International Journal of Statistics and Management System
Vol. 5 No. 1 (January-June, 2020)

Submitted: 4th March 2020 Revised: 23rd April 2020 Accepted: 29th May 2020

Buckely–James–type estimators in classical–cohort studies *

MENGGANG YU† AND QIQING YU ‡

Abstract

We consider the estimation problem with the classical case-cohort data. The classical case–cohort design was first proposed by Prentice (1986). Most studies focus on the Cox regression model. In this paper, we consider the censored linear regression model. We propose several simple estimators which extend the Buckley–James estimator to the classical case-cohort design. We further carry out simulation studies to compare the asymptotic properties of these simple estimators under different sample sizes, underlying distributions and various subcohort sizes. We also perform data analysis to a real data set and compare to existing results in the literature. A proof of the consistency and asymptotic normality is given in Appendix under some simple regularity conditions.

1. Introduction We consider the estimation problem under the classical case-cohort designs and censored linear regression models. Many epidemiological cohort studies and 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 expensive to collect covariates of interest which might involve, for example, biochemical analysis of specimens. In order to lessen this burden

---

Key words and phrases: Case–cohort study, Buckley–James estimator, right–censorship, linear regression model, survival data.

AMS 2000 subject classifications. Primary 62J05; secondary 62G05.

†*Mailing Address:* Department of Medicine/Biostatistics, Indiana University, Indianapolis, IN 46202, USA

‡*Mailing Address:* Department of Mathematical Sciences, SUNY, Binghamton, NY 13902, USA.

without much loss of efficiency, Prentice (1986) proposed the classical case-cohort design, under which, one observes all covariates for each subject experiencing an event and for each from a random sub-sample of the cohort, selected at the beginning of the study (call a subcohort). The classical case-cohort design does not recorded any survival information for censored patients outside subcohort. This needs to be contrasted with the modified case-cohort design (Chen, 2001) under which the censoring times of all censored subjects in the cohort are observed. For some literature review of case-cohort designs, we refer to Yu, Wong and Yu (2005) and Yu and Yu (2006).

Among the four regression techniques for censored data, Miller and Halpern (1982) concluded that the Cox and the Buckley-James estimators are "two most reliable regression estimates" and that "the choice between them should depend on the appropriateness of the proportional hazards model or the linear model for the data." While Cox's model has been studied in the case-cohort designs, the Buckley-James-type of estimator has not been investigated until recently.

The estimation problem under the censored linear regression models with the case-cohort designs can be formulated as follows. Let Y_i and C_i be monotonically transformed failure and censoring 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). For subject i in the full cohort, let $M_i \equiv Y_i \wedge C_i$ and $\delta_i \equiv \mathbf{1}(Y_i \leq C_i)$. Let $\mathbf{X}_i$ be a vector of p-dimensional covariates. The model is $Y_i = \beta'\mathbf{X}_i + \epsilon_i$, $i = 1$, ..., 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). If either subject i is in the subcohort or the event of interest has taken place, then we observe $(M_i, \delta_i, \mathbf{X}_i)$. Otherwise, we do not observe both M_i and $\mathbf{X}_i$ in the classical design and only observe M_i in the modified design.

Yu, Wong and Yu (2005) propose an extension of the Buckley-James estimator (BJE) under the censored linear regression model and the classical case-cohort design. The BJE depends on the estimators of the underlying distributions. They propose to estimate the underlying distribution functions by the generalized maximum likelihood estimator (GMLE) and propose a self-consistent algorithm for the GMLE. No proof of the asymptotic distribution of the BJE is given. While the BJE is obtained by an iterative algorithm, the GMLE depends on the regression parameter and can only be obtained iteratively. Thus their approach is time consuming.

We shall find simpler extensions of the BJE under the classical case-cohort design setting and propose an algorithm for finding such estimates. For convenience, we call all of them the BJE's. We study them via simulation using different underlying distributions. We also construct a proof of asymptotic properties of a BJE under some simple regularity conditions. The idea in our proof can also be used to establish the asymptotic properties of the estimator studied in Yu, Wong and Yu (2006). In an unpublished manuscript, Yu and Yu (2006) also study this estimation problem under the modified case-cohort design. The estimators and the proofs of its asymptotic properties under the modified case-cohort design have some subtle differences from those under the classical case-cohort design due to different model assumptions.

The asymptotic properties of various extensions of the BJE under general continuity assumptions and under both case-cohort designs remain an outstanding problem. This applies to the BJE based on the GMLE, as well as to the BJE based on the simple estimators of the underlying distribution functions. In this paper, we are only able to construct a proof of the asymptotic properties of the extension of the BJE under a simple discrete assumption on the underlying distributions.

The paper is organized as follows. Section 2 introduces how to extend the BJE to the classical case-cohort design. Section 3 discusses various ways of estimating the underlying distribution functions which are needed in the extension of the BJE. Section 4 introduces an algorithm for the BJE and Section 5 discusses how to estimate the covariance matrix. Section 6 presents some simulation results on comparison of the BJE-type estimators and Section 7 presents a data analysis with a real data set. A proof of consistency and asymptotic normality is given in Appendix.

2. Buckley-James Estimators. In this section we propose a way of using the Buckley-James-type of estimators in analyzing classical case-cohort data. Thus, in additional to the random variables introduced for the usual censored linear regression model, we need to introduce a subcohort indicating random variable η_i such that $\eta_i = 1$ if subject i is selected to be in the subcohort and 0 otherwise. Let $T_i = T_i(\mathbf{b}) = M_i - \mathbf{b}\mathbf{X}_i$. Let $(M_i, \delta_i, \mathbf{X}_i, C_i, \epsilon_i, T_i, \eta_i)$, $i = 1, \ldots, n$, be i.i.d. copies of $(M, \delta, \mathbf{X}, C, \epsilon, T, \eta)$. Hereafter, unless we point out, by case-cohort data, we mean the data from the classical case-cohort design. If an individual either is in a subcohort or experiences the event of interest, that is, either $\eta_i = 1$ or $\delta_i = 1$, then $\mathbf{X}_i$ is measured and thus observed. Otherwise, $(M_i, \mathbf{X}_i)$ is missing. It is shown (see *e.g.*, Li and Pu (1999) or Appendix II) that the BJE is somewhat a zero point of a modification of the score function with the censored linear regression data making use of the assumption, $\epsilon \sim N(\mu, \sigma^2)$.

We shall assume that

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

The identifiability assumption made under the simple linear regression with complete data is $P\{\mathbf{X}_1 \neq \mathbf{X}_2\} > 0$, where $\mathbf{X}_1$, $\mathbf{X}_2$, ..., are i.i.d. copies of $\mathbf{X}$. Under our set-up, it becomes,

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

Note that Scheike and Martinussen (2004) replace A1 by the following assumption.

A3. The random variables ϵ, $\mathbf{X}$ and C are independent.

We also make use of the following regularity condition.

A4. $\mathrm{P}\{\epsilon + \beta'\mathbf{X} = C\} = 0$.

It is worth mentioning that under the assumption that all underlying distribution functions are continuous, assumption A4 is satisfied automatically, but not so otherwise. Kong and Yu (2005) construct an example that the BJE is not asymptotically normally distributed under the full cohort censored linear regression model. In general, we assume F_o is arbitrary, which may or may not be continuous.

Notice that the full nonparametric likelihood function at $\mathbf{b} = \beta$ is

$$\begin{aligned}\mathcal{L} = \quad & \prod_{i=1}^n \Big\{f_o(T_i(\beta)) \int_{c \geq M_i, \mathbf{X}=\mathbf{X}_i} dF_{C,\mathbf{X}}(c, \mathbf{x})\Big\}^{\delta_i} \Big\{S_o(T_i(\beta)) f_{C,\mathbf{X}}(M_i, \mathbf{X}_i)\Big\}^{(1-\delta_i)\eta_i} \\ & \times \prod_{i=1}^n \Big\{ \int_{c \in \mathcal{R}, \mathbf{X} \in \mathcal{R}^p} S_o(c - \beta\mathbf{x}) dF_{C,\mathbf{X}}(c, \mathbf{x})\Big\}^{(1-\eta_i)(1-\delta_i)} \times \prod_{i=1}^n q^{\eta_i}(1-q)^{1-\eta_i},\end{aligned} \tag{2.1}$$

where $q = P\{\eta = 1\}$ and $(M_i, \mathbf{X}_i, T_i)'s$ are realizations, $F_{C,\mathbf{X}}$ and $f_{C,\mathbf{X}}$ are the cdf and the density function of $(C, \mathbf{X})$. Yu, Wong and Yu (2005) show that after centering to an estimate of μ_x $(= E(\mathbf{X}))$, say $\widetilde{\mathbf{X}}$, the score function is

$$\begin{aligned}\frac{\partial \ln \mathcal{L}}{\partial \beta} = \quad & \sigma^{-2} \sum_{i \notin K} \Big\{\delta_i T_i(\beta) - (1-\delta_i) \frac{\int_{t > T_i(\beta)} t dS_o(t)}{S_o(T_i(\beta))}\Big\}(\mathbf{X}_i - \widetilde{\mathbf{X}}) \\ & -\sigma^{-2} \sum_{i \in K} \frac{\sum_{c,\mathbf{X}} \int_{t > c - \beta\mathbf{X}} t dS_o(t) f_{C,\mathbf{X}}(c,\mathbf{x})(\mathbf{x} - \widetilde{\mathbf{X}})}{\sum_{c,\mathbf{X}} S_o(c - \beta\mathbf{X}) f_{C,\mathbf{X}}(c,\mathbf{x})} \\ \stackrel{def}{=} \quad & \sigma^{-2} H(\beta, S_o, f_{C,\mathbf{X}}, \widetilde{\mathbf{X}}),\end{aligned} \tag{2.2}$$

where $K = \{i : \ \eta_i + \delta_i = 0\}$.

Mimicking the BJE under the full cohort studies (Buckley and James, 1979), we should now replace S_o, $f_{C,\mathbf{X}}$ in Eq.(2.2) by proper estimators, say $\tilde{S}_\beta$ and $\tilde{f}_{C,\mathbf{X},\beta}$ to obtain an estimating equations. Of course, they all depend on the unknown true parameter β. However for each given $\mathbf{b}$, one can find these estimators pretending $\beta = \mathbf{b}$. A BJE may be defined to be a root of $\widetilde{H}$, where $\widetilde{H}(\mathbf{b}) = H(\mathbf{b}, \tilde{S}_{\mathbf{b}}, \tilde{f}_{C,\mathbf{X},\mathbf{b}}, \widetilde{\mathbf{X}})$. Similar to the full cohort case, $\widetilde{H}(\mathbf{b})$ may not have a root, then a BJE is a point at which $\widetilde{H}(\mathbf{b})$ changes its sign (called a *zero-crossing*) (see James and Smith, 1984). Noticing that for any subject $i \in K$, the last term in (2.2) is the same, we can express our estimating equation as

$$\begin{aligned}\widetilde{H}(\mathbf{b}) = & \sum_{i \notin K} \left\{ \delta_i T_i(\mathbf{b}) - (1-\delta_i) \frac{\int_{t > T_i(\mathbf{b})} t d\tilde{S}_{\mathbf{b}}(t)}{\tilde{S}_{\mathbf{b}}(T_i(\mathbf{b}))} \right\} (\mathbf{X}_i - \widetilde{\mathbf{X}}) \\ & - n_K \frac{\sum_{j,k \notin K} \sum_{t > M_j - \mathbf{b}\mathbf{X}_k} t d\tilde{S}_{\mathbf{b}}(t) \tilde{f}_{C,\mathbf{X},\mathbf{b}}(c, \mathbf{X}_k)(\mathbf{X}_k - \widetilde{\mathbf{X}})}{\sum_{j,k \notin K} \tilde{S}_{\mathbf{b}}(M_j - \mathbf{b}\mathbf{X}_k) \tilde{f}_{C,\mathbf{X},\mathbf{b}}(c, \mathbf{X}_k)},\end{aligned} \tag{2.3}$$

where n_K is the number of subjects in K. Careful readers might notice that in (2.3), we also need to come up with a $\widetilde{\mathbf{X}}$ to estimate $E(\mathbf{X})$. We can use $\widetilde{\mathbf{X}} = \sum_{c,\mathbf{X}} \mathbf{x} \tilde{f}_{C,\mathbf{X}}(c, \mathbf{x})$ or use either the mean of $\mathbf{X}$ among subcohort subjects (i.e. among subjects with $\eta = 1$) or we can use all observed $\mathbf{X}$, then a sensible estimate for the latter is $\widetilde{\mathbf{X}} = n^{-1} \sum_{i=1}^n (\delta_i x_i + \eta_i (1-\delta_i) x_i * n/n_1)$, where n_1 is the size of subcohort. This idea of weighting any censored subject in the subcohort by ratio of full cohort size over subcohort size is natural and also crucial for providing sensible estimates of S_o and $f_{C,\mathbf{X}}$ in Section 3.

Notice that for full cohort studies (i.e. $K = \emptyset$), after replacing $\widetilde{\mathbf{X}}$ by $\overline{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i$ and replacing S_o by the PLE $\hat{S}_{\mathbf{b}}$ based on $(T_i(\mathbf{b}), \delta_i)$'s, we have the original Buckley-James estimating equation

$$\hat{H}(\mathbf{b}) = \sum_{i=1}^n \left\{ \delta_i T_i(\mathbf{b}) - (1-\delta_i) \frac{\int_{t > T_i(\mathbf{b})} t d\hat{S}_{\mathbf{b}}(t)}{\hat{S}_{\mathbf{b}}(T_i(\mathbf{b}))} \right\} (\mathbf{X}_i - \overline{\mathbf{X}}). \tag{2.4}$$

Remark 1. A naive approach is to construct the BJE based on the subcohort data alone. Since the subcohort follows the linear regression model with right-censored data, the asymptotic properties of the BJE is well established, though the BJE is not efficient, because it does not utilize the information not contained in the subcohort. We shall call this estimator *Subcohort BJE* or $\hat{\beta}_1$ in the simulation results of Section 6.

Remark 2. A second naive approach for extension of the BJE is to utilize partial likelihood, or $\widetilde{\mathrm{L}} = \prod_{i \notin K} \left\{ f(T_i(\mathbf{b}))^{\delta_i} S(T_i(\mathbf{b}))^{(1-\delta_i)} \right\}$. Then, a solution is the zero-crossing of $\check{H}(\mathbf{b}) =$

$\sum_{i\notin K}(\mathbf{X}_i - \widetilde{\mathbf{X}})T_i^*(\mathbf{b})$, where $\widetilde{\mathbf{X}} = \sum_{i\notin K}\mathbf{X}_i/n_2$, n_2 is the number of observations with $\eta_i + \delta_i \geq 1$, and $T_i^*(\mathbf{b}) = \begin{cases} T_i(\mathbf{b}) & \text{if (1) } \delta_i = 1 \\ \frac{\sum_{t>T_i(\mathbf{b})} t\hat{f}_{\mathbf{b}}(t)}{\hat{S}_{\mathbf{b}}(T_i(\mathbf{b}))} & \text{if } \delta_i = 0, \eta_i = 1 \end{cases}$. Here $\hat{S}_{\mathbf{b}}$ is the PLE of S based on $\{(T_i^*(\mathbf{b}), \delta_i) : \delta_i + \eta_i \geq 1\}$ and $\hat{f}_{\mathbf{b}}(t) = \hat{S}_{\mathbf{b}}(t-) - \hat{S}_{\mathbf{b}}(t)$. We refer this approach as partial likelihood BJE $\hat{\beta}_2$ in the simulation section. The simulation results suggest that the estimator is not consistent. Verify that the expression $\check{H}$ in Remark 2 is the same as $\hat{H}$ in (2.3).

3. Estimation of Underlying Distributions. A logical approach for estimating S_o and $f_{C,\mathbf{X}}$ is to find the generalized maximum likelihood estimator (GMLE). That is, an estimate of $(S_o, f_{C,\mathbf{X}})$ that maximizes $\mathcal{L}$ in (2.1) over all possible estimates, with given $\mathbf{b}$. Yu, Wong and Yu (2006) study the problem. There is no closed form solution for the GMLE and they propose a numerical algorithm to find it. Their approach may not be feasible for large sample sizes, as the number of parameters may be too large, especially when all the time points are distinct.

We shall now discuss how to estimate S_o and $f_{C,\mathbf{X}}$ with a simple approach which has an explicit expression. To further simplify (2.2), as in Scheike and Martinussen (2004), we assume A3. Under this assumption, we have $f_{C,\mathbf{X}} = f_C f_{\mathbf{X}}$. The advantage is that we only need to estimate the univariate $f_{\mathbf{X}}$ and f_C instead of the bivariate $f_{C,\mathbf{X}}$.

A desirable estimator of S_o and $f_{\mathbf{X}}$ should utilize all available data. A quick way to do this is to estimate $f_{\mathbf{X}}$ using empirical density function and to estimate S_o and F_C using Kaplan-Meier method based on $(T_i(\mathbf{b}), \delta_i), i \notin K$. By reordering, without loss of generality (WLOG), assume that the first n_2 observations do not have missing $\mathbf{X}$. Then $\hat{S}_{\mathbf{b}}(t) = \prod_{T_{(j)}(\mathbf{b})\leq t}\left\{1 - \frac{\delta_{(j)}}{n_2-j+1}\right\}$, where $T_{(1)}(\mathbf{b}) \leq \cdots \leq T_{(n_2)}(\mathbf{b})$ are order statistics of $T_1(\mathbf{b}), ..., T_{n_2}(\mathbf{b})$, $\delta_{(j)}$ is the δ_i associated with $T_{(j)}(\mathbf{b})$; $\hat{f}_{\mathbf{X}}(\mathbf{x}) = \sum_{i\notin K}\mathbf{1}(\mathbf{X}_i = \mathbf{x})/n_2$. Moreover, $\hat{F}_C(c) = 1 - \prod_{M_{(i)}\leq c}\{1 - \frac{1-\delta_{(i)}}{n_2-i+1}\}$, where $M_{(1)} \leq M_{(2)} \leq \cdots \leq M_{(n_2)}$ are order statistics of M_j's. We shall call the extension of the BJE resulted from such $\hat{S}_{\mathbf{b}}$ and $\hat{f}_{\mathbf{X}}$ *naive BJE* or $\hat{\beta}_3$ in our simulation results.

However due to the fact that we are observing $\mathbf{X}$ only if $\delta + \eta > 0$, hence the chance of non-missing for failure group is higher than that for censored group. Naive way of estimating S_o, F_C and $f_{\mathbf{X}}$ without taking this fact into consideration will likely result in inconsistent estimators and our simulation results in Section 7 suggest that this is indeed so. One way to correct for potential bias is through weighting. The specific weight for each observation i is taken to be $w_i = \delta_i + w(1-\delta_i)\eta_i$ where w can be chosen either as n/n_1 or n_3/n_4 where $n_3 = n - \sum_{i=1}^n \delta_i$ is the total number of censored subjects and $n_4 = \sum_{i=1}^n \eta_i(1-\delta_i)$ is the number of censored subjects in the subcohort. In our simulation study, we use n_3/n_4. Our proposed estimators are then

$$\begin{aligned} \tilde{F}_{\mathbf{X}}(\mathbf{x}) &= \tfrac{1}{n}\textstyle\sum_{i=1}^n w_i \mathbf{1}(\mathbf{X}_i \le \mathbf{x}), \\ \tilde{S}_{\mathbf{b}}(t) &= \begin{cases} \prod_{t_j \le t}(1-\frac{d_j}{R_j}) & \text{if } t < \tau_1 + 1, \\ 0 & \text{if } t \ge \tau_1 + 1, \end{cases} \\ \tilde{F}_C(c) &= \begin{cases} 1-\prod_{s_j \le c}(1-\frac{c_j}{U_j}) & \text{if } c < \tau_2, \\ 1 & \text{if } c \ge \tau_2, \end{cases} \end{aligned} \tag{3.1}$$

where $\tau_1 = \max_j t_j$, $d_j = \sum_{i=1}^n \delta_i \mathbf{1}(T_i(\mathbf{b}) = t_j)$, $R_j = \sum_{i \notin K} w_i \mathbf{1}(T_i(\mathbf{b}) \ge t_j)$, $\tau_2 = \max_j M_j$, $c_j = \sum_{i\notin K} w_i \mathbf{1}(M_i = s_j, \delta_i = 0)$, and

$$U_j = \sum_{i\notin K} w_i\big[\mathbf{1}(M_i > s_j, \delta_i = 1) + \mathbf{1}(M_i \ge s_j, \delta_i = 0)\big],$$

with $t_1 < t_2 < \cdots$ are all the distinct values among $T_i(\mathbf{b})$'s with $\delta_i = 1$, $s_1 < s_2 < \cdots$ are all the distinct values among M_i's with $\delta_i = 0$ and $i \notin K$, and in order to avoid degenerated case, we define $w_i = 1$ if n_1 or $n_3 = 0$.

Remark 3. Note that when the largest residual, say $T_{(n_2)}(\mathbf{b})$, is right censored, since $\frac{\sum_{t > T_i(\mathbf{b})} t\hat{f}_{\mathbf{b}}(t)}{\hat{S}_{\mathbf{b}}(T_i(\mathbf{b}))}$ in (2.3) is not defined, Buckley and James treated it as an exact one, then $\hat{S}_{\mathbf{b}}$ puts the tail weight to the $T_{(n_2)}$. To avoid the complexity in the proof, we define in (3.1) that $\hat{S}_{\mathbf{b}}$ puts the tail weight to $T_{(n_2)} + 1$. Both conventions do not affect the asymptotic properties of the BJE of β.

In an obvious way, denote $\tilde{f}_C$, $\tilde{f}_{\mathbf{b}}$ and $\tilde{f}_{\mathbf{X}}$ the density functions corresponding to $\tilde{F}_C$, $\tilde{S}_{\mathbf{b}}$ and $\tilde{F}_{\mathbf{X}}$, respectively. In Lemmas 1, 2 and 3 of Appendix I, we show that under certain regularity conditions, $\tilde{S}_{\mathbf{b}}$, $\tilde{F}_C$, and $\tilde{F}_{\mathbf{X}}$ are consistent. Define $\tilde{f}_{C,\mathbf{X}} = \tilde{f}_C \tilde{f}_{\mathbf{X}}$ and

$$\widetilde{H}(\mathbf{b}) = H(\mathbf{b}, \tilde{S}_{\mathbf{b}}, \tilde{f}_{C,\mathbf{X}}, \widetilde{\mathbf{X}}) \text{ where } \widetilde{\mathbf{X}} = \sum_{\mathbf{x}} \mathbf{x}\tilde{f}_{\mathbf{X}}(\mathbf{x}), \tag{3.2}$$

that is, $\widetilde{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^n w_i \mathbf{X}_i$. The Buckley-James-type of estimator is a zero-crossing of $\widetilde{H}(\mathbf{b})$. We call it the *weighted BJE* or $\hat{\beta}_4$ in our simulation results of Section 6.

In summary, we have considered four estimators under the classical case-cohort design. They are subcohort BJE ($\hat{\beta}_1$), partial likelihood BJE ($\hat{\beta}_2$), naive BJE ($\hat{\beta}_3$), and weighted BJE ($\hat{\beta}_4$). To gain intuition behind these estimators, we bring readers attention that there are two estimating procedures involved. One pertains to the treatment of likelihood and resulting estimating function, and the other concerns estimating the underlying distributions to be plugged into the estimating function. The difference of these four parameters lies in either one or both of these two procedures. The subcohort BJE utilizes the only data in subcohort for likelihood and estimation of underlying distributions. It has a sole effect of reducing sample size. Hence it is easy to implement using existing software for full cohort BJE (see e.g. Stare et. al. 2001). As long as the subcohort is representative of the full cohort, $\hat{\beta}_1$ will be consistent. This is the case when in particular simple random sampling is used to select subcohort members. However there is necessarily efficiency loss. On the other hand, the partial likelihood BJE uses all the available data for likelihood and estimation of underlying distributions. This should increase efficiency. However due to over-sampling of failure, $\hat{\beta}_2$ is inconsistent. The naive BJE attempts to correct partially for such biased sampling by using full nonparametric likelihood. However in estimating underlying distributions, it proceeds like the partial likelihood BJE. This also introduces bias. The weighted BJE avoids this problem by using both full nonparametric likelihood and weighted KM estimate for the underlying distributions. Intuitively, $\hat{\beta}_4$ should have best performance. All these points seem to be confirmed by our simulation study in Section 6.

4. Computation and Algorithms. We present in this section the algorithm for the BJE with respect to $\widetilde{H}$. Since $T_i(\mathbf{b}) = M_i - \mathbf{b}\mathbf{X}_i$ for $i \notin K$, one can verify that $\widetilde{H}$ in §3 becomes

$$\widetilde{H}(\mathbf{b}) = \mathcal{A}(\mathbf{b}) - \mathcal{B}(\mathbf{b})\mathbf{b}, \tag{4.1}$$

where $\mathcal{A} = \sum_i \mathcal{A}_i(\mathbf{b})$, $\mathcal{B} = \sum_i \mathcal{B}_i(\mathbf{b})$, and for $i = 1, ..., n$,

$$\begin{aligned}
\mathcal{A}_i(\mathbf{b}) = & \quad \left\{ M_i\delta_i + (1-\delta_i)\eta_i \sum_{t>T_i(\mathbf{b})} \frac{\tilde{f}_{\mathbf{b}}(t)}{\tilde{S}_{\mathbf{b}}(T_i(\mathbf{b}))} \frac{\sum_{h\notin K} M_h\delta_h \mathbf{1}(T_h(\mathbf{b})=t)}{\sum_{h\notin K} \delta_h \mathbf{1}(T_h(\mathbf{b})=t)} \right\} (\mathbf{X}_i - \widetilde{\mathbf{X}}) \qquad (4.2) \\
& + \mathbf{1}(i\in K) \frac{\sum_{j,k\notin K} \left\{ \sum_{t>M_j-\mathbf{b}\mathbf{X}_k} \tilde{f}_{\mathbf{b}}(t) \frac{\sum_{h\notin K} M_h\delta_h \mathbf{1}(T_h(\mathbf{b})=t)}{\sum_{h\notin K} \delta_h \mathbf{1}(T_h(\mathbf{b})=t)} \right\} \tilde{f}_C(M_j)\tilde{f}_{\mathbf{X}}(\mathbf{X}_k)(\mathbf{X}_k - \widetilde{\mathbf{X}})}{\sum_{j,k\notin K} \tilde{S}_{\mathbf{b}}(M_j - \mathbf{b}\mathbf{X}_k)\tilde{f}_C(M_j)\tilde{f}_{\mathbf{X}}(\mathbf{X}_k)}, \\
\mathcal{B}_i(\mathbf{b}) = & \quad (\mathbf{X}_i - \widetilde{\mathbf{X}}) \left\{ \mathbf{X}_i\delta_i + (1-\delta_i)\eta_i \sum_{t>T_i(\mathbf{b})} \frac{\tilde{f}_{\mathbf{b}}(t)}{\tilde{S}_{\mathbf{b}}(T_i(\mathbf{b}))} \frac{\sum_{h\notin K} \mathbf{X}_h\delta_h \mathbf{1}(T_h(\mathbf{b})=t)}{\sum_{h\notin K} \delta_h \mathbf{1}(T_h(\mathbf{b})=t)} \right\}' \\
& + \mathbf{1}(i\in K) \frac{\sum_{j,k\notin K} \left\{ \tilde{f}_C(M_j)\tilde{f}_{\mathbf{X}}(\mathbf{X}_k)(\mathbf{X}_k - \widetilde{\mathbf{X}}) \sum_{t>M_j-\mathbf{b}\mathbf{X}_k} \tilde{f}_{\mathbf{b}}(t) \frac{\sum_{h\notin K} \mathbf{X}'_h\delta_h \mathbf{1}(T_h(\mathbf{b})=t)}{\sum_{h\notin K} \delta_h \mathbf{1}(T_h(\mathbf{b})=t)} \right\}}{\sum_{j,k\notin K} \tilde{S}_{\mathbf{b}}(M_j - \mathbf{b}\mathbf{X}_k)\tilde{f}_C(M_j)\tilde{f}_{\mathbf{X}}(\mathbf{X}_k)}.
\end{aligned}$$

Verify that $\mathcal{A}$ is a $p \times 1$ dimensional vector and $\mathcal{B}$ is a $p \times p$ dimensional matrix,

Algorithm (for the BJE with case-cohort data). Give an initial value to β, say $\mathbf{b}_0$. For $k \geq 1$, update $\mathbf{b}_{k-1}$ by $\mathbf{b}_h = \left\{\mathcal{B}(\mathbf{b}_{k-1})\right\}^{-1}\mathcal{A}(\mathbf{b}_{k-1})$. Stop either at the convergence (*i.e.*, $||\mathbf{b}_k - \mathbf{b}_{k-1}||$ is very small), or at the case that $\mathbf{b}_k$ 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 4. 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 $\widetilde{H}(\mathbf{b})$ between the oscillating points to find a zero-crossing of $\widetilde{H}(\mathbf{b})$ if $p = 1$.

5. Variance Estimation 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 case-cohort case, though of course the efficient lower bound is different. Based on this belief, we may use the inverse of the Fisher information matrix as the estimate of the covariance matrix of the BJE:

$$\hat{\Sigma}_{\hat{\beta}} = (\hat{\mathcal{I}})^{-1}, \text{ where } \hat{\mathcal{I}} = \sum_{i=1}^{n} \left\{ (\mathcal{A}_i(\hat{\beta}) - \mathcal{B}_i(\hat{\beta})\hat{\beta})(\mathcal{A}_i(\hat{\beta}) - \mathcal{B}_i(\hat{\beta})\hat{\beta})' \right\} \hat{\sigma}^{-4}, \qquad (5.1)$$

$\hat{\sigma}^2$ is an estimator of $Var(\epsilon)$.

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. Our simulation results indicate that even under the exponential distribution the approximation seems quite good.

The asymptotic variances of the estimators we have established in this paper are all different. One can use the delta method to obtain the asymptotic variance for the weighted BJE using the argument given in the proof of Theorem 1 in the Appendix I. However, the expression is quite long.

6. Simulation Studies. In this section, we present some simulation study results for evaluating our proposed estimator $\hat{\beta}_4$ (weighted BJE, see Section 3) under various sample sizes and various distributions. According to the results established in Lai and Ying (1991), we expect our estimator to be efficient under the normal assumption and consistent in general. In Appendix I, we establish consistency and asymptotic normality for the BJE with discrete distribution assumptions. In this section, we shall present three sets of simulation results on our proposed estimators under continuous distributions and under assumption A3. The simulation is performed on a Pentium III workstation. In each simulation study, we had 1000 replications and computed the sample mean and sample standard error (SE) of the 1000 estimates. The computation is quite fast, it only takes a few seconds for a sample size of 800.

As discussed in section 3, different ways of estimations of S_o and $f_{\mathbf{X}}$ can be used and the choices may affect the performance of BJE. To demonstrate this, we use both our proposed weighted estimators $\hat{\beta}_4$ and naive estimators $\hat{\beta}_3$ of S_o and $f_{\mathbf{X}}$ and compare resulting estimators of the regression parameter β. We also compare them to the partial likelihood BJE $\hat{\beta}_2$ (see Remark 2) and the Subcohort BJE $\hat{\beta}_1$ based on the subcohort data alone (see Remark 1). In the literature, all the studies are not based on the linear regression model, in particular, most studies are based on the Cox regression model. It is well known that under the linear regression model if the underlying distribution is normal then the distribution does not follow the Cox regression model. Consequently, in our simulation, it is not appropriate to compare our estimates to the estimates under the Cox regression model, as well as the estimates under other models.

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 3 different cases.

Case 1 (censored-data under a normal distribution).Suppose $\epsilon \sim N(0,1)$ (the normal distribution), $C \sim N(0,1)$ and $X \sim Exp(1.25, 1.25)$. $q =$P$\{\eta = 1\} = 0.2$. $\beta = 1$. Results are listed in Block 1 of Table 1.

Case 2 (censored-data under a normal distribution).Suppose $\epsilon \sim N(1,1)$, $X \sim U(0,1)$

(the uniform distribution) and $C \sim Exp(1,1)$. $q = 0.5$. $\beta = 1$. Results are listed in Block 2 of Table 1.

Case 3 (censored-data under an exponential distribution).Suppose $\epsilon \sim Exp(1,1)$, $X \sim U(0,1)$ and $C \sim Exp(1,1)$. $q = 0.7$. $\beta = 1$. Results are listed in Block 3 of Table 1 and in Table 2. In Table 2, the entries corresponding to $\overline{\hat{\sigma}_{\hat{\beta}}}$ are the sample averages of the estimates of standard deviation of the BJE based on formula (5.1). SE is the standard error of the BJE or the estimate of $\sigma_{\hat{\beta}}$ in 1000 simulations.

The simulation results in Table 1 suggest that the weighted BJE $\hat{\beta}_4$ is consistent under both the normal distribution and the exponential distribution, as β is within the interval $(\overline{\hat{\beta}_4} - 2SE, \overline{\hat{\beta}_4} + 2SE)$ and the SE decreases, as n increases.

Furthermore, they confirms that the weighted BJE $\hat{\beta}_4$ with the full likelihood is better than the naive BJE $\hat{\beta}_3$ and the partial likelihood BJE $\hat{\beta}_2$. This can be viewed as follows.

Even though it is seen from Table 1 that $\hat{\beta}_3$ and $\hat{\beta}_2$ have smaller standard errors than $\hat{\beta}_4$, it is seen from the second block of Table 1 that the $\hat{\beta}_3$ is steadily getting close to 1.2, and is quite different from $\beta = 1$, in fact, in another simulation with sample size $n = 3000$, the sample mean is 1.24 with SE of 0.10, thus the estimate is significantly different from $\beta = 1$, thus it suggests that the naive BJE is inconsistent. Moreover, we can see from Block 1 of Table 1 that $\hat{\beta}_2$ is steadily getting close to 0.82, in particular, when $n = 1600$, its sample average is 0.822 with a SE of 0.064, thus it is significantly different from $\beta = 1$. This suggests that these two estimators are inconsistent. It is seen that $\hat{\beta}_2$ does behavior pretty good in the other two cases. The major difference is $q = 0.2$ in the first block of Table 1 whereas $q \geq 0.5$ in the other two blocks.

On the contrary, $\hat{\beta}_4$ is getting close to β in all the three cases. In all the three cases, it is seen that BJE $\hat{\beta}_4$ is more efficient that the BJE $\hat{\beta}_1$ based on subcohort alone, as one expects. One also expects that the BJE is efficient under the normal assumption, thus it is appropriate to use formula (5.1) as an estimate of the standard deviation of the BJE. It seen from Table 2 that even under the exponential distribution, the estimator of $\sigma_{\hat{\beta}_4}$ by formula (5.1) matches its standard errors quite well.

Table 1. Comparison On Four BJE Methods

Under N(0,1) With $q = 0.2$					
n	β	Naive(SE)	Weighted (SE)	Partial (SE)	Subco (SE)
100	1	1.094(0.332)	0.994(1.500)	0.867(0.280)	1.412(1.822)
200	1	1.107(0.235)	1.019(0.420)	0.836(0.195)	1.130(0.575)
400	1	1.136(0.169)	0.997(0.206)	0.830(0.133)	1.054(0.318)
800	1	1.149(0.124)	0.992(0.128)	0.825(0.092)	1.016(0.208)
Under N(1,1) With $q = 0.5$					
n	β	Naive(SE)	Weighted (SE)	Partial (SE)	Subco (SE)
100	1	1.205(0.957)	0.912(1.042)	0.957(0.500)	0.992(0.669)
200	1	1.205(0.418)	0.964(0.388)	0.951(0.343)	0.979(0.464)
400	1	1.227(0.303)	0.984(0.273)	0.947(0.242)	0.999(0.327)
800	1	1.242(0.206)	0.994(0.187)	0.945(0.169)	1.002(0.229)
Under Exp(1,1) With $q = 0.7$					
n	β	Naive(SE)	Weighted (SE)	Partial (SE)	Subco (SE)
100	1	1.137(0.349)	0.969(0.335)	1.015(0.336)	1.004(0.393)
200	1	1.141(0.238)	0.988(0.223)	1.008(0.222)	1.002(0.263)
400	1	1.146(0.168)	0.998(0.157)	1.006(0.156)	1.002(0.182)
800	1	1.143(0.123)	0.996(0.114)	1.000(0.113)	1.001(0.135)

Table 2. Variance Estimation
Under Exp(1,1) Distribution With $q = 0.7$
SE of BJE $\hat{\beta}_4$ comparing to estimator $\hat{\sigma}_{\hat{\beta}_4}$ (SE)

n	SE of $\hat{\beta}_4$	$\hat{\sigma}_{\hat{\beta}_4}$ (SE)
n=100	0.335	0.417 (0.109)
n=200	0.223	0.277 (0.051)
n=400	0.157	0.189 (0.025)
n=800	0.114	0.130 (0.013)

7. A Real Example. In this section, we carry out data analysis on the Welsh Nickel Refinery Study which has been used frequently in literature to illustrate case-cohort studies (see *e.g.*, Lin and Ying 1993 and Barlow *et. al.* 1999). The original data were full cohort data and published in Breslow and Day (1987). In this study, employees in a nickel refinery in South Wales were investigated to determine the risk of developing nasal cancer. There are 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).

Table 3. Data Analyses of Time from the First Employment to the Nasal Sinus Cancer Death for the Welsh Nickel Refiner Study

Parameters	Full cohort	subcohort	case-cohort
log(EXP+1)			
Est.	−0.189	−0.561	−0.202
SE.	0.002	0.019	0.001
log(AFE-10)			
Est.	−0.617	−0.399	−0.467
SE.	0.003	0.036	0.03

We take a random sample of $q = 20\%$ of the full cohort to be a subcohort group. The estimates are given in Table 3 and they are all significant. Here, our estimate of the standard deviation of the BJE is based on formula (5.1).

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 at least 50.

8. Appendix I Consistency and asymptotic normality are important properties of an estimator. For simplicity, we shall establish these properties for the weighted BJE $\hat{\beta}_4$ under a simple regularity condition. It is clear that the Subcohort BJE $\hat{\beta}_1$ is consistent and asymptotic normal, though not efficient. Simulation results suggest that the other two BJE's $\hat{\beta}_2$ and $\hat{\beta}_3$ are inconsistent. Thus it is not of interests to study their asymptotic properties.

In our simulation study, we assume all underlying distributions are continuous. In the literature, discrete assumption 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 (1980) on the PLE with right-censored data and Turnbull (1976) on the GMLE with interval censored data. We also follow this path here. In some but not all of the proofs, we make use of the following discrete assumption.

A5. ϵ, $\mathbf{X}$ and C all take on finitely many values.

Under A5, WLOG, we can assume that τ_1 and τ_2 do not depend on n. We shall first discuss the consistency of $\tilde{S}_{\mathbf{b}}$ given in (3.1), as it is needed in the main proof.

Lemma 1. *Under assumptions A1, A2 and A5, $\tilde{S}_\beta$ in (3.1) satisfies $\tilde{S}_\beta(t) \to S_*(t)$ a.s. for each t, where $S_*(t) = S_o(t)\mathbf{1}(t \le \tau_1) + S_o(\tau_1)\mathbf{1}(t \in (\tau_1, \tau_1+1)) + 1(t \ge \tau_1+1)$.*

Proof. Denote f_* the density function of S_*. Notice that $n/n_1 \to 1/q$ and $n_4/n_3 \to \frac{P\{\delta=0\}}{P\{\eta=1,\delta=0\}} = 1/q$. That is $\lim_{n\to\infty} n/n_1 = \lim_{n\to\infty} n_4/n_3$. Thus WLOG, we only consider the case $w = \frac{n}{n_1}$. By the definitions in (3.1), R_j, d_j and $\tilde{S}_{\mathbf{b}}$ are all functions of $\mathbf{b}$. When $\mathbf{b} = \beta$, by the strong law of large numbers (SLLN), with probability one (w.p.1),

$$
\begin{array}{rl}
\frac{R_j}{n} = & \frac{\sum_{i=1}^{n}(\delta_i + \frac{n}{n_1}(1-\delta_i)\eta_i)\mathbf{1}(T_i(\mathbf{b})\ge t_j)}{n} \qquad (t_j \text{ is given in (3.1)}) \\
\to & P\{T(\beta) \ge t_j, \delta = 1\} + P\{T(\beta) \ge t_j, \delta = 0, \eta = 1\}/q \text{ (by SLLN)} \\
= & P\{T(\beta) \ge t_j, \delta = 1\} + P\{T(\beta) \ge t_j, \delta = 0\}P\{\eta = 1\}/q \qquad \text{(by A1)} \\
= & P\{T(\beta) \ge t_j\}.
\end{array}
$$

Thus, $\frac{d_j}{R_j} \to \frac{f_*(t_j)P\{C-\beta\mathbf{X}\ge t_j\}}{S_*(t_j-)P\{C-\beta\mathbf{X}\ge t_j\}} = \frac{f_*(t_j)}{S_*(t_j-)}$ and $\tilde{S}_\beta(t) \to \prod_{t_j\le t}(1 - \frac{f_*(t_j)}{S_*(t_j-)}) = S_*(t)$ for each $t \le \tau_1$. The proof of $t > \tau_1$ is trivial by (3.1). ▫

Lemma 2. *$\tilde{F}_{\mathbf{X}}$ is consistent under assumptions A1 and A2.*

Proof.

$$
\begin{array}{rcl}
\lim_{n\to\infty} \tilde{F}_{\mathbf{X}}(\mathbf{x}) & = & \lim_{n\to\infty} \frac{1}{n}\sum_{i\notin K}(\delta_i + \frac{n}{n_1}(1-\delta_i)\eta_i)\mathbf{1}(\mathbf{X}_i \le \mathbf{x}). \\
& = & P\{\delta = 1, \mathbf{X} \le \mathbf{x}\} + P\{\delta = 0, \mathbf{X} \le \mathbf{x}\} \\
& = & P\{\mathbf{X} \le \mathbf{x}\}. \ ▫
\end{array}
$$

Note that the proof of the Lemma 2 does not invoke A3 and A5.

Lemma 3. *Under assumptions A1, A2 and A5, $\tilde{F}_C$ in (3.1) satisfies that $\tilde{F}_C(t) \to F_C^*(t)$ a.s. for each t, where $F_C^*(t) = F_C(t)\mathbf{1}(t < \tau_2) + \mathbf{1}(t \ge \tau_2)$.*

Proof. Denote the density function of F^*_C by f^*_C. With probability one,

$$\begin{aligned}\tfrac{U_j}{n} &= \tfrac{\sum_{i=1}^n(\mathbf{1}(M_i>s_j)\delta_i+\frac{n}{n_1}(1-\delta_i)\eta_i)\mathbf{1}(M_i\ge s_j)}{n} \qquad (s_j \text{ is given in (3.1)})\\ &\to P\{M>s_j,\delta=1\}+P\{M\ge s_j,\delta=0,\eta=1\}/q \text{ (by SLLN)}\\ &= P\{M>s_j,\delta=1\}+P\{M\ge s_j,\delta=0\}P\{\eta=1\}/q \qquad \text{(by A1)}\\ &= P\{M>s_j,\delta=1\}+P\{M\ge s_j,\delta=0\}\\ &= P\{M>s_j\}+P\{C=s_j<\epsilon+\beta\mathbf{X}\}.\end{aligned}$$

Thus,

$$\begin{aligned}1-\tfrac{c_j}{U_j} &\to 1-\tfrac{\frac{1}{q}f^*_C(s_j)P\{\epsilon+\beta\mathbf{X}>s_j,\eta=1\}}{P\{M>s_j\}+f^*_C(s_j)P\{\epsilon+\beta\mathbf{X}>s_j\}}\\ &=1-\tfrac{\frac{1}{P\{\eta=1\}}f^*_C(s_j)P\{\epsilon+\beta\mathbf{X}>s_j\}P\{\eta=1\}}{P\{M>s_j\}+f^*_C(s_j)P\{\epsilon+\beta\mathbf{X}>s_j\}}\\ &=\tfrac{P\{M>s_j\}}{P\{M>s_j\}+f^*_C(s_j)P\{\epsilon+\beta\mathbf{X}>s_j\}}\\ &=\tfrac{1-F^*_C(s_j)}{1-F^*_C(s_j-)}\end{aligned}$$

a.s. and $1-\tilde F_C(t)\to\prod_{s_j\le t}\frac{1-F^*_C(s_j)}{1-F^*_C(s_j-)}=1-F^*_C(t)$ a.s. for $t<\max_j s_j$. The proof for $t\ge\tau_2$ is trivial by (3.1). ▫

We shall give a simple proof for the consistency and asymptotic normality of the weighted BJE $\hat\beta_4$. Assumption A4 is crucial in the proof. Otherwise the asymptotic normality would not hold even under the full cohort case (see Kong and Yu (2005)). Moreover, Yu and Wong (2002) show that the BJE is not unique in the full cohort case, thus it is also not unique under the case-cohort designs either.

Under assumption A3, by (4.1) and (4.2), we have $\widetilde H(\mathbf{b})=\sum_{i=1}^n\mathcal{A}_i(\mathbf{b})-\sum_{i=1}^n\mathcal{B}_i(\mathbf{b})\mathbf{b}=\mathcal{A}(\mathbf{b})-\mathcal{B}(\mathbf{b})\mathbf{b}$. By assumption A5, there are only m distinct values of $\mathbf{O}$ $(=(M,\delta,\eta,\mathbf{X}))$, where m is finite. Let $\mathbf{O}_1$, ..., $\mathbf{O}_n$ be i.i.d. copies of $\mathbf{O}$. By taking n large enough, WLOG, we can assume that the first m such values are all distinct. Hereafter, denote $\overline N_i(\mathbf{b})=\frac{1}{n}\sum_{j=1}^n\mathbf{1}((T_j(\mathbf{b}),\delta_j,\eta_j,\mathbf{X}_j)=(T_i(\mathbf{b}),\delta_i,\eta_i,\mathbf{X}_i))$, $i=1,...,m$, and $\overline{\mathbf{N}}(\mathbf{b})=(\overline N_2(\mathbf{b}),...,\overline N_m(\mathbf{b}))$ (as $\sum_{i=1}^m\overline N_i=1$).

Theorem 1. *Assume that A1 through A5 hold then a solution of the weighted BJE $\hat\beta_4$ in §3 is a rational function of $\overline{\mathbf{N}}(\mathbf{b})$, say $\hat\beta_4=g(\overline{\mathbf{N}}(\mathbf{b}))$. Moreover, if $g'(\mathbf{y})\ne\mathbf{0}$, where $\mathbf{y}=E(\overline{\mathbf{N}}(\mathbf{b}))$, then $\hat\beta_4$, is consistent and asymptotically normally distributed.*

Proof. WLOG, we shall assume that $S_*=S_o$ and $F^*_C=F_C$, Otherwise, replace (S_o,f_o,F_C) by (S_*,f_*,F^*_C) in the proof.

The proof is quite long, we list the main steps. We shall show that

(a) given a random sample, in a neighborhood of β, $(\mathcal{A}(\mathbf{b}), \mathcal{B}(\mathbf{b}))$ is constant in $\mathbf{b}$;

(b) $\hat{\beta}_4 = (\mathcal{B}(\beta))^{-1}\mathcal{A}(\beta)$ if n is large enough. As $(\mathcal{B}(\beta))^{-1}\mathcal{A}(\beta) \to \beta$, we prove the consistency of $\hat{\beta}_4$.

(c) $\tilde{S}_{\mathbf{b}}$, $\tilde{f}_{\mathbf{X}}$ and $\tilde{f}_C$ are rational functions of $\overline{\mathbf{N}}(\mathbf{b})$ for each given $\mathbf{b}$;

(d) $\hat{\beta}_4$ is asymptotically normally distributed.

Step 1 (verify statement (a)). Let $T_{j,k} = T_{j,k}(\mathbf{b}) = M_j - \mathbf{b}\mathbf{X}_k$ for $j, k \notin K$. Note that $K = \{i : \eta_i + \delta_i = 0\}$ is the set of censored subjects not in sub-cohort. This notation extends $T_j = T_{j,j}$ and arises from the last term in expression (4.2). Since the weighted PLE only assigns weights to exact observations $T_i(\mathbf{b})$'s unless $T_i(\mathbf{b})$ is the largest observation, we can define that $T_{j_1,k_1} = T_{j_2,k_2}$. if T_{j_1,k_1} and T_{j_2,k_2} are censored (*i.e.*, $\delta_{j_1} = \delta_{j_2} = 0$), and if there is no exact observation within T_{j_1,k_1} and T_{j_2,k_2}. In order to prove statement (a), we shall show that *if* $\mathbf{b}_1$ *and* $\mathbf{b}_2$ *are two values of* $\mathbf{b}$ *such that for all* $j, k \notin K$,

(i) rank$(T_{j,k}(\mathbf{b}_1))$ *= rank*$(T_{j,k}(\mathbf{b}_2))$ *for each right-censored* $T_{j,k}$ *(i.e.* $\delta_j = 0$*),*

(ii) $T_i(\mathbf{b}_1) < T_{j,k}(\mathbf{b}_1)$ *iff* $T_i(\mathbf{b}_2) < T_{j,k}(\mathbf{b}_2)$, *whenever* $(\delta_i, \delta_j) = (1, 0)$,

(iii) $T_i(\mathbf{b}_1) \neq T_{j,k}(\mathbf{b}_1)$, *whenever* $(\delta_i, \delta_j) = (1, 0)$,

then $\mathcal{A}_i(\mathbf{b}_1) = \mathcal{A}_i(\mathbf{b}_2)$, $\mathcal{B}_i(\mathbf{b}_1) = \mathcal{B}_i(\mathbf{b}_2)$, $\mathcal{A}(\mathbf{b}_1) = \mathcal{A}(\mathbf{b}_2)$ *and* $\mathcal{B}(\mathbf{b}_1) = \mathcal{B}(\mathbf{b}_2)$. When $\delta_i = 1$, by (4.2), $(\mathcal{A}_i(\mathbf{b}), \mathcal{B}_i(\mathbf{b})) = (M_i, \mathbf{X}_i)$ and it is obvious that the above holds.

When $\delta_i = 0$ and $i \notin K$. Then by (4.2),

$$\mathcal{A}_i = \frac{1}{\tilde{S}_{\mathbf{b}}(T_i)} \sum_{j:T_j>T_i} \sum_{h\notin K} \frac{\tilde{f}_{\mathbf{b}}(T_j)}{\sum_{h\notin K} \delta_h \mathbf{1}(T_h = T_j)} M_h \delta_h \mathbf{1}(T_h = T_j)(\mathbf{X}_i - \widetilde{\mathbf{X}}).$$

In order to prove $\mathcal{A}_i(\mathbf{b}_1) = \mathcal{A}_i(\mathbf{b}_2)$ it then suffices to verify the following equations:

(1) $\tilde{S}_{\mathbf{b}_1}(T_i(\mathbf{b}_1)) = \tilde{S}_{\mathbf{b}_2}(T_i(\mathbf{b}_2))$;
(2) $\{j \notin K : T_j(\mathbf{b}_1) > T_i(\mathbf{b}_1)\} = \{j \notin K : T_j(\mathbf{b}_2) > T_i(\mathbf{b}_2)\}$;
(3) $\frac{\tilde{f}_{\mathbf{b}_1}(T_j(\mathbf{b}_1))}{\sum_{h\notin K} \delta_h \mathbf{1}(T_h(\mathbf{b}_1)=T_j(\mathbf{b}_1))} = \frac{\tilde{f}_{\mathbf{b}_2}(T_j(\mathbf{b}_2))}{\sum_{h\notin K} \delta_h \mathbf{1}(T_h(\mathbf{b}_2)=T_j(\mathbf{b}_2))}$ if $\delta_j = 1$.

Equation (1) follows from (3.1) and Equation (2) from conditions (i), (ii) and (iii) with $k = j$. To prove equation (3), we first illustrate under a simple scenario. Let's say there are

2 exact observations between $T_{j_1}(\mathbf{b}_1)$ and $T_{j_2}(\mathbf{b}_1)$, say $T_1(\mathbf{b}_1) < T_2(\mathbf{b}_1)$, when $\mathbf{b}_1$ changes to $\mathbf{b}_2$, there is a tie between the two, i.e. $T_1(\mathbf{b}_2) = T_2(\mathbf{b}_2)$. From the definition of $\tilde{S}_{\mathbf{b}}$ in (3.1), the total weight assigned to the interval between two consecutive censored observations $T_{j_1}(\mathbf{b}_h)$ and $T_{j_2}(\mathbf{b}_h)$ will be the same for $h = 1$ and $h = 2$ under conditions (i), (ii) and (iii), and each of the exact observations $T_i(\mathbf{b}_h)$ between two consecutive censored observations $T_{j_1}(\mathbf{b}_h)$ and $T_{j_2}(\mathbf{b}_h)$ will be assigned equal weight. Hence under $\mathbf{b} = \mathbf{b}_1$,

$$\frac{\tilde{f}_{\mathbf{b}_1}(T_1(\mathbf{b}_1))}{\sum_{h\notin K} \delta_h \mathbf{1}(T_h(\mathbf{b}_1) = T_1(\mathbf{b}_1))} = \frac{\tilde{S}_{\mathbf{b}_1}(T_{j_1}(\mathbf{b}_1)-)}{R_{j_1}(\mathbf{b}_1)},$$

and

$$\frac{\tilde{f}_{\mathbf{b}_1}(T_2(\mathbf{b}_1))}{\sum_{h\notin K} \delta_h \mathbf{1}(T_h(\mathbf{b}_1) = T_2(\mathbf{b}_1))} = \frac{\tilde{S}_{\mathbf{b}_1}(T_{j_1}(\mathbf{b}_1)-)\left\{1 - \frac{1}{R_{j_1}(\mathbf{b}_1)}\right\}}{R_{j_1}(\mathbf{b}_1) - 1} = \frac{\tilde{S}_{\mathbf{b}_1}(T_{j_1}(\mathbf{b}_1)-)}{R_{j_1}(\mathbf{b}_1)},$$

where R_j is the weighted risk set, i.e. $R_j = \sum_{h\notin K} w_i \mathbf{1}(T_h(\mathbf{b}) \geq T_j(\mathbf{b}))$. Under $\mathbf{b} = \mathbf{b}_2$, since $T_1(\mathbf{b}_2) = T_2(\mathbf{b}_2)$, we have

$$\frac{\tilde{f}_{\mathbf{b}_2}(T_j(\mathbf{b}_2))}{\sum_{h\notin K} \delta_h \mathbf{1}(T_h(\mathbf{b}_2) = T_j(\mathbf{b}_2))} = \frac{\tilde{S}_{\mathbf{b}_2}(T_{j_1}(\mathbf{b}_2)-)}{R_{j_1}(\mathbf{b}_2)} = \frac{\tilde{S}_{\mathbf{b}_1}(T_{j_1}(\mathbf{b}_1)-)}{R_{j_1}(\mathbf{b}_1)},$$

which is the same as under $\mathbf{b} = \mathbf{b}_1$. Similarly, we can show more general scenarios for Equation (3) holds and we have $\mathcal{A}_i(\mathbf{b}_1) = \mathcal{A}_i(\mathbf{b}_2)$ if $\delta_i = 0$ and $i \notin K$. The proof for $\mathcal{B}_i(\mathbf{b}_1) = \mathcal{B}_i(\mathbf{b}_2)$ if $\delta_i = 0$ and $i \notin K$ is quite similar and is skipped.

Finally for $i \in K$. Note that $\mathcal{A}_i(\mathbf{b}) =$

$$\frac{\sum_{j,k\notin K} \sum_{l:T_l > T_{j,k}} \sum_{h\notin K} \frac{\tilde{f}_{\mathbf{b}}(T_l)}{\sum_{h\notin K} \delta_h \mathbf{1}(T_h = T_l)} M_h \delta_h \mathbf{1}(T_h = T_l) \tilde{f}_C(M_j) \tilde{f}_{\mathbf{X}}(\mathbf{X}_k)(\mathbf{X}_k - \widetilde{\mathbf{X}})}{\sum_{j,k\notin K} \tilde{S}_{\mathbf{b}}(T_{j,k}) \tilde{f}_C(M_j) \tilde{f}_{\mathbf{X}}(\mathbf{X}_k)}.$$

It suffices to show that for $j, k \notin K$,

(1′) $\tilde{S}_{\mathbf{b}_1}(T_{j,k}(\mathbf{b}_1)) = \tilde{S}_{\mathbf{b}_2}(T_{j,k}(\mathbf{b}_2))$;

(2′) $\{l \notin K : T_j(\mathbf{b}_1) > T_{i,k}(\mathbf{b}_1)\} = \{l \notin K : T_j(\mathbf{b}_2) > T_{i,k}(\mathbf{b}_2)\}$;

(3′) $\frac{\tilde{f}_{\mathbf{b}_1}(T_j(\mathbf{b}_1))}{\sum_{h\notin K} \delta_h \mathbf{1}(T_h(\mathbf{b}_1) = T_j(\mathbf{b}_1))} = \frac{\tilde{f}_{\mathbf{b}_2}(T_j(\mathbf{b}_2))}{\sum_{h\notin K} \delta_h \mathbf{1}(T_h(\mathbf{b}_2) = T_j(\mathbf{b}_2))}$ if $\delta_j = 1$.

The idea of the proof is similar to that in the foregoing case and we skip the proof.

Now under assumptions A4 and A5, it can be verified that

$$\mathbf{b} = \beta \text{ is not a solution to } T_i(\mathbf{b}) = T_{j,k}(\mathbf{b}) \text{ with } (\delta_i, \delta_j) = (1, 0), \text{ and } j \notin K, \tag{8.1}$$

(that is, $T_i = \epsilon_i + \beta\mathbf{X}_i$ and $T_j = C_j$). Otherwise, $T_i(\mathbf{b}) = T_{j,k}(\mathbf{b})$, $\mathbf{b} = \beta$ and $(\delta_i, \delta_j) = (1, 0)$ are equivalent to $\epsilon_i + \beta\mathbf{X}_i - \mathbf{b}\mathbf{X}_i = C_j - \mathbf{b}\mathbf{X}_k$ and $\mathbf{b} = \beta$, that is, $\epsilon_i = C_j - \beta\mathbf{X}_k$. Since ϵ, C and $\mathbf{X}$ are discrete (by A5) and independent (by A3), it implies that $\mathrm{P}\{\epsilon = C - \beta\mathbf{X}\} > 0$, which is impossible by A4.

By A5 and by our assumption, there are at most m_o ($\leq m^2$) distinct $T_{j,k}$'s and thus there are at most m_o^2 many distinct equations of the form $T_{i,j}(\mathbf{b}) = T_{k,h}(\mathbf{b})$, where m_o does not change as n increases. Because (1) the equation $T_i(\mathbf{b}) = T_{j,k}(\mathbf{b})$ with $(\delta_i, \delta_j) = (1, 0)$ is an hyperplane in $\mathcal{R}^p$ and there are only finitely many such hyperplanes; and because (2) β does not belong to any of such hyperplanes in (1) by the discussion in the previous paragraph, and thus the distance from β to each of the hyperplanes in (1) is positive, there is an open neighborhood of β, say $O(\beta, c)$ $(= \{\mathbf{b} : ||\mathbf{b} - \beta|| < c\})$, with $c > 0$ such that the three conditions (i) - (iii) hold for each (i, j, k) and for each pair of $\mathbf{b}_1, \mathbf{b}_2 \in O(\beta, c)$. Thus

$$\mathcal{A}_i(\mathbf{b}), \mathcal{B}_i(\mathbf{b}), \mathcal{A}(\mathbf{b}) \text{ and } \mathcal{B}(\mathbf{b}) \text{are constants in } \mathbf{b} \in O(\beta, c). \tag{8.2}$$

This complete the proof of statement (a).

Step 2 (verify statement (b)). Note that by A5 and A4, for n large enough, the distinct hyperplanes $T_i(\mathbf{b}) = T_j(\mathbf{b})$ with $(\delta_i, \delta_j) = (1, 0)$ remain the same and β does not belong to these hyperplanes, thus $O(\beta, c)$ will remain the same. Moreover, $(\mathcal{A}_i(\mathbf{b}), \mathcal{B}_i(\mathbf{b}))$ will take at most m values, where m is finite and does not depend on n. By reordering, WLOG, we can assume that the first m of $(\mathcal{A}_i(\mathbf{b}), \mathcal{B}_i(\mathbf{b}))$'s are distinct and we can correspond $(\mathcal{A}_i(\mathbf{b}), \mathcal{B}_i(\mathbf{b}))$ to $(M_i, \delta_i, \eta_i, \mathbf{X}_i)$, $i = 1, ..., m$, the m realizations of $(M, \delta, \eta, \mathbf{X})$. By (8.2), we have

$$\frac{\widetilde{H}(\mathbf{b})}{n} = \sum_{i=1}^{m} \overline{N_i}\mathcal{A}_i(\beta) - \sum_{i=1}^{m} \overline{N_i}\mathcal{B}_i(\beta)\mathbf{b} \text{ if } \mathbf{b} \in O(\beta, c).$$

Notice that $\tilde{S}_\beta(t_j) \to S_o(t_j)$, $\tilde{f}_C(M_j) \to f_C(M_j)$ and $\tilde{f}_{\mathbf{X}}(\mathbf{X}_j) \to f_{\mathbf{X}}(\mathbf{X}_j)$ by Lemmas 1, 2 and 3, and that $\mathcal{A}_1(\beta)$, ..., $\mathcal{A}_m(\beta)$, $\mathcal{B}_1(\beta)$, ..., $\mathcal{B}_m(\beta)$, and $\overline{N_j}$ converge by the SLLN. $\overline{N}_i \to p_i$ $(\stackrel{def}{=} P\{(M, \delta, \eta, \mathbf{X}) = (M_i, \delta_i, \eta_i, \mathbf{X}_i)\})$ and by (8.2), for $\mathbf{b} \in O(\beta, c)$ and $i = 1, ..., m$,

$$\mathcal{A}_i(\mathbf{b}) = \mathcal{A}_i(\beta) \to \qquad (\mathbf{X}_i - \mu_x)\Big\{M_i\delta_i + (1 - \delta_i)\eta_i E(M|\epsilon > T_i(\beta))\Big\}$$

$$+(1-\delta_i)(1-\eta_i)\frac{E\Big[(\mathbf{X}-\mu_x)E\Big\{C^*\mathbf{1}(\epsilon^*>C-\beta\mathbf{x})\,|\,C,\mathbf{X}\Big\}\Big]}{E\Big\{S_o(C-\beta\mathbf{X})\Big\}}\ (\stackrel{def}{=}\mathcal{A}_{io}),$$

$$\mathcal{B}_i(\mathbf{b})=\mathcal{B}_i(\beta)\rightarrow\qquad(\mathbf{X}_i-\mu_x)\Big\{\mathbf{X}_i'\delta_i+(1-\delta_i)\eta_i E(\mathbf{X}'|\epsilon>T_i(\beta))\Big\}\tag{8.3}$$

$$+(1-\delta_i)(1-\eta_i)\frac{E\Big[(\mathbf{X}-\mu_x)E\Big\{(\mathbf{X}^*)'\mathbf{1}(\epsilon^*>C-\beta\mathbf{X})|C,\mathbf{X}\Big\}\Big]}{E\Big\{S_o(C-\beta\mathbf{X})\Big\}}\ (\stackrel{def}{=}\mathcal{B}_{io}),$$

with (ϵ^*, $\mathbf{X}^*$, C^*) i.i.d. as (ϵ, $\mathbf{X}$, C). Obviously, for $\mathbf{b}\in O(\beta,c)$,

$$\lim_{n\rightarrow\infty}\frac{\widetilde{H}(\mathbf{b})}{n}=\sum_{i=1}^{m}p_i(\mathcal{A}_{io}-\mathcal{B}_{io}\mathbf{b})\ a.s.\tag{8.4}$$

Next we show that $\sum_{i=1}^m p_i\mathcal{B}_{io}$ is nonsingular and

$$\mathbf{b}=\beta\text{is the unique solution to }\sum_{\mathrm{i=1}}^{\mathrm{m}}\mathrm{p_i}\big(\mathcal{A}_{\mathrm{io}}-\mathcal{B}_{\mathrm{io}}\mathbf{b}\big)=\mathbf{0}.\tag{8.5}$$

The non–singularity of $\sum_{i=1}^m p_i\mathcal{B}_{io}$ can be established by the fact we show below, that is, $\sum_{i=1}^m p_i\mathcal{B}_{io}=E\big\{\delta(\mathbf{X}-\mu_x)(\mathbf{X}-\mu_x)'\big\}$ and the assumption A2. Notice that under the assumption A1, we can write $\mathcal{B}_{io}$ in (8.3) as

$$\begin{aligned}\mathcal{B}_{io}=\ &(\mathbf{X}_i-\mu_x)\Big[\mathbf{X}_i'\delta_i+(1-\delta_i)\eta_i\mu_x'\Big]\\&+(1-\delta_i)(1-\eta_i)\frac{E\Big[E\Big\{(\mathbf{X}-\mu_x)\mathbf{1}(\epsilon^*>C-\beta\mathbf{X})|C,\mathbf{X}\Big\}\Big]\mu_x'}{E\Big\{S_o(C-\beta\mathbf{X})\Big\}}\\=\ &\delta_i(\mathbf{X}_i-\mu_x)(\mathbf{X}_i-\mu_x)'+\Big\{\delta_i+(1-\delta_i)\eta_i\Big\}(\mathbf{X}_i-\mu_x)\mu_x'\\&+(1-\delta_i)(1-\eta_i)\frac{E\Big\{(\mathbf{X}-\mu_x)\mathbf{1}(\epsilon^*>C-\beta\mathbf{X})\Big\}\mu_x'}{E\Big\{S_o(C-\beta\mathbf{X})\Big\}}.\end{aligned}$$

Hence

$$\begin{aligned}\textstyle\sum_{i=1}^m p_i\mathcal{B}_{io}=\ &\textstyle\sum_{i=1}^m p_i\delta_i(\mathbf{X}_i-\mu_x)(\mathbf{X}_i-\mu_x)'+\sum_{i=1}^m p_i\Big\{\delta_i+(1-\delta_i)\eta_i\Big\}(\mathbf{X}_i-\mu_x)\mu_x'\\&\textstyle+\sum_{i=1}^m p_i(1-\delta_i)(1-\eta_i)\frac{E\Big\{(\mathbf{X}-\mu_x)\mathbf{1}(\epsilon^*>C-\beta\mathbf{X})\Big\}\mu_x'}{E\Big\{S_o(C-\beta\mathbf{X})\Big\}}.\end{aligned}$$

Recall that $p_i=P\{(M,\delta,\eta,\mathbf{X})=(M_i,\delta_i,\eta_i,\mathbf{X}_i)\}$. Hence we have $\sum_{i=1}^m(1-\delta_i)(1-\eta_i)p_i=E\big\{S_o(C-\beta\mathbf{X})(1-\eta)\big\}=E\big\{S_o(C-\beta\mathbf{X})\big\}E(1-\eta)$, where the last equality follows from A1. Now the third term in the right hand side of the foregoing equation can be written as $E[(1-\delta)(1-\eta)(\mathbf{X}-\mu_x)\mu_x']$. As a result,

$$\begin{aligned}\sum_{i=1}^m p_i\mathcal{B}_{io} = \quad & E\big\{\delta(\mathbf{X}-\mu_x)(\mathbf{X}-\mu_x)'\big\} + E\Big[\big\{\delta+(1-\delta)\eta\big\}(\mathbf{X}-\mu_x)\mu_x'\Big] \\ & +E\big\{(1-\delta)(1-\eta)(\mathbf{X}-\mu_x)\mu_x'\big\} \\ = \quad & E\big\{\delta(\mathbf{X}-\mu_x)(\mathbf{X}-\mu_x)'\big\}.\end{aligned}$$

For (8.5), verify that

$$\begin{aligned}& \sum_{i=1}^m p_i(\mathcal{A}_{io}-\mathcal{B}_{io}\beta) \\ = \; & \sum_{i=1}^m p_i\Bigg\{(\mathbf{X}_i-\mu_x)\Big[\delta_i T_i(\beta)+(1-\delta_i)\eta_i E\big\{\epsilon|\epsilon>T_i(\beta)\big\}\Big] \\ & \qquad +(1-\delta_i)(1-\eta_i)\frac{E\Big[(\mathbf{X}-\mu_x)E\big\{\epsilon^*\mathbf{1}(\epsilon^*>C-\beta\mathbf{X})|C,\mathbf{X}\big\}\Big]}{E\big\{S_o(C-\beta\mathbf{X})\big\}}\Bigg\} \\ = \; & \sum_{i=1}^m p_i\Bigg\{(\mathbf{X}_i-\mu_x)\Big[\delta_i T_i(\beta)+(1-\delta_i)\eta_i E\big\{\epsilon|\epsilon>T_i(\beta)\big\}\Big] \\ & \qquad +(1-\delta_i)(1-\eta_i)\frac{E\Big[(\mathbf{X}-\mu_x)\epsilon\mathbf{1}(\epsilon>C-\beta\mathbf{X})\Big]}{E\big\{S_o(C-\beta\mathbf{X})\big\}}\Bigg\} \\ = \; & \sum_{i=1}^m \Bigg[E\big\{(\mathbf{X}-\mu_x)\epsilon\,\mathbf{1}(\delta_i=1,\mathbf{X}=\mathbf{X}_i,M_i<C)\big\} \\ & \qquad +E\big\{(\mathbf{X}-\mu_x)\epsilon\,\mathbf{1}(\delta_i=0,\eta_i=1,\mathbf{X}=\mathbf{X}_i,C=M_i)\big\} \\ & \qquad +E\big\{(\mathbf{X}-\mu_x)\epsilon\,\mathbf{1}(\delta_i=0,\eta_i=0)\big\}\Bigg] \\ = \; & E\big\{\epsilon(\mathbf{X}-\mu_x)\big\}=\mathbf{0}.\end{aligned}$$

where the last equality follows from the assumption that ϵ and $\mathbf{X}$ are independent. Moreover, since $\sum_{i=1}^m p_i\mathcal{B}_{io}$ is nonsingular the solution to $\sum_{i=1}^m p_i(\mathcal{A}_{io}-\mathcal{B}_{io}\mathbf{b})=\mathbf{0}$ is unique.

By (8.3) and (8.4),

$$\sum_{i=1}^m \overline{N_i}\mathcal{A}_i(\beta)-\sum_{i=1}^m \overline{N_i}\mathcal{B}_i(\beta)\mathbf{b}=\frac{\widetilde{H}(\mathbf{b})}{n}\begin{cases}\to\mathbf{0} & \text{if } \mathbf{b}=\beta,\\ \not\to\mathbf{0} & \text{if } \mathbf{b}\in O(\beta,c)\setminus\{\beta\}.\end{cases} \tag{8.6}$$

It follows from (8.5) that if n is large enough, $\sum_{i=1}^m \overline{N_i}\mathcal{B}_i(\beta)\to\sum_{i=1}^m p_i\mathcal{B}_{io}$, thus it is nonsingular and the root of $\sum_{i=1}^m \overline{N_i}\mathcal{A}_i(\beta)-\sum_{i=1}^m \overline{N_i}\mathcal{B}_i(\beta)\mathbf{b}$ satisfies that

$$\hat{\mathbf{b}}=\Big\{\sum_{i=1}^m \overline{N_i}\mathcal{B}_i(\beta)\Big\}^{-1}\sum_{i=1}^m \overline{N_i}\mathcal{A}_i(\beta)=(\mathcal{B}(\beta))^{-1}\mathcal{A}(\beta)\to\beta \textit{ as } n\to\infty. \tag{8.7}$$

Though $\hat{\mathbf{b}}$ may not be a BJE for a small n, for a large enough n, $\hat{\mathbf{b}}\in O(\beta,c)$ by (8.7), thus

$$\mathbf{0}=\mathcal{A}(\beta)-\mathcal{B}(\beta)\hat{\mathbf{b}}=\mathcal{A}(\hat{\mathbf{b}})-\mathcal{B}(\hat{\mathbf{b}})\hat{\mathbf{b}}=\widetilde{H}(\hat{\mathbf{b}}). \tag{8.8}$$

As a consequence, $\hat{\beta}_4 = \hat{\mathbf{b}} = (\mathcal{B}(\beta))^{-1}\mathcal{A}(\beta)$ if n is large enough. The consistency of the BJE follows from (8.7) and (8.8).

Step 3 (verify statement (c)). Fix a $\mathbf{b}$. By (3.1), we have $\tilde{S}_{\mathbf{b}}(t) = \prod_{j:\ t_j \geq t}(1 - \frac{d_j}{R_j})$, and d_j's and R_j's are only rational functions of $\overline{\mathbf{N}}(\mathbf{b})$ and w_i, which is again a rational function of $\overline{\mathbf{N}}(\mathbf{b})$. Hence,

$$\tilde{S}_{\mathbf{b}}(T_i(\mathbf{b})) \text{ is only a rational function of } \overline{\mathbf{N}}(\mathbf{b}) \text{ for each } i. \tag{8.9}$$

In view of the expressions of $\tilde{f}_{\mathbf{X}}(\mathbf{X}_i)$ and $\tilde{f}_C(M_i)$ (see (3.1)), they are only rational functions of $\mathbf{X}_i$, M_i, w_i and $\overline{\mathbf{N}}$. $\mathbf{X}$ and C only take on finitely many values, if n is large then one can treat $\{(\mathbf{X}_i, M_i) :\ i = 1, ..., m\}$ as fixed constants. Now it can be verified that

$$\tilde{f}_{\mathbf{X}}(\mathbf{X}_i) \text{ and } \tilde{f}_C(M_i) \text{ are only rational functions of } \overline{\mathbf{N}}(\mathbf{b}) \text{ for each } i. \tag{8.10}$$

(8.9) and (8.10) yield statement (c).

Step 4 (verify statement (d)). Now from the expressions of $\mathcal{A}_i(\beta)$ and $\mathcal{B}_i(\beta)$ in (4.2), it is seen that they are only rational functions of M_i, $\mathbf{X}_i$, $\tilde{f}_{\mathbf{X}}(\mathbf{X}_i)$, $\tilde{f}_C(M_i)$ and $\tilde{f}_\beta(T_i(\beta))$. Again M_i and $\mathbf{X}_i$ can be viewed as constants, thus it follows from (8.2) and (8.3) that $\mathcal{A}_i(\beta)$ and $\mathcal{B}_i(\beta)$ are only rational functions of $\overline{\mathbf{N}}(\beta)$. Since $\mathcal{A}(\beta)$ and $\mathcal{B}(\beta)$ are linear functions of $\mathcal{A}_i$ and $\mathcal{B}_i$,

$$\mathcal{A}(\beta) \text{ and } \mathcal{B}(\beta) \text{ are only rational functions of } \overline{\mathbf{N}}(\beta) \tag{8.11}$$

Since $\hat{\beta}_4 = (\mathcal{B}(\beta))^{-1}\mathcal{A}(\beta)$, under A5 and A4 and by (8.11) and (8.7) $\hat{\beta}_4$ is only a rational function of $\overline{\mathbf{N}}(\beta)$, and $\overline{\mathbf{N}}(\beta)$ is a sample mean of a random vector with finite dimension $(m-1)$. Thus $\hat{\beta}_4 = g(\overline{\mathbf{N}})$, where g is a function on $\mathcal{R}^{m-1}$ with continuous partial derivatives. It follows that the asymptotic normality can now be shown using Slutsky's theorem and the central limit theorem, provided that $g'(\mathbf{y}) \neq \mathbf{0}$, where $\mathbf{y} = E(\overline{\mathbf{N}}(\mathbf{b}))$. Of course, the asymptotic covariance matrix of the BJE can be derived using the delta method as well under A5 and A4. For simplicity, we skip the details. ▫

References

[1] Barlow, W. E., Ichikawa, L., Rosner, D. and Izumi, S. (1999). Analysis of Case-Cohort Designs. *Journal of Clinical Epidemiology* **52** 11651172.

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

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

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

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

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

[7] Kalbfleisch, J. D. and Prentice, R. L. (2002), *The Statistical Analysis of Failure Time Data*, 2nd Ed., New York, John Wiley & Sons, Inc.

[8] Kong, F.H. and Yu, Q.Q (2005). Asymptotic distributions of the Buckley-James estimator under non-standard conditions. *Statistica Sinica* (accepted).

[9] Lai, T.L. and 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.

[10] Li, L.X. and Pu, Z.W. (1999). Regression models with arbitrarily interval-censored observations. *Comm. in Statist., Theory and Methods* 1547–1563.

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

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

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

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

[15] Stare J, Harrell F.E., Heinzl H.B.J. (2001). An S-Plus program to fit linear regression models to censored data using the Buckley-James method. *Computer Methods and Programs in Biomedicine*, **64**(1): 45–52.

[16] Turnbull, B. W. (1976). The empirical distribution function with arbitrary grouped, censored and truncated data. *J. Roy. Statistics Soc. Ser. B* **38** 290–295.

[17] Yu, Q.Q. and Wong, G.Y.C. (2002). How to find all Buckley-James estimates instead of just one? *Journal of Statistical Computation and Simulation.* **72** 451–460.

[18] Yu, Q.Q. and Yu, M.G. (2004). Extensions of Buckley-James estimators to modified case-cohort studies. (Unpublished manuscript).

[19] Yu, Q.Q., Wong, G.Y.C. and Yu, M.G. (2006). Buckley–James–type of estimators under the classical case–cohort design. *Ann. Inst. Statist. Math.* (Accepted).