Statistics and Its Interface Volume 13 (2020) 181–191

Additive hazards regression for case-cohort studies with interval-censored data

Mingyue Du*, Huiqiong Li†‡, and Jianguo Sun§

A large literature has been developed for the analysis of case-cohort studies that are often performed with the aim of reducing the cost on the collection of covariate information. In particular, many authors have discussed their regression analysis under the framework of the additive hazards model, which is often preferred when the risk difference is of main interest. However, all of the existing methods assume or are applicable only to right-censored data. In this paper, we consider the case of interval-censored data, which often occur in practice and include right-censored data as a special case, and propose two estimation approaches, an estimating equation-based method and a maximum likelihood method. The resulting estimators of regression parameters are shown to be consistent and asymptotically normal. Also a simulation study is conducted and suggests that the proposed methods works well in practice, and an application is provided.

Keywords and phrases: Additive hazards model, Case-cohort design, Interval censoring, Sieve estimation.

1. INTRODUCTION

Since its proposal (Prentice, 1986), a large literature has been developed for the case-cohort design that aims to reduce the cost on the collection of covariate information among others. In large epidemiological cohort studies, the assembling or collecting of covariate information on all study subjects may be expensive and some examples of such covariates include chemical exposures in blood samples and genetic information. Instead of collecting the information from all subjects, the case-cohort design selects a random sample or subcohort from the original whole cohort and collects or measures the covariate information only from the subjects in the subcohort or who experience the failure event of interest.

Many estimation methods have been proposed for the analysis of case-cohort studies or the failure time data collected under the case-cohort design (Chen and Lo, 1999; Nan, 2004; Scheike and Martinussen, 2004). For example, Prentice (1986) and Self and Prentice (1988) developed some pseudo likelihood estimation procedures and Barlow (1994) gave a robust variance estimation procedure for the data arising from the proportional hazards models (Cox, 1972). Chen (2001) and Keogh and White (2013) also considered the same problem. The former investigated some estimating equation-based methods and the latter developed some multiple imputation-based methods. Furthermore Kang and Cai (2009), Kang et al. (2013) and Kim et al. (2013) discussed the situation where there exist multivariate failure times and developed some weighted estimating equation-based approaches.

Note that all of the methods described above assume or apply only to right-censored failure time data, where the failure time of interest is either exactly observed or right-censored (Kalbfleisch and Prentice, 2002). In many epidemiological or medical follow-up studies, however, it is common that one can only observe interval-censored failure time data, meaning that the failure time of interest is known or observed only to belong to an interval. It is apparent that interval-censored data include right-censored data as a special case, and there exist several methods for the analysis of case-cohort studies that yield interval-censored data (Li and Nan, 2011). For example, Li et al. (2008) discussed the problem when one faces the situation where the observation process generating censoring intervals is the same for all subjects, and Zhou et al. (2017) developed a sieve semiparametric maximum likelihood estimation approach. However, all of these methods assume the proportional hazards model and it is well-known that this assumption may not hold in practice. Also in many situations such as epidemiological studies, one is often interested in the risk difference and in these cases, the additive hazards model is usually preferred over the proportional hazards model (Lin et al., 1998). In the following, we will consider this latter situation, for which there does not seem to exist an established estimation procedure, and propose two estimation methods.

The remainder of the paper is organized as follows. We will first describe in Section 2 some notation, the model and some assumptions that will be used throughout the paper, and two estimation procedures will be developed in Section 3. One is an estimating equation-based procedure and the other is a maximum likelihood estimation procedure. Furthermore the resulting estimators of regression parameters will be shown to be consistent and follow asymptotically

*Center for Applied Statistical Research, School of Mathematics, Jilin University.

†Department of Statistics, Yunnan University.

‡Corresponding author.

§Department of Statistics, University of Missouri.

a normal distribution. Section 4 presents some results obtained from a simulation study conducted to assess the finite sample performance of the proposed methods and they suggest that both methods work well in practical situations. An application is provided in Section 5 and Section 6 contains some discussion and concluding remarks.

2. NOTATION, MODEL AND ASSUMPTIONS

Consider a cohort study that consists of n independent subjects and for subject i, let T_i denote the associated failure time of interest and Z_i a p-dimensional vector of covariates that may be related to T_i. For the relationship between T_i and Z_i, we will assume that given Z_i, the hazard function of T_i has the form

$$\lambda_i(t|Z_i) = \lambda(t) + \beta' Z_i(t)\,, \tag{1}$$

where $\lambda(t)$ is an unknown baseline hazard function and β a p-dimensional vector of regression parameters. That is, T_i follows the additive hazards model (Lin et al., 1998). In the following, it will be assumed that the focus is on estimation of the covariate effect or β.

For inference about model (1), suppose that for subject i, there exist two examination times denoted by U_i and V_i with $U_i < V_i$. Define the indicator functions $\delta_{1i} = I(T_i \le U_i)$, $\delta_{2i} = I(U_i < T_i \le V_i)$ and $\delta_{3i} = 1 - \delta_{1i} - \delta_{2i}$, indicating if the failure event of interest occurs before or at U_i, during the interval $(U_i, V_i]$, or after V_i, respectively. Also define $O_i = \{\, U_i, V_i, \delta_{1i}, \delta_{2i}, Z_i \,\}$ and assume that the observed full cohort data would be $O = \{\, O_i,\ i = 1, ..., n \,\}$ if the covariate information is available for all subjects. That is, we have interval-censored data on the T_i's (Sun, 2006). Then the corresponding likelihood function would have the form

$$L_n(\beta, \Lambda|O) = \prod_{i=1}^{n} \Big\{ [1 - \exp\{-\Lambda(U_i) - \beta' Z_i^*(U_i)\}]^{\delta_{1i}}$$
$$\times [\exp\{-\Lambda(U_i) - \beta' Z_i^*(U_i)\} - \exp\{-\Lambda(V_i) - \beta' Z_i^*(U_i)\}]^{\delta_{2i}}$$
$$\times [\exp\{-\Lambda(V_i) - \beta' Z_i^*(V_i)\}]^{\delta_{3i}} \Big\}\,,$$

where $\Lambda(t) = \int_0^t \lambda(s) ds$.

Of course, for case-cohort studies, the covariate information is available only for the subjects from the subcohort or who have experienced the failure event of interest. Define $\xi_i = 1$ if the covariate Z_i is known or observed and 0 otherwise, $i = 1, \ldots, n$. Then under the case-cohort design, the observed data have the form

$$O^{\xi} = \{\, O_i^{\xi} = (U_i, V_i, \delta_{1i}, \delta_{2i}, \xi_i Z_i, \xi_i);\ i = 1, \ldots, n \,\}\,.$$

For the selection of the subcohort, by following Zhou et al. (2017), we will focus on the independent Bernoulli sampling with the selection probability $q \in (0, 1)$. Then the probability that the covariate Z_i can be observed is given by

$$\Pr(\xi_i = 1) \overset{d}{=} \pi_q(\delta_{1i}, \delta_{2i}) = \delta_{1i} + \delta_{2i} + (1 - \delta_{1i} - \delta_{2i})q\,,$$

$i = 1, \ldots, n$. Also we will assume that given Z_i, T_i is independent of the examination process or times U_i and V_i. That is, we have the independent censoring mechanism (Sun, 2006).

3. ESTIMATION PROCEDURES

In this section, we will propose two estimation procedures for the regression parameter β. First we will describe an estimating equation-based procedure with the use of the inverse probability weighting technique and then a pseudo likelihood-based procedure.

3.1 Inverse probability weighted estimation

To present the estimating equation-based estimation procedure, for each i, define $N_i^{(1)}(t) = (1 - \delta_{1i}) I(U_i \le t)$. Also conditional on U_i, define $N_i^{(2)}(t) = \delta_{3i} I(V_i \le t)$ if $t \ge U_i$ and 0 if $t < U_i$, and for $k = 0$, 1 and 2, define

$$S_{1,\beta}^{(k)}(t, \beta) = \frac{1}{n} \sum_{i=1}^{n} I(t \le U_i) \exp\{-\beta' Z_i{}^*(t)\} Z_i^*(t)^{\otimes k}\,,$$

and

$$S_{2,\beta}^{(k)}(t, \beta) = \frac{1}{n} \sum_{i=1}^{n} I(U_i < t \le V_i) \exp\{-\beta' Z_i{}^*(t)\} Z_i^*(t)^{\otimes k}$$

with $Z_i^*(t)^{\otimes 0} = 1$, $Z_i^*(t)^{\otimes 1} = Z_i^*(t)$, $Z_i^*(t)^{\otimes 2} = Z_i^*(t) Z_i^*(t)^T$. First note that if the full-cohort data O were available, Wang et al. (2010) suggested to estimate the regression parameter β by solving the estimating function

$$U(\beta) = \sum_{i=1}^{n} \left[\int_0^{\tau_1} \left\{ Z_i^*(t) - \frac{S_{1,\beta}^{(1)}(t, \beta)}{S_{1,\beta}^{(0)}(t, \beta)} \right\} dN_i^{(1)}(t) \right.$$
$$\left. + \int_0^{\tau_2} \left\{ Z_i^*(t) - \frac{S_{2,\beta}^{(1)}(t, \beta)}{S_{2,\beta}^{(0)}(t, \beta)} \right\} dN_i^{(2)}(t) \right]\,, \tag{2}$$

where $Z_i^*(t) = \int_0^t Z_i(s) ds$, $\tau_1 = \sup\{t : \Pr(U \ge t) > 0\}$, and $\tau_2 = \sup\{t : \Pr(U < t \le V) > 0\}$.

For the case of case-cohort studies, it is apparent that the estimating equation above is not available. To address this, for $k = 0$, 1 and 2, define

$$S_{w,1,\beta}^{(k)}(t, \beta) = \frac{1}{n} \sum_{i=1}^{n} \omega_i I(t \le U_i) \exp\{-\beta' Z_i{}^*(t)\} [Z_i^*(t)]^{\otimes k}\,,$$

and

$$S_{w,2,\beta}^{(k)}(t, \beta) = \frac{1}{n} \sum_{i=1}^{n} \omega_i I(U_i < t \le V_i) \exp\{-\beta' Z_i{}^*(t)\} [Z_i^*(t)]^{\otimes k},$$

where the w_i's are some weights given by

$$w_i = \frac{\xi_i}{\delta_{1i} + \delta_{2i} + (1 - \delta_{1i} - \delta_{2i})q} \triangleq \frac{\xi_i}{\pi_q(\delta_{1i}, \delta_{2i})}\,.$$

Then by following Zhou et al. (2017) and motivated by (2), we consider the following inverse probability weighted estimating function

$$U_{IPW}(\beta) = \sum_{i=1}^{n}\left[\int_0^{\tau_1} w_i\left\{Z_i^*(t) - \frac{S^{(1)}_{w,1,\beta}(t,\beta)}{S^{(0)}_{w,1,\beta}(t,\beta)}\right\} dN_i^{(1)}(t) + \int_0^{\tau_2} w_i\left\{Z_i^*(t) - \frac{S^{(1)}_{w,2,\beta}(t,\beta)}{S^{(0)}_{w,2,\beta}(t,\beta)}\right\} dN_i^{(2)}(t)\right],$$

and define the inverse probability weight estimator $\hat{\beta}_{IPW}$ of β as the solution to $U_{IPW}(\beta) = 0$. The following theorem establishes the asymptotic properties of $\hat{\beta}_{IPW}$.

Theorem 1. Suppose that the regularity conditions $(A1)-(A4)$ given in the Appendix hold. Then $\hat{\beta}_{IPW}$ is consistent and as $n \to \infty$, we have that

$$n^{1/2}(\hat{\beta}_{IPW} - \beta_0) \longrightarrow N(0, \mathbf{\Omega}_w^{-1}\Gamma_w\mathbf{\Omega}_w^{-1})$$

in distribution, where $\mathbf{\Omega}_w = B_1 + B_2$ and

$$\Gamma_w = E\left(\frac{1}{\pi_q(\delta_{1i},\delta_{2i})}\left[\left\{\int_0^{\tau_1}\left\{Z_i^*(t) - \frac{s^{(1)}_{w,1,\beta}(t,\beta)}{s^{(0)}_{w,1,\beta}(t,\beta)}\right\} dM_i^{(1)}(t)\right\}^{\otimes 2} + \left\{\int_0^{\tau_2}\left\{Z_i^*(t) - \frac{s^{(1)}_{w,2,\beta}(t,\beta)}{s^{(0)}_{w,2,\beta}(t,\beta)}\right\} dM_i^{(2)}(t)\right\}^{\otimes 2}\right]\right)$$

with

$$B_1 = E\left(\int_0^{\tau_1}\left\{Z_i^*(t) - \frac{s^{(1)}_{w,1,\beta}(t,\beta)}{s^{(0)}_{w,1,\beta}(t,\beta)}\right\}^{\otimes 2} I(U_i \geq t)\exp\left\{-\Lambda(t) - \beta' Z_i^*(t)\right\} dt\right),$$

$$B_2 = E\left(\int_0^{\tau_2}\left\{Z_i^*(t) - \frac{s^{(1)}_{w,2,\beta}(t,\beta)}{s^{(0)}_{w,2,\beta}(t,\beta)}\right\}^{\otimes 2} I(U_i < t \leq V_i)\exp\left\{-\Lambda(t) - \beta' Z_i^*(t)\right\} dt\right),$$

$$M_i^{(1)}(t) = N_i^{(1)}(t) - \int_0^t I(U_i \geq s)\exp\{-\Lambda(s) - \beta' Z_i^*(s))\}ds,$$

$$M_i^{(2)}(t) = N_i^{(2)}(t) - \int_0^t I(U_i < s \leq V_i)\exp\{-\Lambda(s) - \beta' Z_i^*(s)\}ds,$$

$$s^{(k)}_{w,1,\beta}(t,\beta) = E\left(I(t \leq U_i)\exp\{-\beta' Z_i{}^*(t)\}[Z_i^*(t)]^{\otimes k}\right),\ \text{for } k = 0,1,2,$$

and

$$s^{(k)}_{w,2,\beta}(t,\beta) = E\left(I(U_i < t \leq V_i)\exp\{-\beta' Z_i{}^*(t)\}[Z_i^*(t)]^{\otimes k}\right),$$

for $k = 0,1,2$.

The proof of the theorem above is sketched in the Appendix. For the application of the result above, one needs to estimate the asymptotic covariance matrix, and although it is possible to derive a consistent estimator, it would be complicated. Thus instead by following Ma and Kosorok (2005), we suggest to employ the nonparametric weighted bootstrap procedure to estimate the covariance matrix of $\hat{\beta}_{IPW}$ as follows. Specifically, let B be a given integer and for each b $(1 \leq b \leq B)$, let $\{u_1^b, \ldots, u_n^b\}$ denote n independent realizations of a bounded positive random variable u satisfying $E(u) = 1$ and $\text{var}(u) = \varepsilon_0 < \infty$. Define the new weights $w_i^b = u_i^b w_i$ for $i = 1, \ldots, n$ and let $\hat{\beta}^b_{IPW}$ denote the estimator of θ defined above with replacing the w_i's by the w_i^b's. Then one can estimate the asymptotic covariance matrix of $\hat{\beta}_{IPW}$ by the sample covariance matrix of the $\hat{\beta}^b_{IPW}$'s. One can expect that the estimator is consistent (Ma and Kosorok, 2005) and the numerical results below indicate that this method seems to work well.

Note that one advantage of the approach given above is that it does not involve the estimation of the baseline hazard function $\Lambda(t)$ and thus it can be relatively stable or robust as discussed below. On the other hand, it may lose some efficiency and corresponding to this, we will present a pseudo likelihood-based approach in the next subsection.

3.2 Sieve pseudo maximum likelihood estimation

In general, if all information was available, one would prefer to estimate β by maximizing the likelihood function $L_n(\beta, \Lambda)$. Since this is not possible, instead by following Zhou et al. (2017), we suggest to consider the inverse probability weighted pseudo log likelihood function

$$\begin{aligned} l_n^w(\beta,\Lambda) &= \sum_{i=1}^{n} l^w(\beta,\Lambda;O_i^{\xi}) = \sum_{i=1}^{n} w_i l(\beta,\Lambda;O_i) \\ &= \sum_{i=1}^{n} w_i\Big\{\delta_{1i}\log\left[1 - \exp\{-\Lambda(U_i) - \beta' Z_i^*(U_i)\}\right] \\ &+\delta_{2i}\log\left[\exp\{-\Lambda(U_i) - \beta' Z_i^*(U_i)\} - \exp\{-\Lambda(V_i) - \beta' Z_i^*(U_i)\}\right] \\ &+(1-\delta_{1i}-\delta_{2i})\{-\Lambda(V_i) - \beta' Z_i^*(V_i)\}\Big\}. \end{aligned} \tag{3}$$

Of course, as mentioned above, now we have to deal with the estimation of β and the baseline hazard function $\Lambda(t)$ together, which may be difficult. For this, by following Ma et al. (2015) and others, we will approximate $\Lambda(t)$ first by Bernstein polynomials.

Specifically, let

$$\Theta = \Big\{\theta = (\beta, \Lambda) \in \mathcal{B} \otimes \mathcal{M}\Big\}$$

denote the parameter space of θ, where $\mathcal{B} = \{\beta \in \mathbb{R}^p : ||\beta|| \leq M\}$ with M being a positive constant and $\mathcal{M}$ is the collection of all continuous nonnegative and nondecreasing functions over the interval $[r_1, r_2]$. Here, r_1 and r_2 are supposed to be known constants that are usually taken in practice to be the lower and upper bounds of all observation times. Also denote the sieve space

$$\Theta_n = \Big\{\theta_n = (\beta, \Lambda_n) \in \mathcal{B} \otimes \mathcal{M}_n\Big\},$$

where

$$\mathcal{M}_n = \left\{ \Lambda_n(t) = \sum_{k=0}^{m} \phi_k B_k(t,m,r_1,r_2) : \phi_m \geq \cdots \geq \phi_1 \geq \phi_0, \sum_{k=0}^{m} |\phi_k| \leq M_n \right\}.$$

In the above, $B_k(t,m,r_1,r_2)$ denotes the Bernstein basis polynomial of degree $m = o(n^v)$ for some $v \in (0,1)$ given by

$$B_k(t,m,r_1,r_2) = \left(\begin{array}{c} m \\ k \end{array} \right) (\frac{t-r_1}{r_2-r_1})^k (1 - \frac{t-r_1}{r_2-r_1})^{m-k} \quad (k = 0, \ldots, m),$$

and $M_n = O(n^a)$ with a being a positive constant (Lorentz 1986; Shen 1997). We will define the sieve pseudo maximum likelihood estimator $\hat{\theta}_n = (\hat{\beta}_n, \hat{\Lambda}_n)$ of θ to be the value of θ that maximizes the pseudo log likelihood function $l_n^w(\theta)$ over Θ_n.

Note that one advantage of the use of Bernstein polynomials is that it can easily accommodate the non-decreasing property of the baseline cumulative hazard function $\Lambda(t)$. Also the method can be relatively easily implemented. Of course, instead of Bernstein polynomials, one may employ other smooth functions such as B-spline or I-spline functions for the approximation. To establish the asymptotic properties of the proposed estimator $\hat{\theta}_n$, let $G(u,v)$ denote the joint density function of the two random examination times U and V and $g(u,v|z)$ the conditional density of U and V given $Z = z$. Also let $\theta_0 = (\beta_0, \Lambda_0)$ denote the true value of θ and define the distance between $\theta^1 = (\beta^1, \Lambda^1)$ and $\theta^2 = (\beta^2, \Lambda^2)$ as

$$d(\theta^1, \theta^2) = \{ ||\beta^1 - \beta^2||^2 + ||\Lambda^1 - \Lambda^2||_2^2 \}^{1/2},$$

where $||v||$ denotes the Euclidean norm of a vector v and $||\Lambda^1 - \Lambda^2||_2^2 = \int[\{\Lambda^1(u) - \Lambda^2(u)\}^2 + \{\Lambda^1(v) - \Lambda^2(v)\}^2] dG(u,v)$. Then the following theorems establish the asymptotic consistency and normality of the proposed estimators.

Theorem 2. Suppose that the regularity conditions $(A1)$, $(A3) - (A6)$ given in the Appendix hold. Then as $n \to \infty$, we have that $d(\hat{\theta}_n, \theta_0) \to 0$ almost surely and $d(\hat{\theta}_n, \theta_0) = O_p(n^{-min\{(1-v)/2, vr/2\}})$, where $v \in (0,1)$ such that $m = o(n^v)$ and r is defined in the regularity condition $(A5)$.

Theorem 3. Suppose that the regularity conditions $(A1)$, $(A3) - (A6)$ given in the Appendix hold with $r > 2$ in the regularity condition $(A5)$. Then if $v > 1/(2r)$ and as $n \to \infty$, we have that

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = I^{-1}(\beta_0) n^{-1/2} \sum_{i=1}^{n} w_i l^*(\beta_0, \Lambda_0; O_i) + o_p(1) \to N(0, \Sigma)$$

in distribution with

$$\Sigma = I^{-1}(\beta_0) + I^{-1}(\beta_0) E \left\{ \frac{1 - \pi_q(\delta_1, \delta_2)}{\pi_q(\delta_1, \delta_2)} \{ l^*(\beta_0, \Lambda_0; O) \}^{\otimes 2} \right\} \times I^{-1}(\beta_0),$$

where $v^{\otimes 2} = vv^T$ for a vector v, and $I(\beta)$ and $l^*(\beta_0, \Lambda_0; O)$ denote the information matrix and efficient score for β, respectively, based on a single observation and given in the Appendix.

The proof of the theorems above will be sketched in the Appendix. To use the results above, as before, we need to estimate the covariance matrix of $\hat{\beta}_n$, for which it would be difficult to derive a reasonable, consistent estimator. Thus again as before, we suggest to employ the weighted bootstrap procedure of Ma and Kosorok (2005), which is easy to be implemented and works reasonably well as seen below. Specifically, let B, $\{u_1^b, \ldots, u_n^b\}$ and the $w_i^b = u_i^b w_i$'s be defined as before, and $\hat{\theta}_n^b = (\hat{\beta}_n^b, \hat{\Lambda}_n^b)$ denote the estimator of θ defined above based on the new weights w_i^b's. Then one can estimate the asymptotic covariance matrix of $\hat{\beta}_n$ by the sample covariance matrix of the $\hat{\beta}_n^b$'s.

For the implementation of the estimation approach proposed above, also note that there are some restrictions on the parameters due to the nonnegativity and monotonicity of the baseline cumulative hazard function $\Lambda(t)$. However, they can be easily removed by using some reparameterization. For example, a commonly used method is to reparameterize $\{\phi_0, \cdots, \phi_m\}$ by the cumulative sums of $\{\exp(\phi_0^*), \cdots, \exp(\phi_m^*)\}$. On the restriction $\sum_{k=0}^{m} |\phi_k| \leq M_n$, it can be usually ignored since $M_n = O(n^a)$ is needed mainly for technical reasons and can be chosen reasonably large for a fixed sample size in practice. For the determination of $\hat{\theta}_n$, many existing optimization methods can be used, including the interior-point algorithm and the Newton-Raphson method. For the numerical studies below, the interior-point algorithm in Matlab, given in fmincon used. In addition, one needs to choose or determine the degree m of Bernstein polynomials, which controls the smoothness of the approximation. For this, as suggested by other authors, one could consider several different values of m and choose the value based on the AIC criterion that minimizes

$$AIC = -2\, l_n^w(\hat{\theta}_n) + 2(p + m + 1).$$

4. A SIMULATION STUDY

Now we report some results obtained from a simulation study conducted to assess the finite sample performance of the two estimation procedures proposed in the previous sections. In the study, we considered two covariate situations with one being that there exists only one covariate Z_i generated from the Bernoulli distribution with the probability of success 0.5. For the other situation, it was assumed that there exists two covariates, one being discrete like in the first situation and the other being continuous and following the uniform distribution over $(0, 1)$. Given Z_i, the failure time T_i was then generated from model (1) with the cumulative baseline hazard function $\Lambda(t) = 0.5t^2$ or $\Lambda(t) = t$. For the generation of the observed interval-censored data, we mimicked biomedical follow-up studies and assumed that there existed a sequence of observation time points for each subject. Specifically, let $e_1, \ldots, e_k$ denote the k equally space time-points over the interval $(0, \tau)$, where τ represents the stopping time of the study. Then for each subject, define a new sequence of time points $\{e_j^*\}_{j=1}^k$ by setting e_j^* to be e_i plus a random number generated from the uniform distribution over $(-\tau/3(k+1), \tau/3(k+1))$ and assuming that the subject was observed at each e_j^* with probability ς, independent of the examination results at other time-points.

For subject i, given the sequence of real observation time points, if the generated T_i is less than the smallest observation time point, we set U_i to be the smallest observation time and $V_i = \tau$, and if the generated T_i is larger than largest observation time point, we took $U_i = 0$ and V_i to be the largest observation time point. Otherwise, U_i and V_i were taken to be the two consecutive observation time points bracketing the generated T_i. For the results given below, we set $k = 8$ and $\varsigma = 0.8$ and determined τ according to the desired percentages of the subjects with $\delta_1 = 1$ and $\delta_2 = 1$. For the proportion of the observed failure events, we considered the situation with $p = 0.05$ or $p = 0.1$ and for the generation of the subcohort, we adopt the independent Bernoulli sampling with the selection probability $q = 0.2$. The results given below are based on $n = 1000$ or 2000 with 1000 replications.

Tables 1 and 2 present the results obtained on estimation of the regression parameter β with one covariate, the true value of β being 0.2, 0.5 or log(2), $\Lambda(t) = 0.5t^2$, and $p = 0.1$ and 0.05, respectively. The results include the estimated bias (Bias) of the proposed estimators $\hat{\beta}_{IPW}$ and $\hat{\beta}_n$ given by the average of the estimates minus the true value, the sample standard error of the estimates (SSE), the average of the estimated standard errors (ESE) and the 95% empirical coverage probability (CP). Here for the variance estimations of the two proposed estimators, we generated the random sample $\{u_1^b, ..., u_n^b\}$ from the exponential distribution with $B = 100$. One can see from the tables that both proposed estimators seem to be unbiased and the variance estimation procedures also appear to work well. Furthermore the results on the CP indicate that the normal approximation to the distributions of the proposed estimators seem to be appropriate too and as expected, the results became better when the sample size or the proportion of the failure event increased.

In addition, the two tables suggest that as discussed above, the pseudo likelihood-based estimation procedure yielded more efficient estimation than the estimating equation-based estimation procedure. On the other hand, as pointed out before, the former may be less stable or robust than the latter. To see this, we repeated the simulation study above and Table 3 gives the results obtained on estimation of β under the same set-up giving Table 2 except $\Lambda(t) = t$. As one can see, although the estimating equation-based method still gave similar performance as in Table 2, the pseudo likelihood-based method did not seem to perform as well as above, especially for the situation with $n = 1000$, maybe because of the need of estimating the baseline cumulative hazard function $\Lambda(t)$.

The results obtained on estimation of the regression parameter β with two covariates are given in Table 4 with $n = 2000$, the true value of β being $(0.2, 0.5)$, $(0.5, 0.5)$ or $\log(2), 0.5)$ and the other set-up being the same as in Table 2. It can be seen that they gave similar conclusions as above and again indicate that both proposed estimation procedures seem to work well. We also considered other set-ups and obtained similar results and conclusions.

5. AN APPLICATION

To illustrate the methodology proposed in the previous sections, we apply it to an epidemiological follow-up study, the Atherosclerosis Risk in Communities study (Zhou et al., 2017). It started in 1987 and involves the participants with ages between 45 to 64 at the beginning and recruited from four locations in the US, Forsyth County of North Carolina, Jackson of Mississippi, Minneapolis suburbs of Minnesota and Washington County of Maryland. During the study, the participants were scheduled to be examined several times and at each examination, the medical, social and demographic data were collected. As expected, some participants missed some scheduled revisits or were examined at times different from the scheduled times. In consequence, only interval-censored data were observed on the occurrence of a disease such as diabetes. In addition to the location where a participant was recruited, other available covariates include high-density lipoprotein cholesterol level, total cholesterol level, body mass index, smoking status and age. For the analysis below, we will focus on the 2814 white women younger than 55 years at the beginning of the study and among them, 210 were observed to have developed diabetes during the study.

To construct the interval-censored case-cohort sample, we selected a simple random sample of the cohort by the Bernoulli sampling with the selection probability of $q = 0.05$ or 0.1. For the application of the proposed estimation procedures, define the failure time T to be the occurrence time of diabetes and let the covariates Z to include the five covariates described above plus the two dimensional location indicators. For the location, we use Location-F ((0,1)), Location-J ((1,1)), Location-M ((0,0)) and Location-W ((1,0)) to denote four places with treating Location-M as the reference. Tables 5 and 6 present the results on the estimated covariate effects given by the proposed two estimation methods

Table 1. Simulation results on estimation of β with $p = 0.05$ and $\Lambda(t) = 0.5t^2$

n	True value	Parameter	Bias	SSE	ESE	CP
1000	$\beta = 0.2$	$\hat{\beta}_{IPW}$	0.0044	0.5824	0.5685	0.9500
		$\hat{\beta}_n$	0.0005	0.0696	0.0667	0.9400
	$\beta = 0.5$	$\hat{\beta}_{IPW}$	-0.0253	0.8894	0.8758	0.9530
		$\hat{\beta}_n$	0.0046	0.1086	0.1087	0.9370
	$\beta = \log 2$	$\hat{\beta}_{IPW}$	-0.0325	1.0405	1.0528	0.9540
		$\hat{\beta}_n$	0.0012	0.1326	0.1233	0.8900
2000	$\beta = 0.2$	$\hat{\beta}_{IPW}$	0.0015	0.3962	0.3972	0.9430
		$\hat{\beta}_n$	0.0012	0.0459	0.0470	0.9590
	$\beta = 0.5$	$\hat{\beta}_{IPW}$	0.0059	0.6028	0.6129	0.9480
		$\hat{\beta}_n$	0.0043	0.0747	0.0763	0.9550
	$\beta = \log 2$	$\hat{\beta}_{IPW}$	-0.0313	0.7324	0.7353	0.9570
		$\hat{\beta}_n$	-0.0018	0.0932	0.0923	0.9400

Table 2. Simulation results on estimation of β with $p = 0.1$ and $\Lambda(t) = 0.5t^2$

n	True value	Parameter	Bias	SSE	ESE	CP
1000	$\beta = 0.2$	$\hat{\beta}_{IPW}$	-0.0130	0.3571	0.3611	0.9430
		$\hat{\beta}_n$	-0.0011	0.0621	0.0636	0.9480
	$\beta = 0.5$	$\hat{\beta}_{IPW}$	-0.0633	0.4330	0.4645	0.9640
		$\hat{\beta}_n$	0.0051	0.0938	0.0964	0.9560
	$\beta = \log 2$	$\hat{\beta}_{IPW}$	-0.0710	0.5255	0.5447	0.9560
		$\hat{\beta}_n$	0.0042	0.1161	0.1106	0.9160
2000	$\beta = 0.2$	$\hat{\beta}_{IPW}$	-0.0153	0.2460	0.2519	0.9510
		$\hat{\beta}_n$	-0.0003	0.0454	0.0445	0.9440
	$\beta = 0.5$	$\hat{\beta}_{IPW}$	-0.0396	0.3313	0.3234	0.9440
		$\hat{\beta}_n$	0.0049	0.0674	0.0675	0.9470
	$\beta = \log 2$	$\hat{\beta}_{IPW}$	-0.0614	0.3724	0.3806	0.9370
		$\hat{\beta}_n$	0.0039	0.0813	0.0816	0.9450

along with the estimated standard errors (ESE) and the p-values for testing no covariate effect for each of the covariates for the selection probability $q = 0.05$ and 0.1, respectively.

One can see from the two tables above that the pseudo likelihood-based method with both q values indicates that none of the covariates had significant effects on the development of diabetes, while the estimating equation-based method with $q = 0.05$ gave similar conclusions but the same method with $q = 0.1$ suggests that total cholesterol, age and Location-W may have some significant effects on the onset of diabetes. To further investigate this, we repeated the analysis above by employing the pseudo likelihood-based approach with the use of the full cohort and give the estimation results in Table 7 with both $B = 100$ and 1000 in order to assess the effect of the bootstrap sample size B on the estimation. It is apparent that the results here are consistent with those given in Tables 5 and 6 and again suggest that none of the covariates was significantly related to the onset of diabetes. In contrast, the estimating equation-based procedure seems to be sensitive to q or the size of the subcohort.

6. DISCUSSION AND CONCLUDING REMARKS

In this paper, we considered the estimation of the additive hazards model, one of the most commonly used regression models in failure time data analysis, when one observes interval-censored case-cohort data. As discussed above, there exists a great deal of literature on either the analysis of interval-censored data or the analysis of case-cohort studies but it did not seem to exist an established estimation procedure when one faces both interval censoring and case-cohort design together. For the problem, two estimation procedures were proposed and the asymptotic properties of the resulting estimators of regression parameters were established. In addition, the numerical studies were performed for the assessment of their performance in practice and suggested that both methods seem to work well. On the other hand, one can also see through the numerical studies that the estimating equation-based approach is generally less efficient than the pseudo likelihood-based approach but unlike the former, the latter involves or requires estimation of the cumulative baseline hazard func-

Table 3. Simulation results on estimation of β with $p = 0.1$ and $\Lambda(t) = t$

n	True value	Parameter	Bias	SSE	ESE	CP
1000	$\beta = 0.2$	$\hat{\beta}_{IPW}$	0.0194	1.2670	1.2332	0.9510
		$\hat{\beta}_n$	0.0462	0.2188	0.1944	0.8790
	$\beta = 0.5$	$\hat{\beta}_{IPW}$	-0.0296	1.4888	1.4918	0.9400
		$\hat{\beta}_n$	0.0058	0.2820	0.2352	0.8050
	$\beta = \log 2$	$\hat{\beta}_{IPW}$	-0.0772	1.4720	1.4785	0.9460
		$\hat{\beta}_n$	-0.0418	0.2715	0.2449	0.8520
2000	$\beta = 0.2$	$\hat{\beta}_{IPW}$	0.0059	0.8681	0.8650	0.9420
		$\hat{\beta}_n$	0.0055	0.1759	0.1815	0.9550
	$\beta = 0.5$	$\hat{\beta}_{IPW}$	-0.0469	1.0128	1.0351	0.9580
		$\hat{\beta}_n$	0.0052	0.2169	0.1984	0.9010
	$\beta = \log 2$	$\hat{\beta}_{IPW}$	-0.0597	1.0297	1.0319	0.9450
		$\hat{\beta}_n$	-0.0058	0.2019	0.1883	0.8650

Table 4. Simulation results on estimation of β with $p = 0.1$ and $\Lambda(t) = 0.5t^2$

True value	Parameter	Bias	SSE	ESE	CP
$\beta = (0.2, 0.5)$	$\hat{\beta}_{1,IPW}$	-0.0201	0.3919	0.3780	0.9400
	$\hat{\beta}_{1,n}$	-0.0044	0.0721	0.0705	0.9240
	$\hat{\beta}_{2,IPW}$	-0.0591	0.6439	0.6495	0.9500
	$\hat{\beta}_{2,n}$	-0.0201	0.1148	0.1130	0.9460
$\beta = (0.5, 0.5)$	$\hat{\beta}_{1,IPW}$	-0.0513	0.4980	0.4804	0.9400
	$\hat{\beta}_{1,n}$	-0.0028	0.0978	0.0976	0.9530
	$\hat{\beta}_{2,IPW}$	-0.0518	0.8264	0.8232	0.9460
	$\hat{\beta}_{2,n}$	-0.0284	0.1350	0.1298	0.9270
$\beta = (\log 2, 0.5)$	$\hat{\beta}_{1,IPW}$	-0.0875	0.5258	0.5245	0.9480
	$\hat{\beta}_{1,n}$	0.0046	0.1081	0.1096	0.9300
	$\hat{\beta}_{2,IPW}$	-0.0561	0.8837	0.9037	0.9620
	$\hat{\beta}_{2,n}$	-0.0250	0.1392	0.1361	0.9410

tion and thus may need large sample sizes for better performance.

As mentioned above, the focus here has been on interval-censored data generated from case-cohort studies that can be described by two observation time points. In practice, two other types of interval-censored data may arise and it would be useful to generalize the proposed estimation procedures to these situations (Sun, 2006). One is the so-called current status data, meaning that each subject is observed only once and thus the failure time of interest is either left- or right-censored. The other is the case K interval-censored data where there exists a sequence of observation times for each subject. Although the proposed methods can be adopted for the case K interval-censored data, one may prefer to develop some approaches that could make full use of the information on the observation process. Another direction for future research is that instead of the additive hazards model (1), sometimes one may prefer a different model such as the proportional odds model or the linear transformation model. It is apparent that it would be helpful to derive some estimation procedures for these or other models.

APPENDIX A. PROOFS OF THEOREMS 1–3

In this Appendix, we will sketch the proof of Theorems 1, 2 and 3. For this, we need the following regularity conditions.

(A1) Assume that $\Lambda(\tau_1) < \infty$, $\Lambda(\tau_2) < \infty$, and there exists a positive constant η such that $P(V - U > \eta) > 0$. The union of the supports of U and V is contained in the interval $[r_1, r_2]$ with $0 < r_1 < r_2 < +\infty$.

(A2) The following matrics

$$\begin{aligned} B_1 &= E\Bigg(\int_0^{\tau_1} \left\{ Z_i^*(t) - \frac{s^{(1)}_{w,1\beta}(t,\beta)}{s^{(0)}_{w,1,\beta}(t,\beta)} \right\}^{\otimes 2} \\ &\qquad I(U_i \geq t) \exp\left\{-\Lambda(t) - \beta' Z_i^*(t)\right\} dt \Bigg), \\ B_2 &= E\Bigg(\int_0^{\tau_2} \left\{ Z_i^*(t) - \frac{s^{(1)}_{w,2,\beta}(t,\beta)}{s^{(0)}_{w,2,\beta}(t,\beta)} \right\}^{\otimes 2} \\ &\qquad I(U_i < t \leq V) \exp\left\{-\Lambda(t) - \beta' Z_i^*(t)\right\} dt \Bigg) \end{aligned}$$

Table 5. Estimated covariate effects on the occurrence time of diabetes with $q = 0.05$

Covariate	$\hat{\beta}_n$	ESE	*p-value*	$\hat{\beta}_{IPW}$	ESE	*p-value*
High-density lipoprotein cholesterol	-0.0331	0.1822	0.8558	-0.0733	0.1694	0.6654
Total cholesterol	0.0209	0.1939	0.9142	0.2058	0.2672	0.4412
Body mass index	0.0687	0.2288	0.7640	0.0631	0.1752	0.7187
Current smoking status	0.0040	0.0522	0.9393	0.0063	0.0214	0.7669
Age	-0.0512	0.3618	0.8876	-1.0749	0.6827	0.1154
Location-F	0.0021	0.0407	0.9587	-0.0201	0.0314	0.5233
Location-W	-0.0030	0.0177	0.8638	-0.0028	0.0188	0.8806

Table 6. Estimated covariate effects on the occurrence time of diabetes with $q = 0.1$

Covariate	$\hat{\beta}_n$	ESE	*p-value*	$\hat{\beta}_{IPW}$	ESE	*p-value*
High-density lipoprotein cholesterol	-0.0480	0.1999	0.8100	-0.0303	0.1294	0.8150
Total cholesterol	0.0322	0.2393	0.8897	0.2531	0.1395	0.0697
Body mass index	0.0682	0.2293	0.7660	0.1843	0.1589	0.2462
Current smoking status	0.0007	0.0615	0.9915	0.0197	0.0142	0.1658
Age	-0.0552	0.4192	0.8953	-1.3952	0.4822	0.0038
Location-F	0.0002	0.0486	0.9964	-0.0281	0.0191	0.1396
Location-W	-0.0032	0.0291	0.9115	-0.0299	0.0140	0.0327

are positive define, where $s^{(k)}_{w,1,\beta}$ and $s^{(k)}_{w,2,\beta}$ denote the limits of $S^{(k)}_{w,1,\beta}$ and $S^{(k)}_{w,2,\beta}$, respectively, $k = 0, 1, 2$.

(A3) The distribution of Z has bounded support and is not concentrated on any proper subspace of $\mathbb{R}^p$. Also, $E\{var(Z|U)\}$ and $E\{var(Z|V)\}$ are positive definite.

(A4) There exists a constant q such that $0 < q \leq \pi_q(\delta_1, \delta_2) \leq 1$.

(A5) The function $\Lambda_0 \in M$ is continuously differentiable up to order r in $[r_1, r_2]$, with the first derivative being strictly positive, and satisfies $\alpha^{-1} < \Lambda_0(r_1) < \Lambda_0(r_2) < \alpha$ for some positive constant α. Also, β_0 is an interior point of $\mathcal{B} \in \mathbb{R}^p$.

(A6) The conditional density $g(u, v|z)$ of (u, v) given z has bounded partial derivatives with respect to u and v, and the bounds of these partial derivatives do not depend on (u, v, z).

Proof of Theorem 1.

Consistency of $\hat{\beta}_{IPW}$: Note that according to Theorem 5.9 in Van der Vaart (1998), any sequence of estimators $\hat{\beta}_{IPW}$ such that $U_{IPW}(\hat{\beta}_{IPW}) = 0$ converges in probability to β_0. This suggests that we only need to verify that

(1)

$$\sup_{\beta \in \mathbb{B}} ||U_{IPW}(\beta) - \mathrm{E}(U_{IPW}(\beta))|| \xrightarrow{p} 0,$$

where $\mathbb{B}$ is a compact neighborhood of the true parameter β_0.

(2)

$$\inf_{\beta : |\beta - \beta_0| \geq \epsilon} ||\mathrm{E}(U_{IPW}(\beta))|| > 0.$$

Obviously, $\{\omega_i,\, i = 1, \ldots, n\}$ and $\{Z_i{}^*(t),\, i = 1, \ldots, n\}$ are Euclidean classes. By Lemma 5 in Sherman (1994) and Lemma 2.14 in Pakes and Pollard (1989), it is easy to see that

$$\begin{aligned}
\epsilon_{11} &= \left\{ \int_0^{\tau_1} \omega_i Z_i{}^*(t) dN_i^{(1)}(t),\, i = 1, \ldots, n \right\}, \\
\epsilon_{12} &= \left\{ \int_0^{\tau_1} \omega_i \frac{S^{(1)}_{w,1,\beta}(t, \beta)}{S^{(0)}_{w,1,\beta}(t, \beta)} dN_i^{(1)}(t),\, i = 1, \ldots, n \right\}, \\
\epsilon_{21} &= \left\{ \int_0^{\tau_2} \omega_i Z_i{}^*(t) dN_i^{(2)}(t),\, i = 1, \ldots, n \right\}, \\
\epsilon_{22} &= \left\{ \int_0^{\tau_2} \omega_i \frac{S^{(1)}_{w,2,\beta}(t, \beta)}{S^{(0)}_{w,2,\beta}(t, \beta)} dN_i^{(2)}(t),\, i = 1, \ldots, n \right\}
\end{aligned}$$

are Euclidean classes for their envelop functions, $(\tau_1 \sup |Z|)/\epsilon_0$, $\{\sup_{t,\beta}\{\frac{s^{(1)}_{w,1,\beta}(t, \beta)}{s^{(0)}_{w,1,\beta}(t, \beta)}\}\}\frac{\tau_1}{\epsilon_0}$, $(\tau_2 \sup |Z|)/\epsilon_0$, $\{\sup_{t,\beta}\{\frac{s^{(1)}_{w,2,\beta}(t, \beta)}{s^{(0)}_{w,2,\beta}(t, \beta)}\}\}\frac{\tau_2}{\epsilon_0}$, respectively. Therefore,

$$\epsilon_1 = \left\{ \left[\int_0^{\tau_1} \omega_i \left\{ Z_i{}^*(t) - \frac{S^{(1)}_{w,1,\beta}(t, \beta)}{S^{(0)}_{w,1,\beta}(t, \beta)} \right\} dN_i^{(1)}(t) \right.\right.$$

$$\left.\left. + \int_0^{\tau_2} \omega_i \left\{ Z_i{}^*(t) - \frac{S^{(1)}_{w,2,\beta}(t, \beta)}{S^{(0)}_{w,2,\beta}(t, \beta)} \right\} dN_i^{(2)}(t) \right],\, \beta \in \mathbb{B} \right\}$$

is a Euclidean class with an integrable envelope function.

Table 7. Estimated covariate effects on the occurrence time of diabetes based on the full cohort

	$B = 100$			$B = 1000$		
Covariate	$\hat{\beta}_n$	ESE	*p-value*	$\hat{\beta}_n$	ESE	*p-value*
High-density lipoprotein cholesterol	-0.0395	0.1364	0.7723	-0.0395	0.1435	0.7833
Total cholesterol	0.0272	0.1560	0.8614	0.0272	0.1619	0.8664
Body mass index	0.0765	0.1761	0.6641	0.0765	0.1865	0.6819
Current smoking status	0.0016	0.0622	0.9801	0.0016	0.0647	0.9809
Age	-0.0639	0.3690	0.8625	-0.0639	0.3637	0.8605
Location-F	0.0011	0.0420	0.9787	0.0011	0.0478	0.9812
Location-W	-0.0029	0.0262	0.9108	-0.0029	0.0372	0.9371

Hence, we have

$$\sup_{\beta\in\mathbb{B}} ||U_{IPW}(\beta) - \mathrm{E}(U_{IPW}(\beta))|| \xrightarrow{p} 0.$$

Meanwhile, we have that

$$\begin{aligned}
&\inf_{\beta:|\beta-\beta_0|\geq\epsilon} ||\mathrm{E}(U_{IPW}(\beta))|| \\
&= \inf_{\beta:|\beta-\beta_0|\geq\epsilon} ||\mathrm{E}(U_{IPW}(\beta)) - \mathrm{E}(U_{IPW}(\beta_0))|| \\
&= \inf_{\beta:|\beta-\beta_0|\geq\epsilon} \Bigg\| \mathrm{E}\Bigg\{ \int_0^{\tau_1} \omega_i \Bigg\{ \frac{S^{(1)}_{w,1,\beta}(t,\beta_0)}{S^{(0)}_{w,1,\beta}(t,\beta_0)} \\
&- \frac{S^{(1)}_{w,1,\beta}(t,\beta)}{S^{(0)}_{w,1,\beta}(t,\beta)} \Bigg\} dN_i^{(1)}(t) \\
&+ \int_0^{\tau_2} \omega_i \Bigg\{ \frac{S^{(1)}_{w,2,\beta}(t,\beta_0)}{S^{(0)}_{w,2,\beta}(t,\beta_0)} - \frac{S^{(1)}_{w,2,\beta}(t,\beta)}{S^{(0)}_{w,2,\beta}(t,\beta)} \Bigg\} dN_i^{(2)}(t) \Bigg\} \Bigg\| \\
&= \inf_{\beta:|\beta-\beta_0|\geq\epsilon} \Bigg\| \mathrm{E}\Bigg\{ \int_0^{\tau_1} \omega_i \Bigg\{ \frac{S^{(1)}_{w,1,\beta}(t,\beta_0)}{S^{(0)}_{w,1,\beta}(t,\beta_0)} \\
&- \left(\frac{S^{(1)}_{w,1,\beta}(t,\beta_0)}{S^{(0)}_{w,1,\beta}(t,\beta_0)} \right)^{\otimes 2} \Bigg\} dN_i^{(1)}(t) \\
&+ \int_0^{\tau_2} \omega_i \Bigg\{ \frac{S^{(2)}_{w,2,\beta}(t,\beta_0)}{S^{(0)}_{w,2,\beta}(t,\beta_0)} - \left(\frac{S^{(1)}_{w,2,\beta}(t,\beta_0)}{S^{(0)}_{w,2,\beta}(t,\beta_0)} \right)^{\otimes 2} \Bigg\} dN_i^{(2)}(t) \Bigg\} \Bigg\| \\
&||\beta - \beta_0|| + o(1) \xrightarrow{p} \mathbf{\Omega}_w \epsilon > 0.
\end{aligned}$$

This proves the consistency of $\hat{\beta}_{IPW}$.

Asymptotic normality of $\hat{\beta}_{IPW}$: First note that by the Taylor series expansions of $U_{IPW}(\hat{\beta}_{IPW})$ around β_0, we have that

$$\begin{aligned}
U_{IPW}(\hat{\beta}_{IPW}) = U_{IPW}(\beta_0) + \frac{\partial U_{IPW}(\beta)}{\partial \beta'}|_{\beta=\beta_0}(\hat{\beta}_{IPW} - \beta_0) \\
+ o_p(1)\,.
\end{aligned}$$

Thus, by the consistency of β_{IPW} for any β satisfying $|\hat{\beta}_{IPW} - \beta| = o_p(1)$ and Taylor expansion, we have

$$\sqrt{n}(\hat{\beta}_{IPW} - \beta_0) = (\hat{\mathbf{\Omega}}_w)^{-1} n^{-\frac{1}{2}} U_{IPW}(\beta_0),$$

where

$$\begin{aligned}
&n^{-\frac{1}{2}} U_{IPW}(\beta_0) \\
&\quad = \sum_{i=1}^{n} \int_0^{\tau_1} \omega_i \left\{ Z_i{}^*(t) - \frac{s^{(1)}_{w,1,\beta}(t,\hat{\beta}_{IPW})}{s^{(0)}_{w,1,\beta}(t,\hat{\beta}_{IPW})} \right\} dM_i^{(1)}(t) \\
&\qquad + \int_0^{\tau_2} \omega_i \left\{ Z_i{}^*(t) - \frac{s^{(1)}_{w,2,\beta}(t,\hat{\beta}_{IPW})}{s^{(0)}_{w,2,\beta}(t,\hat{\beta}_{IPW})} \right\} dM_i^{(2)}(t) + o_p(1),
\end{aligned}$$

with

$$M_i^{(1)}(t) = N_i^{(1)}(t) - \int_0^t I(U_i \geq s)\exp\{-\Lambda(s) - \beta' Z_i^*(s))\}ds,$$

$$M_i^{(2)}(t) = N_i^{(2)}(t) - \int_0^t I(U_i < s \leq V_i)\exp\{-\Lambda(s) - \beta' Z_i^*(s)\}ds.$$

Here, $s^{(k)}_{w,l,\beta}(t,\beta_0)$ denote the limits of $S^{(k)}_{w,l,\beta}(t,\beta_0)$ for $l = 1, 2,\ k = 0, 1, 2$.

Therefore, the asymptotic covariance matrix of $\sqrt{n}(\hat{\beta}_{IPW} - \beta_0)$ can be consistently estimated by $\hat{\mathbf{\Omega}}_w^{-1} \hat{\Gamma}_w \hat{\mathbf{\Omega}}_w^{-1}$.

Proof of Theorem 2.

In this following, we will prove Theorems 2 and 3 by employing the empirical process theory and nonparametric techniques. First define $\mathrm{P}f = \int f(x)dP(x)$, and $\mathrm{P}_n f = n^{-1} \sum_{i=1}^{n} f(X_i)$ for a function f, a probability function P and a sample $X_1, \ldots, X_n$.

To prove Theorem 2, we will first define the covering number of the class $\mathcal{L}_n = \{l^w(\theta; O^\xi) = wl(\theta; O) : \theta \in \Theta_n\}$ and establish two lemmas. For $\epsilon > 0$, define the covering number $N(\epsilon, \mathcal{L}_n, L_1(P_n))$ as the smallest value of κ for which there exists $\{\theta^{(1)}, \ldots, \theta^{(\kappa)}\}$ such that

$$\min_{j\in\{1,\ldots,\kappa\}} \frac{1}{n} \sum_{i=1}^{n} |l^w(\theta, O^\xi) - l^w(\theta^{(j)}, O^\xi)| < \epsilon$$

for all $\theta \in \Theta_n$, where $\theta^{(j)} = (\beta^{(j)\prime}, \Lambda^{(j)})' \in \Theta_n$, $j = 1, \ldots, \kappa$. We will define $N(\epsilon, \mathcal{L}_n, L_1(P_n)) = \infty$ if no such κ exists.

Lemma 1. Assume that Conditions (A1), (A3)–(A6) hold. Then the covering number of the class $\mathcal{L}_n = \{l^w(\theta; O^\xi) : \theta \in \Theta_n\}$ satisfies

$$N(\epsilon, \mathcal{L}_n, L_1(P_n)) \le K M_n^{(m+1)} \epsilon^{-(p+m+1)}.$$

Lemma 2. Assume that Conditions (A1), (A3)–(A6) hold. Then,

$$\sup_{\theta\in\Theta_n} |P_n l^w(\theta; O^\xi) - P l^w(\theta; O^\xi)| \to 0$$

Proof of Lemma 1 and 2. The proof is similar to that of Zhou et al. (2017) and Hu et al. (2017) and thus omitted. □

Let $M(\theta, O^\xi) = -l(\theta, O^\xi)$ and define $K_\epsilon = \{\theta : d(\theta, \theta_0) \ge \epsilon, \theta \in \Theta_n\}$ for any $\epsilon > 0$ and

$$\begin{aligned} \zeta_{1n} &= \sup_{\theta\in\Theta_n} |P_n M(\theta, O^\xi)) - PM(\theta, O^\xi)|, \\ \zeta_{2n} &= P_n M(\theta_0, O^\xi)) - PM(\theta_0, O^\xi). \end{aligned}$$

Then we can show that

$$\begin{aligned} &\inf_{K_\epsilon} PM(\theta, O^\xi) = \inf_{K_\epsilon}\{PM(\theta, O^\xi) - P_n M(\theta, O^\xi) \\ &+P_n M(\theta, O^\xi)\} \le \zeta_{1n} + \inf_{K_\epsilon} P_n M(\theta, O^\xi), \end{aligned}$$

if $\hat{\theta}_n \in K_\epsilon$, then we have

$$\begin{aligned} &\inf_{K_\epsilon} P_n M(\theta, O^\xi) = P_n M(\hat{\theta}_n, O^\xi) \le P_n M(\theta_0, O^\xi) \\ &= \zeta_{2n} + PM(\theta_0, O^\xi). \end{aligned}$$

Define $\delta_\epsilon = \inf_{K_\epsilon} PM(\theta, O^\xi)) - PM(\theta_0, O^\xi)$. Then according to Lemma 5 of Hu et al. (2017), we have

$$\begin{aligned} &\inf_{K_\epsilon} PM(\theta, O^\xi) \le \zeta_{1n} + \zeta_{2n} + PM(\theta_0, O^\xi) \\ &= \zeta_n + PM(\theta_0, O^\xi) \end{aligned}$$

with $\zeta_n = \zeta_{1n} + \zeta_{2n}$. Hence we can obtain that $\zeta_n \ge \delta_\epsilon$ and furthermore $\hat{\theta}_n \in K_\epsilon$ implies $\zeta_n \ge \delta_\epsilon$. By Lemma 2 and the Strong Law of Large Numbers we have $\zeta_{1n} = o(1)$ and $\zeta_{2n} = o(1)$ almost surely. Therefore, $\cup_{k=1}^\infty \cap_{n=k}^\infty \{\hat{\theta}_n \in K_\epsilon\} \subseteq \cup_{k=1}^\infty \cap_{n=k}^\infty \{\zeta_n \ge \delta_\epsilon\}$, which proves that $d(\theta_n, \theta_0) \to 0$ almost surely.

To establish the convergence rate, for any $\eta > 0$, define the class $\mathcal{F}_\eta = \{l^w(\theta_{n0}, O^\xi) - l^w(\theta, O^\xi) : \theta \in \Theta_n, d(\theta, \theta_{n0}) \le \eta\}$ with $\theta_{n0} = (\beta_0, \Lambda_{n0})$. Following the calculation of Shen and Wong (1994, P.597), we can establish that $\log N_{[]}(\epsilon, \mathcal{F}_\eta, \|\,.\,\|_2) \le CN\log(\eta/\epsilon)$ with $N = m+1$, where $N_{[]}(\epsilon, \mathcal{F}_\eta, d)$ denotes the bracketing number (see the Definition 2.1.6 in Van Der Vaart and Wellner, 1996) with respect to the metric or semi-metric d of a function class $\mathcal{F}$. Moreover, some algebraic calculations lead to $||l^w(\theta_{n0}, O^\xi) - l^w(\theta, O^\xi)||_2^2 \le C\eta^2$ for any $l^w(\theta_{n0}, O^\xi) - l^w(\theta, O^\xi) \in \mathcal{F}_\eta$. Therefore, by Lemma 3.4.2 of Van Der Vaart and Wellner (1996), we obtain

$$\begin{aligned} (S) \quad & E_p\|n^{1/2}(P_n - P)\|_{\mathcal{F}_\eta} \\ & \le CJ_\eta(\epsilon, \mathcal{F}_\eta, \|\,.\,\|_2)\left\{1 + \frac{J_\eta(\epsilon, \mathcal{F}_\eta, \|\,.\,\|_2)}{\eta^2 n^{1/2}}\right\}, \end{aligned}$$

where $J_{[]}(\eta, \mathcal{F}_\eta, \|\,.\,\|_2) = \int_0^\eta \{1 + \log N_{[]}(\epsilon, \mathcal{F}_\eta, \|\,.\,\|_2)\}^{1/2} d\epsilon \le CN^{1/2}\eta$. The right-hand side of (S) yield $\phi_n(\eta) = C(N^{1/2}\eta + N/n^{1/2})$. It is easy to see that $\phi_n(\eta)/\eta$ decreases in η, and $r_n^2\phi_n(1/r_n) = r_n N^{1/2} + r_n^2 N/n^{1/2} < 3n^{1/2}$, where $r_n = N^{-1/2}n^{1/2} = n^{(1-v)/2}$ with $0 < v < 0.5$. Hence, $n^{(1-v)/2}d(\hat{\theta} - \theta_{n0}) = O_p(1)$ by Theorem 3.2.5 of Van Der Vaart and Wellner (1996). This, together with $d(\theta_{n0}, \theta_0) = O_p(n^{-rv})$ (Lemma A1 in Lu et al. (2007)), yields that $d(\hat{\theta}, \theta_0) = O_p(n^{-(1-v)/2} + n^{-rv})$.

Proof of Theorem 3.

Now we prove the asymptotic normality of $\hat{\beta}_n$. Notice that $w = \frac{\xi}{\pi_q(\delta_1, \delta_2)}$ is bounded and does not depend on the parameters (β, Λ) and $E\{w|O\} = 1$. Following the proof of Theorem 2 in Zhang et al. (2010) and Zhou et al. (2017), one can obtain that

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = I^{-1}(\beta_0) n^{-1/2} \sum_{i=1}^n w_i l^*(\beta_0, \Lambda_0; O_i) + o_p(1)$$

where $l^*(\beta_0, \Lambda_0; O)$ and $I(\beta)$, the efficient score and information for β based on $O = \{U, V, \delta_1, \delta_2, Z\}$, are defined in Zhang et al. (2010, p. 344), with our parameters (β, Λ) corresponding to their $\{\theta, exp(\phi)\}$. Note that

$$\begin{aligned} & var\{wl^*(\beta_0, \Lambda_0; O)\} = var[E\{wl^*(\beta_0, \Lambda_0; O)|O\}] \\ & +E[var\{wl^*(\beta_0, \Lambda_0; O)|O\}] \\ = \quad & var\{l^*(\beta_0, \Lambda_0; O)\} + E\left[var(\xi|O)\frac{l^*(\beta_0, \Lambda_0; O)^{\otimes 2}}{\pi_q^2(\delta_1, \delta_2)}\right] \\ = \quad & I(\beta_0) + E\left[\frac{1 - \pi_q(\delta_1, \delta_2)}{\pi_q(\delta_1, \delta_2)}\{l^*(\beta_0, \Lambda_0; O)\}^{\otimes 2}\right]. \end{aligned}$$

Thus we have

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \to N(0, \Sigma), \ \ n \to \infty$$

in distribution, where

$$\begin{aligned} \Sigma = I^{-1}(\beta_0) + I^{-1}(\beta_0) E\left\{\frac{1 - \pi_q(\delta_1, \delta_2)}{\pi_q(\delta_1, \delta_2)}\{l^*(\beta_0, \Lambda_0; O)\}^{\otimes 2}\right\} \\ \times I^{-1}(\beta_0). \end{aligned}$$

This completes the proof of Theorem 3.

ACKNOWLEDGEMENT

The authors wish to thanks the Editor, Dr. Ming-Hui Chen, and two reviewers for their helpful comments and suggestions. The research of Huiqiong Li was partially supported by a grant from the Natural Science Foundation of China (11561075) and a grant from reserve talents for academic and technical leaders of middle-aged and young people in Yunnan Province (2017HB004).

Received 11 May 2019

REFERENCES

Barlow, W. E. (1994). Robust variance estimation for the case-cohort design. *Biometrics* **50**, 1064–1072.

Chen, K. (2001). Generalized case-control sampling. *Journal of the Royal Statistical Society, Series B* **63**, 791–809. MR1872067

Chen, K. and Lo, S. H. (1999). Case-cohort and case-control analysis with Cox's model. *Biometrika* **86**, 755–764. MR1741975

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187–220. MR0341758

Hu, T., Zhou, Q. and Sun, J. (2017). Regression analysis of bivariate current status data under the proportional hazards model. *The Canadian Journal of Statistics* **45**, 410–424. MR3729978

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed. Wiley, New York. MR0570114

Kang, S. and Cai, J. (2009). Marginal hazards model for case-cohort studies with multiple disease outcomes. *Biometrika* **96**, 887–901. MR2767277

Kang, S., Cai, J. and Chambless, L. (2013). Marginal additive hazards model for case-cohort studies with multiple disease outcomes: an application to the Atherosclerosis risk in communities study. *Biostatistics* **14**, 28–41.

Keogh, R. H. and White, I. R. (2013). Using full-cohort data in nested case-control and case-cohort studies by multiple imputation. *Statistics in Medicine* **32**, 4021–4043. MR3102432

Kim, S., Cai, J. and Lu, W. (2013). More efficient estimators for case-cohort studies. *Biometrika* **100**, 695–708. MR3094446

Li, Z., Gilbert, P. and Nan, B. (2008). Weighted likelihood method for grouped survival data in case-cohort studies with application to HIV vaccine trials. *Biometrics* **64**, 1247–1255. MR2522274

Li, Z. and Nan, B. (2011). Relative risk regression for current status data in case-cohort studies. *The Canadian Journal of Statistics* **39**, 557–577. MR2860827

Lin, D. Y., Oakes, D. and Ying, Z. (1998). Additive hazards regression with current status data. *Biometrika* **85**, 289–298. MR1649115

Lorentz, G. G. (1986). *Bernstein Polynomials*. Chelsea Publishing Co., New York. MR0864976

Lu, M., Zhang, Y. and Huang, J. (2007). Estimation of the mean function with panel count data using monotone polynomial splines. *Biometrika* **94**, 705–718. MR2410018

Ma, S. and Kosorok, M. R. (2005). Robust semiparametric M-estimation and the weighted bootstrap. *Journal of Multivariate Analysis* **96**, 190–217. MR2202406

Ma, L., Hu, T. and Sun, J. (2015). Sieve maximum likelihood regression analysis of dependent current status data. *Biometrika* **85**, 649–658. MR3394289

Nan, B. (2004). Efficient estimation for case-cohort studies. *The Canadian Journal of Statistics* **32**, 403–419. MR2125853

Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica* **57**, 1027–1057. MR1014540

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

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. MR2066254

Shen, X. (1997). On methods of sieves and penalization. *Annals of Statistics* **25**, 2555–2591. MR1604416

Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *Annals of Statistics* **22**, 580–615. MR1292531

Sherman, R. P. (1994). Maximal inequalities for degenerate u-process with application to optimization estimators. *The Annal of Statistics* **22**, 439–459. MR1272092

Self, S. G. and Prentice, R. L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *Annals of Statistics* **16**, 64–81. MR0924857

Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*. Springer, New York. MR2287318

Van Der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press. MR1652247

Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag, New York. MR1385671

Wang, L., Sun, J. and Tong, X. (2010). Regression analysis of case II interval-censored failure time data with the additive hazards model. *Statistica Sinica* **20**, 1709–1723. MR2777342

Zhou, Q., Zhou, H. and Cai, J. (2017). Case-cohort studies with interval-censored failure time data. *Biometrika* **104**, 17–29. MR3626480

Zhang, Y., Hua, L. and Huang, J. (2010). A spline-based semiparametric maximum likelihood estimation method for the Cox model with interval-censored data. *Scandinavian Journal of Statistics* **37**, 338–354. MR2682304

Mingyue Du
Center for Applied Statistical Research, School of Mathematics
Jilin University
Changchun
P.R. China
E-mail address: dumy17@mails.jlu.edu.cn

Huiqiong Li
Department of Statistics
Yunnan University
Kunming 650091
P.R. China
E-mail address: lihuiqiong@ynu.edu.cn

Jianguo Sun
Department of Statistics
University of Missouri
Columbia, MO, 65211
USA
E-mail address: sunj@missouri.edu