Biometrika (1998), **85**, 2, *pp.* 289–298
Printed in Great Britain

Additive hazards regression with current status data

BY D. Y. LIN
Department of Biostatistics, Box 357232, University of Washington, Seattle, Washington 98195, U.S.A.
danyu@biostat.washington.edu

DAVID OAKES
Department of Biostatistics, University of Rochester Medical Center, 610 Elmwood Avenue, Box 630, Rochester, New York 14642, U.S.A.
oakes@metro.bst.rochester.edu

AND ZHILIANG YING
Department of Statistics, Hill Center, Busch Campus, Rutgers University, Piscataway, New Jersey 08855, U.S.A.
zying@stat.rutgers.edu

SUMMARY

Current status data arise when the only knowledge about the failure time of interest is whether the failure occurs before or after a random monitoring time. We propose to analyse such data by the semiparametric additive hazards model, which specifies that the hazard function for the failure time associated with a set of possibly time-dependent covariates is the sum of an arbitrary baseline hazard function and a regression function of covariates. Under certain conditions on the monitoring time, one can make inferences about the regression parameters of the additive hazards model by using the familiar asymptotic theory and software for the proportional hazards model with right censored data. An application to a carcinogenicity experiment is provided.

Some key words: Additive risk model; Counting process; Failure time; Interval censoring; Martingale; Partial likelihood; Proportional hazards; Time-dependent covariate.

1. INTRODUCTION

Current status data arise when the failure time of interest T cannot directly be observed, but can only be determined to lie below or above a random monitoring time C. Such data are commonly encountered in biomedicine, economics, sociology and other scientific areas. For example, in carcinogenicity experiments, animals are randomly assigned to various doses of a suspected carcinogen and are examined at sacrifice or death time for evidence of a malignancy. The time to tumour onset is of interest, but not directly observable. Rather, one knows only the age at death and whether or not the tumour is present at that time.

There is much less information about the failure time T in current status data than in the familiar right-censored data. In the latter, one observes $\min(T, C)$ and $I(C \leqslant T)$, where

$I(.)$ is the indicator function; the probability of observing T exactly is positive. In the former, however, one only observes C and $I(C \leqslant T)$; the exact value of T is never observed.

A number of authors, including Ayer et al. (1955), Peto (1973), Turnbull (1976) and Groeneboom & Wellner (1992, § 2.3), have introduced algorithms for the nonparametric maximum likelihood estimation of the distribution function of T with current status data. Groeneboom & Wellner (1992, §§ 4.1, 5.1) showed that the estimator is consistent and converges at $n^{1/3}$-rate to a complicated limiting distribution.

Semiparametric regression methods have also been studied. Finkelstein (1986) developed a method for fitting the proportional hazards model (Cox, 1972) to current status data, but the properties of her estimator are still unknown. Diamond & McDonald (1991) and Shiboski & Jewell (1992) translated the proportional hazards model to a semiparametric binary regression model, and proposed some ad hoc estimation procedures. Huang (1996) recently showed that, profiled over the cumulative baseline hazard function, the maximum likelihood estimator for the regression parameter under the proportional hazards model is asymptotically normal with the $n^{\frac{1}{2}}$-convergence rate. Klein & Spady (1993) and Rabinowitz, Tsiatis & Aragon (1995) considered the linear regression model. All the aforementioned regression methods are fairly complicated, requiring the estimation of the distribution function, or even the density, of T. Furthermore, none of them handles time-dependent covariates.

In the present paper, we study semiparametric methods for analysing current status data under the additive hazards regression model. This model specifies that the hazard function of T at time t, given the history of a p-dimensional covariate process $Z(.)$ up to t, has the form

$$\lambda(t \mid Z) = \lambda_0(t) + \beta_0' Z(t), \tag{1·1}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function, and β_0 is a p-vector of unknown regression parameters. Some related models are discussed in § 5. Normally, $Z(.)$ is restricted to external covariates (Kalbfleisch & Prentice, 1980, p. 123).

There are two major motivations for using model (1·1). First, the additive hazards model describes a different aspect of the association between the failure time and covariates than the proportional hazards model, and is more plausible than the latter in many applications (Aranda-Ordaz, 1983; Buckley, 1984; Cox & Oakes, 1984, p. 74; Thomas, 1986; Breslow & Day, 1987, pp. 122–31, 182; Lin & Ying, 1994). Secondly, we have discovered that there exist surprisingly simple methods for making inferences about β_0 under model (1·1) with current status data.

In the next section, we present the new methodology for the special case in which the monitoring time C is independent of T and Z. In § 3, we extend the results of § 2 to allow the dependence of C on Z through the proportional hazards model. In § 4, we report the results of our simulation studies and provide an illustration with real data taken from a carcinogenicity experiment. In § 5, we discuss the extension of our approach to more general interval censored data and give some concluding remarks.

2. INFERENCE PROCEDURES WITH INDEPENDENT MONITORING

In this section, we assume that C is independent of T and Z. Let $\{T_i, C_i, Z_i(.)\}$ $(i = 1, \ldots, n)$ be independent replicates of $\{T, C, Z(.)\}$. The observations consist of C_i, $\delta_i := I(C_i \leqslant T_i)$ and $Z_i(t)$ $(t \leqslant C_i)$.

For $i = 1, \ldots, n$, define the counting process $N_i(t) = \delta_i I(C_i \leqslant t)$, which jumps by unity

whenever subject i is monitored at time t and found still to be failure-free. The process $N_i(t)$ is regarded as censored when subject i is monitored and found to have experienced failure. Subject i is at risk at t if and only if $N_i(t)$ is still uncensored and has not taken its jump by time t, which happens if and only if $C_i \geqslant t$. Consider now the hazard $dH_i(t)$, say, of a jump of $N_i(t)$ for a subject that has not yet been monitored. Two things have to happen for $dN_i(t) = 1$, namely (i) $C_i = t$, and (ii) the resulting monitoring reveals that subject i has been failure-free up to t. Let us denote the hazard of event (i) by $d\Lambda_c(t)$. Under model (1·1), event (ii) has the conditional probability

$$\mathrm{pr}\{T_i \geqslant t \mid Z_i(s), s \leqslant t\} = e^{-\Lambda_0(t) - \beta_0' Z_i^*(t)},$$

where

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

The product of the two probabilities for events (i) and (ii) is of the form

$$dH_i(t) = e^{-\beta_0' Z_i^*(t)}\, dH_0(t), \tag{2·1}$$

where $dH_0(t) = e^{-\Lambda_0(t)}\, d\Lambda_c(t)$.

Equation (2·1) is of course the well-known Cox proportional hazards model. Letting $Y_i(t) = I(C_i \geqslant t)$, we show in the Appendix that the compensated counting processes

$$M_i(t) := N_i(t) - \int_0^t Y_i(s) e^{-\beta_0' Z_i^*(s)}\, dH_0(s) \quad (i = 1, \ldots, n)$$

are martingales with respect to the σ-filtration

$$\mathscr{F}_t := \sigma\{N_i(s), Y_i(s), Z_i(s) : s \leqslant t, i = 1, \ldots, n\}.$$

Thus, one can make inferences about β_0 by applying the partial likelihood principle to model (2·1) with the data $\{N_i(.), Y_i(.), Z_i^*(.)\}$ $(i = 1, \ldots, n)$. Specifically, the partial likelihood score function and observed information matrix for β_0 are

$$U(\beta) := \sum_{i=1}^n \int_0^\infty \left\{ Z_i^*(t) - \frac{S^{(1)}(\beta, t)}{S^{(0)}(\beta, t)} \right\} dN_i(t), \tag{2·2}$$

$$\mathscr{I}(\beta) := \sum_{i=1}^n \int_0^\infty \left\{ \frac{S^{(2)}(\beta, t)}{S^{(0)}(\beta, t)} - \frac{S^{(1)}(\beta, t)^{\otimes 2}}{S^{(0)}(\beta, t)^2} \right\} dN_i(t),$$

where

$$S^{(k)}(\beta, t) = \sum Y_j(t) e^{-\beta' Z_j^*(t)} Z_j^*(t)^{\otimes k} \quad (k = 0, 1, 2),$$

and $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$ and $a^{\otimes 2} = aa'$. The maximum partial likelihood estimator $\hat{\beta}$ is the solution to $U(\beta) = 0$. By the counting-process martingale arguments (Andersen & Gill, 1982), the random vectors $n^{-\frac{1}{2}} U(\beta_0)$ and $n^{\frac{1}{2}}(\hat{\beta} - \beta_0)$ converge in distribution to zero-mean normal random vectors with covariance matrices Ω and Ω^{-1}, respectively, where Ω is the limit of $n^{-1}\mathscr{I}(\beta_0)$ and can be consistently estimated by $n^{-1}\mathscr{I}(\beta_0)$ or $n^{-1}\mathscr{I}(\hat{\beta})$. As usual, we assume that Ω is nonsingular.

The above asymptotic results can be used to test nonparametrically the null hypothesis that $\beta_0 = 0$ and to make semiparametric inferences about the regression parameters. These procedures are much simpler than the existing regression methods for current status data,

not involving any complicated nonparametric estimation of the distribution of T. No special software is required to implement the proposed methods. In fact, one can just input the data $\{C_i, \delta_i, Z_i^*(.)\}$ $(i=1,\ldots,n)$ into standard software for fitting the proportional hazards model with right-censored data. Incidentally, if model (1·1) were replaced by a proportional hazards model, then $dH_i(t)$ would not be of the proportional hazards form so that partial likelihood methods would not be applicable.

Interestingly, the distribution of the censoring time for $N_i(t)$ also involves β_0 so that the censoring times are informative about β_0. This implies that $\hat{\beta}$ may not be fully efficient. It is possible to improve efficiency by utilising the extra information about β_0 provided by the censoring times, as will be discussed at the end of § 3.

When the monitoring time distribution is degenerate, $U(\beta)$ has a close connection with the parametric likelihood score function. To see this, suppose that $C_i = c$, a positive constant. We then have a simple regression problem with binary responses $I(T_i \geqslant c)$ $(i=1,\ldots,n)$. Under the model (1·1), $\mathrm{pr}(T_i \geqslant c \mid Z_i) = e^{-\alpha_0 - \beta_0'\tilde{Z}_i}$, where $\alpha_0 = \Lambda_0(c)$ and $\tilde{Z}_i = Z_i^*(c)$. The likelihood score equations for the unknown parameters α_0 and β_0 are

$$\sum_{i=1}^{n} \begin{bmatrix} 1 \\ \tilde{Z}_i \end{bmatrix} (1 - e^{-\alpha-\beta'\tilde{Z}_i})^{-1} \{I(T_i \geqslant c) - e^{-\alpha-\beta'\tilde{Z}_i}\} = 0.$$

Removing the weights $(1 - e^{-\alpha-\beta'\tilde{Z}_i})^{-1}$ from the above equations and eliminating the nuisance parameter α, we obtain the following equation for β_0:

$$\sum_{i=1}^{n} \left(\tilde{Z}_i - \frac{\sum_{j=1}^{n} e^{-\beta'\tilde{Z}_j} \tilde{Z}_j}{\sum_{j=1}^{n} e^{-\beta'\tilde{Z}_j}} \right) I(T_i \geqslant c) = 0.$$

The left-hand side of the above equation is exactly what (2·2) reduces to in this special case. Since the maximum likelihood estimator is asymptotically efficient, we expect $\hat{\beta}$ to have high efficiency at least when C is not too dispersed.

3. Inference procedures with dependent monitoring

When the monitoring time C is not independent of the covariate vector Z, we formulated their relationship through the proportional hazards model. Specifically, the hazard function of C at time t conditional on the covariate history up to t is given by

$$d\Lambda_c(t \mid Z) = e^{\gamma_0' Z(t)}\, d\Lambda_{c,0}(t), \tag{3·1}$$

where $\Lambda_{c,0}(t)$ is an unspecified baseline cumulative hazard function, and γ_0 is a p-vector of unknown regression parameters. We require that C and T be independent conditional on Z.

By the arguments leading to model (2·1), we can show that, under models (1·1) and (3·1),

$$dH_i(t) = e^{-\beta_0' Z_i^*(t) + \gamma_0' Z_i(t)}\, dH_0(t), \tag{3·2}$$

where $dH_0(t) = e^{-\Lambda_0(t)}\, d\Lambda_{c,0}(t)$. This is also a proportional hazards model. Furthermore, as shown in the Appendix, the compensated counting processes

$$M_i(t) := N_i(t) - \int_0^t Y_i(s) e^{-\beta_0' Z_i^*(s) + \gamma_0' Z_i(s)}\, dH_0(s) \quad (i=1,\ldots,n) \tag{3·3}$$

are martingales with respect to the σ-filtration $\mathscr{F}_t$. Here and in the sequel, we adopt the

notation N_i, M_i, H_i, H_0, etc. used in § 2. This should not create any ambiguity since § 2 may be regarded as a special case of the present section with $\gamma_0 = 0$.

As in § 2, one could apply the partial likelihood principle to model (3·2) with the augmented data $\{N_i(.), Y_i(.), Z_i^*(.), Z_i(.)\}$ $(i = 1, \ldots, n)$ to make inferences about β_0 and γ_0. However, it will be more efficient to estimate γ_0 by applying the partial likelihood theory directly to model (3·1) since the C_i are always observed. Let $\hat{\gamma}$ be the maximum partial likelihood estimator for γ_0 based on the data $\{C_i, Z_i(.)\}$ $(i = 1, \ldots, n)$, which is the root of the score function

$$U_\gamma(\gamma) := \sum_{i=1}^{n} \left\{ Z_i - \frac{\sum_{j=1}^{n} Y_j(C_i) e^{\gamma' Z_j(C_i)} Z_j(C_i)}{\sum_{j=1}^{n} Y_j(C_i) e^{\gamma' Z_j(C_i)}} \right\}.$$

Given $\hat{\gamma}$, we estimate β_0 by the estimating function $U_\beta(\beta; \hat{\gamma})$, where

$$U_\beta(\beta; \gamma) := \sum_{i=1}^{n} \int_0^\infty \left\{ Z_i^*(t) - \frac{S^{(1)}(\beta, \gamma, t)}{S^{(0)}(\beta, \gamma, t)} \right\} dN_i(t) \tag{3·4}$$

is the partial likelihood score function for β_0 under model (3·2), and

$$S^{(k)}(\beta, \gamma, t) = \sum Y_j(t) e^{-\beta' Z_j^*(t) + \gamma' Z_j(t)} Z_j^*(t)^{\otimes k} \quad (k = 0, 1).$$

Denote the resulting estimator by $\tilde{\beta}$.

Note that $S^{(k)}(\beta, \gamma, t)$ $(k = 0, 1)$ differ from $S^{(k)}(\beta, t)$ only in that the former involve an extra factor $e^{\gamma' Z(t)}$, which is free of β. Thus, the numerical propertiess of $U_\beta(\beta; \hat{\gamma})$ as a function of β are very similar to those of (2·2), and $\tilde{\beta}$ can be obtained by the usual Newton–Raphson algorithm.

Let

$$\hat{\Omega}_\beta(\beta; \gamma) = n^{-1} \frac{\partial U_\beta(\beta; \gamma)}{\partial \beta'}, \quad \hat{\Omega}_{\beta\gamma}(\beta; \gamma) = -n^{-1} \frac{\partial U_\beta(\beta; \gamma)}{\partial \gamma'}, \quad \hat{D}_\gamma(\gamma) = -n^{-1} \frac{\partial U_\gamma(\gamma)}{\partial \gamma'},$$

and let Ω_β, $\Omega_{\beta\gamma}$ and D_γ denote their limits at $\beta = \beta_0$ and $\gamma = \gamma_0$. Assume that Ω_β and D_γ are nonsingular. We show in the Appendix that the random vectors $n^{-\frac{1}{2}} U_\beta(\beta_0; \hat{\gamma})$ and $n^{\frac{1}{2}}(\tilde{\beta} - \beta_0)$ converge in distribution to zero-mean normal random vectors with covariance matrices $\Omega_\beta - \Omega_{\beta\gamma} D_\gamma^{-1} \Omega_{\beta\gamma}'$ and $V_\beta := \Omega_\beta^{-1} - \Omega_\beta^{-1} \Omega_{\beta\gamma} D_\gamma^{-1} \Omega_{\beta\gamma}' \Omega_\beta^{-1}$, respectively, and the limiting covariance matrices can be consistently estimated by replacing Ω_β, $\Omega_{\beta\gamma}$ and D_γ by their respective sample estimators.

If γ_0 is known, then Ω_β^{-1} will be the limiting covariance matrix for the maximum partial likelihood estimator of β_0 under model (3·2). Since $\Omega_\beta^{-1} \Omega_{\beta\gamma} D_\gamma^{-1} \Omega_{\beta\gamma}' \Omega_\beta^{-1}$ is positive semidefinite, V_β is always smaller than Ω_β^{-1} in that $\Omega_\beta^{-1} - V_\beta$ is positive semidefinite; i.e. it will be more efficient to estimate γ_0 from the data even if γ_0 is known. Thus, we can gain efficiency by using the two-stage method even when we know that the monitoring time is independent of the covariates, that is, $\gamma_0 = 0$. On the other hand, standard likelihood theory indicates that the covariance matrix for the maximum partial likelihood estimator of β_0 under model (3·2) with unknown γ_0 is always larger than Ω_β^{-1}. These results confirm that the proposed two-step estimation procedure is indeed more efficient than the simultaneous partial-likelihood estimation of β_0 and γ_0 under model (3·2) based on $\{N_i(.), Y_i(.), Z_i^*(.), Z_i(.)\}$ $(i = 1, \ldots, n)$. If for some reason one component of $Z(t)$ is the integral of another component of $Z(t)$, so that $Z(t)$ and $Z^*(t)$ share a common component, the parameters β_0 and γ_0 in model (3·2) will not be identifiable. In this case, only the two-step procedure can be used.

4. NUMERICAL RESULTS

4·1. *Simulation studies*

A series of simulation studies were conducted to assess the performance of the methods developed in the previous two sections. The failure times were generated from model (1·1) with $\lambda_0(.) = 1$, $\beta_0 = 0{\cdot}5$ and Z being a uniform random variable on $(0, \sqrt{(12)})$. The monitoring times were generated from the exponential distribution with hazard rate $\lambda_{c,0} e^{\gamma_0 Z}$, where $\lambda_{c,0} = 0{\cdot}5$, 1·0 or 1·5, and $\gamma_0 = 0$ and 0·5 for the cases of independent and dependent monitoring times, respectively. Sample sizes of 100 and 200 were considered. There were 10 000 simulation samples for each combination of the simulation parameters.

Table 1 displays the Monte Carlo estimates for the sampling means and standard errors of the estimators $\hat{\beta}$ and $\tilde{\beta}$, the sampling means of their standard error estimators $\mathscr{I}^{-\frac{1}{2}}(\hat{\beta})$ and $n^{-\frac{1}{2}} \hat{V}_{\beta}^{\frac{1}{2}}(\tilde{\beta}; \hat{\gamma})$, and the empirical coverage probabilities of the 95% Wald confidence intervals. These results suggest that the biases of the parameter estimators and their standard error estimators are negligible, at least for $n > 100$. In addition, the 95% confidence intervals have proper coverage probabilities. As expected, $\tilde{\beta}$ is more efficient than $\hat{\beta}$ in the case of independent monitoring time.

Table 1. *Summary statistics for the simulation studies*

(*a*) *Independent monitoring time*

	$n = 100$			$n = 200$		
	$\lambda_{c,0} = 0{\cdot}5$	1·0	1·5	$\lambda_{c,0} = 0{\cdot}5$	1·0	1·5
Mean of $\hat{\beta}$	0·56	0·54	0·53	0·53	0·52	0·51
Stand. error of $\hat{\beta}$	0·45	0·40	0·40	0·29	0·26	0·27
Mean of $\mathscr{I}^{-1/2}(\hat{\beta})$	0·42	0·38	0·39	0·28	0·26	0·27
Cov. prob. of 95% CI	0·96	0·96	0·95	0·96	0·95	0·95
Mean of $\tilde{\beta}$	0·56	0·54	0·53	0·53	0·52	0·51
Stand. error of $\tilde{\beta}$	0·42	0·36	0·35	0·27	0·24	0·23
Mean of $n^{-1/2} \hat{V}_{\beta}^{1/2}(\tilde{\beta}; \hat{\gamma})$	0·40	0·35	0·34	0·26	0·23	0·23
Cov. prob. of 95% CI	0·96	0·96	0·95	0·95	0·95	0·95

(*b*) *Dependent monitoring time*

	$n = 100$			$n = 200$		
	$\lambda_{c,0} = 0{\cdot}5$	1·0	1·5	$\lambda_{c,0} = 0{\cdot}5$	1·0	1·5
Mean of $\tilde{\beta}$	0·54	0·53	0·52	0·52	0·52	0·51
Stand. error of $\tilde{\beta}$	0·38	0·42	0·50	0·25	0·29	0·33
Mean of $n^{-1/2} \hat{V}_{\beta}^{1/2}(\tilde{\beta}; \hat{\gamma})$	0·38	0·41	0·49	0·25	0·28	0·33
Cov. prob. of 95% CI	0·95	0·96	0·95	0·95	0·95	0·95

4·2. *A real example*

We now provide an illustration with real data taken from a carcinogenicity experiment. The study was originally reported by Hoel & Walburg (1972); the data were reproduced by Finkelstein & Wolfe (1985). The purpose of the study was to compare the time until the onset of lung tumour for untreated mice in a germ-free versus conventional environment. A total of 144 RFM mice were involved in the study. At the times of death/sacrifice, lung tumours were found in 27 out of the 96 mice assigned to the conventional environment, as compared with 35 out of the 48 mice assigned to the germ-free environment. Since lung tumours are nonlethal in RMF mice, it is reasonable to regard the tumour

onset time T to be independent of the death time C. However, the survival experiences may be different between the conventional and germ-free environments. Therefore, the methods of § 3 seem to be the most appropriate.

Let $Z=0$ for conventional mice and $Z=1$ for germ-free mice. The maximum partial likelihood estimate $\hat{\gamma}$ for the death time is $-1{\cdot}96$, with estimated standard error of 0·24, which indicates that germ-free mice live much longer than conventional mice. The risk difference estimate $\tilde{\beta}$ for the onset of lung tumour is 0·00071, with estimated standard error of 0·00041. The normalised test statistic is 1·73, resulting in a two-sided P-value of 0·084. These results suggest that lung tumours tend to occur earlier in the germ-free environment than in the conventional environment.

The above data were previously analysed by Finkelstein (1986) and Huang (1996) under the proportional hazards model. Finkelstein reported a standard-normal statistic of 1·64 for testing no group difference in tumour incidence, with an associated P-value of 0·10. Huang obtained a log relative risk estimate of 0·55 with an estimated standard error of 0·29, corresponding to a P-value of 0·054. Incidentally, our estimate $\tilde{\beta}$ is not directly comparable to Huang's estimate since the former pertains to $\{\lambda(t\,|\,Z=1)-\lambda(t\,|\,Z=0)\}$ and the latter to $\log\{\lambda(t\,|\,Z=1)/\lambda(t\,|\,Z=0)\}$.

5. Remarks

Given β_0, the values of $Z(t)$ are constrained so that the right-hand side of (1·1) is nonnegative. This constraint may sometimes be achieved by replacing $Z(t)$ in (1·1) by $Z(t)q(t)$, where q is a prespecified function. One may also remove the constraint by using a nonnegative link function, such as $\lambda_0(t)+e^{\beta_0' Z(t)}$. The basic results presented in this paper hold for any link functions, although the linear form of (1·1) gives rise to the simplest formulae.

As with the case of right-censored data, it is essential to assume the conditional independence of T and C given Z, without which the problem is non-identifiable. The proportional hazards modelling of C is somewhat restrictive. However, model (3·1) is fully testable since the C_i are always observed. In many applications, especially in controlled experiments, the structures for the monitoring time imposed in §§ 2 and 3 are appropriate. In such cases, the simplicity of the inference procedures seems to justify their restrictive assumptions.

In the case of independent monitoring, that is $\gamma_0=0$, any functional of the covariate history may be used in model (3·1), and the resulting two-stage estimator is more efficient than $\hat{\beta}$. It can be shown by the Cauchy–Schwarz inequality that the most efficient functional is $Z^*(t)e^{-\beta_0' Z^*(t)-\Lambda_0(t)}$. One may replace β_0 and $\Lambda_0(t)$ involved in this functional by any preliminary estimators or guesses. Note that $\Lambda_0(t)$ can be estimated by solving the integral equation $\hat{H}_0(t)=\int_0^t e^{-\Lambda_0(u)}\,d\hat{\Lambda}_{c,0}(u)$, where $\hat{H}_0$ and $\hat{\Lambda}_{c,0}$ are the Aalen–Breslow type estimators of H_0 and $\Lambda_{c,0}$ given at the end of the Appendix. The estimator $\tilde{\beta}$ corresponding to the most efficient functional does not achieve the semiparametric information bound. One may derive semiparametric efficient estimators of β_0 for both independent and dependent monitoring along the lines of Huang (1996), but the resulting inference procedures will be much more complicated than those of §§ 2 and 3.

In some applications including cancer screening trials, there is more than one monitoring time for each study subject, which gives rise to general interval censored data. Suppose that there are K monitoring times. For each of the K monitoring times, we may construct a modification of $U(\beta)$ or $U_\beta(\beta;\hat{\gamma})$. Then a linear combination of these K functions will

yield a single estimating function for β_0. Inference procedures similar to those of §§ 2 and 3 can be obtained. These new theoretical developments, along with their applications, will be reported elsewhere.

ACKNOWLEDGEMENT

We are grateful to the Editor and a referee for useful comments and to Professor Cun-Hui Zhang for helpful discussions. This research was supported by the U.S. National Institutes of Health, National Science Foundation and National Security Agency.

APPENDIX

Some technical results

It is helpful to introduce some notation in connection with the estimation of model (3·1). Let $N_i^c(t) = I(C_i \leqslant t)$ and

$$M_i^c(t) = N_i^c(t) - \int_0^t Y_i(s) e^{\gamma_0' Z_i(s)}\, d\Lambda_{c,0}(s) \quad (i = 1, \ldots, n).$$

It is well known that $M_i^c(t)$ $(i = 1, \ldots, n)$ are martingales with respect to the σ-filtration

$$\sigma\{N_i^c(s), Z_i(s) : s \leqslant t, i = 1, \ldots, n\}.$$

Furthermore,

$$n^{\frac{1}{2}}(\hat{\gamma} - \gamma_0) = D_\gamma^{-1} n^{-\frac{1}{2}} \sum_{i=1}^n \int_0^\infty \{Z_i(t) - e(\gamma_0, t)\}\, dM_i^c(t) + o_p(1), \tag{A·1}$$

where $e(\gamma, t) = E\{Y_1(t) e^{\gamma' Z_1(t)} Z_1(t)\} / E\{Y_1(t) e^{\gamma' Z_1(t)}\}$ (Andersen & Gill, 1982).

We now prove that $M_i(t)$ defined in (3·3) are martingales with respect to the σ-filtration $\mathscr{F}_t$. It suffices to show that, for any $\Delta t > 0$, $E\{M_i(t + \Delta t) | \mathscr{F}_t\} = M_i(t)$ or

$$E\{N_i(t + \Delta t) - N_i(t) | \mathscr{F}_t\} = \int_{(t, t+\Delta t]} \mathrm{pr}(C_i \geqslant s | \mathscr{F}_t) e^{-\beta_0' Z_i^*(s) + \gamma_0' Z_i(s)}\, dH_0(s). \tag{A·2}$$

Clearly, both sides of (A·2) are zero if $C_i \leqslant t$. Thus, we only need to establish the equality on the set $\{C_i > t\}$. However,

$$\begin{aligned} E\{N_i(t + \Delta t) - N_i(t) | \mathscr{F}_t, C_i > t\} &= \mathrm{pr}(t < C_i \leqslant t + \Delta t, T_i \geqslant C_i | C_i > t, Z_i) \\ &= \frac{\int_{(t, t+\Delta t]} e^{-\Lambda_0(s) - \beta_0' Z_i^*(s)} \mathrm{pr}(C_i \geqslant s | Z_i) e^{\gamma_0' Z_i(s)}\, d\Lambda_{c,0}(s)}{\mathrm{pr}(C_i > t | Z_i)} \\ &= \int_{(t, t+\Delta t]} \mathrm{pr}(C_i \geqslant s | \mathscr{F}_t, C_i > t) e^{-\beta_0' Z_i^*(s) + \gamma_0' Z_i(s)}\, dH_0(s). \end{aligned}$$

Hence, (A·2) holds.

Simple algebraic manipulation shows that

$$U_\beta(\beta_0; \gamma_0) = \sum_{i=1}^n \int_0^\infty \left\{ Z_i^*(t) - \frac{S^{(1)}(\beta_0, \gamma_0, t)}{S^{(0)}(\beta_0, \gamma_0, t)} \right\} dM_i(t),$$

which is a martingale integral. A standard calculation of predictable variation yields

$$n^{-\frac{1}{2}} U_\beta(\beta_0; \gamma_0) = n^{-\frac{1}{2}} \sum_{i=1}^n \int_0^\infty \{Z_i^*(t) - e(\beta_0, \gamma_0, t)\}\, dM_i(t) + o_p(1), \tag{A·3}$$

where $e(\beta, \gamma, t) = E\{S^{(1)}(\beta, \gamma, t)\} / E\{S^{(0)}(\beta, \gamma, t)\}$.

By Taylor series expansion,

$$n^{-\frac{1}{2}}U_\beta(\beta_0;\hat{\gamma}) = n^{-\frac{1}{2}}U_\beta(\beta_0;\gamma_0) - \hat{\Omega}_{\beta\gamma}(\beta_0;\gamma^*)n^{\frac{1}{2}}(\hat{\gamma}-\gamma_0),$$

where γ^* is on the line segment between $\hat{\gamma}$ and γ_0. It then follows from (A·1), (A·3), the law of large numbers and the consistency of $\hat{\gamma}$ that

$$n^{-\frac{1}{2}}U_\beta(\beta_0;\hat{\gamma}) = n^{-\frac{1}{2}}\sum_{i=1}^{n}\int_0^\infty \{Z_i^*(t)-e(\beta_0,\gamma_0,t)\}\,dM_i(t)$$
$$-\Omega_{\beta\gamma}D_\gamma^{-1}n^{-\frac{1}{2}}\sum_{i=1}^{n}\int_0^\infty \{Z_i(t)-e(\gamma_0,t)\}\,dM_i^c(t)+o_p(1). \tag{A·4}$$

Since the right-hand side of (A·4) is a sum of n independent and identically distributed zero-mean random vectors, a direct application of the Lindeberg central limit theorem shows that $n^{-\frac{1}{2}}U_\beta(\beta_0;\hat{\gamma})$ is asymptotically normal. Standard calculations of predictable variation for martingale integrals yield Ω_β and $\Omega_{\beta\gamma}D_\gamma^{-1}\Omega'_{\beta\gamma}$ as the limiting variances for the first and second terms on the right-hand side of (A·4), respectively. Note now that

$$dM_i(t) = I(T_i \geqslant t)\,dM_i^c(t) + \{I(T_i\geqslant t)-\mathrm{pr}(T_i\geqslant t\,|\,Z_i)\}Y_i(t)e^{\gamma_0'Z_i(t)}\,d\Lambda_{c,0}(t).$$

Since $\int\{Z_i(t)-e(\gamma_0,t)\}\,dM_i^c(t)$ does not involve T_i, its covariance with $\{I(T_i\geqslant t)-\mathrm{pr}(T_i\geqslant t\,|\,Z_i)\}$ must be zero. Thus,

$$\mathrm{cov}\left[\int_0^\infty \{Z_i^*(t)-e(\beta_0,\gamma_0,t)\}\,dM_i(t),\ \int_0^\infty \{Z_i(t)-e(\gamma_0,t)\}'\,dM_i^c(t)\right]$$
$$=\mathrm{cov}\left[\int_0^\infty I(T_i\geqslant t)\{Z_i^*(t)-e(\beta_0,\gamma_0,t)\}\,dM_i^c(t),\ \int_0^\infty \{Z_i(t)-e(\gamma_0,t)\}'\,dM_i^c(t)\right]$$
$$=E\left[\int_0^\infty \{Z_i^*(t)-e(\beta_0,\gamma_0,t)\}\{Z_i(t)-e(\gamma_0,t)\}'I(T_i\geqslant t)\,dN_i^c(t)\right]$$
$$=E\left[\int_0^\infty \{Z_i^*(t)-e(\beta_0,\gamma_0,t)\}\{Z_i(t)-e(\gamma_0,t)\}'\,dN_i(t)\right],$$

which is exactly $\Omega_{\beta\gamma}$. Thus, the covariance between the two terms on the right-hand side of (A·4) equals the variance of the second term. Consequently, the limiting covariance matrix for $n^{-\frac{1}{2}}U_\beta(\beta_0;\hat{\gamma})$ is $\Omega_\beta-\Omega_{\beta\gamma}D_\gamma^{-1}\Omega'_{\beta\gamma}$.

The proof for the consistency of $\tilde{\beta}$ is similar to that of the maximum partial likelihood estimator since the slope matrix $\hat{\Omega}_\beta(\beta;\hat{\gamma})$ is positive semidefinite and its limit is assumed to be positive definite at β_0. Given the consistency, the desired asymptotic normality for $\tilde{\beta}$ follows from the Taylor series expansion of $U_\beta(\tilde{\beta};\hat{\gamma})$ at $U_\beta(\beta_0;\hat{\gamma})$, together with the asymptotic normality of $n^{-\frac{1}{2}}U_\beta(\beta_0;\hat{\gamma})$. The consistency of the limiting covariance matrix estimators follows from the arguments for establishing the consistency of the covariance matrix estimator for the maximum partial likelihood estimator (Andersen & Gill, 1982).

The variance–covariance formulae given in §§ 2 and 3, as well as their derivations given above, implicitly assumed that Λ_c is continuous. In the presence of discontinuities, a correction is needed for the predictable variation (Fleming & Harrington, 1991, Theorem 2.6.2). Consequently, the limiting covariance matrix for $n^{-\frac{1}{2}}U_\beta(\beta_0;\hat{\gamma})$ is $\Psi := \Omega_\beta^*-\Omega_{\beta\gamma}D_\gamma^{-1}D_\gamma^*D_\gamma^{-1}\Omega'_{\beta\gamma}$, and that of $n^{\frac{1}{2}}(\tilde{\beta}-\beta_0)$ is $\Omega_\beta^{-1}\Psi\Omega_\beta^{-1}$, where

$$\Omega_\beta^* = E\left[\int_0^\infty \{Z_1^*(t)-e(\beta_0,\gamma_0,t)\}^{\otimes 2}\{1-e^{-\beta_0'Z_1^*(t)+\gamma_0'Z_1(t)}\,\Delta H_0(t)\}Y_1(t)e^{-\beta_0'Z_1^*(t)+\gamma_0'Z_1(t)}\,dH_0(t)\right],$$

$$D_\gamma^* = E\left[\int_0^\infty \{Z_1(t)-e(\gamma_0,t)\}^{\otimes 2}\{1-e^{\gamma_0'Z_1(t)}\,\Delta\Lambda_{c,0}(t)\}Y_1(t)e^{\gamma_0'Z_1(t)}\,d\Lambda_{c,0}(t)\right].$$

Consistent estimators can again be obtained by inserting proper sample estimators for the unknown quantities in the covariance matrix expressions. Note that $H_0(t)$ and $\Lambda_{c,0}(t)$ can be estimated by the Aalen–Breslow type estimators

$$\hat{H}_0(t) = \sum_{i=1}^{n} \int_0^t \frac{dN_i(s)}{\sum_{j=1}^{n} Y_j(s) e^{-\tilde{\beta}' Z_j^*(s) + \hat{\gamma}' Z_j(s)}}, \quad \hat{\Lambda}_{c,0}(t) = \sum_{i=1}^{n} \int_0^t \frac{dN_i^c(s)}{\sum_{j=1}^{n} Y_j(s) e^{\hat{\gamma}' Z_j(s)}}.$$

REFERENCES

ANDERSEN, P. K. & GILL, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10**, 1100–20.

ARANDA-ORDAZ, F. J. (1983). An extension of the proportional hazards model for grouped data. *Biometrics* **39**, 109–17.

AYER, M., BRUNK, H. D., EWING, G. M., REID, W. T. & SILVERMAN, E. (1955). An empirical distribution function for sampling with incomplete information. *Ann. Math. Statist.* **26**, 641–7.

BRESLOW, N. E. & DAY, N. E. (1987). *Statistical Methods in Cancer Research*, **2**, *The Design and Analysis of Cohort Studies.* Lyon: IARC.

BUCKLEY, J. D. (1984). Additive and multiplicative models for relative survival rates. *Biometrics* **40**, 51–62.

COX, D. R. (1972). Regression models and life-tables (with Discussion). *J. R. Statist. Soc.* B **34**, 187–220.

COX, D. R. & OAKES, D. (1984). *Analysis of Survival Data.* London: Chapman & Hall.

DIAMOND, I. D. & MCDONALD, J. W. (1991). Analysis of current-state data. In *Demographic Applications of Event History Analysis*, Ed. J. Trussell, R. Hankinson and J. Tilton, pp. 231–52. Oxford: Oxford University Press.

FINKELSTEIN, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* **42**, 845–54.

FINKELSTEIN, D. M. & WOLFE, R. A. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics* **41**, 993–1045.

FLEMING, T. R. & HARRINGTON, D. P. (1991). *Counting Processes and Survival Analysis.* New York: Wiley.

GROENEBOOM, P. & WELLNER, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation.* Basel: Birkhäuser Verlag.

HOEL, D. G. & WALBURG, H. E. (1972). Statistical analysis of survival experiments. *J. Nat. Cancer Inst.* **49**, 361–72.

HUANG, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Ann. Statist.* **24**, 540–68.

KALBFLEISCH, J. D. & PRENTICE, R. L. (1980). *The Statistical Analysis of Failure Time Data.* New York: Wiley.

KLEIN, R. W. & SPADY, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica* **61**, 387–421.

LIN, D. Y. & YING, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika* **81**, 61–71.

PETO, R. (1973). Experimental survival curves for interval-censored data. *Appl. Statist.* **22**, 86–91.

RABINOWITZ, D., TSIATIS, A. & ARAGON, J. (1995). Regression with interval censored data. *Biometrika* **82**, 501–13.

SHIBOSKI, S. C. & JEWELL, N. P. (1992). Statistical analysis of the time dependence of HIV infectivity based on partner study data. *J. Am. Statist. Assoc.* **87**, 360–72.

THOMAS, D. C. (1986). Use of auxiliary information in fitting nonproportional hazards models. In *Modern Statistical Methods in Chronic Disease Epidemiology*, Ed. S. H. Moolgavkar and R. L. Prentice, pp. 197–210. New York: Wiley.

TURNBULL, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. R. Statist. Soc.* B **38**, 290–5.

[*Received September* 1996. *Revised September* 1997]