AISM (2007) 59:675–695
DOI 10.1007/s10463-006-0086-0

Buckley–James-type of estimators under the classical case cohort design

Qiqing Yu · George Y. C. Wong · Menggang Yu

Received: 4 January 2005 / Revised: 6 April 2006 /
Published online: 24 November 2006

Abstract We consider the estimation problem with classical case-cohort data. The case-cohort design was first proposed by Prentice (Biometrics 73:1–11, 1986). Most studies focus on the Cox regression model. In this paper, we consider the linear regression model. We propose an estimator which extends the Buckley–James estimator to the classical case-cohort design. In order to derive the BJE, there is an additional problem of finding the generalized maximum likelihood estimator (GMLE) of the underlying distribution functions. We propose a self-consistent algorithm for the GMLE. We also justify that the GMLE is consistent and asymptotically normally distributed under certain regularity conditions. We further present some simulation results on the asymptotic properties of the BJE and apply our procedure to a data set used in the literature.

Keywords Case-cohort study · Buckley–James estimator · Right-censorship · Linear regression model · Self-consistent algorithm · Survival data

1 Introduction

We consider the estimation problem under the classical case-cohort design and under linear regression models. Many epidemiological cohort studies and

Q. Yu (✉)
Department of Mathematical Sciences, SUNY, Binghamton, NY 13902, USA
e-mail: qyu@math.binghamton.edu

G. Y. C. Wong
Strang Cancer Prevention Center, 428 E 72nd Street, New York, NY 10021, USA

M. Yu
Department of Medicine/Biostatistics, Indiana University, Indianapolis, IN 46202, USA

disease prevention trials try to investigate the effects of certain covariates for relatively rare disease. As a result, the cohort must be large to provide informative conclusion about the covariate effects. It is often that it is expensive to collect covariates of interest which might involve, for example, biochemical analysis of specimens. In order to lessen this burden without much loss of efficiency, Prentice (1986) proposed the classical case-cohort design, under which, one observes covariates for each subject experiencing an event and for each from a random sample of the cohort, selected at the beginning of the study (call a subcohort).

A pseudo-likelihood method assuming Cox's regression model (1972) was proposed and was later studied in details in a slightly modified version by Self and Prentice (1988). In the classical case-cohort design, the censoring times for subjects not in the subcohort do not need to be recorded and the pseudo-likelihood method does not utilize this information. Chen (2001) considered transformation regression models for modified case-cohort designs, under which the censoring time of the subjects not in the subcohort are also available. Scheike and Martinussen (2004) suggested an estimation method based on maximizing likelihood under Cox's regression model for modified case-cohort studies. Models other than Cox model have been studied more recently. Kulich and Lin (2000) proposed an additive hazards model for classical case-cohort studies that allows estimation of absolute risk parameters. Kong et al. (2004) used weighted estimating equation approach for transformation models under the classical case-cohort design.

We consider an extension of Buckley–James-type (1979) of estimator under linear regression models with the classical case-cohort design. In particular, let Y_i's be monotonically transformed failure times obtained from a known transformation. The log transformation is often used in practice to give the accelerated failure time model (see, e.g., Kalbfleisch and Prentice 2002). Let $\mathbf{X}_i$ be a vector of p-dimensional covariates. The model is $Y_i = \beta'\mathbf{X}_i + \epsilon_i, i = 1, \cdots, n$, where β' is the transpose of a regression coefficient vector β. We shall further simplify notation and write $\beta\mathbf{X} = \beta'\mathbf{X}$. In general, we assume ϵ_i has an unknown cdf F_o. $E(\epsilon_i)$ may or may not be zero, which is not important, as in general $E(\epsilon_i)$ is not identifiable under right censoring (Lai and Ying 1991).

In the case of complete data, the least squares estimator (LSE) is the standard approach. Under right censoring, the Buckley–James (1979) estimator (BJE) is an extension of the LSE. It is well known that the Cox and the Buckley–James estimators are "two most reliable regression estimates to use with censored data" and that "the choice between them should depend on the appropriateness of the proportional hazards model or the linear model for the data" (see, e.g., Miller and Halpern 1982; Hillis 1993). Lai and Ying (1991) showed that the BJE is consistent and asymptotically normally distributed under certain regularity conditions, and is efficient if $\epsilon \sim N(\mu, \sigma^2)$.

As far as we known, in the literature, the Buckley–James-type of estimator has not been investigated in the classical case-cohort design setting. The focus of the paper is to define an extension of the Buckley–James estimator under the classical case-cohort design setting and to propose an algorithm for finding such

an estimate. As explained in the paper, our derivation involves the estimation of the underlying distribution F_o and the estimation of the joint distribution function of the covariates and the censoring variable.

The paper is organized as follows. In Sect. 2, we introduce the notations and an approach of the BJE, which is a score equation based approach which leads to an extension of the BJE to case-cohort studies. This extension of the BJE involves estimation of underlying distribution functions. In Sect. 3, we propose to utilize the generalized maximum likelihood estimators (GMLE) of the underlying distribution functions and provide a feasible algorithm for computing the GMLE. In Sect. 4, we propose an algorithm for obtaining a BJE. Section 5 deals with the estimation of the covariance matrix of the estimator. Section 6 presents some simulation results. We apply our procedure to a real data example in Sect. 7. The detailed proofs of some statements in Sects 2–4 are relegated to Appendices.

2 Buckley–James-type of estimator under the case-cohort design

Consider the linear regression model $Y = \beta\mathbf{X} + \epsilon$, where ϵ has an unknown distribution F_o and $\mathbf{X}$ is a $p \times 1$ dimensional covariate vector. Let C be a censoring variable. Denote $\delta = \mathbf{1}_{(Y \le C)}$, the indicator function of the event $\{Y \le C\}$. Let $M = \min\{Y, C\}$. Under the classical case-cohort design, in addition to the random variables introduced for the usual linear regression model, we need to introduce another random variable, namely, the indicator function that the individual is selected to be in the sub-cohort, denoted by η. Under the classical case-cohort design, if an individual is in the subcohort or if an individual is not in the subcohort but the event of interest has taken place, then $(M, \mathbf{X})$ is recorded, otherwise, $\mathbf{X}$ is not measured and M is not recorded. the observation about $(M, \delta, \mathbf{X}, \eta)$ can be represented by $\mathbf{O}$, where

$$\mathbf{O} = \begin{cases} (M, \delta, \mathbf{X}, \eta) & \text{if } \delta = 1 \text{ or } \eta = 1, \\ (\text{missing}, \delta, \text{ missing}, \eta) & \text{if } \delta = 0 \text{ and } \eta = 0. \end{cases}$$

Note that $\mathbf{O}$ is an extended random vector in the sense that its components can take a "value" called "missing".

Let $T = T(\mathbf{b}) = M - \mathbf{b}\mathbf{X}$. Let $(M_i, \delta_i, \mathbf{X}_i, C_i, \epsilon_i, T_i, \eta_i, \mathbf{O}_i)$, $i = 1, \ldots, n$, be i.i.d. copies of $(M, \delta, \mathbf{X}, C, \epsilon, T, \eta, \mathbf{O})$. We assume that

A1 η, ϵ and $(C, \mathbf{X})$ are independent and $\mathrm{P}\{\eta = 1\} > 0$.

The identifiability assumption made under the uncensored simple linear regression case is $P\{\mathbf{X}_1 \ne \mathbf{X}_2\} > 0$. The identifiability condition under the censored simple linear regression model is $P\{\delta_1 = \delta_2 = 1 \text{ and } \mathbf{X}_1 \ne \mathbf{X}_2\} > 0$. Under multiple linear regression it becomes

$$\mathbf{A2}\ P\left\{\delta_1 = \delta_2 = \cdots = \delta_{p+1} = 1,\ \text{rank}\begin{pmatrix} 1 & \cdots & 1 \\ \mathbf{X}_1 & \cdots & \mathbf{X}_{p+1} \end{pmatrix} = p + 1\right\} > 0.$$

The BJE was first proposed by replacing Y in the score function by its conditional expectation. However, it is not clear how to extend this idea to the case cohort design. It is well known in the literature (see Ritov 1990) that the BJE is somewhat a zero point of a modified score function with the censored linear regression data under the normal distribution assumption, namely, $\epsilon \sim N(\mu, \sigma^2)$, taking advantage of the expressions

$$\frac{f_o'}{f_o}(x) = -x \quad \text{if} \quad \epsilon \sim N(0,1), \quad \text{and } f(x) = -\int\limits_{t>x} \frac{f'}{f}(t) f(t) \mathrm{d}t. \tag{1}$$

This approach provides a valid method to extend the BJE to the case-cohort design.

We now modify the BJE with case-cohort data through a score function with plugging-in method. Abusing notations, we treat $(M_i, \mathbf{X}_i)$ as random variables as well as realizations of the random variables, whenever it is appropriate. Denote $F_{C,\mathbf{X}}$ the cdf of $(C, \mathbf{X})$. Notice that the full nonparametric likelihood function is

$$\begin{aligned} \mathcal{L} = \prod_{i=1}^{n} & \left\{ f_o(T_i(\mathbf{b})) \int\limits_{c \geq M_i, \mathbf{x} = \mathbf{X}_i} \mathrm{d}F_{C,\mathbf{X}}(c, \mathbf{x}) \right\}^{\delta_i} \\ & \times \{ S_o(T_i(\mathbf{b})) f_{C,\mathbf{X}}(M_i, \mathbf{X}_i) \}^{(1-\delta_i)\eta_i} \\ & \times \left\{ \int\limits_{c \in \mathcal{R}, \mathbf{x} \in \mathcal{R}^p} S_o(c - \mathbf{b}\mathbf{x}) \mathrm{d}F_{C,\mathbf{X}}(c, \mathbf{x}) \right\}^{(1-\eta_i)(1-\delta_i)}, \end{aligned} \tag{2}$$

where the three factors in $\mathcal{L}$ correspond to the observations which are observed failures, censored times in the subcohort and missing censored times outside the subcohort, respectively. In our later approach, we need to find the non-parametric or generalized maximum likelihood estimator (GMLE) of $(S_o, F_{C,\mathbf{X}})$, say $(\hat{S}_{\mathbf{b}}, \hat{F}_{C,\mathbf{X},\mathbf{b}})$, which depends on $\mathbf{b} \in \mathcal{R}^p$. The GMLE maximizes the likelihood $\mathcal{L}$ over all possible values of $(S_o, F_{C,\mathbf{X}})$ and over all possible values of $\mathbf{b}$. Under the nonparametric set-up, it suffices to assume that $(C, \mathbf{X})$ is discrete and thus $\mathcal{L}$ in (2) becomes

$$\begin{aligned} \mathcal{L} = & \left[\prod_{i=1}^{n} (f_o(T_i(\mathbf{b})))^{\delta_i} (S_o(T_i(\mathbf{b})))^{\eta_i(1-\delta_i)} \right] \left(\sum_{c,\mathbf{x}} S_o(c - \mathbf{b}\mathbf{x}) f_{C,\mathbf{X}}(c, \mathbf{x}) \right)^{n_1} \\ & \times \prod_{i=1}^{n} \left(\sum_{c \geq M_i} f_{C,\mathbf{X}}(c, \mathbf{X}_i) \right)^{\delta_i} (f_{C,\mathbf{X}}(M_i, \mathbf{X}_i))^{(1-\delta_i)\eta_i}, \end{aligned} \tag{3}$$

where $\sum_{c,\mathbf{x}}$ means the summation over all possible values of the random vector $(C, \mathbf{X})$ and $n_1 = \sum_{i\in K} 1$ and $K = \{i:\ \delta_i + \eta_i = 0\}$. Taking derivative of $\ln \mathcal{L}$ with respect to $\mathbf{b}$ yields

$$\begin{aligned}\frac{\partial \ln \mathcal{L}}{\partial \mathbf{b}} &= \sum_{i=1}^{n}\left\{\delta_i \frac{f_o'}{f_o}(T_i(\mathbf{b})) - (1-\delta_i)\eta_i \frac{f_o}{S_o}(T_i(\mathbf{b}))\right\}(-\mathbf{X}_i)\sigma^{-2} \\ &\quad - n_1 \frac{\sum_{c,\mathbf{x}} f_o(c-\mathbf{b}\mathbf{x}) f_{C,\mathbf{X}}(c,\mathbf{x})(-\mathbf{x})}{\sum_{c,\mathbf{x}} S_o(c-\mathbf{b}\mathbf{x}) f_{C,\mathbf{X}}(c,\mathbf{x})}\sigma^{-2} \\ &= \sum_{i\notin K}\left\{\delta_i T_i(\mathbf{b}) + (1-\delta_i)\frac{\int_{t>T_i(\mathbf{b})} t f_o(t)\mathrm{d}t}{S_o(T_i(\mathbf{b}))}\right\}\mathbf{X}_i\sigma^{-2} \quad \text{(by (1))} \\ &\quad + n_1 \frac{\sum_{c,\mathbf{x}} \int_{t>c-\mathbf{b}\mathbf{x}} t f_o(t)\mathrm{d}t\, f_{C,\mathbf{X}}(c,\mathbf{x})\mathbf{x}}{\sum_{c,\mathbf{x}} S_o(c-\mathbf{b}\mathbf{x}) f_{C,\mathbf{X}}(c,\mathbf{x})}\sigma^{-2}.\end{aligned}$$

After centering to $E(\mathbf{X})$ and multiplying σ^2, the foregoing score function becomes

$$\begin{aligned}H(\mathbf{b}, S_o, f_{C,\mathbf{X}}) &= \sum_{i\notin K}\left\{\delta_i T_i(\mathbf{b}) - (1-\delta_i)\frac{\int_{t>T_i(\beta)} t\ \mathrm{d}S_o(t)}{S_o(T_i(\mathbf{b}))}\right\}(\mathbf{X}_i - E(\mathbf{X})) \\ &\quad - n_1 \frac{\sum_{c,\mathbf{x}} \int_{t>c-\mathbf{b}\mathbf{x}} t\ \mathrm{d}S_o(t) f_{C,\mathbf{X}}(c,\mathbf{x})(\mathbf{x}-E(\mathbf{X}))}{\sum_{c,\mathbf{x}} S_o(c-\mathbf{b}\mathbf{x}) f_{C,\mathbf{X}}(c,\mathbf{x})}.\end{aligned}$$

Since $S_o, f_{C,\mathbf{X}}$ and $E(\mathbf{X})$ are unknown, one can replace them by their estimators. A logical choice is their GMLEs, say $\hat{S}_{\mathbf{b}}, \hat{f}_{C,\mathbf{X},\mathbf{b}}$ and $\bar{\mathbf{X}}$ $(= \sum_{c,\mathbf{x}} \mathbf{x}\hat{f}_{C,\mathbf{X},\mathbf{b}}(c,\mathbf{x}))$, where $\hat{f}_{C,\mathbf{X},\mathbf{b}}$ is the pdf of the GMLE of $F_{C,\mathbf{X}}$.

Ideally, a solution to $\hat{H}(\mathbf{b}) = 0$ should be called an extension of the BJE, where

$$\hat{H}(\mathbf{b}) = H(\mathbf{b}, \hat{S}_{\mathbf{b}}, \hat{f}_{C,\mathbf{X},\mathbf{b}}). \tag{4}$$

If $\hat{H}(\mathbf{b})$ does not have a root, then a BJE is a point at which $\hat{H}(\cdot)$ changes its sign (called a *zero-crossing*) (see James and Smith 1981). A standard extension of the BJE under the case-cohort design is then a zero-crossing of $\hat{H}(\mathbf{b})$. This basically is an estimation equation approach, as in Ritov (1990) for obtaining various estimators of β with interval-censored data under linear regression models.

3 The GMLE of the underlying cdf's

We shall now discuss how to find the GMLE of $(S_o, f_{C,\mathbf{X}})$. For each given $\mathbf{b}$, a GMLE of $(F_o, F_{C,\mathbf{X}})$ is a pair of distribution functions such that it maximizes $\mathcal{L}$

over all possible values of $(F_o, F_{C,\mathbf{X}})$, say (F, G). From the GMLE of $(F_o, F_{C,\mathbf{X}})$, we can obtain the GMLE of $(S_o, f_{C,\mathbf{X}})$ needed in $\hat{H}$ (see (4)).

Verify that an observation on $(C, \mathbf{X})$ can be represented by an observable $1 + p$ dimensional rectangle I_i, where I_i is of the form

$$I_i = \begin{cases} [M_i, \infty) \times \{\mathbf{X}_i\} & \text{if } \delta_i = 1, \text{ where } \{\mathbf{x}\} \text{ is a singleton set,} \\ \{M_i\} \times \{\mathbf{X}_i\} & \text{if } \delta_i = 0 \text{ and } \eta_i = 1, \\ \mathcal{R}^{p+1} & \text{otherwise.} \end{cases}$$

An observation on ϵ can be represented by an observed interval which is of the form

$$B_i = \begin{cases} \{T_i(\mathbf{b})\} & \text{if } \delta_i = 1, \\ (T_i(\mathbf{b}), \infty) & \text{if } \delta_i = 0 \text{ and } \eta_i = 1, \\ (-\infty, \infty) & \text{otherwise.} \end{cases}$$

Using an arguement similar to that in Wong and Yu (1999), it can be shown that in order to maximize $\mathcal{L}(\mathbf{b}, F, G)$, it suffices to put mass of G to all maximum intersections (MI) $A_1, \ldots, A_{m_g}$ induced by I_i's, where an MI A induced by sets $I_1, \ldots, I_n$ is a nonempty intersection of these I_is such that $A \cap I_i$ equals either A or $\emptyset$ for each i. Notice that A_j is either of the form $\{M_i\} \times \{\mathbf{X}_i\}$, or of the form $[\mathbf{M}_i, \infty) \times \{\mathbf{X}_i\}$. It is well known that if an MI is not a singleton set then the mass assigned by a GMLE to the MI is not uniquely defined (see Yu et al. 2000). Thus if $A_j = [M_i, \infty) \times \{\mathbf{X}_i\}$ is an MI, then we replace it by the set $\{M_i\} \times \{\mathbf{X}_i\}$.

Moreover, it can be shown that in order to maximize $\mathcal{L}(\mathbf{b}, F, G)$ it suffices to put mass of F to all distinct maximal intersection induced by the observed intervals B_i's on ϵ. Denote n_2 the number of observed $\mathbf{X}_i$'s and $T_{(n_2)}$ the largest among $T_i, i \notin K$. It can be verified that each exact observation is a maximal intersection and if the largest $T_{(n_2)}(\mathbf{b})$ is right censored, then $V = (T_i(\mathbf{b}), \infty)$ is also a maximal intersection. For convenience, denote $t_1 < \cdots < t_{m_f - 1}$ the first $m_f - 1$ smallest distinct exact observations and denote t_{m_f} the largest exact observation if $T_{(n_2)}$ is not right-censored and denote $t_{m_f} = T_{(n_2)} + 1$ if $T_{(n_2)}$ is right-censored. We shall use $\{t_{m_f}\}$ to replace the role of the maximal intersection V.

For convenience, denote $A_j = \{c_j\} \times \{\mathbf{x}_j\}$, $\phi_{ij} = \mathbf{1}(A_j \subset I_i)$ and $\psi_{ij} = \mathbf{1}(t_j \in B_i)$. Let f_j be the mass assigned by F on t_j and g_j be the mass assigned by G on A_j. By (3),

$$\ln \mathcal{L} = \sum_{i \notin K} \ln \sum_{j=1}^{m_f} \psi_{ij} f_j + n_1 \ln \sum_{j=1}^{m_g} g_j \sum_{k: t_k > c_j - \mathbf{b}\mathbf{x}_j} f_k + \sum_{i \notin K} \ln \sum_{j=1}^{m_g} \phi_{ij} g_j. \tag{5}$$

In order to find the GMLE, it suffices to maximize the foregoing expression over all possible values of $(f_1, \ldots, f_{m_f}, g_1, \ldots, g_{m_g})$ subject to the constraints that $f_i \geq 0$, $g_i \geq 0$, $\sum_i f_i = \sum_i g_i = 1$.

One may choose to find the GMLE with the Newton–Raphson method. However, there maybe too many parameters involved in implementing the

Newton–Raphson method if n_2 is moderate or large. Thus, it may cause a problem computationally in computing the inverse matrix of the Fisher information matrix, which is needed in the method.

In order to design a feasible algorithm for computing the GMLE, we shall establish in Appendices self-consistent equations for the case-cohort data as follows:

$$f_k = \frac{1}{n}\sum_{i\notin K}\frac{\psi_{ik}f_k}{\sum_{h=1}^{m_f}\psi_{ih}f_h} + \frac{n_1 f_k}{n}\frac{\sum_{j=1}^{m_g} g_j \mathbf{1}(t_k > c_j - \mathbf{b}\mathbf{x}_j)}{\sum_{j=1}^{m_g} g_j \sum_{m:\ t_m > c_j - \mathbf{b}\mathbf{x}_j} f_m}, \quad k = 1, \ldots, m_f, \quad (6)$$

$$g_k = \frac{1}{n}\sum_{i\notin K}\frac{\phi_{ik}g_k}{\sum_{h=1}^{m_g}\phi_{ih}g_h} + \frac{n_1 g_k}{n}\frac{\sum_{t_h > c_k - \mathbf{b}\mathbf{x}_k} f_h}{\sum_{j=1}^{m_g} g_j \sum_{m:\ t_m > c_j - \mathbf{b}\mathbf{x}_j} f_m}, \quad k = 1, \ldots, m_g. \quad (7)$$

Verify that if $\mathrm{P}\{\eta = 1\} = 1$ then Eq. (6) becomes

$$f_k = \frac{1}{n}\sum_{i}\frac{\psi_{ik}f_k}{\sum_{h=1}^{m_f}\psi_{ih}f_h}, \quad (8)$$

which is of the same form as the self-consistent equation proposed by Turnbull (1976) for computing the GMLE with univariate interval-censored data in order to overcome the same type of computational difficulty. However, form (8) is quite different from form (6).

Equations (6) and (7) lead to a feasible self-consistent algorithm for deriving the GMLE of F_o and $F_{C,\mathbf{X}}$ as follows.

Self-consistent algorithm:

Step 1. Denote $f_{k,1} = 1/m_f$, $k = 1, \ldots, m_f$ and $g_{k,1} = 1/m_g$, $k = 1, \ldots, m_g$.

Step v ($v \geq 2$). For $k = 1, \ldots, m_f$, compute

$$f_{k,v} = \frac{1}{n}\sum_{i\notin K}\frac{\psi_{ik}f_{k,v-1}}{\sum_{h=1}^{m_f}\psi_{ih}f_{h,v-1}} + \frac{n_1 f_{k,v-1}}{n}\frac{\sum_{j=1}^{m_g} g_{j,v-1}\mathbf{1}(t_k > c_j - \mathbf{b}\mathbf{x}_j)}{\sum_{j=1}^{m_g} g_{j,v-1}\sum_{m:\ t_m > c_j - \mathbf{b}\mathbf{x}_j} f_{m,v-1}}. \quad (9)$$

For $k = 1, \cdots, m_g$, compute

$$g_{k,v} = \frac{1}{n}\sum_{i\notin K}\frac{\phi_{ik}g_{k,v-1}}{\sum_{h=1}^{m_g}\phi_{ih}g_{h,v-1}} + \frac{n_1 g_{k,v-1}}{n}\frac{\sum_{h:\ t_h > c_k - \mathbf{b}\mathbf{x}_k} f_{h,v-1}}{\sum_{j=1}^{m_g} g_{j,v-1}\sum_{m:\ t_m > c_j - \mathbf{b}\mathbf{x}_j} f_{m,v-1}}. \quad (10)$$

Stop if $|f_{k,v} - f_{k,v-1}| < c$ and $|g_{k,v} - g_{k,v-1}| < c$ for all possible k, where c is a predetermined tolerance. Otherwise, go to next step.

Denote the resulting estimator by $\hat{f}_k$ and $\hat{g}_k$. Then the GMLE's of S_o and $f_{C,\mathbf{X}}$ for the given $\mathbf{b}$ are $\hat{S}_{\mathbf{b}}(t) = \sum_{t_k > t}\hat{f}_k$ and $\hat{f}_{C,\mathbf{X},\mathbf{b}}(c, \mathbf{x}) = \hat{g}_k$ if $(c, \mathbf{x}) \in A_k$.

A solution to self-consistent equations (6) and (7) is called a self-consistent estimator (SCE) of f_k and g_k. The GMLE is an SCE but an SCE may not be the GMLE. However, we shall establish the following statement in Appendices.

Proposition 1 *If the foregoing self-consistent algorithm converges, then the solution is the GMLE. Moreover, the GMLE satisfies that* $\hat{f}_k, \hat{g}_k \in (0,1)$ *for all possible k.*

The key to the proof of the proposition is that the initial point in the algorithm satisfies $f_{k,1} \neq 0$ and $g_{k,1} \neq 0 \; \forall \; k$.

Though in practice, we have not encountered the case in which the algorithm does not converge, there is still difficulty in proving that in general the self-consistent algorithm under the case-cohort design does converge. However, this is also true for the self-consistent algorithm with interval-censored data proposed by Turnbull (1976) and the standard numerical algorithm using the NR method. In all the three cases, it has only been shown that the algorithm converges under the assumption that the initial point is very close to a true local maximum point. Since the proof of the foregoing statement under the case-cohort design is almost exactly the same as the arguement in the paragraph corresponding to Equation (3.12) in Turnbull (1976), we shall not repeat here.

A desirable estimator should be consistent and efficient if $\mathbf{b} = \beta$. In Theorems 1 and 2, we show that the GMLE satisfies such properties under a simple assumption:

A3 $(\epsilon, \mathbf{X}, C)$ takes on finitely many values.

In the literature, A3 has been utilized in many pioneering papers on a new procedure in order to justify its nice properties without going through the lengthy investigation, see for example, Miller (1981) on the PLE with right-censored data and Turnbull (1976) on the GMLE with interval-censored data. We also follow this path here.

Theorem 1 *Under* A1, A2 *and* A3, *the GMLE* $(\hat{F}_\beta, \hat{F}_{C,\mathbf{X},\beta})$ *satisfies that with probability one, we have* $\lim_{n\to\infty} \hat{F}_\beta(t) = F_*(t)$, *where*

$$F_*(t) = \begin{cases} F_o(t) & \text{if for } t \leq \tau_1, \\ F_o(\tau_1) & \text{if } t \in (\tau_1, \tau_1 + 1), \\ 1 & \text{if } t \geq \tau_1 + 1, \end{cases} \quad \text{and} \quad \tau_1 = \max_j T_j(\beta),$$

and $\lim_{n\to\infty} \hat{F}_{C,\mathbf{X},\beta}(c, \mathbf{x}) = F^*_{C,\mathbf{X}}(c, \mathbf{x})$ *pointwise, where*

$$F^*_{C,\mathbf{X}}(c, \mathbf{x}) = \begin{cases} F_{C,\mathbf{X}}(c, \mathbf{x}) & \text{if } c < \tau_{\mathbf{x}}, \\ F_{C,\mathbf{X}}(\infty, \mathbf{x}) & \text{if } c \geq \tau_{\mathbf{x}}, \end{cases} \quad \text{and} \quad \tau_{\mathbf{x}} = \max\{M_j : \mathbf{X}_j = \mathbf{x}, j \notin K\}.$$

.

Theorem 2 *Under* A1, A2 *and* A3, *the GMLE* $(\hat{F}_\beta, \hat{F}_{C,\mathbf{X},\beta})$ *is asymptotically efficient.*

The proofs of these two theorems are given in Appendices.

4 Computation of the BJE

If p is 1 or 2, one can graph the function $\hat{H}$ by plotting $(\mathbf{b}, \hat{H}(\mathbf{b}))$ to find out the zero-crossings directly. If $p > 2$, we shall introduce an algorithm for computing the BJE. Notice that

$$\begin{aligned}\hat{H}(\mathbf{b}) &= H(\mathbf{b}, \hat{S}_{\mathbf{b}}, \hat{f}_{C,\mathbf{X},\mathbf{b}}) \\ &= \sum_{i\notin K}\left\{\delta_i T_i(\mathbf{b}) + (1-\delta_i)\frac{\sum_{t>T_i(\mathbf{b})} t\hat{f}_{\mathbf{b}}(t)}{\hat{S}_{\mathbf{b}}(T_i(\mathbf{b}))}\right\}(\mathbf{X}_i - \bar{\mathbf{X}}) \\ &\quad + n_1\frac{\sum_{c,\mathbf{x}}\sum_{t>c-\mathbf{b}\mathbf{x}} t\hat{f}_{\mathbf{b}}(t)\hat{f}_{C,\mathbf{X},\mathbf{b}}(c,\mathbf{x})(\mathbf{x}-\bar{X})}{\sum_{c,\mathbf{x}}\hat{S}_{\mathbf{b}}(c-\mathbf{b}\mathbf{x})\hat{f}_{C,\mathbf{X},\mathbf{b}}(c,\mathbf{x})}.\end{aligned}$$

Since $T_i(\mathbf{b}) = M_i - \mathbf{b}\mathbf{X}_i$ for $i \notin K$, one can verify that $\hat{H}$ becomes

$$\hat{H}(\mathbf{b}) = A(\mathbf{b}) - B(\mathbf{b})\mathbf{b}, \tag{11}$$

where

$$\begin{aligned}A(\mathbf{b}) &= \sum_{i\notin K}\left\{M_i\delta_i + (1-\delta_i)\sum_{t>T_i(\mathbf{b})}\frac{\hat{f}_{\mathbf{b}}(t)}{\hat{S}_{\mathbf{b}}(T_i(\mathbf{b}))}\frac{\sum_{j\notin K} M_j \mathbf{1}_{(T_j(\mathbf{b})=t,\delta_j=1)}}{\sum_{k\notin K}\mathbf{1}_{(T_k(\mathbf{b})=t,\delta_k=1)}}\right\}(\mathbf{X}_i - \bar{\mathbf{X}}) \\ &\quad + n_1\frac{\sum_{c,\mathbf{x}}\sum_{t>c\mathbf{b}\mathbf{x}}\hat{f}_{\mathbf{b}}(t)\frac{\sum_{h\notin K} M_h\mathbf{1}_{(T_h(\mathbf{b})=t,\delta_h=1)}}{\sum_{m\notin K}\mathbf{1}_{(T_m(\mathbf{b})=t,\delta_m=1)}}\hat{f}_{C,\mathbf{X},\mathbf{b}}(c,\mathbf{x})(\mathbf{x}-\bar{\mathbf{X}})}{\sum_{c,\mathbf{x}}\hat{S}_{\mathbf{b}}(c-\mathbf{b}\mathbf{x})\hat{f}_{C,\mathbf{X},\mathbf{b}}(c,\mathbf{x})}, \\ B(\mathbf{b}) &= \sum_{i\notin K}(\mathbf{X}_i - \bar{\mathbf{X}})\left\{\mathbf{X}_i\delta_i + (1-\delta_i)\sum_{t>T_i(\mathbf{b})}\frac{\hat{f}_{\mathbf{b}}(t)}{\hat{S}_{\mathbf{b}}(T_i(\mathbf{b}))}\frac{\sum_{j\notin K}\mathbf{X}_j\mathbf{1}_{(T_j(\mathbf{b})=t,\delta_j=1)}}{\sum_{k\notin K}\mathbf{1}_{(T_k(\mathbf{b})=t,\delta_k=1)}}\right\}' \\ &\quad + n_1\frac{\sum_{c,\mathbf{x}}(\mathbf{x}-\bar{\mathbf{X}})\sum_{t>c\mathbf{b}\mathbf{x}}\hat{f}_{\mathbf{b}}(t)\frac{\sum_{h\notin K}\mathbf{X}_h'\mathbf{1}_{(T_h(\mathbf{b})=t,\delta_h=1)}}{\sum_{m\notin K}\mathbf{1}_{(T_m(\mathbf{b})=t,\delta_m=1)}}\hat{f}_{C,\mathbf{X},\mathbf{b}}(c,\mathbf{x})}{\sum_{c,\mathbf{x}}\hat{S}_{\mathbf{b}}(c-\mathbf{b}\mathbf{x})\hat{f}_{C,\mathbf{X},\mathbf{b}}(c,\mathbf{x}).}\end{aligned}$$

Ideally, a BJE is a solution to the equation $A(\mathbf{b}) + B(\mathbf{b})\mathbf{b} = 0$ (see (11). It yields

$$\mathbf{b} = \{B(\mathbf{b})\}^{-1}A(\mathbf{b}). \tag{12}$$

Algorithm (for the BJE with case-cohort data)

1. Give an initial value to β, say $\mathbf{b}_0$.
2. Obtain $(A(\mathbf{b}_0), B(\mathbf{b}_0))$'s with the given $\mathbf{b}_0$.

3. In view of (12), update $\mathbf{b}_0$ by

$$\mathbf{b}_1 = \{B(\mathbf{b}_0)\}^{-1} A(\mathbf{b}_0).$$

4. For $k \geq 2$, repeat Steps 2 and 3 iteratively, with values of $\mathbf{b}_0$ and $\mathbf{b}_1$ updated by $\mathbf{b}_{k-1}$ and $\mathbf{b}_k$, respectively, until $\mathbf{b}_k$ converges (*i.e.*, $|\mathbf{b}_k - \mathbf{b}_{k-1}|$ is very small), or oscillates between two or more values. In the latter case, take the midpoint of the last two values, say $\mathbf{b}_k$ and $\mathbf{b}_{k-1}$, as an estimate of β.

Remark 1 It is well known that in the case when the algorithm oscillates, the algorithm may not result in a solution of the BJE. However, if the two oscillating points are close, the estimate resulted from the algorithm can be viewed as an approximation of the BJE. Finally, if the two oscillating points are far apart, then one can graph the function $\hat{H}(\mathbf{b})$ between the oscillating points to find a zero crossing of $\hat{H}(\mathbf{b})$.

5 Estimation of the variance of the BJE

Under the assumption that ϵ has a normal distribution, since the BJE is efficient in the censored regression data case, we expect that the BJE is also efficient in the subcohort case, though of course the efficient lower bound is different.

Based on the belief that the BJE is efficient under the normal assumption, we estimate the covariance matrix of the BJE by $\hat{\Sigma}_{\hat{\beta}} = (\hat{\mathcal{I}})^{-1}$, where $\mathcal{I}$ is the Fisher information matrix. Under certain regularity, it can be estimated by

$$\hat{\mathcal{I}} = \sum_{i \notin K} \left\{ \delta_i T_i(\hat{\beta}) + (1-\delta_i) \frac{\int_{t>T_i(\hat{\beta})} t \mathrm{d}\hat{S}_{\hat{\beta}}(t)}{\hat{S}_{\hat{\beta}}(T_i(\hat{\beta}))} \right\}^2 (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'(\hat{\sigma}^{-2})^2 + n_1 \frac{UU'(\hat{\sigma}^{-2})^2}{(\sum_{c,\mathbf{x}} \hat{S}_{\hat{\beta}}(c - \hat{\beta}\mathbf{x}) \hat{f}_{C,\mathbf{X},\hat{\beta}}(c,\mathbf{x}))^2}.$$

where $U = \sum_{c,\mathbf{x}} \int_{t>c-\hat{\beta}\mathbf{x}} t \, \mathrm{d}\hat{S}_{\hat{\beta}}(t) \hat{f}_{C,\mathbf{X},\hat{\beta}}(c,\mathbf{x})(\mathbf{x} - \bar{X})$, where $\hat{\sigma}^2$, $\hat{S}_{\hat{\beta}}$ and $\hat{f}_{C,\mathbf{X},\hat{\beta}}$ are estimates of σ^2, S_o and $f_{C,\mathbf{X}}$, respectively. The expression is based on the formula that for $n = 1$, under certain regularity conditions,

$$-E\Big(\frac{\partial^2 \ln \mathcal{L}}{\partial\beta\partial\beta'}\Big) = E\Big[\frac{\partial \ln \mathcal{L}}{\partial\beta} \frac{\partial \ln \mathcal{L}}{\partial\beta}'\Big] = E\Big[\{A(\beta) - \beta B(\beta)\}\{A(\beta) - \beta B(\beta)\}'\Big]\sigma^{-4}.$$

For each i, denote $v_i = M_i - \hat{\beta}\mathbf{X}_i$ and $m = \sum_i \delta_i$. The parameter σ can be estimated in two ways. If the largest v_i is not censored, then one can estimate it by $\hat{\sigma}^2 = \sum_i v_i^2 \hat{f}_{\hat{\beta}}(v_i) - (\sum_i v_i \hat{f}_{\hat{\beta}}(v_i))^2$. Otherwise, we can use the least squares method as follows. We can find the quantiles of $\hat{F}(v_i)$ under $N(0,1)$, say q_i's. Then we find the least squares estimate of (μ, σ) that minimizes

$$\sum_{i=1}^{m} \delta_i \left(\frac{v_i - \mu}{\sigma} - q_i \right)^2.$$

It can be shown that the LSE is $\hat{\sigma} = \frac{\bar{vq} - \bar{v} \cdot \bar{q}}{\bar{q^2} - (\bar{q})^2}$, where $\bar{v} = \frac{1}{m} \sum_{i=1}^{m} \delta_i v_i$, $\bar{q} = \frac{1}{m} \sum_{i=1}^{m} \delta_i q_i$, $\bar{q^2} = \frac{1}{m} \sum_{i=1}^{m} \delta_i q_i^2$ and $\bar{vq} = \frac{1}{m} \sum_{i=1}^{m} \delta_i v_i q_i$.

If F_o is not normal, then the estimator $(\hat{\mathcal{I}})^{-1}$ is no longer valid, as the BJE is not efficient in this case. At the moment, we propose to bootstrap the BJE to derive an empirical estimator of the covariance matrix.

6 Simulation studies

In this section, we shall present four sets of simulation results on our proposed estimators for evaluating its propoties under various sample sizes and various continuous distributions. In each simulation study, we had 1,000 replications and computed the sample mean and sample standard error (SE) of the 1,000 estimates. The computation was quite fast, it only took a few seconds for a sample size of 800.

We also compare the estimator to the BJE based on the subcohort data alone (called subcohort BJE). The estimator is easy to compute.

For simplicity, we assume $p = 1$ and the random vector (C, X) is a discrete uniform distribution on the set $\{(0,0),(0,1),(1,1)\}$. Hereafter $Exp(\mu, \sigma)$ denotes an exponential distribution with the pdf $f(x) = \frac{1}{\sigma} e^{-[\frac{x-\mu}{\sigma}+1]} \mathbf{1}_{(x>\mu-\sigma)}$. We consider 4 different cases.

Case 1 (censored-data under a normal distribution). Suppose $\epsilon \sim N(2,1)$ (the normal distribution), $q = \mathrm{P}\{\eta = 1\} = 0.2$ and 0.5. $\beta = 1$. Censoring rate is 0.984. Results are listed in Blocks 1 and 2 of Table 1.

Case 2 (censored-data under a normal distribution). Suppose $\epsilon \sim N(0,1)$, $q = 0.2$ and 0.5. $\beta = 1$. Censoring rate is 0.614. Results are listed in Blocks 3 and 4 of Table 1.

Case 3 (censored-data under an exponential distribution). Suppose $\epsilon \sim \mathrm{Exp}(0.9, 1)$. $q = 0.2$ and 0.5. $\beta = 1$. Censoring rate is 0.937. Results are listed in Blocks 1 and 2 of Table 2.

Case 4 (censored-data under an exponential distribution). Suppose $\epsilon \sim \mathrm{Exp}(0, 1)$. $q = 0.2$ and 0.5. $\beta = 1$. Censoring rate is 0.579. Results are listed in Blocks 3 and 4 of Table 2.

In Cases 1 and 2, ϵ has a normal distribution. We adjust the mean so that there is a severe censoring rate in one case but not in the other. Similarly, in Cases 3 and 4, ϵ has an exponential distribution. We adjust the mean so that one has a severe censoring rate and the other does not. The case cohort design is proposed for epidemiological studies for rare diseases, thus one expects that

Table 1 Comparison on two BJE methods

n	β	GMLE (SE)	Subcohort (SE)
Under N(2,1) with censoring rate 0.984			
with $q = 0.2$			
100	1	0.779 (21.302)	SE too large
200	1	1.191 (1.038)	SE too large
400	1	1.607 (0.581)	SE too large
800	1	1.226 (0.527)	SE too large
1600	1	1.045 (0.288)	SE too large
with $q = 0.5$			
100	1	0.792 (4.872)	SE too large
200	1	0.980 (0.472)	SE too large
400	1	0.972 (0.411)	SE too large
800	1	0.974 (0.288)	SE too large
1600	1	0.987 (0.165)	0.993 (0.309)
Under N(0,1) with censoring rate 0.614			
with $q = 0.2$			
100	1	1.035 (0.385)	SE too large
200	1	1.008 (0.262)	SE too large
400	1	0.994 (0.178)	1.000 (0.296)
800	1	0.993 (0.126)	0.997 (0.198)
with $q = 0.5$			
100	1	1.002 (0.288)	0.991 (0.380)
200	1	0.999 (0.207)	0.997 (0.259)
400	1	0.994 (0.142)	0.993 (0.180)
800	1	0.995 (0.103)	0.996 (0.129)

Table 2 Comparison on two BJE methods

n	β	GMLE (SE)	Subcohort (SE)
Under exp(0.9,1) with censoring rate 0.937			
with $q = 0.2$			
100	1	1.522 (0.501)	SE too large
200	1	1.235 (0.358)	SE too large
400	1	1.017 (0.109)	SE too large
800	1	1.002 (0.034)	SE too large
with $q = 0.5$			
100	1	0.999 (0.091)	SE too large
200	1	0.993 (0.043)	SE too large
400	1	0.998 (0.010)	SE too large
800	1	0.998 (0.008)	1.000 (0.024)
Under exp(0,1) with censoring rate 0.579			
with $q = 0.2$			
100	1	1.062 (0.194)	SE too large
200	1	1.033 (0.134)	1.008 (0.229)
400	1	1.016 (0.087)	1.006 (0.156)
800	1	1.010 (0.060)	1.006 (0.105)
with $q = 0.5$			
100	1	1.020 (0.171)	1.010 (0.198)
200	1	1.014 (0.119)	1.007 (0.135)
400	1	1.009 (0.086)	1.004 (0.093)
800	1	1.007 (0.062)	1.005 (0.063)

the censoring rate is quite high. Also, in a typical data analysis with case cohort design the proportion of subcohort q is chosen to be 0.2 (see, e.g., Lin and Ying 1993). This is the motivation we choose those parameters. The case that $q > 0.5$ and the censoring rate is greater than 0.5 is not that interesting in applications of case cohort designs. Thus we do not present any simulation results in such cases.

In both tables, the column corresponding to GMLE is the average of 1000 BJE estimates based on the GMLE method and SE is the sample standard deviation of these 1000 estimates. The column corresponding to Subcohort is the average of 1000 BJE estimates based on subcohort alone. In the tables, the phrase "SE too large" means that the sample standard deviation is larger than 1,000. The advantage is not that significant when the censoring rate decreases.

The simulation results in Table 1 suggest that the BJE $\hat{\beta}$ is consistent under both the normal distribution and the exponential distribution, as β is within the interval $(\bar{\hat{\beta}} - 2SE, \bar{\hat{\beta}} + 2SE)$ and the SE decreases, as n increases.

Furthermore, they confirm that the BJE $\hat{\beta}$ with the full likelihood is better than the subcohort BJE under the normal assumption and under the exponential distribution, in the sense that the SE of the BJE based on the GMLE is smaller than that of the BJE based on subcohort. Our simulation results suggest that the BJE based on subcohort alone is very unstable as we expect, because there are not much events taken place within the subcohort. From Table 1 and the first half of Table 2, it is seen that our new methodology has an obvious advantage over the naive method, especially when the censoring rate is high and the size of subcohort is small.

7 A real example

In this section, we carry out data analysis on the Welsh Nickel Refinery Study which has been used frequently in the literature to illustrate case-cohort studies (see e.g., Lin and Ying 1993; Barlow et al. 1999). In this study, employees in a nickel refinery in South Wales were investigated to determine the risk of developing nasal cancer. There were 56 cancer cases among the 679 workers employed before 1925. The variables used in our analysis are exposure (EXP) level and age at first employment (AFE). Exposure level is log transformed to log(EXP +1) and age at first employment is transformed to log(AFE-10). We chose exactly the same subcohort as in Barlow et al. (1999) which has 135 subjects including 9 cases. The remaining 47 cases are outside the subcohort.

One can group $\mathbf{O}_i$'s according to the following three forms: (1) $\mathbf{O}_i$'s with $\delta_i = \eta_i = 0$, (2) $\mathbf{O}_k$'s of the form $[M_k, \infty) \times \{\mathbf{X}_k\}$. (3) $\mathbf{O}_j$'s of form $\{M_j\} \times \{\mathbf{X}_j\}$. There are 497, 56 and 126 $\mathbf{O}_i$'s in Groups 1, 2 and 3, respectively. All observations M_is in Groups 2 and 3 are distinct. Thus all observations in Group 2 correspond to distinct MIs induced by observations on $(C, \mathbf{X})$. Most but not all observations in Group 3 correspond to MIs as well. This is different from the cases in our simulation studies, due to the simple assumption on $(C, \mathbf{X})$ there.

Thus there are 236 parameters in searching for the GMLE of $F_{C,\mathbf{X}}$, in addition to the parameters for estimating F_o and the BJE.

The analysis using our procedure results in estimates of $(-0.30, -1.15)$ for (log(EXP +1), log(AFE-10)) with standard error $(0.01, 0.01)$. The negative estimates make sense since both are believed to reduce the time to cancer occurrence. We see that both effect are significant.

In comparison to the data analysis based on the Cox regression model carried out by Lin and Ying (1993), our estimates are more significant than theirs. Their z-scores are 3 and 4, respectively, while ours z-scores are much higher.

In our data analysis, we actually replace $\bar{\mathbf{X}}$ in Sect. 2 by another simpler estimator

$$\bar{\mathbf{X}} = \frac{1}{n}\sum_{i\notin K} w_i \mathbf{X}_i,$$

where $w_i = \delta_i + n_4(1-\delta_i)\eta_i/n_3$, $n_3 = \sum_{i=1}^n \eta_i(1-\delta_i)$ is the number of censored subjects in the subcohort and $n_4 = n - \sum_{i=1}^n \delta_i$ is the total number of censored subjects. It does not need to be computed in each iteration.

Appendices

We shall prove Eqs. (6) and (7) and Proposition 1 in Appendix 1, and the consistency and asymptotic normality of the GMLE under a simple assumption in Appendix 2. Hereafter, let D_f be the set of the probability mass $(f_2, \ldots, f_{m_f})$, where $f_i \geq 0$ and $\sum_{i=1}^{m_f} f_i = 1$; let D_g be the convex set of the probability mass $(g_2, \ldots, g_{m_g})$, where $g_i \geq 0$ and $\sum_{j=1}^{m_g} g_j = 1$. Let D be the product set $D_f \times D_g$.

Appendix 1

In this appendix, we shall prove Proposition 1 and Eqs. (6) and (7). We shall first prove a lemma.

Lemma 1 *Under the model assumptions, we have*

(1) *for each* $(g_2, \ldots, g_{m_g})$, $-\ln\mathcal{L}$ *is strictly convex on* D_f*;*
(2) *for each* $(f_2, \ldots, f_{m_f})$, $-\ln\mathcal{L}$ *is strictly convex on* D_g*;*
(3) *there is a unique maximum point of* $\mathcal{L}$ *on* D *and it is an interior point of* D*.*

Proof By (5), it is obvious that statements (1) and (2) hold.

We shall assume that statement (3) is false and reach a contradiction. Notice that D is close. Thus the maximum point is on the boundary of D, with $f_k = 0$ or $g_k = 0$ for some k.

If $f_k = 0$, then either the MI induced by B_i's is $\{t_k\}$ and $t_k = T_i$ for some i with $\delta_i = 1$, or the MI is (T_i, ∞) with $\delta_i = 0$ and T_i is the largest observation among T_h's. In either case, $\psi_{ij} = \begin{cases} 1 \text{ if } j = k, \\ 0 \text{ otherwise} \end{cases}$. Hence, $0 \leq \mathcal{L} \leq f_k = 0$. Note

that $\mathcal{L} > 0$ if $f_j = 1/m_f$ and $g_j = 1/m_g$ for each possible j. It contradicts the assumption that $\mathcal{L}$ reaches its maximum value at a point with $f_k = 0$.

If $g_k = 0$, then the corresponding MI induced by I_h's is either $A_j = \{M_i\} \times \{\mathbf{X}_i\}$, or $A_j = [M_i, \infty) \times \{\mathbf{X}_i\}$, Thus, there is always an i such that $\phi_{ij} = \begin{cases} 1 \text{ if } j = k, \\ 0 \text{ otherwise.} \end{cases}$ As a consequence, $0 \le \mathcal{L} \le g_k = 0$. On the other hand, $\mathcal{L} > 0$ if $f_j = 1/m_f$ and $g_j = 1/m_g$ for each possible j. It contradicts the assumption that $\mathcal{L}$ reaches its maximum value at a point with $g_k = 0$. This completes the proof of the lemma. □

Proof of the self-consistent equations in (6) *and* (7) Fix a k. For each j, replace f_j in $\mathcal{L}$ (see (5)) by

$$f_j^w = \begin{cases} \frac{f_k + w}{1+w} \left(= \frac{f_k - 1}{1+w} + 1 \right) & \text{if } j = k \\ \frac{f_j}{1+w} & \text{otherwise} \end{cases} = \frac{f_j}{1+w} + \mathbf{1}_{(j=k)} \left(1 - \frac{1}{1+w} \right). \quad (13)$$

Then $\mathcal{L}$ is a function of $\mathbf{f}^w$ and $\mathbf{g}$, denoted by $\mathcal{L}(\mathbf{f}^w, \mathbf{g})$, where $\mathbf{f}^w = (f_1^w, \ldots, f_{m_f}^w)$ and $\mathbf{g} = (g_1, \ldots, g_{m_g})$. It suffices to take derivative of $\mathcal{L}(\mathbf{f}^w, \mathbf{g})$ with respect to w. Note that

$$\frac{\partial f_j^w}{\mathrm{d}w} = \frac{\mathbf{1}_{(j=k)} - f_j}{(1+w)^2}.$$

Taking derivative of $\ln\mathcal{L}$ with respect to w yields

$$\begin{aligned} \left. \frac{\partial \ln\mathcal{L}}{\partial w} \right|_{w=0} &= \sum_{i \notin K} \frac{\psi_{ik} - \sum_{j=1}^{m_f} \psi_{ij} f_j}{\sum_{h=1}^{m_f} \psi_{ih} f_h} + n_1 \frac{\sum_{j=1}^{m_g} g_j \left(\mathbf{1}_{(t_k > c_j - \beta \mathbf{x}_j)} - \sum_{h:\ t_h > c_j - \beta \mathbf{x}_j} f_h \right)}{\sum_{j=1}^{m_g} g_j \sum_{m:\ t_m > c_j - \beta \mathbf{x}_j} f_m} \\ &= \sum_{i \notin K} \left(\frac{\psi_{ik}}{\sum_{h=1}^{m_f} \psi_{ih} f_h} - 1 \right) + n_1 \left(\frac{\sum_{j=1}^{m_g} g_j \mathbf{1}_{(t_k > c_j - \beta \mathbf{x}_j)}}{\sum_{j=1}^{m_g} g_j \sum_{m:\ t_m > c_j - \beta \mathbf{x}_j} f_m} - 1 \right), \end{aligned} \quad (14)$$

$k = 1, \ldots, m_f$.

It follows from Lemma 1 that there is a unique maximum point of $\mathcal{L}$ and it is an interior point. Thus, if $\mathcal{L}$ reaches it maximum at a point $\mathbf{u} = (f_2, \ldots, f_{m_f}, g_2, \ldots, g_{m_g})$ then it satisfies $\frac{\partial \ln\mathcal{L}(\mathbf{f}^w, \mathbf{g})}{\partial w}\big|_{w=0} = 0$ at that point. Setting $\frac{\partial \ln\mathcal{L}}{\partial w}\big|_{w=0} = 0$ yields

$$\sum_{i \notin K} \frac{\psi_{ik}}{\sum_{h=1}^{m_f} \psi_{ih} f_h} + n_1 \frac{\sum_{j=1}^{m_g} g_j \mathbf{1}_{(t_k > c_j - \beta \mathbf{x}_j)}}{\sum_{j=1}^{m_g} g_j \sum_{m:\ t_m > c_j - \beta \mathbf{x}_j} f_m} = n.$$

Multiplying f_k/n on both side of the foregoing equation yields the self-consistent equations in (6):

$$f_k = \frac{1}{n}\sum_{i\notin K}\frac{\psi_{ik}f_k}{\sum_{h=1}^{m_f}\psi_{ih}f_h} + \frac{n_1 f_k}{n}\frac{\sum_{j=1}^{m_g} g_j \mathbf{1}_{(t_k>c_j-\beta\mathbf{x}_j)}}{\sum_{j=1}^{m_g} g_j \sum_{m:\ t_m>c_j-\beta\mathbf{x}_j} f_m}.$$

The derivation of (7) is similar. Fix a k and replace g_j in $\mathcal{L}$ (see (5)) by

$$g_j^w = \begin{cases} \frac{g_k+w}{1+w} \left(= \frac{g_k-1}{1+w}+1\right) & \text{if } j=k \\ \frac{g_j}{1+w} & \text{otherwise} \end{cases} = \frac{g_j}{1+w} + 1_{(j=k)}\left(1-\frac{1}{1+w}\right). \quad (15)$$

Let $\mathbf{f} = (f_1,\ldots,f_{m_f})$ and $\mathbf{g}^w = (g_1^w,\ldots,g_{m_g}^w)$. Taking derivative of $\ln\mathcal{L}$ with respect to w yields

$$\begin{aligned}\left.\frac{\partial \ln\mathcal{L}}{\partial w}\right|_{w=0} &= \sum_{i\notin K}\frac{\phi_{ik}-\sum_{j=1}^{m_g}\phi_{ij}g_j}{\sum_{h=1}^{m_g}\phi_{ih}g_h} + n_1\frac{-\sum_{j=1}^{m_g}(g_j-\mathbf{1}_{(j=k)})\sum_{h:\ t_h>c_j-\beta\mathbf{x}_j} f_h}{\sum_{j=1}^{m_g} g_j\sum_{m:\ t_m>c_j-\beta\mathbf{x}_j} f_m} \\ &= \sum_{i\notin K}\frac{\phi_{ik}}{\sum_{h=1}^{m_g}\phi_{ih}g_h} + n_1\frac{\sum_{h:\ t_h>c_k-\beta\mathbf{x}_k} f_h}{\sum_{j=1}^{m_g} g_j\sum_{m:\ t_m>c_j-\beta\mathbf{x}_j} f_m} - n. \quad (16)\end{aligned}$$

It follows from Lemma 1 that there is a unique maximum point of $\mathcal{L}$ and it is an interior point. Thus if $\mathcal{L}$ reaches it maximum at a point $\mathbf{u} = (f_2,\ldots,f_{m_f}, g_2,\ldots,g_{m_g})$, then it satisfies that $\frac{\partial \ln\mathcal{L}(\mathbf{f},\mathbf{g}^w)}{\partial w}\big|_{w=0} = 0$ at that point. Setting $\frac{\partial \ln\mathcal{L}(\mathbf{f},\mathbf{g}^w)}{\partial w}\big|_{w=0} = 0$ yields

$$1 = \frac{1}{n}\sum_{i\notin K}\frac{\phi_{ik}}{\sum_{h=1}^{m_g}\phi_{ih}g_h} + \frac{n_1}{n}\frac{\sum_{h:\ t_h>c_k-\beta\mathbf{x}_k} f_h}{\sum_{j=1}^{m_g} g_j\sum_{m:\ t_m>c_j-\beta\mathbf{x}_j} f_m}.$$

It follows that

$$g_k = \frac{1}{n}\sum_{i\notin K}\frac{\phi_{ik}g_k}{\sum_{h=1}^{m_g}\phi_{ih}g_h} + \frac{n_1 g_k}{n}\frac{\sum_{h:\ t_h>c_k-\beta\mathbf{x}_k} f_h}{\sum_{j=1}^{m_g} g_j\sum_{m:\ t_m>c_j-\beta\mathbf{x}_j} f_m}.$$

Proof of Proposition 1 If the self-consistent algorithm converges, the solution satisfies the self-consistent equations (6) and (7). Thus it is an SCE.

We shall assume that the proposition is false, that is, the $\mathcal{L}$ reaches its maximum point at the boundary of D, that is $f_k = 0$ or $g_k = 0$, and then it leads to a

contradiction. Verify that

$$\frac{1}{n}\sum_{i\notin K}\frac{\psi_{ik}}{\sum_{h=1}^{m_f}\psi_{ih}f_h}+\frac{n_1}{n}\frac{\sum_{j=1}^{m_g}g_j\mathbf{1}_{(t_k>c_j-\beta\mathbf{x}_j)}}{\sum_{j=1}^{m_g}g_j\sum_{m:\ t_m>c_j-\beta\mathbf{x}_j}f_m}>0,$$
$$\frac{1}{n}\sum_{i\notin K}\frac{\phi_{ik}}{\sum_{h=1}^{m_g}\phi_{ih}g_h}+\frac{n_1}{n}\frac{\sum_{h:\ t_h>c_k-\beta\mathbf{x}_k}f_h}{\sum_{j=1}^{m_g}g_j\sum_{m:\ t_m>c_j-\beta\mathbf{x}_j}f_m}>0,$$

whenever $\mathbf{f}\in D_f$ and $\mathbf{g}\in D_g$, as all the coefficients in the expressions are nonnegative. Since the initial point $f_{k,1}=\frac{1}{m_f}>0$ and $g_{k,1}=\frac{1}{m_g}>0$

$$f_{k,\nu-1}>0\quad\text{and}\quad g_{k,\nu-1}>0\text{ for }\nu\geq 2. \tag{17}$$

In view of (5), one can write $\mathcal{L}=\mathcal{L}(\mathbf{f},\mathbf{g})$, where $\mathbf{f}=(f_1,\ldots,f_{m_f})$ and $\mathbf{g}=(g_1,\ldots,g_{m_g})$. By the three statements in Lemma 1 on the boundary of D

$$\frac{\partial\ln\mathcal{L}(\mathbf{f}^w,\mathbf{g})}{\partial w}|_{w=0}>0\quad\text{and}\quad\frac{\partial\ln\mathcal{L}(\mathbf{f},\mathbf{g}^w)}{\partial w}|_{w=0}>0,$$

where $\mathbf{f}^w=(f_1^w,\ldots,f_{m_f}^w)$ (see (13)) and $\mathbf{g}^w=(g_1^w,\ldots,g_{m_g}^w)$(see (15)). It follows from the foregoing inequalities, Eqs. (14) and (16) that

$$\frac{1}{n}\sum_{i\notin K}\frac{\psi_{ik}}{\sum_{h=1}^{m_f}\psi_{ih}f_h}+\frac{n_1}{n}\frac{\sum_{j=1}^{m_g}g_j\mathbf{1}_{(t_k>c_j-\beta\mathbf{x}_j)}}{\sum_{j=1}^{m_g}g_j\sum_{m:\ t_m>c_j-\beta\mathbf{x}_j}f_m}>1,$$
$$\frac{1}{n}\sum_{i\notin K}\frac{\phi_{ik}}{\sum_{h=1}^{m_g}\phi_{ih}g_h}+\frac{n_1}{n}\frac{\sum_{h:\ t_h>c_k-\beta\mathbf{x}_k}f_h}{\sum_{j=1}^{m_g}g_j\sum_{m:\ t_m>c_j-\beta\mathbf{x}_j}f_m}>1.$$

Thus if $(f_{2,\nu-1},\ldots,f_{m_f,\nu-1},g_{2,\nu-1},\ldots,g_{m_g,\nu-1})$ belongs to the neighborhood of the boundary of D, then the foregoing inequalities together with (9) and (10) yield

$$f_{k,\nu}=\frac{1}{n}\sum_{i\notin K}\frac{\psi_{ik}f_{k,\nu-1}}{\sum_{h=1}^{m_f}\psi_{ih}f_{h,\nu-1}}$$
$$+\frac{n_1}{n}\frac{\sum_{j=1}^{m_g}g_{j,\nu-1}\mathbf{1}_{(t_k>c_j-\mathbf{b}\mathbf{x}_j)}f_{k,\nu-1}}{\sum_{j=1}^{m_g}g_{j,\nu-1}\sum_{m:\ t_m>c_j-\mathbf{b}\mathbf{x}_j}f_{m,\nu-1}}>f_{k,\nu-1}>0,$$

by (17);

$$g_{k,\nu}=\frac{1}{n}\sum_{i\notin K}\frac{\phi_{ik}g_{k,\nu-1}}{\sum_{h=1}^{m_g}\phi_{ih}g_{h,\nu-1}}+\frac{n_1}{n}\frac{\sum_{h:\ t_h>c_k-\mathbf{b}\mathbf{x}_k}f_{h,\nu-1}g_{k,\nu-1}}{\sum_{j=1}^{m_g}g_{j,\nu-1}\sum_{m:\ t_m>c_j-\mathbf{b}\mathbf{x}_j}f_{m,\nu-1}}>g_{k,\nu-1}>0$$

by (17). Thus it will not converge to 0, contradicting the assumption that it converges to 0. It follows that the SCE must be the GMLE too. Moreover, the GMLE $\hat{f}_k, \hat{g}_k > 0$ for each k. □

Remark 2 We'd like to point out that if the initial step in the self-consistent algorithm does not satisfy that $f_{k,1} > 0$ and $g_{k,1} > 0$, then an SCE may not be a GMLE, that is, the conclusion in the proposition no longer holds.

7.1 Proof of Theorem 1

WLOG, we can assume that $F_o = F_*$ and $F_{C,\mathbf{X}} = F^*_{C,\mathbf{X}}$.

By assumption A3 there are finitely many distinct values of $\mathbf{O}$. By taking n large enough, without loss of generality (WLOG), we can assume that the first m $\mathbf{O}_i$'s are all the possible distinct ones, and $(\eta_1, \delta_1) = (0, 0)$ (i.e., $1 \in K$). Let $N_j = \sum_{i=1}^n \mathbf{1}(\mathbf{O}_i = \mathbf{O}_j)$. Let $Q(\mathbf{O}_j)$ be the probability of the event $\mathbf{O} = \mathbf{O}_j$, corresponding to an arbitrary pair (F, G) and $Q_o(\mathbf{O}_j)$ corresponding to $(F_o, F_{C,\mathbf{X}})$. Let $\hat{Q}$ be the GMLE induced by $(\hat{F}_\beta, \hat{F}_{C,\mathbf{X},\beta})$. By taking a subsequence and by A3, WLOG, we can assume that $\lim_{n\to\infty} \hat{Q}(\mathbf{O}_j) = Q^*(\mathbf{O}_j)$ for each j, where Q^* is again a probability measure.

The normalized log likelihood is

$$\mathcal{N}(Q) = \frac{1}{n}\ln\mathcal{L} = \tfrac{1}{n}\sum_{j=1}^{m} N_j \ln Q(\mathbf{O}_j).$$

By the strong law of large number, with probability one, we have

$$\lim_{n\to\infty} \mathcal{N}(Q_o) \to E(\mathcal{N}(Q_o)).$$

Since $\hat{Q}$ is the GMLE,

$$\mathcal{N}(\hat{Q}) \geq \mathcal{N}(Q_o).$$

It follows that

$$\underline{\lim}_{n\to\infty} \mathcal{N}(\hat{Q}) \geq \lim_{n\to\infty} \mathcal{N}(Q_o) = E(\mathcal{N}(Q_o)).$$

By Fatou's lemma, with probability one,

$$\overline{\lim}_{n\to\infty} \mathcal{N}(\hat{Q}) \leq \sum_{i=1}^{m} Q_o(\mathbf{O}_j) \ln \overline{\lim}_{n\to\infty} \hat{Q}(\mathbf{O}_j) \ \left(= E\left(\mathcal{N}\left(\overline{\lim}_{n\to\infty} \hat{Q}\right)\right)\right).$$

That is, with probability one,

$$E(\mathcal{N}(Q_o)) \leq E\left(\mathcal{N}\left(\overline{\lim}_{n\to\infty} \hat{Q}\right)\right).$$

By our assumption in the first paragraph, $Q^* = \overline{\lim_{n\to\infty}} \hat{Q}$ is again a probability measure. As a consequence, with probability one,

$$E(\mathcal{N}(Q_o)) \le E\left(\mathcal{N}\left(\overline{\lim_{n\to\infty}} \hat{Q}\right)\right) = E(\mathcal{N}(Q^*)).$$

It follows from the Shannon–Kolmogorov inequality that $E(\mathcal{N}(Q)) < E(\mathcal{N}(Q_o))$ for all probability measure $Q \neq Q_o$. As a consequence

$$\lim_{n\to\infty} \hat{Q} = Q^* = Q_o.$$

Let Q be the probability measure corresponding to $\mathbf{s}\,(= (f_2, \ldots, f_{m_f}, g_2, \ldots, g_{m_g})$. We shall show in Lemma 2 that $\mathbf{s}^o = (f_2^o, \ldots, f_{m_f}^o, g_2^o, \ldots, g_{m_g}^o)$ uniquely maximizes $E(\mathcal{N}(Q))$ over all $\mathbf{s}$. Thus $Q^* = Q_o$ implies that $\mathbf{s} = \mathbf{s}^o$. This completes the proof of the theorem. □

Lemma 2 $\mathbf{s}^o = (f_2^o, \ldots, f_{m_f}^o, g_2^o, \ldots, g_{m_g}^o)$ *uniquely maximizes* $E(\mathcal{N}(Q))$ *over all* $\mathbf{s}$.

Let D_f, D_g and D be defined as in the proof of Proposition 1. Verify that

$$E(\mathcal{N}(Q)) = \sum_{i\notin K, i\le m} Q_o(\mathbf{O}_i) \ln \sum_{j=1}^{m_f} \psi_{ij} f_j + Q_o(\mathbf{O}_1) \ln \sum_{j=1}^{m_g} g_j \sum_{k: t_k > c_j - \beta \mathbf{x}_j} f_k$$
$$+ \sum_{i\notin K, i\le m} Q_o(\mathbf{O}_i) \ln \sum_{j=1}^{m_g} \phi_{ij} g_j.$$

Moreover, verify that (1) for each fixed $(g_2, \ldots, g_{m_g})$, $-E(\mathcal{N}(Q))$ is strictly convex on D_f; (2) for each fixed $(f_2, \ldots, f_{m_f})$, $-E(\mathcal{N}(Q))$ is strictly convex on D_g. Thus either (a) there is unique maximum point of $E(\mathcal{N}(Q))$ and it is an interior point, or (b) the maximum point is on the boundary of D, with $f_k = 0$ or $g_k = 0$ for some k.

However, case (b) is impossible, otherwise, $E(\mathcal{N}(Q)) = -\infty$, which can be viewed as follows. For each k, there is an observation B_i such that either $\delta_i = 1$ and $T_i(\beta) = t_k$ or $\delta_i = 0$ and $(T_i(\beta), \infty)$ is an MI induced by $B_1, \ldots, B_n$. In either case, there is an i such that $\psi_{ij} = \begin{cases} 1 \text{ if } j = k \\ 0 \text{ otherwise} \end{cases}$. Similarly, one can verify that there is always an i such that $\phi_{ij} = \begin{cases} 1 \text{ if } j = k \\ 0 \text{ otherwise} \end{cases}$. As a consequence,

$-\infty \le E(\mathcal{N}(Q)) \le Q_o(\mathbf{O}_i) \ln f_k = -\infty$ (here we define $\ln 0 = -\infty$), or
$-\infty \le E(\mathcal{N}(Q)) \le Q_o(\mathbf{O}_i) \ln g_k = -\infty$.

It follows that $E(\mathcal{N}(Q)) = -\infty$. On the other hand, since $Q_o(\mathbf{O}_j) > 0$ for each j by A3, $E(\mathcal{N}(Q_o)) > -\infty$. Thus a point on the boundary of D cannot be a maximum point of $E(\mathcal{N}(Q))$. □

Remark 3 The consistency proof actually can be extended to the case that $(T(\beta), \mathbf{X})$ takes on countably many values. However, under assumption A3, the proof is shorter.

7.2 Proof of Theorem 2

WLOG, we can assume that $F_o = F_*$ and $F_{C,\mathbf{X}} = F^*_{C,\mathbf{X}}$.

Recall $\mathcal{N} = \frac{1}{n}\ln\mathcal{L}(Q)$. For convenience, denote $f^o_k = f_o(t_k)$ and $g^o_k = f_{C,\mathbf{X}}(c_k, \mathbf{x}_k)$. Let $\mathbf{s} = (s_1, \ldots, s_{m_f+m_g-2})' = (g_2, \ldots, g_{m_g}, f_2, \ldots, f_{m_f})'$, corresponding to a probability measure Q, $\mathbf{s}^o = (s^o_1, \ldots, s^o_{m_f+m_g-2})' = (g^o_2, \ldots, g^o_{m_g}, f^o_2, \ldots, f^o_{m_f})'$, corresponding to the probability measure Q_o. Verify that $E(\mathcal{N}(Q))$ is a function of $\mathbf{s}$, say $l(\mathbf{s}) = E(\mathcal{N}(Q))$. Let

$$J = \left(\frac{\partial^2 E(\mathcal{N}(Q))}{\partial\mathbf{s}\partial\mathbf{s}'}\right)\Bigg|_{\mathbf{s}=\mathbf{s}^o}.$$

Under A3, the estimation problem is essentially a multinomial distribution problem and thus it can be verified that

$$\left(\frac{\partial^2 E(\mathcal{N}(Q))}{\partial\mathbf{s}\partial\mathbf{s}'}\right)\Bigg|_{\mathbf{s}=\mathbf{s}^o} = E\left(\frac{\partial^2 \mathcal{N}(Q)}{\partial\mathbf{s}\partial\mathbf{s}'}\right)\Bigg|_{\mathbf{s}=\mathbf{s}^o}.$$

In view of (5), it can be verified that $\frac{\partial^2 \ln\mathcal{L}}{\partial s_i \partial s_j}$ is continuous in $\mathbf{s}$ and thus $\frac{\partial^2 \ln\mathcal{L}}{\partial s_j \partial s_i} = \frac{\partial^2 \ln\mathcal{L}}{\partial s_i \partial s_j}$ for each pair (i, j). As a consequence, $\frac{\partial^2 E(\ln\mathcal{L})}{\partial\mathbf{s}\partial\mathbf{s}'}$ is symmetric. That is, $-J = U'\mathrm{Diag}(\lambda_i)U$, where $\mathrm{Diag}(\lambda_i)$ is a $d \times d$ dimensional diagonal matrix with $d = m_f + m_g - 2$ and diagonal elements $\lambda_1 \geq \cdots \geq \lambda_d$, and $U' = U^{-1}$.

It can be shown by a similar arguement as in Wong and Yu (1999) that $-J$ is actually positive definite. For simplicity, we skip the details.

It is easy to verify that

$$\frac{\partial^2 \mathcal{N}(\hat{Q})}{\partial\mathbf{s}\partial\mathbf{s}'} \to E\left(\frac{\partial^2 \mathcal{N}(\hat{Q})}{\partial\mathbf{s}\partial\mathbf{s}'}\right) = -J.$$

It thus follows that

$$\frac{\partial \mathcal{N}(\hat{Q})}{\partial\mathbf{s}} = \frac{\partial \mathcal{N}(Q_o)}{\partial\mathbf{s}} + J\triangle_n + o_p(||\triangle_n||),$$

where $\triangle_n$ is the d-dimensional column vector with entries $\hat{s}_i - s_i^o$. Let $\Omega_n = \{\inf_{i \le d} \hat{s}_i = 0\}$, Verify that

$$0 = \frac{\partial \mathcal{N}(\hat{Q})}{\partial \mathbf{s}} \quad \text{except on the event } \Omega_n,$$

and by Theorem 1, $\mathrm{P}(\Omega_n) \to 0$ as $n \to \infty$. It follows from the central limit theorem that $\sqrt{n}\frac{\partial \mathcal{N}(\hat{Q})}{\partial \mathbf{s}}$ is asymptotically normal with mean 0 and dispersion matrix J. This shows that $\triangle_n = J^{-1}\frac{\partial \mathcal{N}(Q_o)}{\partial \mathbf{s}} + o_p(n^{-1/2})$ and the theorem is proved. □

Acknowledgements The authors thank the Editor and two referees for helpful comments. Ms. Cuixian also pointed out several types in the paper.

References

Barlow, W. E., Ichikawa, L., Rosner, D., Izumi, S. (1999). Analysis of case-cohort designs. *Journal of Clinical Epidemiology*, *52*, 1165–1172.

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

Chen, H. Y. (2001). Fitting semiparametric transformation regression models to data from a modified case-cohort design. *Biometrika*, *88*, 255–268.

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

Hillis, S. L. (1993). A comparison of 3 Buckley–James variance estimators. *Communication in Statistics*, Series B, *22*, 955–973.

James, I. R., Smith, P. J. (1981). Consistency results for linear regression with censored data. *Annals of Statistics*, *12*, 590–600.

Kalbfleisch, J. D., Prentice, R. L. (2002). *The statistical analysis of failure time data* (2nd Ed.) New York: Wiley.

Kong, L., Cai, J., Sen, P. K. (2004). Weighted estimating equations for semiparametric transformation models with censored data from a case-cohort design. *Biometrika*, *91*, 305–319.

Kulich, M., Lin, D. Y. (2000). Additive hazards regression for case-cohort studies. *Biometrika*, *87*, 73–87.

Lai, T. L., Ying, Z. L. (1991). Large sample theory of a modified Buckley–James estimator for regression-analysis with censored data. *Annals of Statistics*, *19*, 1370–1402.

Lin, D. Y., Ying, Z. (1993). Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association*, *88*, 1341–1349.

Miller, R. G. (1981). *Survival analysis* (Vol. 50). New York: Wiley.

Miller, R. G., Halpern, J. (1982). Regression with censored data. *Biometrika*, *69*, 521–531.

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrics*, *73*, 1–11.

Ritov, Y. (1990). Estimation in a linear regression model with censored data. *Annals of Statistcs*, *18*, 303–328.

Scheike, T. H., Martinussen, T. (2004). Maximum likelihood estimation for Cox's regression model under case-cohort sampling. *Scandinavial Journal of Statistics*, *31*, 283–293.

Self, S., Prentice, R. L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *Annals of Statistics, 16*, 64–81.

Turnbull, B. W. (1976). The empirical distribution function with arbitrary grouped, censored and truncated data. *Journal of Royal Statistics Society*, *Series B, 38*, 290–295.

Wong, G. Y. C., Yu, Q. Q. (1999). Generalized MLE of a joint distribution function with multivariate interval-censored data. *Journal of Multivariate Analysis*, *69*, 155–166.

Yu, Q. Q, Wong, G. Y. C., He, Q. M. (2000). Estimation of a joint distribution function with multivariate interval-censored data when the nonparametric MLE is not unique. *Biometrical Journal*, *42*, 747–763.