\beta_0, \hat{\gamma})$, Section 3.
For $i = 1, \ldots, n$, define

$$
\begin{aligned}
M_i^{(1)}(t) &= N_i^{(1)}(t) - \int_0^t I(s \le U_i)\, \lambda_1^*(s)\, e^{-\beta_0' Z_i^*(s) + \gamma_0' Z_i(s)}\, ds\,, \\
M_i^{(2)}(t) &= N_i^{(2)}(t) - \int_0^t I(U_i < s \le V_i)\, \lambda_2^*(s)\, e^{-\beta_0' Z_i^*(s) + \gamma_0' Z_i(s)}\, ds\,, \\
\tilde{M}_i^{(1)}(t) &= \tilde{N}_i^{(1)}(t) - \int_0^t I(s \le U_i)\, \lambda_1(s)\, e^{\gamma_0'\, Z_i(s)}\, ds\,, \\
\tilde{M}_{2i}^{(2)}(t) &= \tilde{N}_i^{(2)}(t) - \int_0^t I(U_i < s \le V_i)\, \lambda_2(s)\, e^{\gamma_0'\, Z_i(s)}\, ds\,,
\end{aligned}
$$

where $\lambda_1^*(t) = \lambda_1(t)\, e^{-\Lambda_0(t)}$ and $\lambda_2^*(t) = \lambda_2(t)\, e^{-\Lambda_0(t)}$. Then $M_i^{(1)}$, and $\tilde{M}_i^{(1)}$ are martingales starting at 0, and $M_i^{(2)}$ and $\tilde{M}_i^{(2)}$ are martingales starting at the

observed monitoring time U_i. Also define

$$A_1 = E\Big(\int_0^\infty \{Z_1^*(t) - \frac{s_{1,\beta}^{(1)}(t,\beta_0,\gamma_0)}{s_{1,\beta}^{(0)}(t,\beta_0,\gamma_0)}\}^{\otimes 2} I(U_1 \geq t)\lambda_1^*(t)\, e^{-\beta_0' Z_1^*(t)+\gamma_0' Z_1(t)}\, dt \Big),$$

$$A_2 = E\Big(\int_0^\infty \{Z_1^*(t) - \frac{s_{2,\beta}^{(1)}(t,\beta_0,\gamma_0)}{s_{2,\beta}^{(0)}(t,\beta_0,\gamma_0)}\}^{\otimes 2} I(U_1 < t \leq V_1)\lambda_2^*(t) e^{-\beta_0' Z_1^*(t)+\gamma_0' Z_1(t)} dt \Big),$$

$$\tilde{A}_1 = E\Big(\int_0^\infty \{Z_1(t) - \frac{s_{1,\gamma}^{(1)}(t,\gamma_0)}{s_{1,\gamma}^{(0)}(t,\gamma_0)}\}^{\otimes 2} I(U_1 \geq t)\lambda_1(t)\, e^{\gamma_0' Z_1(t)}\, dt \Big),$$

$$\tilde{A}_2 = E\Big(\int_0^\infty \{Z_1(t) - \frac{s_{2,\gamma}^{(1)}(t,\gamma_0)}{s_{2,\beta}^{(0)}(t,\gamma_0)}\}^{\otimes 2} I(U_1 < t \leq V_1)\lambda_2(t)\, e^{\gamma_0' Z_1(t)}\, dt \Big),$$

where $s_{l,\gamma}^{(j)}(t,\gamma)$ and $s_{l,\beta}^{(j)}(t,\beta,\gamma)$ denote the limits of $S_{l,\gamma}^{(j)}(t,\gamma)$ and $S_{l,\beta}^{(j)}(t,\beta,\gamma)$, respectively, for $l=1,2$ and $j=0,1$. Let $A_\gamma = A_1 + A_2$ and $B = \tilde{A}_1 + \tilde{A}_2$, and assume that both A_γ and B are positive definite. Also let $\hat{A}_\gamma(\beta,\gamma) = n^{-1}\partial U_\beta(\beta,\gamma)/\partial\gamma$ and $\hat{B}(\gamma) = -n^{-1}\,\partial U_\gamma(\gamma)/\partial\gamma$. Then A_γ and B are the limits of $\hat{A}_\gamma(\beta,\gamma)$ and $\hat{B}(\gamma)$ at β_0 and γ_0, respectively.

To investigate the asymptotic normality of $n^{-1/2}\, U_\beta(\beta_0,\hat{\gamma})$, first note that a Taylor series expansions of $U_\beta(\beta_0,\hat{\gamma})$ and $U_\gamma(\hat{\gamma})$ around γ_0 has

$$n^{-1/2}\, U_\beta(\beta_0,\hat{\gamma}) = n^{-1/2}\, U_\beta(\beta_0,\gamma_0) + A_\gamma\, B^{-1}\, \{\, n^{-1/2}\, U_\gamma(\gamma_0)\,\} + o_p(1)\,.$$

Following Lin, Oakes, and Ying (1998), it can be shown that

$$\begin{aligned} n^{-1/2}\, U_\beta(\beta_0,\gamma_0) &= n^{-1/2} \sum_{i=1}^n \{\, a_{1i}(\beta_0,\gamma_0) + a_{2i}(\beta_0,\gamma_0)\,\} + o_p(1) \\ n^{-1/2}\, U_\gamma(\gamma_0) &= n^{-1/2} \sum_{i=1}^n \{\, b_{1i}(\gamma_0) + b_{2i}(\gamma_0)\,\} + o_p(1)\,, \end{aligned}$$

where

$$\begin{aligned} a_{1i}(\beta,\gamma) &= \int_0^\infty \Big\{ Z_i^*(t) - \frac{s_{1,\beta}^{(1)}(t,\beta,\gamma)}{s_{1,\beta}^{(0)}(t,\beta,\gamma)} \Big\} dM_i^{(1)}(t)\,, \\ a_{2i}(\beta,\gamma) &= \int_0^\infty \Big\{ Z_i^*(t) - \frac{s_{2,\beta}^{(1)}(t,\beta,\gamma)}{s_{2,\beta}^{(0)}(t,\beta,\gamma)} \Big\} dM_i^{(2)}(t)\,, \\ b_{1i}(\gamma) &= \int_0^\infty \Big\{ Z_i(t) - \frac{s_{1,\gamma}^{(1)}(t,\gamma)}{s_{1,\gamma}^{(0)}(t,\gamma)} \Big\} d\tilde{M}_i^{(1)}(t), \end{aligned}$$

$$b_{2i}(\gamma) = \int_0^\infty \Big\{ Z_i(t) - \frac{s_{2,\gamma}^{(1)}(t,\gamma)}{s_{2,\gamma}^{(0)}(t,\gamma)} \Big\} d\tilde{M}_i^{(2)}(t),$$

which are all martingales with mean zero. Then

$$n^{-1/2}\, U_\beta(\beta_0, \hat{\gamma}) = n^{-1/2} \sum_{i=1}^{n} \alpha_i(\beta_0, \gamma_0) + o_p(1) \ ,$$

where $\alpha_i(\beta,\gamma) = a_{1i}(\beta,\gamma) + a_{2i}(\beta,\gamma) + A_\gamma\, B^{-1} \{\, b_{1i}(\gamma) + b_{2i}(\gamma) \,\}$. It thus follows from the multivariate Central Limit Theorem or U-statistic theory (Lee (1990) that $n^{-1/2}\, U_\beta(\beta_0, \hat{\gamma})$ converges in distribution to a zero-mean normal random vector.

The asymptotic covariance matrix of $n^{-1/2}\, U_\beta(\beta_0, \hat{\gamma})$ can be consistently estimated by

$$\hat{\Gamma} = \frac{1}{n} \sum_{i=1}^{n} \hat{\alpha}_i(\hat{\beta}, \hat{\gamma})\ \hat{\alpha}_i'(\hat{\beta}, \hat{\gamma}) \ ,$$

with $\hat{\alpha}_i(\hat{\beta}, \hat{\gamma}) = \hat{a}_{1i}(\hat{\beta}, \hat{\gamma}) + \hat{a}_{2i}(\hat{\beta}, \hat{\gamma}) + \hat{A}_\gamma(\hat{\beta}, \hat{\gamma})\, \hat{B}(\hat{\gamma}) \{\, \hat{b}_{1i}(\hat{\gamma}) + \hat{b}_{2i}(\hat{\gamma}) \,\}$, where $\hat{a}_{1i}$, $\hat{a}_{2i}$, $\hat{b}_{1i}$ and $\hat{b}_{2i}$ are the estimates of a_{1i}, a_{2i}, b_{1i} and b_{2i}, respectively, with the corresponding martingale replaced by its estimate in each expression.

References

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

Cox, D. R. (1972). Regression analysis and life tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34**, 187-220.

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

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Ann. Statist.* **24**, 540-568.

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

Kelly, P. J. and Lim, L. L-Y. (2000). Survival analysis for recurrent event data: an application to childhood infectious diseases. *Statist. Medicine* **19**, 13-33.

Lee, A. J. (1990). *U-statistics: Theory and Practice.* Marcel Dekker, New York.

Lin, D. Y. (1994). Cox regression analysis of multivariate failure time data: the marginal approach. *Statist. Medicine* **13**, 2233-2247.

Lin, D. Y., Oakes, D. and Ying, Z. (1998). Additive hazards regression with current status data. *Biometrika* **85**, 289-298.

Martinussen, T. and Scheike, T. H. (2002). Efficient estimation in additive hazards regression with current status data. *Biometrika* **89**, 649-658.

Prentice, R. L., Williams, B. J. and Peterson, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika* **68**, 373-379.

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

Wei, L. J., Lin, D. Y. and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J. Amer. Statist. Assoc.* **84**, 1065-1073.

Yang, Y. and Ying, Z. (2001). Marginal proportional hazards models for multiple event-time data. *Biometrika* **88**, 581-586.

Zeng, D., Cai, J. and Shen, Y. (2006). Semiparametric additive risks model for interval-censored data. *Statist. Sinica* **16**, 287-302.

Zhang, Z., Sun, J. and Sun, L. (2005). Statistical analysis of current status data with informative observational times. *Statist. Medicine* **24**, 1399-1407.

Department of Statistics, University of South Carolina, 209C, LeConte College, Columbia, SC 29208, USA.

E-mail: wang99@mailbox.sc.edu

Department of Statistics, University of Missouri, 134E Middlebush Hall, Missouri 65211, USA.

E-mail: sunj@missouri.edu

Department of Statistics and Financial Mathematics, Beijing Normal University, Beijing 100875, China.

E-mail: xweitong@bnu.edu.cn

(Received June 2007; accepted May 2009)