Biometrika (2009), **96**, 3, *pp.* 601–615
© 2009 Biometrika Trust
Printed in Great Britain

doi: 10.1093/biomet/asp026
Advance Access publication 24 June 2009

Pseudo-partial likelihood for proportional hazards models with biased-sampling data

By WEI YANN TSAI

Department of Biostatistics, Columbia University, New York, New York 10032, U.S.A.
wt5@columbia.edu

Summary

We obtain a pseudo-partial likelihood for proportional hazards models with biased-sampling data by embedding the biased-sampling data into left-truncated data. The log pseudo-partial likelihood of the biased-sampling data is the expectation of the log partial likelihood of the left-truncated data conditioned on the observed data. Asymptotic properties of the estimator that maximize the pseudo-partial likelihood are derived. Applications to length-biased data, biased samples with right censoring and proportional hazards models with missing covariates are discussed.

Some key words: EM algorithm; Left-truncation; Length-biased data; Missing covariate; Right censoring.

1. Introduction

The partial likelihood function of Cox (1975) has been mainly used for proportional hazards models with censored data (Cox, 1972). For more complicated incomplete data, no unified method exists to find a partial likelihood for inference on the parameters of the proportional hazards models. Dempster et al. (1977) developed the EM algorithm to obtain maximum likelihood estimators for incomplete data. Originally used for fully parametric models, the EM algorithm was subsequently extended successfully to many nonparametric problems. In survival analysis, there is substantial literature generalizing the EM algorithm to frailty models (Andersen et al., 1993, § 9), missing covariates (Paik & Tsai, 1997; Qi et al., 2005) and interval-censored data (Betensky et al., 1999).

In semiparametric models, one usually obtains, through a conditioning argument, an objective function with finitely many parameters of interest to which the EM algorithm can be readily applied. The present paper gives an analogous pseudo-partial likelihood for proportional hazards models with biased-sampling data that can be used without intensive computation.

Under the proportional hazards models for biased-sampling data, the conditional probability density function of an observed nonnegative random variable T, given covariates $z(t)$ and x, can be expressed as

$$h(t \mid x, z) = W(t, x) f\{t \mid z(\cdot)\}/\alpha(x, z), \tag{1}$$

where $W(t, x)$ is a completely known nonnegative weight function, $z(t) = \{z_1(t), \ldots, z_p(t)\}^{\mathrm{T}}$ is a p-dimensional time-dependent covariate, $x = (x_1, \ldots, x_q)^{\mathrm{T}}$ is a q-dimensional time-independent covariate, $f(t \mid z)$ denotes a population conditional density function given $z(s)$ for $s \leqslant t$ and $\alpha(x, z)$ is a normalization constant making $h(\cdot \mid x, z)$ a genuine probability density function.

Furthermore, we will assume a proportional hazards model with $f(t \mid z)/S(t \mid z) = \lambda(t \mid z) = \lambda_0(t)e^{\beta^{\mathrm{T}} z(t)}$, where $S(t \mid z) = \int_t^\infty f(s \mid z)ds$ is the conditional survival function.

Biased-sampling data arise naturally in complex surveys. For example, in large-scale population-based surveys with multi-stage sampling, the complex design results in a set of probability weights for each subject. The weight function $W(T_i, x_i)$ represents the probability that the ith observation (T_i, x_i, z_i) was sampled from the population. Binder (1992) and Lin (2000) have proposed and studied a method for estimating the parameters of proportional hazards models from such survey data. For $W(t, x) = t$, the data are referred to as length-biased data. Wang (1996) proposed statistical inference for length-biased data based on Cox's model. For $W(t, x) = I(x \leqslant t)$, density (1) becomes a conditional probability density for left-truncated data. This problem has been extensively studied in the literature. Wang et al. (1986) used a classical approach to study the properties of the nonparametric maximum likelihood estimator. Keiding & Gill (1990) used counting process techniques to study the properties of the same estimator.

The following four real datasets illustrate different types of biased-sampling data.

Example 1. *Shrub data.* Muttlak & McDonald (1990) presented widths of 46 shrubs. Wang (1996) assumed that the probability of observing a shrub is proportional to the shrub's width, so that the sampling is length-biased. Wang (1996) analyzed the data with a proportional hazards model.

Example 2. *Channing House data.* Channing House is a retirement centre in Palo Alto, California. Hyde (1980) reported ages at entry and at death of 462 retirees, 365 females and 97 males, who were in residence between January 1964 and July 1975. The individuals who left Channing House or were still in the centre at the end of the study were censored. The data can be viewed as left-truncated with right censoring since the individual's death age must be greater than the entry age. The entry age serves as the left-truncation time.

Example 3. *Stanford heart transplant data.* Crowley & Hu (1977) gave information on 103 potential heart transplant recipients who were enrolled in the Stanford heart transplant programme from October 1967 to April 1974. The data include age, waiting time to transplantation, survival or censoring time from acceptance to the programme, and three mismatch scores. Among the 103 potential heart transplant recipients, there were 69 patients who underwent the heart transplant operation. Later, Miller & Halpern (1982) updated the data by reporting the survival or censoring times and ages of 184 patients who were enrolled in the same programme and had received heart transplants from October 1967 to February 1980. If we are only interested in analyzing the transplant patients, then the data of Crowley & Hu are left-truncated and right-censored data, with transplant waiting time as a random left-truncation variable. However, because Miller & Halpern did not report the transplant waiting times, their data can be viewed as biased-sampling data with right censoring. The weight function is the distribution of the transplant waiting time random variable.

Example 4. *Mouse leukaemia data.* Kalbfleisch & Prentice (2002) reported the survival of 204 mice. The mice were followed up for two years for mortality due to thymic or nonthymic leukaemia. The two covariates of interest were the GPD1 phenotype and the level of endogenous murine leukaemia virus. There were 175 mice whose levels of endogenous murine leukaemia virus were recorded. The GPD1 phenotype was determined only on a subgroup of the 100 mice that survived 400 days; thus, the probability of missing covariates clearly depends on the follow-up time. The complete-case analysis, which uses only the mice with complete information, is clearly biased since the selection probability depends on the survival time outcome variable. In fact, the complete cases comprise biased-sampling data with the weight function equal to the probability of selecting complete cases.

2. Partial likelihood

2·1. *The general approach*

Let χ be a sample space, and let $x \in \chi$ be a realization of the random vector X with density $f_X(x; \phi)$ depending on a vector parameter $\phi = (\beta, \eta)$, in which β is of interest and η is a nuisance parameter. In some applications, the dimension of η may increase with the sample size and the application of maximum likelihood estimation may lead to spurious results. However, suppose that $x = (c_1, x_1, \ldots, c_n, x_n)$ and that the full likelihood factorizes into

$$f_X(x; \phi) = \prod_{i=1}^{n} f_\phi(c_i \mid d_i) \prod_{i=1}^{n} f_\beta(x_i \mid e_i), \tag{2}$$

where $d_i = (c_1, x_1, \ldots, c_{i-1}, x_{i-1})$ and $e_i = (c_1, x_1, \ldots, c_{i-1}, x_{i-1}, c_i)$. The second product on the right-hand side of (2) is the partial likelihood of β based on $(x_1, \ldots, x_n)$ in the sequence $(c_i, x_i)(i = 1, \ldots, n)$. Cox (1975) argued that inference based only on the partial likelihood would be acceptable if the information about β, contained in the first factor, was small.

A complication that sometimes occurs is that one observes a function $Y(x) = y \in \mathcal{Y}$, instead of observing $x \in \chi$. Therefore, inferences about β must be based on y. The following algorithm is a simple generalization of the EM algorithm for partial likelihood. The EM algorithm applied to gamma frailty models discussed in Andersen et al. (1993, § 9) is a special case.

For incomplete data, the EM algorithm finds maximum likelihood estimates for β and η through the iterative maximization of

$$\sum_{i=1}^{n} E\{\log f_\phi(c_i \mid d_i) \mid \beta^{(c)}, \eta^{(c)}, y\} + \sum_{i=1}^{n} E\{\log f_\beta(x_i \mid e_i) \mid \beta^{(c)}, \eta^{(c)}, y\},$$

where $\beta^{(c)}$ and $\eta^{(c)}$ denote the current estimates. Therefore,

$$l_{\mathrm{p}}(\beta, \eta) = \sum_{i=1}^{n} E\{\log f_\beta(x_i \mid e_i) \mid \beta, \eta, y\}, \quad U_{\mathrm{p}}(\beta, \eta) = \sum_{i=1}^{n} E\{U_i(\beta) \mid \beta, \eta, y\}$$

are, respectively, a log pseudo-partial likelihood function and a pseudo-partial score function of β for the observed data y, where $U_i(\beta) = \partial \log f_\beta(x_i \mid e_i)/\partial\beta$ is the score function of the partial likelihood for complete data. Unfortunately, U_{p} still involves the nuisance parameter η in many applications. For example, in frailty models U_{p} is also a function of the frailty parameters and, therefore, cannot be used directly. However, U_{p} is a function of β alone in some applications. In other situations, the maximization of $l_{\mathrm{p}}(\beta, \eta, y)$ with respect to β and η has a simple solution. In § 3, we show two pseudo-partial likelihood functions for Cox's models with biased-sampling data in these two situations.

2·2. *Partial likelihood for left-truncated and right-censored data*

Let the lifetime T_i^0 have distribution function F_i, and let the truncation time and censoring time (V_i, C_i) have joint distribution function G_i and joint probability density function g_i. It is assumed that T_i^0 and (V_i, C_i) are mutually independent. Moreover, we assume that there is a positive probability that $T_i^0 \geqslant V_i$ and $C_i \geqslant V_i$. We do not sample from the joint distribution, but from the conditional distribution given the event $\{T^0 \geqslant V, C \geqslant V\}$. Let (V_i, T_i^0, C_i) $(i = 1, \ldots, n)$ be a sample of n independent triples from this conditional distribution. Then, our left-truncated and right-censored sample is $(V_1, T_1, D_1), \ldots, (V_n, T_n, D_n)$, where $T_i = \min(T_i^0, C_i)$, $D_i = I(T_i = T_i^0)$ and $I(\cdot)$ is an indicator function. Note that $T_i \geqslant V_i$ $(i = 1, \ldots, n)$. We let $N_i(t) = I(T_i \leqslant t, D_i = 1)$ and $Y_i^v(t) = I(V_i \leqslant t)Y_i(t)$ be, respectively, the indicator of whether or not the ith individual failed before time t and the indicator of whether or not the ith individual is at risk just before time t, where $Y_i(t) = I(T_i \geqslant t)$. Furthermore, for given time-dependent covariates $z_i(t) = \{z_{1i}(t), z_{2i}(t), \ldots, z_{pi}(t)\}^{\mathrm{T}}$, we assume that T_i^0 follows a proportional hazards model, i.e.

$$\lambda(t \mid z_i) = \lambda_0(t)e^{\beta^{\mathrm{T}} z_i(t)},$$

where $\lambda(t \mid z_i)$ is the conditional hazard function of T_i^0 given $z_i(t)$, β is a $p \times 1$ vector of unknown regression coefficients and $\lambda_0(t)$ is the underlying or baseline hazard function. For convenience of notation, if it is unambiguous, the dependence of z_i on t will be suppressed. As shown by Andersen et al. (1993, §§ 3.3 and 3.4), $(N_1, \ldots, N_n)$ is a multivariate counting process that has an intensity process $\{\lambda(t \mid z_1)Y_1^v(t), \ldots, \lambda(t \mid z_n)Y_n^v(t)\}$ with respect to the filtration $\mathcal{F}(t)$, which is defined in Andersen et al. (1993, p. 153). Let $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ be the cumulative underlying hazard function. The log partial likelihood, which is the conditional likelihood given V, can be written as

$$l_{\mathrm{c}1}(\beta, d\Lambda_0) = \sum_t \sum_{i=1}^n \Delta N_i(t)\big[\log\{d\Lambda_0(t)\} + \beta^{\mathrm{T}} z_i(t)\big] - \int_0^\infty nS^{(0)}(\beta, \mu)d\Lambda_0(\mu), \qquad (3)$$

see equations (3.3.3) and (7.2.2′) of Andersen et al. (1993), where

$$S^{(0)}(\beta, t) = \frac{1}{n}\sum_{i=1}^n e^{\beta^{\mathrm{T}} z_i(t)} Y_i^v(t),$$

and $\Delta H(t) = H(t+) - H(t-)$ for any function $H(t)$. The partial derivative of (3) with respect to $\Delta\Lambda_0(t)$ is $\{\Delta N(t)/\Delta\Lambda_0(t)\} - nS^{(0)}(\beta, t)$, where $N(t) = \sum_{i=1}^n N_i(t)$ is the number of observed failures up to time t. Therefore, for a fixed value of β, we would estimate $\Lambda_0(t)$ by the Nelson–Aalen estimator

$$\hat{\Lambda}_0(t, \beta) = \int_0^t \frac{J^v(\mu)}{nS^{(0)}(\beta, \mu)} dN(\mu), \qquad (4)$$

where $J^v(\mu) = I\{\sum_{i=1}^n Y_i^v(\mu) > 0\}$. Inserting (4) into (3), we obtain the log profile partial likelihood $l_{\mathrm{c}2}(\beta) + \sum_t \Delta N(t)\log\{\Delta N(t)\} - N(\infty)$. Here,

$$l_{\mathrm{c}2}(\beta) = \sum_t \sum_{i=1}^n \Delta N_i(t)[\beta^{\mathrm{T}} z_i(t) - \log\{nS^{(0)}(\beta, t)\}] \qquad (5)$$

is the generalized log partial likelihood, originally derived by Cox (1972, 1975) for the case of censored survival data. Thus, the score function of the generalized Cox partial likelihood is $\sum_{i=1}^n \int_0^\infty \{z_i(t) - S^{(1)}(\beta, t)/S^{(0)}(\beta, t)\}dN_i(t)$, where $S^{(1)}(\beta, t) = \partial S^{(0)}(\beta, t)/\partial\beta$.

We will treat equations $l_{\mathrm{c}1}$ and $l_{\mathrm{c}2}$ as our working log partial likelihood for the complete data. Two log pseudo-partial likelihoods for biased-sampling data will be derived, respectively, based on $l_{\mathrm{c}1}$ and $l_{\mathrm{c}2}$ in the next section. The notation $N_i(t)$, $N(t)$ and $Y_i(t)$ will be used throughout with obvious adjustments for data with no censoring and/or no truncation. A parameter with subscript zero will denote the true value.

3. Data from biased sampling

3·1. *Embedding the data into the left-truncation model*

First we assume that $W(\cdot, x)$ is a distribution function for every fixed x; this assumption will be relaxed later. Let V and T^0 be nonnegative random variables with conditional distribution functions $\text{pr}(V < t \mid x) = W(t, x)$ and $\text{pr}\{T^0 < t \mid z(\cdot)\} = 1 - S(t \mid z)$, respectively. We assume that, conditional on x and z, V and T^0 are independent. We observe (V, x, T^0, z) only if $T^0 \geqslant V$. Therefore, the conditional density of observing (V, T^0) given (x, z) is proportional to $I(t \geqslant v)w(v, x)f(t \mid z)$, where $w(t, x) = \partial W(t, x)/\partial t$. Hence, the marginal density of observed T^0, given z and x, is proportional to $\int I(t \geqslant v)w(v, x)f(t \mid z)dv = W(t, x)f(t \mid z)$, which is proportional to the conditional probability density function given in (1). Consequently, we may treat (V, x, T^0, z) as the complete data vector and (x, T^0, z) as the incomplete data with the truncation time V completely missing.

3·2. *Pseudo-partial likelihoods*

According to the argument in § 2·1 and from the Cox log partial likelihood l_{c2} of equation (5) in § 2·2, the following function, which is the conditional expectation of l_{c2} given the observed data, can be considered as a log pseudo-partial likelihood for the observed data (x_i, T_i^0, z_i) $(i = 1, \ldots, n)$:

$$\tilde{l}(\beta) = \sum_{i=1}^{n} \beta^{\mathrm{T}} z_i - \sum_{i=1}^{n} E\left\{\log \sum_{j=1}^{n} Y_j^v(T_i^0)e^{\beta^{\mathrm{T}} z_j} \,\middle|\, x_1, T_1^0, z_1, \ldots, x_n, T_n^0, z_n\right\}. \tag{6}$$

If we condition on $T^0 = t$, $X = x$, then the random variable V has conditional distribution $W\{\min(v, t), x\}/W(t, x)$. Therefore, the second term of (6) is a function of $(T_1^0, x_1, z_1, \ldots, T_n^0, x_n, z_n)$ and β, which does not involve the nuisance parameter $\lambda_0(t)$. We may, therefore, use the Monte Carlo method to compute it. For example, let V_{jk} $(j = 1, \ldots, n, k = 1, \ldots, m)$ be random samples from the distribution $W\{\min(v, T_j^0), x_j\}/W(T_j^0, x_j)$. Then the second term of (6) can be approximated by

$$\sum_{i=1}^{n} \frac{1}{m} \sum_{k=1}^{m} \log\left\{\sum_{j=1}^{n} I\left(V_{jk} \leqslant T_i^0\right) Y_j\left(T_i^0\right) e^{\beta^{\mathrm{T}} z_j(T_i^0)}\right\}. \tag{7}$$

We substitute (7) into the second term of (6) and obtain an approximate loglikelihood

$$\tilde{l}_{(m)}(\beta) = \sum_{i=1}^{n} \beta^{\mathrm{T}} z_i - \sum_{i=1}^{n} \frac{1}{m} \sum_{k=1}^{m} \log\left\{\sum_{j=1}^{n} I\left(V_{jk} \leqslant T_i\right) Y_j(T_i) e^{\beta^{\mathrm{T}} z_j(T_i)}\right\}. \tag{8}$$

In particular, for length-biased data, i.e. $W(t, x) = t$, the maximum approximate loglikelihood estimator based on (8) is asymptotically equivalent to the improved estimator proposed by Wang (1996). In addition, for $m = 1$, the approximate loglikelihood $\tilde{l}_{(1)}$ is identical to the log-pseudolikelihood described by Wang (1996).

The disadvantages of using l_{c2} as the working log partial likelihood for complete data are as follows: we must assume that $W(t, x)$ is a nondecreasing function in t for every fixed x; the underlying cumulative hazard function must be estimated by other methods; it requires intensive computation to obtain the log pseudo-partial likelihood $\tilde{l}(\beta)$; and the loglikelihood l_{c2} contains less information about the parameters than the loglikelihood l_{c1}. Therefore, we may take l_{c1} as our working log partial likelihood and apply the same procedure to (3). The resulting log pseudo-partial likelihood can be written as

$$\sum_{i=1}^{n} \Delta N(T_i^0)\{\log d\Lambda_0(T_i^0) + \beta^{\mathrm{T}} z_i(T_i^0)\} - \int_0^\infty n\bar{S}^{(0)}(\beta, \mu) d\Lambda_0(\mu), \tag{9}$$

where

$$\bar{S}^{(0)}(\beta, t) = E\left\{\frac{1}{n}\sum_{i=1}^{n} e^{\beta^{\mathrm{T}} z_i(t)} Y_i(t) I(V_i \leqslant t) \mid \text{data}\right\} = \frac{1}{n}\sum_{i=1}^{n} e^{\beta^{\mathrm{T}} z_i(t)} Y_i(t) \frac{W(t, x_i)}{W(T_i, x_i)}.$$

For a fixed value of β, maximization of (9) with respect to $\Delta\Lambda_0(t)$ leads to $\Delta\Lambda_0(t, \beta) = \Delta N(t)/\{n\bar{S}^{(0)}(\beta, t)\}$. Therefore, for a fixed value of β, we would estimate $\Lambda_0(t)$ by the Nelson–Aalen estimator

$$\hat{\Lambda}_0(t, \beta) = \int_0^t \frac{J(\mu)}{n\bar{S}^{(0)}(\beta, \mu)} dN(\mu), \tag{10}$$

where $J(\mu) = I\{\sum_{i=1}^n Y_i(\mu) > 0\}$. Inserting (10) into (8), we obtain the following log profile pseudo-partial likelihood depending only on β:

$$l(\beta) = \sum_t \sum_{i=1}^{n} \Delta N_i(t)[\beta^{\mathrm{T}} z_i(t) - \log\{n\bar{S}^{(0)}(\beta, t)\}], \tag{11}$$

which is also a generalized Cox log partial likelihood. The vector of score statistics is given as

$$U(\beta) = \partial l/\partial\beta = \sum_{i=1}^{n} \int_0^\infty \left\{z_i(t) - \frac{\bar{S}^{(1)}(\beta, t)}{\bar{S}^{(0)}(\beta, t)}\right\} dN_i(t), \tag{12}$$

where $\bar{S}^{(1)}(\beta, t) = \partial\bar{S}^{(0)}(\beta, t)/\partial\beta$. We shall base our estimator of β on (12), and the value of β which maximizes (11) will be denoted by $\hat{\beta}$. As $\mathrm{pr}\{W(T_i^0, x_i) = 0\} = 0$, $W(t, x_i)/W(T_i^0, x_i)$ is well defined.

The log pseudo-partial likelihood (11) is identical to the usual log partial likelihood of the models $\lambda(t \mid z_i, z_i^*) = \lambda_0(t)\exp\{\beta^{\mathrm{T}} z_i + z_i^*(t)\}$, where $z_i^*(t) = \log\{W(t, x_i)/W(T_i^0, x_i)\}$. In particular, when $W(t, x) = W(t)$, $z_i^*(t)$ can be simplified to $-\log\{W(T_i^0)\}$. Standard statistical software, such as SAS, can be used to obtain the estimate $\hat{\beta}$ of β by setting the regression coefficient of z_i^* to be 1, using the option offset$=z^*$ in procedure PHREG. The robust sandwich covariance matrix estimator from SAS is consistent whereas the model-based covariance matrix is inconsistent because the score (12) is not a martingale. For computing estimates of parameters and variances for general weight functions, readers can use the author's R subroutine downloadable from www.columbia.edu/∼wt5/.

3·3. *Censoring*

We may assume that the data are subject not only to biased sampling, but also to right censoring. Let V_i, T_i^0 and C_i be, respectively, truncation time, survival time and censoring time. Recall that when we define the left-truncated and right-censored data, we assume that (V_i, C_i) and T_i^0 are mutually independent. Hence, the joint probability density function of (T_i^0, V_i, C_i) can be expressed as $f_i(t)g_i(v, c)$, where f_i is the probability density function of T_i^0 and g_i is the joint probability density function of (V_i, C_i). We only identify two types of censoring mechanism based on different censoring and truncation mechanisms and assumptions about g_i. However, there are possible applications to other types of censoring.

The first type of censoring assumes that the given covariates (x_i, z_i), original truncated time, survival time and censoring time are mutually independent. However, we observe the data (V_i, T_i, D_i)

only if $V_i \leqslant T_i$, where $T_i = \min(T_i^0, C_i)$ and $D_i = I(T_i^0 \leqslant C_i)$. The biased censored data (T_i, D_i) are obtained by applying the censoring mechanism to the data before the data are sampled with bias. In the embedding, we first apply the censoring mechanism to the survival time and then apply the truncation mechanism to the observed censored data. This type of censoring is equivalent to assuming that

$$g_i(v, c) = w_i(v) g_{2i}(c) I(v \leqslant c) \Big/ \int_0^\infty \int_0^c w_i(v) g_{2i}(c) dv\, dc,$$

where $g_{2i}(t)$ is the probability density function of the censoring time C_i. We may also say that V_i and C_i are quasi-independent in the region $\{(v, c) \mid v \leqslant c\}$. The observed data (V_i, T_i, D_i) comprise a special case of the standard left-truncated and right-censored data as defined in § 2·2. Hence, in the first type of censoring, the conditional probability density function of the observed data (V_i, T_i, D_i), given $(x_i, z_i) = (x, z)$, is

$$h_2(v, t, d \mid x, z) = \frac{w(v, x)\{f(t \mid z)\mathcal{G}_{2i}(t)\}^d \{S(t \mid z) g_{2i}(t)\}^{1-d}}{\int_0^\infty w(v, z)\mathcal{G}_{2i}(v) S(v \mid z) dv},$$

where $\mathcal{G}_{2i}(t) = \int_t^\infty g_{2i}(s) ds$ is the survival function of the censoring time. Therefore, $E\{I(V_i \leqslant t) I(T_i \geqslant t) | T_i, D_i, x_i, z_i\} = W(t, x_i) I(T_i \geqslant t)/W(T_i, x_i)$. The procedures proposed in §§ 3·2 and 3·3 are still valid because $E\{I(V_i \leqslant t) Y_i(t) \mid T_i, D_i, x_i, z_i\}$ still equals $Y_i(t) W(t, x_i)/W(T_i, x_i)$. The formulae of the log pseudo-partial likelihood will be the same for this type of censoring with T_i^0 replaced by T_i and $N_i(t)$ defined by $N_i(t) = I(T_i \leqslant t, D_i = 1)$.

The second type of censoring comprises the censoring of residual lifetime after the data are sampled with bias. Let $R_i^0 = T_i^0 - V_i$ and $R_{ci} = C_i - V_i$ be, respectively, the residual lifetime and the residual censoring time of the ith individual. Given covariates (x_i, z_i) and $C_i \geqslant V_i$, we assume that R_{ci} and (R_i^0, V_i) are independent. We observe (V_i, T_i, D_i) only if $V_i \leqslant T_i$, where $T_i = V_i + R_i$, $R_i = \min(R_i^0, R_{ci})$ and $D_i = I(R_i^0 \leqslant R_{ci})$. The censoring time C_i and truncation time V_i are not independent in this type of censoring. Let $g_{3i}(t)$ be the probability density function of the residual censoring time R_{ci}. The second type of censoring is equivalent to assuming that

$$g_i(v, c) = w_i(v) g_{3i}(c - v) I(c \geqslant v),$$

where $g_{3i}(t)$ is the probability density function of the residual censoring time R_{ci}. The conditional density function of the observed data (V_i, T_i, D_i) given $(x_i, z_i) = (x, z)$ is

$$h_3(v, t, d \mid x, z) = \frac{w(v, x)\{f(t \mid z)\mathcal{G}_{3i}(t - v)\}^d \{S(t \mid z) g_{3i}(t - v)\}^{1-d}}{\int_0^\infty w(v, x) S(v \mid z) dv},$$

where $\mathcal{G}_{3i}(t) = \int_t^\infty g_{3i}(s) ds$ is the survival function of the residual censoring time. Hence,

$$E\{I(V_i \leqslant t) I(T_i \geqslant t) \mid T_i, D_i, x_i, z_i\} = \frac{\int_0^t h_3(v, T_i, D_i \mid x_i, z_i) dv}{\int_0^{T_i} h_3(v, T_i, D_i \mid x_i, z_i) dv}$$

$$= \left\{\frac{\int_0^t w(v, x_i) g_{3i}(T_i - v) dv}{\int_0^{T_i} w(v, x_i) g_{3i}(T_i - v)}\right\}^{D_i} \left\{\frac{\int_0^t w(v, x_i)\mathcal{G}_{3i}(T_i - v) dv}{\int_0^{T_i} w(v, x_i)\mathcal{G}_{3i}(T_i - v) dv}\right\}^{1-D_i} Y_i(t).$$

Unfortunately, under the second type of censoring, the conditional expectation $E\{I(V_i \leqslant t) Y_i(t) \mid \text{data}\}$ is a function of the censoring distribution. If the truncation time V is observable, then we may use a Kaplan–Meier-type estimator of $\mathcal{G}_{3i}(t)$. Consider the one-sample problem such that $\beta = 0$ and $w(v, x) = w(v)$. The nonparametric maximum likelihood estimator of $\mathcal{G}_3$ is the

Kaplan–Meier estimator based on the censored residual lifetime (R_i, D_i); that is

$$\hat{\mathcal{G}}_3(t) = \prod_{R_i \leqslant t} \left\{ 1 - (1 - D_i) \Big/ \sum_{j=1}^{n} I(R_j \geqslant R_i) \right\}.$$

Hence, the maximum pseudo-partial likelihood estimator of S is

$$\hat{S}_3(t) = \prod_{T_i \leqslant t} \{1 - D_i / \hat{Y}(T_i)\},$$

where

$$\hat{Y}(t) = \sum_{i=1}^{n} \left\{ \frac{\int_0^t w(v) d_v \hat{\mathcal{G}}_3(T_i - v)}{\int_0^{T_i} w(v) d_v \hat{\mathcal{G}}_3(T_i - v)} \right\}^{D_i} \left\{ \frac{\int_0^t w(v) \hat{\mathcal{G}}_3(T_i - v) dv}{\int_0^{T_i} w(v) \hat{\mathcal{G}}_3(T_i - v) dv} \right\}^{1-D_i} Y_i(t).$$

The estimator $\hat{S}_3$ is not the nonparametric maximum conditional likelihood estimator proposed and studied by Tsai et al. (1987), nor is it the nonparametric maximum likelihood estimator. However, if there is no censoring, $\hat{S}_3$ is the nonparametric maximum likelihood estimator. It is straightforward to prove that $\hat{Y}(t)/n$ will converge to $\int_0^t w(v)\mathcal{G}(t - v) dv S(t) / \int_0^\infty w(v) S(v) dv$ and $\sum_{i=0}^n D_i I(T_i \leqslant t)/n$ will converge to $\int_0^t \int_0^s w(v)\mathcal{G}(s - v) dv f(s) ds / \int_0^\infty w(v) S(v) dv$. As a result, under some regularity conditions, $\hat{S}_3(t)$ will converge to the survival function $S(t)$ of the survival time.

If the censoring time depends on the covariates, we need to use a smooth type of Kaplan–Meier estimator of $\mathcal{G}_i$. If the truncation time V_i cannot be observed, we may use the nonparametric maximum likelihood estimator of $\mathcal{G}_i$. More research is needed in order to understand the properties of the proposed method.

The Stanford heart transplant dataset of Miller & Halpern (1982) is a prospective cohort study. The event time of interest is the survival time after entry. The censoring time is the duration between calendar entry date of the patients and February 1980. If we assume that there is no loss of follow-up, then C = February 1980 $-E$; and V = transplant calendar date $-E$ = transplant waiting time, where C, V and E, respectively, are the censoring time, truncation time and calendar entry date of the patient. If V, the transplant waiting time, is independent of E, the calendar entry date, then C and V are independent for all patients in the cohort: the transplant waiting times of the patients who died or were censored before transplantation cannot be observed. Therefore, the censoring time C is quasi-independent of the truncation time V for the transplant patients in the Stanford heart transplant data. Consequently, the censoring is of the first type. The Channing House dataset is a retrospective cohort study. The event time of interest is the death age. The censoring time C for patients who did not leave the centre before July 1975 is the time between the birth date B and July 1975. The truncation time, i.e. entry age, V is $E - B$, where E is the calendar entry date. The residual censoring time R_c is $C - V$ = July 1975 $-E$. If entry age is independent of calendar entry date, i.e. V and E are independent, then the residual censoring time R_c is independent of the truncation time V. This censoring is of the second type.

Most applications can be classified as either the first or second type of censoring. For example, the censoring mechanism of the proportional hazards model with missing covariates, see § 4·3, and the Stanford heart transplant data are of the first type; the censoring mechanisms of the renewal process (Vardi, 1989), cross-sectional survival data (Wang, 1991) and the Channing House data are of the second type.

Table 1. *Monte Carlo simulation for length-biased data. One hundred length-biased observations were generated from Group* 0 *with population density* $f_0(t) = t\exp(-t)I(t > 0)$ *and* 100 *length-biased observations were generated from Group* 1 *with population density* $f_1(t) = t\exp(\beta - e^{\beta}t)I(t > 0)$ *for* $\beta = 0, 1, 2$. *Here* $\hat{\beta}$ *maximizes the loglikelihood* $l(\beta)$, $\tilde{\beta}$ *maximizes the loglikelihood* $\tilde{l}_{(1)}$ *and* $\hat{\beta}_{\text{IPW}}$ *was proposed by Binder* (1992) *and Lin* (2000). *Estimates are based on* 1000 *replications*

	Bias			Sample variance			Mean of estimated variance		
β	$\hat{\beta}$	$\tilde{\beta}$	$\hat{\beta}_{\text{IPW}}$	$\hat{\beta}$	$\tilde{\beta}$	$\hat{\beta}_{\text{IPW}}$	$\hat{\beta}$	$\tilde{\beta}$	$\hat{\beta}_{\text{IPW}}$
0	0·002	0·003	−0·001	0·010	0·020	0·049	0·010	0·020	0·036
1	0·009	0·012	0·014	0·018	0·030	0·070	0·017	0·030	0·050
2	0·032	0·033	0·060	0·058	0·080	0·139	0·053	0·075	0·103

The asymptotic properties of the maximum pseudo-partial likelihood estimators $\hat{\beta}$ and $\hat{\Lambda}$ for censored biased-sampling data of general nonnegative weight functions are established and discussed in the Appendix. We assume that the censoring is of the first type for the rest of the paper.

4. Applications

4·1. *Length-biased data*

The techniques developed in the previous sections can be applied to length-biased data by using $W(t) = t$. We illustrate the method with a simulation study and an analysis of the shrub dataset. Consider a two-sample proportional hazards model with covariate $z = 0$ representing Group 0 and $z = 1$ representing Group 1. We generate 100 length-biased samples from Group 0 with population density $f_0(t) = t\exp(-t)I(t > 0)$ and 100 length-biased samples from Group 1 with population density $f_1(t) = t\exp(\beta - e^{\beta}t)I(t > 0)$ for $\beta = 0, 1, 2$ in three different scenarios. The relative hazard between Group 0 and Group 1 is equal to e^{β} and, thus, the log hazard ratio is $\beta = 0, 1, 2$. We also calculate the estimator $\tilde{\beta}$ obtained by maximizing the approximate likelihood $\tilde{l}_{(1)}$, which was also proposed by Wang (1996). The variance estimator proposed by Wang (1996) is identical to the estimator from SAS. We use equation (A1) in Theorem A2 provided in the Appendix to obtain the variance estimator for the estimator $\hat{\beta}$. For comparison we also include the inverse probability weighted estimator $\hat{\beta}_{\text{IPW}}$ of Binder (1992) and Lin (2000); see also Horvitz & Thompson (1952) and Qi et al. (2005). For the definition of the inverse probability weighted estimator, see equation (13) in § 4·3. Table 1, based on 1000 replicates, which shows that $\hat{\beta}$ is more efficient than $\tilde{\beta}$ and $\hat{\beta}_{\text{IPW}}$ and that the variance estimator of $\hat{\beta}_{\text{IPW}}$ underestimates the true sample variance of $\hat{\beta}_{\text{IPW}}$.

We denote the width of an observed shrub from the shrub dataset by T_i. Wang (1996) assumed that the probability of including T_i in the dataset is proportional to T_i itself. We use the proportional hazards model,

$$\lambda(t \mid z_1, z_2) = \lambda_0(t)\exp(\beta_1 z_1 + \beta_2 z_2),$$

where $z_1 = I(T$ belongs to transect I) and $z_2 = I(T$ belongs to transect II) are two indicator covariates. Use of SAS with offset $-\log(T_i)$ provides $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2) = (0{\cdot}84, 0{\cdot}075)$ with model-based standard error estimates $\{\text{SE}(\hat{\beta}_1), \text{SE}(\hat{\beta}_2)\} = (0{\cdot}49, 0{\cdot}47)$ and $\text{corr}(\hat{\beta}_1, \hat{\beta}_2) = 0{\cdot}74$. The estimates are very similar to the results of Wang (1996), but the model-based standard errors overestimate the true standard errors of $\hat{\beta}$. Use of equation (A1) in the Appendix, which is identical to the robust sandwich covariance estimate from SAS, gives $\{\text{SE}(\hat{\beta}_1), \text{SE}(\hat{\beta}_2)\} = (0{\cdot}32, 0{\cdot}27)$ and $\text{corr}(\hat{\beta}_1, \hat{\beta}_2) = 0{\cdot}56$. For comparison $\hat{\beta}_{\text{IPW}} = (1{\cdot}29, 0{\cdot}33)$ with estimated standard errors (0·33, 0·31).

4·2. *Biased samples with right censoring*

Miller & Halpern (1982) compared four regression techniques on the updated Stanford heart transplant data without acknowledging that the transplant patient's survival time was sampled with bias. As mentioned in § 1, the survival times can be treated as a biased sample with a weight function equal to the distribution of the transplant waiting time. Miller & Halpern (1982) did not provide the transplant waiting times but Crowley & Hu (1977) did. The Weibull distribution fits the transplant waiting times very well. The R^2 of the fit of $\log[-\log\{\hat{S}_w(t)\}]$ to $\log(t)$ is 0·97, where $\hat{S}_w$ is the product-limit estimate of the transplant waiting time survival function based on Crowley & Hu's (1977) 103 patients. The conditional maximum likelihood estimate for the Weibull survival function of transplant waiting time is $\exp(-0{\cdot}027t^{0{\cdot}925})$. Hence, the weight function, $W(t) = 1 - \exp(-0{\cdot}027t^{0{\cdot}925})$, will be used to obtain the parameters of the proportional hazards model. We assume that the hazard rate of the transplant patient's survival is proportional to $\exp\{\beta_1\text{age} + \beta_2(\text{age})^2\}$. Miller & Halpern (1982) deleted 27 patients lacking the T5 mismatch score and 5 patients with survival times less than 10 days from the total of 184 patients in one of their data analyses. Based on 152 Stanford heart transplant patients, the pseudo-partial likelihood estimate $(\hat{\beta}_1, \hat{\beta}_2)$ is $(-0{\cdot}13, 0{\cdot}0021)$ with $\{\text{SE}(\hat{\beta}_1), \text{SE}(\hat{\beta}_2)\} = (0{\cdot}051, 0{\cdot}0006)$ and $\hat{\beta}_{\text{IPW}} = (-0{\cdot}17, 0{\cdot}0026)$ with $\text{SE}(\hat{\beta}_{\text{IPW}}) = (0{\cdot}061, 0{\cdot}0007)$. In calculating the variances of the estimates, we assume that the weight function is known without error. Both methods show a strong relationship between survival time and age.

4·3. *Missing covariates*

It is assumed that, for each i, (T_i, D_i, z_i, R_i) are independent and identically distributed random vectors, where $R_i = 1$ if z_i is fully observed and is zero otherwise. Let $W(t, z) = \text{pr}(R = 1 \mid T = t, Z = z)$ be the conditional probability of observing the full covariates data given the covariates z and follow-up time $T = t$. We assume that $W(t, z)$ is either completely known or can be estimated from other methods; see Qi et al. (2005) and an unpublished Harvard School of Public Health technical report by M. Pugh, J. Robins, S. Lipsitz and D. Harrington. Copas & Farewell (2001), Qi et al. (2005) and Pugh et al.'s report proposed the following weighted complete-case pseudolikelihood score function for inference of the proportional hazards models with missing covariates:

$$U_{\text{IPW}}(\beta) = \sum_{i=1}^{n} W_i^{-1} \int R_i\{z_i(t) - \bar{z}_{\text{IPW}}(\beta, t)\} dN_i(t), \tag{13}$$

where $\bar{z}_{\text{IPW}}(\beta, t) = \sum_j \{R_j z_j Y_j(t) e^{\beta^{\text{T}} z_j} / W_j\} / \sum_j \{R_j Y_j(t) e^{\beta^{\text{T}} z_j} / W_j\}$ and $W_i = W\{T_i, z_i(T_i)\}$. The inverse probability weighted estimator $\hat{\beta}_{\text{IPW}}$ is the solution of $U_{\text{IPW}}(\beta) = 0$. This pseudo-score U_{IPW} was also proposed and studied by Binder (1992) and Lin (2000) in the survey-sampling literature. We may treat the complete case, with $R_i = 1$, as a biased sample from the population with selection probability proportional to $W(t, z)$. The conditional density of (T, D) given the covariates z and $R = 1$ is proportional to $W(t, z) f^D(t \mid z) S^{(1-D)}(t \mid z)$. Since the censoring was applied to the data before the cases with missing covariates were dropped,

Table 2. *Analysis of mouse leukaemia data using the Cox regression model with various methods. Here $\hat{\beta}$ is the pseudo-partial likelihood estimator and $\hat{\beta}_{\text{IPW}}$ is the inverse probability weighted estimator*

	Thymic leukaemia Coefficient estimate (SE)		Thymic and nonthymic leukemia Coefficient estimate (SE)	
Approach	GPD1	Virus	GPD1	Virus
Complete-case	−1·44(0·60)	1·44(0·72)	−1·46(0·57)	1·22(0·65)
$\hat{\beta}$	−1·15(0·64)	1·51(0·71)	−1·19(0·59)	1·28(0·62)
$\hat{\beta}_{\text{IPW}}$	−1·47(0·65)	1·48(0·75)	−1·41(0·63)	1·27(0·67)

SE: estimated standard error.

the censoring is of the first type considered in § 3·3. The pseudo-partial score function becomes $U_{\text{mc}}(\beta) = \sum_{i=1}^{n} U_{\text{mc}i}$, where $U_{\text{mc}i} = \int R_i\{z_i(\mu) - \bar{z}_{\text{p}}(\beta, \mu)\} dN_i(\mu)$, and

$$\bar{z}_{\text{p}}(\beta, t) = \sum_j \{R_j W(t, z_j) z_j Y_j(t) e^{\beta^{\text{T}} z_j} / W_j\} \Big/ \sum_j \{R_j W(t, z_j) Y_j(t) e^{\beta^{\text{T}} z_j} / W_j\}.$$

Let $\hat{\beta}$ denote the solution of $U_{\text{mc}}(\beta) = 0$. If $W(t, z) = W(t)$, then $\bar{z}_{\text{IPW}} = \bar{z}_{\text{p}}$. Hence, $U_{\text{IPW}} = \sum_{i=1}^{n} W_i^{-1} U_{\text{mc}i}$ is a weighted mean of $U_{\text{mc}1}, \ldots, U_{\text{mc}n}$, with weight proportional to $1/W_i$, while U_{mc} is the unweighted mean. When some of the W_i are close to zero, the equation $U_{\text{IPW}} = 0$ becomes unstable. Therefore, in order to prove the asymptotic normality of the estimator $\hat{\beta}_{\text{IPW}}$, Pugh et al.'s report and Qi et al. (2005) had to assume that $W_i > \varepsilon > 0$ for some positive ε. We need weaker assumption to prove the asymptotic properties of the estimator $\hat{\beta}$; see the Appendix.

We now analyze the mouse leukaemia data of Kalbfleisch & Prentice (2002) by the pseudo-partial likelihood method and the inverse probability weighted method. As in Kalbfleisch & Prentice (2002), we dichotomized virus level into a binary variable with zero representing values below 10^4 and one otherwise. There are two analyses, corresponding to the endpoints death by thymic leukaemia and death by thymic or nonthymic leukaemia. Logistic regression was used to model the conditional probability $W(t, z)$ of observing the full covariates data, based on 156 mice that survived for at least 400 days; the observation probabilities were set to zero for the remaining 48 mice that died or were censored before 400 days. In order to make the analysis comparable with that of Wang & Chen (2001), we included only the survival time and its quadratic term as two predictors in our logistic regression model. Both analyses used 204 mice and treated both virus level and GPD1 phenotype as the missing covariates. Table 2 shows the results from applying the pseudo-partial likelihood method and the inverse probability weighted method. The pseudo-partial likelihood method shows that the virus level has a significant relationship with both endpoints, while the GPD1 has a significant relationship with death of thymic or nonthymic leukaemia and GPD1 has a moderately significant relationship with thymic leukaemia death. The inverse probability weighted method shows that GPD1 has a significant relationship with both endpoints and virus level has a moderately significant relationship with both endpoints. The inclusion of death by nonthymic leukaemia changes the estimates for GPD1 phenotype slightly, but moderately reduces the estimates for the virus level. A similar phenomenon was also found by Qi et al. (2005). As in the Stanford heart transplant data analysis, the weight function is treated as known. If the weight function were treated as unknown, then the variance estimator in Theorem A2 would overestimate the true sample variance of $\hat{\beta}$.

If $W(\cdot)$ is a function of D, then $\hat{\beta}_{\text{IPW}}$ is still a consistent estimator. Generally, however, under the same conditions, $\hat{\beta}$ is not consistent, since $E\{U_{\text{mc}}(\beta_0)\} \neq 0$. For generalization to the Cox model with missing covariates and a detailed simulation comparison of $\hat{\beta}$ and $\hat{\beta}_{\text{IPW}}$, see Luo et al. (2009).

5. DISCUSSION

The procedures developed in this paper can be easily extended to other types of incomplete data. The basis of our method is the conditional expectation of the partial score function given the data. For the proportional hazards model, if $N(t)$ is completely known, the conditional expectation involves two terms; see equation (3). One is the conditional expectation of z, while the other is the conditional expectation of $S^{(0)}(\beta, t)$. If z is completely known, we only have to consider the conditional expectation of $S^{(0)}$. In other types of incomplete data, we may also have to compute the conditional expectation of z or $\Delta N(t)$ or both. For example, in the proportional hazards models with missing covariates or with covariates with measurement errors, the covariates z are partially missing. Paik & Tsai (1997) applied a similar idea and proposed two estimators of β for the proportional hazards models with missing covariates.

The estimators $\hat{\beta}$ and $\hat{\Lambda}_0$, proposed in § 3, are not nonparametric maximum likelihood estimators and, therefore, we generally do not expect these estimators to be optimal. However, as a special case, when $W(t, x) = W(t)$ and $\beta_0 = 0$, the estimator $\hat{\Lambda}_0$ is the nonparametric maximum likelihood estimator and is the most efficient estimator. Since $\hat{\beta}$ and $\hat{\Lambda}_0$ maximize the pseudo-partial likelihood, we expect the efficiency of these two estimators to be quite high. Empirical evidence from our limited simulation and real-data experiments suggests that $\hat{\beta}$ is more efficient than $\hat{\beta}_{\text{IPW}}$.

Another special case is given by $W(t, x) = t^x$, $f(t \mid z) = f_0(t)$ and $p = \text{pr}(x = 0) = 1 - \text{pr}(x = 1)$. The nonparametric maximum likelihood estimator $\tilde{F}_0(t)$ of $F_0(t) = \int_0^t f_0(s)ds$ was studied by Vardi (1982). We computed the asymptotic relative efficiency of $\exp\{-\hat{\Lambda}_0(t)\}$ with respect to $1 - \tilde{F}_0(t)$, when F_0 is a uniform distribution on [0,1], for t and $p \in \{0{\cdot}2, 0{\cdot}4, 0{\cdot}6, 0{\cdot}8\}$. When $p = 0$ or 1, the product limit estimator based on $\hat{\Lambda}_0$ is identical to the nonparametric maximum likelihood estimator, so that the asymptotic relative efficiency equals 1 for $p = 0$ and $p = 1$. The lowest asymptotic relative efficiency we obtained was 0·985. Since the estimator $\hat{\Lambda}_0(t)$ is much easier to calculate than the nonparametric maximum likelihood estimator, $\hat{\Lambda}_0(t)$ is a preferred estimator.

ACKNOWLEDGEMENT

The author wishes to thank an associate editor and reviewers for suggestions leading to an overall improvement. He also appreciates Dr. Bruce Levin's comments and inputs. The paper was partially completed during the author's visit to the Department of Statistics, National Cheng Kung University, Taiwan.

APPENDIX

Large sample properties

In deriving equation (12), we explicitly assume that the weight function $W(t, x)$ is a distribution function for any given x. However, after the likelihood (12) is obtained, we only need a weaker assumption, that the weight be nonnegative, to prove the asymptotic properties. Here, we assume that the weight function $W(t, x)$ satisfies Assumption 1.

Assumption 1. For every fixed x, there exists a constant $a(x)$ such that $\{t \mid W(t, x) > 0\} = (a(x), \infty)$ or $[a(x), \infty)$.

Assumption 1 does not require that $W(\cdot, x)$ be a nondecreasing function for any fixed x. The following four theorems describe the asymptotic properties of the estimators $\hat{\beta}$ and $\hat{\Lambda}_0(t, \hat{\beta})$.

Theorem A1. *If Assumption 1 holds and the matrix $I(\beta)$ is positive definite, $\hat{\beta}$ converges to β_0 in probability as $n \to \infty$, where $I(\beta) = \lim_{n\to\infty} \mathcal{I}_n(\beta)$ and $\mathcal{I}_n(\beta) = -n^{-1}\partial U(\beta)/\partial\beta$.*

Proof. If Assumption 1 holds, we have

$$s^{(k)}(\beta_0, t) = E\left[E\left\{I(T \geqslant t)\frac{W(t, x)}{W(T, x)}e^{\beta_0^{\mathrm{T}} z(t)} z^k(t) \mid x, z(s), s \leqslant t\right\}\right]$$
$$= E\left\{W(t, x)e^{\beta_0^{\mathrm{T}} z(t)} z^k(t) S(t \mid z)\mathcal{G}(t \mid z)/\alpha(x, z)\right\} \quad (k = 0, 1),$$

where $\mathcal{G}(t \mid Z) = \mathrm{pr}(C \geqslant t \mid Z)$. Then $n^{-1}\int_0^\infty \bar{S}^{(1)}(\beta_0, t)/\bar{S}^{(0)}(\beta_0, t)dN(t)$ converges in probability to

$$E\left\{D_i s^{(1)}(\beta_0, T_i)/s^{(0)}(\beta_0, T_i)\right\}$$
$$= \int_0^\infty \frac{s^{(1)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} E\left\{W(t, x)\lambda_0(t)e^{\beta_0^{\mathrm{T}} z(t)} S(t \mid z)\mathcal{G}(t \mid z)/\alpha(x, z)\right\}dt$$
$$= \int_0^\infty s^{(1)}(\beta_0, t)\lambda_0(t)dt = \int_0^\infty E\{W(t, x)z(t)f(t \mid z)\mathcal{G}(t \mid z)/\alpha(x, z)\}dt$$
$$= E\{D_i z_i(t)\}.$$

Hence, $U(\beta_0) \to 0$ in probability as $n \to \infty$ and, therefore, Theorem A1 holds by a standard argument.

For asymptotic normality, we need to introduce more notation. Define

$$\xi_i = D_i\left\{z_i(T_i) - \frac{s^{(1)}(\beta_0, T_i)}{s^{(0)}(\beta_0, T_i)}\right\} - \int_0^\infty \frac{e^{\beta_0^{\mathrm{T}} z_i(t)} Y_i(t) W(t, x_i)}{s^{(0)}(\beta_0, t) W(T_i, x_i)}\left\{z_i(t) - \frac{s^{(1)}(\beta_0, t)}{s^{(0)}(\beta_0, t)}\right\} dF_1(t),$$

where $s^{(k)}(\beta, t) = E\{\bar{S}^{(k)}(\beta, t)\}$, $k = 0, 1$, and $F_1(t) = \mathrm{pr}\{T_i < t, I(D_i = 1)\}$. Furthermore, we define

$$\hat{\xi}_i = D_i\left\{z_i(T_i) - \frac{\bar{S}^{(1)}(\hat{\beta}, T_i)}{\bar{S}^{(0)}(\hat{\beta}, T_i)}\right\} - \int_0^\infty \frac{e^{\hat{\beta}^{\mathrm{T}} z_i(t)} Y_i(t) W(t, x_i)}{\bar{S}^{(0)}(\hat{\beta}, t) W(T_i, x_i)}\left\{z_i(t) - \frac{\bar{S}^{(1)}(\hat{\beta}, t)}{\bar{S}^{(0)}(\hat{\beta}, t)}\right\} d\frac{N(t)}{n}.$$

Note that $\hat{\xi}_i$ is obtained by substituting $s^{(0)}$, $s^{(1)}$, $F_1(t)$ and β_0 by $\bar{S}^{(0)}$, $\bar{S}^{(1)}$, $N(t)/n$ and $\hat{\beta}$, respectively, in ξ_i. □

Theorem A2. *Under the same conditions as in Theorem* A1, *$n^{1/2}(\hat{\beta} - \beta_0)$ converges weakly to a normal distribution with zero-mean and covariance matrix $\sum = I^{-1}(\beta_0)\Xi I^{-1}(\beta_0)$, where $\Xi = E(\xi^{\otimes 2})$.*

Proof. The score function $n^{-1/2}U(\beta_0)$ can be expressed as

$$n^{-1/2}\sum_{i=1}^n D_i\left\{z_i(T_i) - \frac{s^{(1)}(\beta_0, T_i)}{s^{(0)}(\beta_0, T_i)}\right\} - n^{1/2}\int_0^\infty \left\{\frac{\bar{S}^{(1)}(\beta_0, t)}{\bar{S}^{(0)}(\beta_0, t)} - \frac{s^{(1)}(\beta_0, t)}{s^{(0)}(\beta_0, t)}\right\} dF_1(t) + o_p(1).$$

By Taylor series expansion, the second term in the above equation can be written as

$$-n^{1/2}\int_0^\infty \frac{\bar{S}^{(1)}(\beta_0, t)s^{(0)}(\beta_0, t) - s^{(1)}(\beta_0, t)\bar{S}^{(0)}(\beta_0, t)}{s^{(0)}(\beta_0, t)s^{(0)}(\beta_0, t)} dF_1(t) + o_p(1)$$
$$= -n^{-1/2}\sum_{i=1}^n \int_0^\infty \frac{e^{\beta_0^{\mathrm{T}} z_i(t)} W(t, x_i)}{s^{(0)}(\beta_0, t) W(T_i, x_i)}\left\{z_i(t) - \frac{s^{(1)}(\beta_0, t)}{s^{(0)}(\beta_0, t)}\right\} dF_1(t) + o_p(1).$$

Hence, $n^{-1/2}U(\beta_0) = n^{-1/2}\sum_{i=1}^n \xi_i + o_p(1)$. By the multivariate central limit theorem and its corollary, $n^{-1/2}U(\beta_0)$ converges to a multivariate normal distribution, yielding Theorem A2. □

Note that $I(\beta_0)$ can be consistently estimated by $\mathcal{I}_n(\hat\beta)$ and Ξ can be consistently estimated by $\hat\Xi = n^{-1}\sum_{i=1}^n \hat\xi_i^{\otimes 2}$. Thus, the covariance matrix $\sum$ can be consistently estimated by

$$\hat V = \mathcal{I}_n^{-1}(\hat\beta)\hat\Xi\mathcal{I}_n^{-1}(\hat\beta). \tag{A1}$$

We now study the large sample properties of the estimated baseline integrated hazard function $\hat\Lambda_0(t, \hat\beta)$.

THEOREM A3. *Let M be a large positive number such that* $\mathrm{pr}(T \geqslant M)$ *is strictly positive. Under the same conditions as in Theorem* A1, *for $t < M$ the process $n^{1/2}\{\hat\Lambda_0(t, \hat\beta) - \Lambda_0(t)\}$ converges weakly to a Gaussian process with zero-mean and covariance function $E\{\xi_{0i}(s)\xi_{0i}(t)\}$, which can be consistently estimated by $n^{-1}\sum_{i=1}^n \hat\xi_{0i}(s)\hat\xi_{0i}(t)$, where*

$$\begin{aligned}
\xi_{0i}(t) &= \xi_i^{\mathrm{T}}\{I(\beta_0)\}^{-1}\int_0^t \frac{s^{(1)}(\beta_0, \mu)}{\{s^{(0)}(\beta_0, \mu)\}^2} dF_1(\mu) + \xi_{1i}(t) + \xi_{2i}(t),\\
\xi_{1i}(t) &= \int_0^t \frac{s^{(0)}(\beta_0, \mu) - e^{\beta_0^{\mathrm{T}} z_i(\mu)} Y_i(\mu) W(\mu, x_i)/W(T_i, x_i)}{\{s^{(0)}(\beta_0, \mu)\}^2} dF_1(\mu),\\
\xi_{2i}(t) &= s^{(0)}(\beta_0, T_i)^{-1} D_i I(T_i \leqslant t) - \Lambda_0(t),\\
\hat\xi_{0i}(t) &= \hat\xi_i^{\mathrm{T}}\mathcal{I}^{-1}(\hat\beta)\int_0^t \frac{\bar S^{(1)}(\hat\beta, \mu)}{\{\bar S^{(0)}(\hat\beta, \mu)\}^2} dN(\mu)/n + \hat\xi_{1i}(t) + \hat\xi_{2i}(t),\\
\hat\xi_{1i}(t) &= \int_0^t \frac{\bar S^{(0)}(\hat\beta, \mu) - e^{\hat\beta^{\mathrm{T}} z_i(\mu)} Y_i(\mu) W(\mu, x_i)/W(T_i, x_i)}{\{\bar S^{(0)}(\hat\beta, \mu)\}^2} d\frac{N(\mu)}{n},\\
\hat\xi_{2i}(t) &= \bar S^{(0)}(\hat\beta, T_i)^{-1} D_i I(T_i \leqslant t) - \hat\Lambda_0(t, \hat\beta).
\end{aligned}$$

THEOREM A4. *Under the conditions of Theorem* A1, *the asymptotic covariance of $n^{1/2}(\hat\beta - \beta_0)$ and $n^{1/2}\{\hat\Lambda_0(t, \hat\beta) - \Lambda_0(t)\}$ can be consistently estimated by $n^{-1}\sum_{i=1}^n \hat\xi_i\hat\xi_{0i}(t)$.*

Proofs of Theorem A3 *and Theorem* A4. We may write $n^{1/2}\{\hat\Lambda_0(t, \hat\beta) - \Lambda_0(t)\}$ as

$$\begin{aligned}
&n^{1/2}\int_0^t \left\{\frac{1}{\bar S^{(0)}(\hat\beta, \mu)} - \frac{1}{\bar S^{(0)}(\beta_0, \mu)} + \frac{1}{\bar S^{(0)}(\beta_0, \mu)} - \frac{1}{s^{(0)}(\beta_0, \mu)}\right\} dF_1(\mu)\\
&\quad + n^{1/2}\int_0^t s^{(0)}(\beta_0, \mu)^{-1} d\left\{\frac{N(\mu)}{n} - F_1(\mu)\right\}\\
&\quad + n^{1/2}\int_0^t \left\{\frac{1}{\bar S^{(0)}(\hat\beta, \mu)} - \frac{1}{\bar S^{(0)}(\beta_0, \mu)} + \frac{1}{\bar S^{(0)}(\beta_0, \mu)} - \frac{1}{s^{(0)}(\beta_0, \mu)}\right\} d\left\{\frac{N(\mu)}{n} - F_1(\mu)\right\}.
\end{aligned}$$

Here, the last term can be shown to converge to zero in probability. By Theorem A2, $n^{1/2}(\hat\beta - \beta_0) = n^{-1/2} I^{-1}(\beta_0)\sum_{i=1}^n \xi_i + o_p(1)$, and by Taylor series expansion around β_0, the first term can be approximated by

$$\begin{aligned}
&-n^{1/2}(\hat\beta - \beta_0)^{\mathrm{T}}\int_0^t \frac{s^{(1)}(\beta_0, \mu)}{\{s^{(0)}(\beta_0, \mu)\}^2} dF_1(\mu) + n^{1/2}\int_0^t \frac{s^{(0)}(\beta_0, \mu) - \bar S^{(0)}(\beta_0, \mu)}{\{s^{(0)}(\beta_0, \mu)\}^2} dF_1(\mu)\\
&\qquad = -n^{-1/2}\int_0^t \frac{s^{(1)}(\beta_0, \mu)}{\{s^{(0)}(\beta_0, \mu)\}^2} dF_1(\mu) I^{-1}(\beta_0)\sum_{i=1}^n \xi_i + n^{-1/2}\sum_{i=1}^n \xi_{1i}(t) + o_p(1)\\
&\qquad = A_n(t) + B_n(t) + o_p(1).
\end{aligned}$$

The second term can be expressed as

$$n^{-1/2}\sum_{i=1}^{n}\{s^{(0)}(\beta_0, T_i)^{-1}I(T_i \leqslant t)D_i - \Lambda_0(t)\} = n^{-1/2}\sum_{i=1}^{n}\xi_{2i}(t) = C_n(t).$$

Therefore, a simple application of the multivariate central limit theorem implies that the finite-dimensional distribution of $\{A_n(t), B_n(t), C_n(t)\}$ is a multivariate normal. As in the proof in Tsiatis (1981), the sequence of distributions induced by A_n, B_n and C_n is tight. □

References

Andersen, P. K., Borgan, O., Gill, R. D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes.* New York: Springer.

Betensky, R. A., Lindsey, J. C., Ryan, L. M. & Wand, M. P. (1999). Local EM estimation of the hazard function for interval-censored data. *Biometrics* **55**, 238–45.

Binder, D. A. (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika* **79**, 139–47.

Copas, A. J. & Farewell, V. T. (2001). Incorporating retrospective data into an analysis of time to illness. *Biostatistics* **2**, 1–12.

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

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–76.

Crowley, J. & Hu, M. (1977). Covariance analysis of heart transplant survival data. *J. Am. Statist. Assoc.* **72**, 27–36.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with Discussion). *J. R. Statist. Soc.* B. **39**, 1–38.

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Am. Statist. Assoc.* **47**, 663–85.

Hyde, J. (1980). Survival analysis with incomplete observations. In *Biostatistics Casebook*, Ed. R. G. Miller, B. Efron, B. W. Brown and L. E. Moses, pp. 31–46. New York: John Wiley & Sons.

Kalbfleisch, J. D. & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed. New York: Wiley.

Keiding, N. & Gill, R. D. (1990). Random truncation models and Markov processes. *Ann. Statist.* **18**, 582–602.

Lin, D. Y. (2000). On fitting Cox's proportional hazards models to survey data. *Biometrika* **87**, 37–47.

Luo, X., Tsai, W.-Y. & Xu, Q. (2009). Pseudo-partial likelihood estimators for Cox regression with missing covariates. *Biometrika* **96**, 617–33.

Miller, R. & Halpern, J. (1982). Regression with censored data. *Biometrika* **69**, 521–31.

Muttlak, H. A. & McDonald, L. L. (1990). Ranked set sampling with size-biased probability of selection. *Biometrics* **46**, 435–46.

Paik, M. C. & Tsai, W.-Y. (1997). On using Cox proportional hazard models with missing covariate. *Biometrika* **84**, 579–93.

Qi, L., Wang, C. Y. & Prentice, R. L. (2005). Weighted estimators for proportional hazards regression with missing covariates. *J. Am. Statist. Assoc.* **100**, 1250–63.

Tsai, W.-Y., Jewell, N. P. & Wang, M.-C. (1987). A note on the product limit estimate of a survival curve under right-censoring and left-truncation. *Biometrika* **74**, 883–6.

Tsiatis, A. A. (1981). A large sample study of Cox's regression model. *Ann. Statist.* **9**, 91–108.

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *Ann. Statist.* **10**, 616–20.

Vardi, Y. (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density. *Biometrika* **76**, 751–61.

Wang, C. Y. & Chen, H. Y. (2001). Augmented inverse probability weighted estimator for Cox missing covariate regression. *Biometrics* **57**, 414–9.

Wang, M.-C. (1991). Nonparametric estimation from cross-sectional survival data. *J. Am. Statist. Assoc.* **86**, 130–43.

Wang, M.-C. (1996). Hazards regression analysis for length-biased data. *Biometrika* **83**, 343–54.

Wang, M.-C., Jewell, N. P. & Tsai, W.-Y. (1986). Asymptotic properties of the product limit estimate under random truncation. *Ann. Statist.* **14**, 1597–605.

[*Received July* 2006. *Revised November* 2008]