Research Article

The Adjustment of Covariates in Cox's Model under Case-Cohort Design

Guocai Rong,¹ Luwei Tang,² Wenting Luo,² Qing Li,¹ and Lifeng Deng¹

¹College of Mathematics and Systems Science, Shandong University of Science and Technology, Qingdao 266590, Shandong, China

²School of Mathematics and Statistics, Nanning Normal University, Nanning 530001, Guangxi, China

Correspondence should be addressed to Lifeng Deng; laiji1234@163.com

Received 27 August 2020; Revised 18 November 2020; Accepted 8 December 2020; Published 24 December 2020

Academic Editor: Songbai Guo

Copyright © 2020 Guocai Rong et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Case-cohort design is a biased sampling method. Due to its cost-effective and theoretical significance, this design has extensive application value in many large cohort studies. The case-cohort data includes a subcohort sampled randomly from the entire cohort and all the failed subjects outside the subcohort. In this paper, the adjustment for the distorted covariates is considered to case-cohort data in Cox's model. According to the existing adjustable methods of distorted covariates for linear and nonlinear models, we propose estimating the distorting functions by nonparametrically regressing the distorted covariates on the distorting factors; then, the estimators for the parameters are obtained using the estimated covariates. The proof of consistency and being asymptotically normal is completed. For calculating the maximum likelihood estimates of the regression coefficients subject in Cox's model, a minorization-maximization (MM) algorithm is developed. Simulation studies are performed to compare the estimations with the covariates undistorted, distorted, and adjusted to illustrate the proposed methods.

1. Introduction

Many survival data have the characteristics of large sample size and high censoring rate. The cost is very high when all variables of each individual in the data are measured. To reduce the cost and improve the efficiency of cohort studies, Prentice [1] firstly proposed the case-cohort design and gave a pseudolikelihood method to estimate regression parameters. Since the publication of the landmark article [1], case-cohort design has been applied more and more, especially with the development of big data in recent years. For example, some scholars have applied this design to Life Science [2–5] and natural disasters [6], and some other scholars have further improved this design [7, 8], respectively.

In practice, some subjects may be interfered by other factors due to their characteristics, so their corresponding real covariates cannot be observed; only the distorted covariates and the distorting factors can be observed. For

example, for the Modification of Diet in Renal Disease (MDRD) study [9, 10], glomerular filtration rate (GFR) and serum creatinine (SCr) data are distorted by the body surface area (BSA). To show the relationship of the two covariates, we need to adjust them by correcting the distorting effect of BSA. The covariate-adjusted regression was introduced for situations where both predictors and response in a regression model are not directly observable, but are contaminated with a multiplicative factor that is determined by the value of an observable factor [11]. For the nonlinear regression model, how to estimate the distorting functions was proposed by nonparametric regression the predictors and response on the distorting covariates [12]. Then, the covariate-adjusted method was extended to other models [13–15]. The data studied in these documents is all complete data. In this paper, the studied data is survival data with right censored data, in which only distorted covariates and distorting factors are included, and the nonparametric method [12] to

estimate the distorting functions to obtain the adjusted value of the distorted covariates on the distorting factors, where the nonparametric method is a kernel smoothing method.

In this article, inspired by the idea of Ding et al. [16] and Deng et al. [17], we construct a surrogate function with a diagonal Hessian matrix by using the convexity of the exponential function and the negative logarithm function and then maximize this surrogate function with a diagonal Hessian matrix. This algorithm is minorization-maximization (MM) algorithm. That is, the first M is to construct a minor surrogate function and the second M is to maximize the function.

The rest of the article is organized as follows. In Section 2, we firstly fit data from case-cohort design to Cox's model with distorting factors and secondly adjust the distorted covariates by using the kernel function. In Section 3, the convergence of the adjusted covariates and the asymptotic properties for the proposed estimators is completed. In Section 4, we propose a MM algorithm for implementation of the estimation and present a cross-validation (CV) to obtain the optimal bandwidth and a nonparametric bootstrap approach to get the standard error estimation. Several simulation studies are conducted to compare the estimations without distorting factors, with distorting factors, and adjusted with distorting factors in Section 5. Real data analysis of heart failure data is in Section 6. Discussion is stated in Section 7. All proofs are given in Appendix.

2. Design and Estimation

2.1. Model and Design. Suppose that there are n independent subjects in the studies cohort, where $\tilde{T}_i$ is the failure time, C_i is the censoring time, $T_i = \min(\tilde{T}_i, C_i)$ is the observe time for the i th subject, $\Delta_i = I(\tilde{T}_i \leq C_i)$ is the right-censoring indicator for failure, $Y_i(t) = I(T_i \geq t)$ is the at risk process, and $N_i(t) = \Delta_i I(T_i \leq t)$ is the counting process, where $I(\cdot)$ is an indicator function, τ is the end of study time, and Z_i is a p th covariate for subject i ; we focus on the time-independent covariate.

Let $\beta = (\beta_1, \beta_2, \dots, \beta_p)$ be the unknown p -dimensional vector of regression coefficients. $\tilde{T}_i$ arises from Cox's model as the following form:

$$\lambda(t|Z) = \lambda_0(t) \exp\{\beta^\top Z_i(t)\}, \quad (1)$$

where $\lambda_0(t)$ is an unspecified baseline hazard function. The corresponding partial likelihood function is widely used for the inference of β [18] as the following form:

$$L_F(\beta) = \prod_{i=1}^n \left[\frac{\exp\{Z_i(T_i)\beta^\top\}}{\sum_{l=1}^n Y_l(T_i) \exp\{Z_l(T_i)\beta^\top\}} \right]^{\Delta_i}. \quad (2)$$

Under the case-cohort design, let $\tilde{A}$ denote the subcohort, which is selected from the full cohort by simple random sampling, and ξ_i be an indicator, equaling 1 if the i th subject is selected into the subcohort and 0 otherwise. Let A denote the case-cohort sample which contained the subjects

from the subcohort $\tilde{A}$ and the case outside the subcohort. Therefore, the observed data structure can be summarized as follows: $\{T_i, \Delta_i, (\xi_i + \Delta_i(1 - \xi_i))Z_i(t), 0 \leq t \leq T_i\}$.

The pseudolikelihood function [1] was proposed, and the corresponding pseudolikelihood function takes the following form:

$$L_p(\beta) = \sum_{i=1}^n \omega_i \Delta_i \left[\beta^\top Z_i(T_i) - \log \sum_{l=1}^n Y_l(T_i) \omega_l \exp\{\beta^\top Z_l(T_i)\} \right], \quad (3)$$

where $\omega_i = \Delta_i + (1 - \Delta_i)\xi_i$.

2.2. The Adjustment of Covariates. We now study that the covariates are distorted by some distorting factors. The distorted observed data structure is

$$\{T_i, \Delta_i, (\xi_i + \Delta_i(1 - \xi_i))\tilde{Z}_i(t), U_i, 0 \leq t \leq T_i\}, \quad (4)$$

where U_i is the i th distorting factor and $\tilde{Z}_i(t)$ is the i th distorted covariate. Here, $Z_i(t)$ is unobservable, and $\phi(U)$ is denoted as the unknown distorting functions of observable factor U .

Firstly, we give some basic assumptions [12] as follows:

- (i) $\tilde{Z}_r = \phi_r(U)Z_r$
- (ii) $E[\phi_r(U)] = 1, r = 1, 2, \dots, p$
- (iii) (Z_r, U) are independent of each other, $r = 1, 2, \dots, p$

Under these assumptions, the working-likelihood function (3) can be rewritten as the following form:

$$\begin{cases} \tilde{L}_p(\beta) = \sum_{i=1}^n \omega_i \Delta_i \left[\beta^\top \tilde{Z}_i(T_i) - \log \sum_{l=1}^n Y_l(T_i) \omega_l \exp\{\beta^\top \tilde{Z}_l(T_i)\} \right], \\ \tilde{Z}_{ri} = \phi_r(U_i)Z_{ri}. \end{cases} \quad (5)$$

Now, our object is to estimate the unknown coefficients β based on observation (4) and function (5). From condition (A3), we have

$$\phi_r(U) = \frac{E[\tilde{Z}_r|U]}{E[Z_r]}, \quad 1 \leq r \leq p. \quad (6)$$

Many scholars have exposed some methods to estimate $\phi_r(U)$ [19]. So, we construct the function $\phi_r(U)$ by using the classic kernel method, for $1 \leq r \leq p$:

$$\hat{\phi}_r(U) = \frac{1/(nh) \sum_{i=1}^n K((u - U_i)/h) \tilde{Z}_{ir}}{1/(nh) \sum_{i=1}^n K((u - U_i)/h)} \times \frac{1}{\bar{\bar{Z}}_r}, \quad (7)$$

where $\bar{\bar{Z}}_r = 1/n \sum_{i=1}^n \tilde{Z}_{ri}$, $K(\cdot)$ is a kernel function, and h is bandwidth.

Let $\widehat{Z}_{ri} = \widetilde{Z}_{ri}/\widehat{\phi}_r(U_i)$ denote the adjusted covariate. That is, we obtain the adjusted covariates by removing the distorting factors from the distorted covariates. Then, function (4) can be abbreviated as

$$\widehat{l}_p(\beta) = \sum_{i=1}^n \omega_i \Delta_i \left[\widehat{Z}_i(T_i) \beta^\top - \log \sum_{l=1}^n Y_l(T_i) \omega_l \exp\{\widehat{Z}_l(T_i) \beta^\top\} \right]. \quad (8)$$

Under the case-cohort design, the estimate $\widehat{\beta}$ about the adjustment of covariates is defined as follows:

$$\widehat{\beta} = \arg \max_{\beta} \widehat{l}_p(\beta). \quad (9)$$

3. Consistency and Asymptotic Normality

To present the asymptotic results, we will introduce some notations. Let β_0 denote the true value of β . For $d = 0, 1, 2$, define

$$\begin{aligned} S^d(\beta, t) &= \frac{1}{n} \sum_{i \in A} Y_i(t) \exp\{\mathbf{Z}_i^\top \beta\} \mathbf{Z}_i^{\otimes d}, \\ \widetilde{S}^d(\beta, t) &= \frac{1}{\widetilde{n}} \sum_{i \in \widetilde{A}} Y_i(t) \exp\{\mathbf{Z}_i^\top \beta\} \mathbf{Z}_i^{\otimes d}, \\ Q^d(\beta, t, \omega) &= \frac{1}{n} \sum_{i \in A} Y_i(t) Y_i(\omega) \exp\{2\mathbf{Z}_i^\top \beta\} \mathbf{Z}_i^{\otimes d}, \\ \widetilde{Q}^d(\beta, t, \omega) &= \frac{1}{\widetilde{n}} \sum_{i \in \widetilde{A}} Y_i(t) Y_i(\omega) \exp\{2\mathbf{Z}_i^\top \beta\} \mathbf{Z}_i^{\otimes d}, \end{aligned} \quad (10)$$

where $\mathbf{a}^{\otimes 0} = 1$, $\mathbf{a}^{\otimes 1} = \mathbf{a}$, and $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^\top$ for a vector $\mathbf{a}$. Define

$$E(\beta, t) = \frac{S^{(1)}(\beta, t)}{S^{(0)}(\beta, t)}, \quad (11)$$

$$\widetilde{E}(\beta, t) = \frac{\widetilde{S}^{(1)}(\beta, t)}{\widetilde{S}^{(0)}(\beta, t)}.$$

We impose the following conditions throughout the paper. Note that convergence properties involve $n \rightarrow \infty$ and $\|\cdot\|$ refers to the Euclidean norm in these conditions:

- (A1) $\int_0^\tau \lambda_0(t) dt < \infty$.
- (A2) There exists some $\delta > 0$ such that $n^{-1/2} \sup_{i, t \in [0, \tau]} Y_i(t) |\mathbf{Z}_i| I\{\mathbf{Z}_i^\top \beta_0 > -\delta |\mathbf{Z}_i|\} \xrightarrow{P} 0$.
- (A3) There exists a neighborhood β of β_0 , and functions $s^d(\beta, t)$, $d = 0, 1, 2$, defined on $\beta \times [0, \tau]$, which satisfy
  - (i) $\sup_{\beta \in \mathcal{B}, t \in [0, \tau]} \|S^d(\beta, t) - s^d(\beta, t)\| \xrightarrow{P} 0$, $d = 0, 1, 2$.

- (ii) $s^d(\beta, t)$, $d = 0, 1, 2$, are continuous of β on $\mathcal{B}$ uniformly in $t \in [0, \tau]$; $s^d(\beta, t)$, $d = 0, 1, 2$, are bounded on $\mathcal{B} \times [0, \tau]$, and $s^{(0)}(\beta, t)$ is bounded away from zero; and $s^{(1)}(\beta, t) = \nabla_{\beta} s^{(0)}(\beta, t)$ and $s^{(2)}(\beta, t) = \nabla_{\beta}^2 s^{(0)}(\beta, t)$ for all $\beta \in \mathcal{B}$, $t \in [0, \tau]$.
- (iii) The matrix

$$\Sigma(\beta_0) = \int_0^\tau \nu(\beta_0, t) s^{(0)}(\beta_0, t) \lambda_0(t) dt \quad (12)$$

is positive definite, where $\nu(\beta_0, t) = ((s^{(2)}(\beta_0, t))/(s^{(0)}(\beta_0, t))) - \{(s^{(1)}(\beta_0, t))/(s^{(0)}(\beta_0, t))\}^{\otimes 2}$.

- (B1) $\widetilde{n}/n \rightarrow \alpha$ for some $\alpha \in (0, 1)$.
- (B2) The sequence of distributions of $n^{1/2}\{\widetilde{E}(\beta_0, t) - nE(\beta_0, t)\}$ is tight on $\mathcal{B} \times [0, \tau]$ of left-continuous functions with right-hand limits equipped with the product Skorohod topology.
- (B3) The existing functions $q^d(\beta, t, \omega)$, $d = 0, 1, 2$, defined on $\mathcal{B} \times [0, \tau]^2$ which satisfy
  - (i) $\sup_{\beta \in \mathcal{B}, (t, \omega) \in [0, \tau]^2} \|Q^{(d)}(\beta, t, \omega) - q^{(d)}(\beta, t, \omega)\| \xrightarrow{P} 0$, $d = 0, 1, 2$.
  - (ii) $q^{(d)}(\beta, t, \omega)$, $d = 0, 1, 2$, are continuous of β on $\mathcal{B}$ uniformly in $(t, \omega) \in [0, \tau]^2$, and $q^{(d)}(\beta, t, \omega)$, $d = 0, 1, 2$, are bounded on $\mathcal{B} \times [0, \tau]^2$.
  - (iii) $\sup_{n \geq 1} E\{Q^{(d)}(\beta, t, \omega)\}$, $d = 0, 1, 2$, are bounded sequences.
- (B4) For $d = 0, 1, 2$,

$$\begin{aligned} \sup_{\beta \in \mathcal{B}, t \in [0, \tau]} \|\widetilde{S}^{(d)}(\beta, t) - s^{(d)}(\beta, t)\| &\xrightarrow{P} 0, \\ \sup_{\beta \in \mathcal{B}, (t, \omega) \in [0, \tau]^2} \|\widetilde{Q}^{(d)}(\beta, t, \omega) - q^{(d)}(\beta, t, \omega)\| &\xrightarrow{P} 0. \end{aligned} \quad (13)$$

- (C) All $g_r(U) = \phi_r(U)p(U)$, $1 \leq r \leq p$, and $\phi_r(U)$ and $p(U)$ are greater than a positive constant. They are differential and their derivatives satisfy the condition that there exists a neighborhood of the origin. For example, Δ and a constant $c > 0$ such that, for any $\xi \in \Delta |g_r^3(U + \xi) - g_r^3(U)| \leq c|\xi|$, $1 \leq r \leq p$, $|p^3(U + \xi) - p^3(U)| \leq c|\xi|$.
- (D1) The continuous kernel function $K(\cdot)$ satisfies the following properties:
  - (i) The support of $K(\cdot)$ is the interval $[-1, 1]$.
  - (ii) $K(\cdot)$ is symmetric about zero.
  - (iii) $\int_{-1}^1 K(U) dU = 1$ and $\int_{-1}^1 U^d k(U) dU = 0$, $d = 1, 2, 3$.
- (D2) As $n \rightarrow \infty$, the bandwidth h is in the range from $O(n^{-1/4} \log(n))$ to $O(n^{-1/8})$.

(E) $E[\mathbf{Z}_r]$ is bounded away from 0, $E[\mathbf{Z}_r]^2 < \infty$. These conditions are mild and can be satisfied in most circumstances.

Conditions (A1)–(A3) are the regularity conditions for the asymptotic results of Cox's model [20]. Conditions (B1)–(B4) are the regular conditions under the case-cohort design [21]. Conditions (C) are related to smoothness of the function $g_r(\cdot)$ and the density function $p(\cdot)$ of U . Conditions (D1)–(D2) are commonly assumed for the root n consistency of kernel-based estimation [22, 23]. Condition (E) is special for this problem [12, 22].

Under these conditions, we have the following result with detailed proofs given in Appendix.

Theorem 1. (asymptotic properties of $\hat{\beta}$). Under conditions (A1)–(A3), (B1)–(B4), (C), (D1)–(D2), and (E), $\hat{\beta}$ converges in probability to β_0 , and

$$\sqrt{n}(\hat{\beta} - t\beta_0) \xrightarrow{d} N(0, \Omega(\beta_0)), \quad (14)$$

where the asymptotic variance matrix $\Omega(\beta_0) = \Sigma_0(\beta_0) + \Sigma_1(\beta_0)$, where

$$\begin{aligned} \Sigma_1(\beta_0) &= \int_0^\tau \int_0^\tau \Psi(\beta_0, t, \omega) s^{(0)}(\beta_0, t) s^{(0)}(\beta_0, \omega) \lambda_0(t) \lambda_0(\omega) dt d\omega, \\ \Psi(\beta_0, t, \omega) &= (1 - \alpha) \alpha^{-1} \{s^{(0)}(\beta_0, t) s^{(0)}(\beta_0, \omega)\}^{-1} \\ &\quad \cdot \{q^{(2)}(\beta_0, t, \omega) - q^{(1)}(\beta_0, \omega, t) e(\beta_0, \omega)' - e(\beta_0, t) q^{(1)}(\beta_0, t, \omega)' \\ &\quad + q^{(0)}(\beta_0, t, \omega) e(\beta_0, t) e(\beta_0, \omega)'\}, \end{aligned} \quad (15)$$

$$e(\beta_0, t) = \frac{s^{(1)}(\beta_0, t)}{s^{(0)}(\beta_0, t)},$$

$$e(\beta_0, \omega) = \frac{s^{(1)}(\beta_0, \omega)}{s^{(0)}(\beta_0, \omega)}.$$

4. The Implementation of Algorithm

In Section 3, we have finished the asymptotic properties for the adjusted estimation $\hat{\beta}$. To obtain $\hat{\beta}$ by solution of function (9), we can use the Newton–Raphson algorithm. However, in the process of calculation, we find that its Hessian matrix is complicated and easily irreversible. So, we proposed the MM algorithm [16] to get $\hat{\beta}$ for the adjustment of distorted covariates under case-cohort design.

4.1. Construct Surrogate. The key of the MM algorithm to construct surrogate function: combine the ideas of [16]; we expose the surrogate function $\hat{Q}(\beta | \beta^{(m)})$ for $\hat{l}_p(\beta)$ by using the convexity of the exponential function e^x and the negative logarithm function $-\log(x)$:

$$\hat{Q}(\beta | \beta^{(m)}) = \hat{c}_0 + \sum_{i=1}^n \omega_i \Delta_i \left[\beta^\top \hat{Z}_i(T_i) - \frac{\sum_{l=1}^n \sum_{r=1}^p Y_l(T_i) \omega_l \hat{\lambda}_{lr} \hat{g}_{lr}(\beta_r | \beta^{(m)})}{\sum_{l=1}^n Y_l(T_i) \omega_l \exp\{\hat{Z}_l^\top(T_i) \beta^{(m)}\}} \right], \quad (16)$$

where $\beta^{(m)}$ is the m th approximate of the adjusted maximum likelihood estimation β defined in (9), and

$$\begin{aligned}\hat{c}_0 &= \sum_{i=1}^n \omega_i \Delta_i \left\{ 1 - \log \left[\sum_{l=1}^n Y_l(T_i) \omega_l \exp \{ \hat{Z}_l(T_i)^\top \beta^{(m)} \} \right] \right\}, \\ \hat{g}_{lr}(\beta_r | \beta^{(m)}) &= \exp \left[\hat{\lambda}_{lr}^{-1} \hat{Z}_{lr}(T_i) (\beta_r - \beta_r^{(m)}) + \hat{Z}_l^\top(T_i) \beta^{(m)} \right], \\ \hat{\lambda}_{lr} &= \frac{|\hat{Z}_{lr}(T_i)|}{\sum_{r=1}^p |\hat{Z}_{lr}(T_i)|},\end{aligned}\tag{17}$$

with $\hat{\lambda}_{lr}^{-1} = 0$ if $\hat{Z}_l^\top(T_i) = 0$, where $\beta_r, \beta_r^{(m)}$, and $\hat{Z}_{lr}$ are the r th components of $\beta, \beta^{(m)}$, and $\hat{Z}_l$, respectively.

With the surrogate function (16), we can transfer the optimization problem (9) into the problem as follows:

$$\beta^{(m+1)} = \arg \max_{\beta} \hat{Q}(\beta | \beta^{(m)}). \tag{18}$$

Following the construction process of the surrogate function (16), we can immediately conclude that

$$\hat{l}_p(\beta) \geq \hat{Q}(\beta | \beta^{(m)}), \tag{19}$$

where $\hat{l}_p(\beta) = \hat{Q}(\beta | \beta^{(m)})$ if and only if $\beta = \beta^{(m)}$.

$$\begin{aligned}\text{Then, with } \hat{Q}(\beta^{(m+1)} | \beta^{(m)}) &\geq \hat{Q}(\beta^{(m)} | \beta^{(m)}), \text{ we have} \\ \hat{l}_p(\beta^{(m+1)}) &= \hat{l}_p(\beta^{(m+1)}) - \hat{Q}(\beta^{(m+1)} | \beta^{(m)}) + \hat{Q}(\beta^{(m+1)} | \beta^{(m)}) \\ &\geq \hat{l}_p(\beta^{(m)}) - \hat{Q}(\beta^{(m)} | \beta^{(m)}) + \hat{Q}(\beta^{(m)} | \beta^{(m)}) \\ &= \hat{l}_p(\beta^{(m)}),\end{aligned}\tag{20}$$

where $\hat{l}_p(\beta)$ is strictly ascending when $\beta^{(m+1)} \neq \beta^{(m)}$.

With the surrogate function (16), we obtain the derivatives with respect to β if $\beta = \beta^{(m)}$ as follows.

The score vector:

$$\frac{\partial \hat{Q}(\beta^{(m)} | \beta^{(m)})}{\partial \beta} = \sum_{i=1}^n \Delta_i \left[\hat{Z}(T_i) - \frac{\sum_{l=1}^n Y_l(T_i) \omega_l \hat{Z}_l(T_i) \exp \{ \hat{Z}_l^\top(T_i) \beta^{(m)} \}}{\sum_{l=1}^n Y_l(T_i) \omega_l \exp \{ \hat{Z}_l^\top(T_i) \beta^{(m)} \}} \right]. \tag{21}$$

The negative Hessian matrix:

$$\begin{aligned}\frac{\partial^2 \hat{Q}(\beta^{(m)} | \beta^{(m)})}{\partial \beta \partial \beta^\top} &= \text{diag} \left\{ -\frac{\partial^2 \hat{Q}(\beta^{(m)} | \beta^{(m)})}{\partial^2 \beta_1}, \dots, \right. \\ &\quad \left. -\frac{\partial^2 \hat{Q}(\beta^{(m)} | \beta^{(m)})}{\partial^2 \beta_p} \right\},\end{aligned}\tag{22}$$

where

$$\begin{aligned}-\frac{\partial^2 \hat{Q}(\beta^{(m)} | \beta^{(m)})}{\partial \beta_r^2} &= \sum_{i=1}^n \Delta_i \omega_i \frac{\sum_{l=1}^n Y_l(T_i) \omega_l \hat{Z}_{lr}^2 / \hat{\lambda}_{lr} \exp \{ \hat{Z}_l^\top(T_i) \beta^{(m)} \}}{\sum_{l=1}^n Y_l(T_i) \omega_l \exp \{ \hat{Z}_l^\top(T_i) \beta^{(m)} \}}, \\ r &= 1, 2, \dots, p.\end{aligned}\tag{23}$$

Here, the Hessian matrix is diagonal, the process of calculation changes simply and is reversible by using Newton-Raphson algorithm.

4.2. Cross-Validation and Bootstrap

4.2.1. Cross-Validation. In the process of the adjustment for the distorted covariates, we use the cross-validation (CV) method to obtain the optimal bandwidth h to construct the

kernel smoothing methods. The idea of the CV method is to calculate the coverage of error between some naive responses and the estimated responses obtained from the model fitted by the other predictors. The specific expression of the function in this article is given as follows:

$$\begin{aligned}CV(h) &= \sum_{j=1}^n \left[\frac{1}{nh} \sum_{i=1}^n K\left(\frac{u_j - u_i}{h}\right) \right]^2 \Delta u \\ &\quad - \frac{2}{n} \sum_{j=1}^n \left[\frac{2}{nh} \sum_{j \neq i} K\left(\frac{u_j - u_i}{h}\right) \right],\end{aligned}\tag{24}$$

where u_{-i} is the set of removed i th element, $\Delta u = u_j - u_{j-1}$.

Then, the optimal bandwidth is defined as follows:

$$\hat{h} = \arg \min_h CV(h). \tag{25}$$

4.2.2. Bootstrap. Bootstrap has been widely used since it was first proposed [24–26]. We adopt this method to calculate the standard error of β ; the basic idea of the nonparametric approach is to construct an empirical distribution function by repeatedly sampling from the observed data. The steps are summarized as follows:

Step 1: $\{\hat{X}_1, \hat{X}_2, \dots, \hat{X}_n\}$ are the adjusted observations under the case-cohort design, where $\hat{X}_i = \{(T_i, \Delta_i, U_i, (\xi_i + (1 - \xi_i)\Delta_i)\hat{Z}_i, 1 \leq i \leq n\}$. We use the samples

with $\Delta_i = 1$ as the failure cases and the samples with $\Delta_i = 0$ as the censoring objects to constitute the bootstrap case-cohort sample $\{\hat{X}_1^*, \hat{X}_2^*, \dots, \hat{X}_n^*\}$.

Step 2: a bootstrap replication $\hat{\beta}^*$ can be obtained based on $\{\hat{X}_1^*, \hat{X}_2^*, \dots, \hat{X}_n^*\}$.

Step 3: repeating step 1 and step 2 B times, we can obtain B bootstrap estimate $\{\hat{\beta}_1^*, \dots, \hat{\beta}_B^*\}$. Therefore, the standard error of the r th component of $\hat{\beta}$ can be estimated by the following form:

$$\widehat{se}(\hat{\beta}_r) = \sqrt{\frac{1}{B-1} \sum_{b=1}^B \left[\hat{\beta}_r^*(b) - \frac{1}{B} \sum_{b=1}^B \hat{\beta}_r^*(b) \right]^2}, \quad (26)$$

$$r = 1, 2, \dots, p.$$

5. Simulation

As we mentioned before, if the distorting factors on covariates were ignored, the result of inference may be misled [11, 12]. In this article, we will construct several simulation studies to compare the three estimators of regression coefficient $\hat{\beta}$ for the covariates without distorting factors, with distorting factors, and adjusted with distorting factors in Cox's model. Under the case-cohort design, where $\hat{\beta}_W$ denotes the estimator calculated based on the covariates without distorting factors, $\hat{\beta}_C$ denotes the estimator calculated based on the covariates with distorting factors, and $\hat{\beta}_A$ denotes the estimator calculate based on the adjusted covariates with distorting factors, given (Z_1, Z_2) , we consider the hazard function of the failure time $\tilde{T}_i$ as follows:

$$\lambda(t|Z_1, Z_2) = \lambda_0(t) \exp(\beta_1 Z_1 + \beta_2 Z_2). \quad (27)$$

We set the true values of the parameters β to $(-0.25, 0.5)$, $(-0.25, 0.693)$, and $(0, 0.5)$. Z_1 is generated from a normal distribution with mean Z and variance 1.44. Z_2 is generated from a uniform distribution $U((7/3) - (\sqrt{19}/2), (7/3) + (\sqrt{19}/2))$. The distorting factor U is generated from a uniform distribution $U \sim U(4 - \sqrt{7}, 4 + \sqrt{7})$. The baseline hazard function $\lambda_0(t)$ is set to be 1 and $2t$. Thus, the failure time $\tilde{T}$ satisfies an exponential distribution with failure rate $\exp(\beta_1 Z_1 + \beta_2 Z_2)$ and scale parameter $[\exp(\beta_1 Z_1 + \beta_2 Z_2)]^{-1/2}$, respectively. The censoring time C is generated from a uniform distribution $U(0, c)$ with c being chosen to desire the censoring rate $\rho = 80\%$, 85% , and 90% .

Under the case-cohort design, 1000 full cohort samples are generated from function (27), and the corresponding distorting factor is generated. The two parts are merged to constitute the observed sample. Then, we randomly selected $\tilde{n} = 150, 200$, and 300 subcohort from the full cohort.

We compare the sample bias (Bias), the sample MSE (SMSE), the sample standard deviations of estimates (SD), the means of estimated standard errors (SE), and the coverage probabilities of 95% nominal confidence intervals (CP) of the three estimators to $\hat{\beta}_C, \hat{\beta}_W$, and $\hat{\beta}_A$ based on 1000

independent simulated datasets and especially apply the nonparametric bootstrap approach in Section 4.2.2 with $B = 500$. We set only the covariate Z_2 to be distorted and adjusted. The distribution of the distorting factor U is set to be $\Phi(u) = ((u + 10)^2 / 194.9160)$. The kernel function is set to be an Epanechnikov kernel [27] $K(t) = (15/32)(3 - 7t^2)(1 - t^2)(|t| \leq 1)$. The optimal bandwidth h is selected by the CV method in Section 4.2.1. The criteria are stopped with $\varepsilon = 0.0001$. The simulation results are summarized in Table 1 with $\beta = (-0.25, 0.5)$, Table 2 with $\beta = (-0.25, 0.693)$, and Table 3 with $\beta = (0, 0.5)$, respectively.

According to the simulation results, the estimators $\hat{\beta}_W$ and $\hat{\beta}_A$ for β_2 are all unbiased, the three values of SMSE, SD, and SE are all closed, and the CP values are reasonable. However, the estimators $\hat{\beta}_C$ for β_2 are all biased, the difference between the two values of SMSE and SD is large, and the CP value is low. These facts confirm that ignoring the distorting factors may lead to the wrong results and conclusion and also explain that the proposed method about the adjustment of covariates is effective. For example, the estimator $\hat{\beta}_C$ and $\hat{\beta}_A$ for $\beta = (-0.25, 0.5)$, $\rho = 80\%$, $\tilde{n} = 150$, for β_2 , the bias is -0.1178 , SMSE are 0.1521 and 0.1095 , and CP are 0.735 and 0.947 , respectively. Furthermore, we can conclude that the estimators for β_1 are all reasonable.

6. Real Data Analysis

The dataset is about the medical records of heart failure patients, from "Machine Learning Repository" (<http://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records>). The original dataset version was collected by Tanvir Ahmad et al. (Government College University, Faisalabad, Pakistan) and made available by them on FigShare under the Attribution 4.0 International copyright in July 2017. The current version of the dataset was elaborated by Davide Chicco et al. and donated to the University of California Irvine Machine Learning Repository under the same Attribution 4.0 International copyright in January 2020.

This dataset contains the medical records of 299 patients who had heart failure, collected during their follow-up period, where each patient profile has 13 clinical features, including age, anaemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, smoking, time, and death event. The goal is to assess the association between clinical features and heart failure and its happened time. This real data is survival data which meets the requirements of the article. Death event and time are regarded as heart failure and its happened time, respectively. For the remaining 11 clinical features, through pairwise correlation analysis, we conclude that ejection fraction and serum sodium and anaemia and Creatinine phosphokinase are significantly related. Referring to relevant medical knowledge, we know that ejection fraction may distort serum sodium. And in the same way, anaemia may distort creatinine phosphokinase. The main purpose of our research is to evaluate the

TABLE 1: $\beta = (-0.25, 0.5)$, $Z_1 \sim N(2, 1.44)$, $Z_2 \sim U((7/3) - (\sqrt{19}/2), (7/3) + (\sqrt{19}/2))$, $u \sim U(4 - \sqrt{7}, 4 + \sqrt{7})$.

$\lambda_0(t) = 1$													
$\bar{n}$	ρ (%)	Type	$\beta_1 = -0.25$					$\beta_2 = 0.5$					
			Bias	SMSE	SD	SE	CP	Bias	SMSE	SD	SE	CP	
150	80	$\hat{\beta}_W$	-0.0102	0.0965	0.0960	0.0992	0.965	0.0146	0.1094	0.1085	0.1118	0.961	
		$\hat{\beta}_C$	-0.0074	0.0961	0.0958	0.0978	0.959	-0.1158	0.1499	0.0951	0.0975	0.742	
		$\hat{\beta}_A$	-0.0100	0.0960	0.0955	0.0990	0.964	0.0050	0.1105	0.1104	0.1126	0.955	
	85	$\hat{\beta}_W$	-0.0145	0.1051	0.1042	0.1053	0.950	0.0115	0.1167	0.1162	0.1189	0.960	
		$\hat{\beta}_C$	-0.0003	0.0968	0.0968	0.1030	0.956	-0.1199	0.1575	0.1022	0.1024	0.724	
		$\hat{\beta}_A$	-0.0014	0.0971	0.0971	0.1041	0.964	-0.0046	0.1201	0.1201	0.1197	0.950	
	90	$\hat{\beta}_W$	-0.0098	0.1114	0.1110	0.1153	0.961	0.0237	0.1306	0.1285	0.1328	0.957	
		$\hat{\beta}_C$	-0.0094	0.1096	0.1093	0.1141	0.962	-0.1129	0.1563	0.1081	0.1116	0.802	
		$\hat{\beta}_A$	-0.0096	0.1113	0.1109	0.1152	0.963	0.0106	0.1292	0.1288	0.1326	0.954	
	200	80	$\hat{\beta}_W$	-0.0090	0.0815	0.0810	0.0878	0.969	0.0084	0.0991	0.0988	0.0992	0.943
			$\hat{\beta}_C$	-0.0051	0.0807	0.0806	0.0868	0.971	-0.1227	0.1497	0.0858	0.0852	0.669
			$\hat{\beta}_A$	-0.0091	0.0818	0.0814	0.0877	0.968	0.0006	0.1012	0.1013	0.0994	0.944
85		$\hat{\beta}_W$	-0.0049	0.0899	0.0898	0.0926	0.951	0.0005	0.1007	0.1008	0.1055	0.963	
		$\hat{\beta}_C$	-0.0027	0.0928	0.0928	0.0914	0.940	-0.1274	0.1560	0.0900	0.0897	0.666	
		$\hat{\beta}_A$	-0.0048	0.0950	0.0949	0.0923	0.941	-0.0026	0.1063	0.1063	0.1061	0.949	
90		$\hat{\beta}_W$	-0.0094	0.1035	0.1031	0.1036	0.954	0.0117	0.1174	0.1169	0.1196	0.962	
		$\hat{\beta}_C$	-0.0089	0.1038	0.1035	0.1027	0.949	-0.1229	0.1563	0.0965	0.0986	0.720	
		$\hat{\beta}_A$	-0.0085	0.1034	0.1031	0.1034	0.953	-0.0005	0.1151	0.1152	0.1188	0.960	
300		80	$\hat{\beta}_W$	-0.0029	0.0732	0.0732	0.0737	0.951	0.0060	0.0848	0.0847	0.0850	0.954
			$\hat{\beta}_C$	-0.0008	0.0725	0.0726	0.0730	0.941	-0.1241	0.1432	0.0714	0.0712	0.561
			$\hat{\beta}_A$	-0.0023	0.0727	0.0727	0.0736	0.950	-0.0046	0.0844	0.0843	0.0848	0.952
	85	$\hat{\beta}_W$	-0.0054	0.0804	0.0803	0.0795	0.948	0.0069	0.0902	0.0900	0.0929	0.959	
		$\hat{\beta}_C$	-0.0025	0.0768	0.0768	0.0792	0.953	-0.1325	0.1532	0.0768	0.0761	0.568	
		$\hat{\beta}_A$	-0.0036	0.0764	0.0763	0.0798	0.954	-0.0114	0.0927	0.0921	0.0921	0.947	
	90	$\hat{\beta}_W$	-0.0034	0.0912	0.0912	0.0910	0.958	0.0062	0.1105	0.1103	0.1078	0.950	
		$\hat{\beta}_C$	-0.0023	0.0902	0.0902	0.0904	0.961	-0.1308	0.1575	0.0877	0.0861	0.632	
		$\hat{\beta}_A$	-0.0037	0.0908	0.0907	0.0909	0.957	-0.0069	0.1093	0.1091	0.1063	0.947	
	$\lambda_0(t) = 2t$												
	150	80	$\hat{\beta}_W$	-0.0067	0.1107	0.1105	0.1092	0.948	0.0103	0.1205	0.1201	0.1230	0.958
			$\hat{\beta}_C$	-0.0037	0.1110	0.1110	0.1078	0.943	-0.1170	0.1571	0.1049	0.1082	0.773
$\hat{\beta}_A$			-0.0060	0.1098	0.1097	0.1091	0.948	0.0001	0.1198	0.1199	0.1238	0.958	
85		$\hat{\beta}_W$	-0.0040	0.1111	0.1111	0.1143	0.966	0.0161	0.1316	0.1306	0.1307	0.952	
		$\hat{\beta}_C$	-0.0001	0.1104	0.1105	0.1127	0.958	-0.1183	0.1606	0.1087	0.1139	0.798	
		$\hat{\beta}_A$	-0.0036	0.1109	0.1109	0.1141	0.964	0.0069	0.1303	0.1302	0.1314	0.951	
90		$\hat{\beta}_W$	-0.0075	0.1242	0.1240	0.1254	0.954	0.0102	0.1445	0.1442	0.1433	0.954	
		$\hat{\beta}_C$	-0.0062	0.1228	0.1227	0.1238	0.954	-0.1259	0.1737	0.1197	0.1216	0.784	
		$\hat{\beta}_A$	-0.0066	0.1234	0.1233	0.1251	0.952	-0.0037	0.1434	0.1434	0.1433	0.947	
200		80	$\hat{\beta}_W$	0.0016	0.0911	0.0911	0.0944	0.951	0.0076	0.1022	0.1020	0.1074	0.957
			$\hat{\beta}_C$	0.0056	0.0916	0.0915	0.0932	0.952	-0.1213	0.1521	0.0917	0.0934	0.710
			$\hat{\beta}_A$	0.0021	0.0918	0.0918	0.0943	0.954	-0.0018	0.1024	0.1025	0.1078	0.959
	85	$\hat{\beta}_W$	-0.0062	0.0971	0.0970	0.1011	0.965	0.0116	0.1139	0.1133	0.1160	0.956	
		$\hat{\beta}_C$	-0.0033	0.0973	0.0973	0.1000	0.959	-0.1198	0.1542	0.0971	0.0997	0.747	
		$\hat{\beta}_A$	-0.0060	0.0970	0.0969	0.1010	0.964	0.0001	0.1141	0.1142	0.1164	0.946	
	90	$\hat{\beta}_W$	-0.0063	0.1060	0.1059	0.1124	0.956	0.0117	0.1278	0.1273	0.1297	0.958	
		$\hat{\beta}_C$	-0.0050	0.1061	0.1060	0.1113	0.957	-0.1212	0.1608	0.1056	0.1085	0.746	
		$\hat{\beta}_A$	-0.0069	0.1064	0.1062	0.1123	0.959	-0.0007	0.1253	0.1254	0.1296	0.956	
	300	80	$\hat{\beta}_W$	-0.0008	0.0763	0.0764	0.0790	0.964	0.0036	0.0880	0.0879	0.0909	0.963
			$\hat{\beta}_C$	0.0024	0.0758	0.0758	0.0783	0.961	-0.1238	0.1458	0.0770	0.0774	0.604
			$\hat{\beta}_A$	-0.0003	0.0766	0.0767	0.0789	0.963	-0.0048	0.0886	0.0886	0.0909	0.959
85		$\hat{\beta}_W$	-0.0033	0.0871	0.0871	0.0854	0.948	0.0120	0.0991	0.0985	0.0989	0.951	
		$\hat{\beta}_C$	0.0002	0.0862	0.0862	0.0847	0.949	-0.1227	0.1479	0.0827	0.0825	0.660	
		$\hat{\beta}_A$	-0.0028	0.0870	0.0870	0.0853	0.948	0.0018	0.0979	0.0979	0.0987	0.942	
90		$\hat{\beta}_W$	-0.0127	0.0935	0.0927	0.0978	0.969	0.0117	0.1149	0.1144	0.1141	0.945	
		$\hat{\beta}_C$	-0.0110	0.0933	0.0927	0.0970	0.961	-0.1272	0.1575	0.0929	0.0930	0.676	
		$\hat{\beta}_A$	-0.0119	0.0929	0.0922	0.0977	0.967	-0.0014	0.1125	0.1126	0.1131	0.946	

Note: $\hat{\beta}_W$ denotes the estimator calculated based on the covariates without distorting factors, $\hat{\beta}_C$ denotes the estimator calculated based on the covariates with distorting factors, and $\hat{\beta}_A$ denotes the estimator calculated based on the adjusted covariates with distorting factors. Algorithm: MM algorithm.

TABLE 2: $\beta = (-0.25, 0.693)$, $Z_1 \sim N(2, 1.44)$, $Z_2 \sim U((7/3) - (\sqrt{19}/2), (7/3) + (\sqrt{19}/2))$, $u \sim U(4 - \sqrt{7}, 4 + \sqrt{7})$.

$\lambda_0(t) = 1$													
$\bar{n}$	ρ (%)	Type	$\beta_1 = -0.25$					$\beta_2 = 0.693$					
			Bias	SMSE	SD	$\hat{\beta}_A$ SE	CP	Bias	SMSE	SD	SE	CP	
150	80	$\hat{\beta}_W$	-0.0048	0.0998	0.0998	0.1038	0.965	0.0147	0.1190	0.1181	0.1228	0.968	
		$\hat{\beta}_C$	0.0005	0.0987	0.0988	0.1016	0.969	-0.1921	0.2170	0.1009	0.1042	0.513	
		$\hat{\beta}_A$	-0.0038	0.1005	0.1005	0.1038	0.963	-0.0010	0.1237	0.1237	0.1236	0.944	
	85	$\hat{\beta}_W$	-0.0141	0.1082	0.1074	0.1108	0.959	0.0110	0.1320	0.1316	0.1318	0.955	
		$\hat{\beta}_C$	-0.0091	0.1042	0.1038	0.1087	0.964	-0.1980	0.2258	0.1086	0.1088	0.517	
		$\hat{\beta}_A$	-0.0137	0.1078	0.1070	0.1106	0.959	-0.0058	0.1351	0.1351	0.1322	0.947	
	90	$\hat{\beta}_W$	-0.0060	0.1127	0.1126	0.1207	0.974	0.0110	0.1462	0.1459	0.1446	0.953	
		$\hat{\beta}_C$	-0.0038	0.1100	0.1100	0.1185	0.975	-0.2056	0.2363	0.1165	0.1160	0.537	
		$\hat{\beta}_A$	-0.0057	0.1139	0.1138	0.1204	0.967	-0.0159	0.1453	0.1445	0.1443	0.946	
	200	80	$\hat{\beta}_W$	-0.0105	0.0941	0.0936	0.0914	0.949	0.0105	0.1089	0.1085	0.1092	0.953
			$\hat{\beta}_C$	-0.0017	0.0929	0.0929	0.0897	0.954	-0.1969	0.2150	0.0863	0.0906	0.382
			$\hat{\beta}_A$	-0.0095	0.0937	0.0933	0.0913	0.948	-0.0072	0.1111	0.1109	0.1097	0.949
85		$\hat{\beta}_W$	-0.0083	0.0973	0.0970	0.0979	0.958	0.0115	0.1088	0.1083	0.1177	0.964	
		$\hat{\beta}_C$	-0.0035	0.0955	0.0955	0.0960	0.954	-0.1990	0.2178	0.0886	0.0949	0.446	
		$\hat{\beta}_A$	-0.0079	0.0979	0.0977	0.0977	0.956	-0.0087	0.1125	0.1122	0.1172	0.953	
90		$\hat{\beta}_W$	-0.0046	0.1064	0.1064	0.1084	0.956	0.0122	0.1264	0.1259	0.1326	0.971	
		$\hat{\beta}_C$	-0.0022	0.1056	0.1057	0.1067	0.953	-0.2018	0.2278	0.1057	0.1040	0.461	
		$\hat{\beta}_A$	-0.0036	0.1060	0.1060	0.1082	0.957	-0.0140	0.1281	0.1273	0.1312	0.959	
300		80	$\hat{\beta}_W$	-0.0025	0.0776	0.0776	0.0765	0.953	0.0088	0.0932	0.0928	0.0928	0.955
			$\hat{\beta}_C$	0.0038	0.0766	0.0766	0.0752	0.955	-0.2000	0.2139	0.0758	0.0745	0.260
			$\hat{\beta}_A$	-0.0011	0.0773	0.0774	0.0763	0.958	-0.0090	0.0957	0.0953	0.0923	0.945
	85	$\hat{\beta}_W$	-0.0037	0.0833	0.0833	0.0832	0.947	0.0084	0.1041	0.1038	0.1024	0.947	
		$\hat{\beta}_C$	0.0011	0.0822	0.0822	0.0819	0.945	-0.2078	0.2223	0.0791	0.0791	0.257	
		$\hat{\beta}_A$	-0.0027	0.0838	0.0838	0.0831	0.951	-0.0145	0.1051	0.1042	0.1011	0.934	
	90	$\hat{\beta}_W$	-0.0035	0.0959	0.0959	0.0951	0.954	0.0026	0.1153	0.1153	0.1181	0.949	
		$\hat{\beta}_C$	0.0001	0.0961	0.0961	0.0940	0.949	-0.2135	0.2302	0.0861	0.0885	0.326	
		$\hat{\beta}_A$	-0.0021	0.0956	0.0957	0.0950	0.953	-0.0195	0.1186	0.1170	0.1160	0.936	
	$\lambda_0(t) = 2t$												
	150	80	$\hat{\beta}_W$	-0.0090	0.1123	0.1120	0.1137	0.952	0.0131	0.1275	0.1269	0.1338	0.954
			$\hat{\beta}_C$	0.0006	0.1110	0.1111	0.1115	0.951	-0.1890	0.2209	0.1144	0.1151	0.579
$\hat{\beta}_A$			-0.0073	0.1111	0.1109	0.1135	0.953	-0.0032	0.1302	0.1303	0.1350	0.951	
85		$\hat{\beta}_W$	-0.0109	0.1207	0.1203	0.1208	0.949	0.0177	0.1396	0.1386	0.1434	0.962	
		$\hat{\beta}_C$	-0.0085	0.1191	0.1188	0.1180	0.953	-0.1864	0.2225	0.1216	0.1205	0.589	
		$\hat{\beta}_A$	-0.0106	0.1198	0.1194	0.1205	0.946	-0.0012	0.1387	0.1388	0.1439	0.950	
90		$\hat{\beta}_W$	-0.0124	0.1287	0.1281	0.1323	0.961	0.0158	0.1631	0.1624	0.1565	0.949	
		$\hat{\beta}_C$	-0.0138	0.1277	0.1270	0.1296	0.955	-0.1939	0.2305	0.1248	0.1292	0.621	
		$\hat{\beta}_A$	-0.0151	0.1294	0.1286	0.1317	0.956	-0.0072	0.1557	0.1556	0.1572	0.955	
200		80	$\hat{\beta}_W$	-0.0115	0.0994	0.0988	0.0986	0.951	0.0162	0.1203	0.1193	0.1184	0.959
			$\hat{\beta}_C$	-0.0025	0.0941	0.0941	0.0963	0.956	-0.1909	0.2154	0.0999	0.0991	0.492
			$\hat{\beta}_A$	-0.0109	0.0988	0.0983	0.0983	0.953	-0.0011	0.1214	0.1215	0.1188	0.939
	85	$\hat{\beta}_W$	-0.0007	0.1009	0.1010	0.1050	0.960	0.0091	0.1260	0.1258	0.1269	0.954	
		$\hat{\beta}_C$	0.0036	0.1008	0.1008	0.1031	0.957	-0.1955	0.2218	0.1048	0.1042	0.489	
		$\hat{\beta}_A$	-0.0009	0.1012	0.1012	0.1050	0.955	-0.0081	0.1280	0.1278	0.1272	0.949	
	90	$\hat{\beta}_W$	-0.0051	0.1105	0.1105	0.1176	0.962	0.0157	0.1484	0.1477	0.1429	0.943	
		$\hat{\beta}_C$	-0.0011	0.1082	0.1082	0.1155	0.970	-0.1996	0.2300	0.1144	0.1142	0.540	
		$\hat{\beta}_A$	-0.0045	0.1103	0.1103	0.1173	0.961	-0.0111	0.1434	0.1430	0.1416	0.941	
	300	80	$\hat{\beta}_W$	-0.0067	0.0799	0.0797	0.0818	0.952	0.0056	0.0934	0.0933	0.0989	0.962
			$\hat{\beta}_C$	0.0008	0.0789	0.0790	0.0805	0.948	-0.1969	0.2114	0.0769	0.0807	0.322
			$\hat{\beta}_A$	-0.0058	0.0800	0.0798	0.0817	0.947	-0.0112	0.0962	0.0956	0.0988	0.952
85		$\hat{\beta}_W$	-0.0083	0.0860	0.0857	0.0894	0.957	0.0081	0.1063	0.1061	0.1082	0.961	
		$\hat{\beta}_C$	-0.0004	0.0854	0.0855	0.0879	0.965	-0.2018	0.2194	0.0861	0.0864	0.352	
		$\hat{\beta}_A$	-0.0072	0.0867	0.0864	0.0893	0.959	-0.0115	0.1079	0.1073	0.1078	0.950	
90		$\hat{\beta}_W$	-0.0082	0.0966	0.0963	0.1007	0.961	0.0125	0.1223	0.1217	0.1259	0.963	
		$\hat{\beta}_C$	-0.0034	0.0948	0.0948	0.0991	0.951	-0.2012	0.2219	0.0935	0.0964	0.428	
		$\hat{\beta}_A$	-0.0060	0.0964	0.0963	0.1006	0.955	-0.0112	0.1206	0.1201	0.1239	0.954	

Note: $\hat{\beta}_W$ denotes the estimator calculated based on the covariates without distorting factors, $\hat{\beta}_C$ denotes the estimator calculated based on the covariates with distorting factors, and $\hat{\beta}_A$ denotes the estimator calculate based on the adjusted covariates with distorting factors. Algorithm: MM algorithm.

TABLE 3: $\beta = (0, 0.5)$, $Z_1 \sim N(2, 1.44)$, $Z_2 \sim U((7/3) - (\sqrt{19}/2), (7/3) + (\sqrt{19}/2))$, $u \sim U(4 - \sqrt{7}, 4 + \sqrt{7})$.

$\bar{n}$	ρ (%)	Type	$\lambda_0(t) = 1$										
			$\beta_1 = 0$					$\beta_2 = 0.5$					
			Bias	SMSE	SD	SE	CP	Bias	SMSE	SD	SE	CP	
150	80	$\hat{\beta}_W$	0.0016	0.0899	0.0899	0.0930	0.961	0.0106	0.1032	0.1027	0.1084	0.968	
		$\hat{\beta}_C$	0.0016	0.0900	0.0901	0.0920	0.963	-0.1184	0.1490	0.0906	0.0954	0.759	
		$\hat{\beta}_A$	0.0017	0.0903	0.0903	0.0929	0.957	0.0014	0.1058	0.1058	0.1091	0.962	
	85	$\hat{\beta}_W$	-0.0013	0.1010	0.1010	0.0999	0.959	0.0079	0.1107	0.1104	0.1151	0.964	
		$\hat{\beta}_C$	-0.0007	0.0991	0.0991	0.0989	0.961	-0.1191	0.1532	0.0964	0.0998	0.728	
		$\hat{\beta}_A$	-0.0011	0.1012	0.1012	0.0997	0.958	-0.0052	0.1114	0.1113	0.1154	0.952	
	90	$\hat{\beta}_W$	-0.0003	0.1078	0.1079	0.1083	0.954	0.0123	0.1242	0.1236	0.1282	0.965	
		$\hat{\beta}_C$	-0.0004	0.1083	0.1084	0.1074	0.946	-0.1218	0.1576	0.1000	0.1076	0.772	
		$\hat{\beta}_A$	-0.0003	0.1078	0.1079	0.1082	0.956	-0.0002	0.1225	0.1225	0.1282	0.965	
	200	80	$\hat{\beta}_W$	0.0016	0.0807	0.0807	0.0823	0.963	0.0073	0.0973	0.0971	0.0963	0.951
			$\hat{\beta}_C$	0.0029	0.0797	0.0796	0.0813	0.962	-0.1239	0.1488	0.0824	0.0827	0.633
			$\hat{\beta}_A$	0.0017	0.0803	0.0804	0.0822	0.966	-0.0018	0.0984	0.0985	0.0965	0.943
85		$\hat{\beta}_W$	0.0007	0.0862	0.0863	0.0881	0.962	0.0071	0.1024	0.1022	0.1036	0.953	
		$\hat{\beta}_C$	0.0010	0.0845	0.0846	0.0873	0.960	-0.1230	0.1515	0.0884	0.0878	0.672	
		$\hat{\beta}_A$	0.0005	0.0865	0.0865	0.0881	0.965	-0.0048	0.1032	0.1031	0.1037	0.948	
90		$\hat{\beta}_W$	0.0037	0.0940	0.0940	0.0986	0.954	0.0091	0.1148	0.1145	0.1168	0.959	
		$\hat{\beta}_C$	0.0044	0.0945	0.0944	0.0977	0.958	-0.1270	0.1580	0.0941	0.0960	0.688	
		$\hat{\beta}_A$	0.0034	0.0945	0.0945	0.0985	0.956	-0.0036	0.1158	0.1158	0.1163	0.953	
300		80	$\hat{\beta}_W$	0.0016	0.0706	0.0706	0.0703	0.951	0.0012	0.0803	0.0804	0.0827	0.964
			$\hat{\beta}_C$	0.0010	0.0705	0.0706	0.0696	0.942	-0.1294	0.1461	0.0679	0.0692	0.526
			$\hat{\beta}_A$	0.0019	0.0706	0.0706	0.0702	0.945	-0.0081	0.0816	0.0813	0.0826	0.955
	85	$\hat{\beta}_W$	-0.0003	0.0767	0.0767	0.0761	0.947	-0.0006	0.0868	0.0868	0.0908	0.960	
		$\hat{\beta}_C$	0.0001	0.0765	0.0765	0.0756	0.947	-0.1321	0.1504	0.0719	0.0745	0.545	
		$\hat{\beta}_A$	-0.0002	0.0767	0.0767	0.0761	0.946	-0.0111	0.0884	0.0878	0.0904	0.950	
	90	$\hat{\beta}_W$	0.0007	0.0855	0.0855	0.0865	0.960	0.0095	0.1046	0.1042	0.1054	0.954	
		$\hat{\beta}_C$	0.0013	0.0845	0.0845	0.0861	0.969	-0.1274	0.1521	0.0831	0.0837	0.638	
		$\hat{\beta}_A$	0.0010	0.0856	0.0857	0.0865	0.959	-0.0027	0.1038	0.1038	0.1040	0.950	
	$\lambda_0(t) = 2t$												
	150	80	$\hat{\beta}_W$	0.0005	0.1012	0.1013	0.1035	0.957	0.0091	0.1144	0.1140	0.1195	0.953
			$\hat{\beta}_C$	0.0012	0.0998	0.0998	0.1024	0.959	-0.1191	0.1570	0.1024	0.1057	0.755
$\hat{\beta}_A$			0.0010	0.1007	0.1007	0.1034	0.956	0.0020	0.1158	0.1158	0.1201	0.954	
85		$\hat{\beta}_W$	-0.0020	0.1059	0.1059	0.1094	0.965	0.0132	0.1210	0.1203	0.1272	0.971	
		$\hat{\beta}_C$	-0.0031	0.1048	0.1048	0.1080	0.965	-0.1185	0.1590	0.1062	0.1109	0.798	
		$\hat{\beta}_A$	-0.0020	0.1057	0.1058	0.1091	0.965	0.0016	0.1198	0.1198	0.1279	0.966	
90		$\hat{\beta}_W$	-0.0038	0.1157	0.1157	0.1189	0.962	0.0069	0.1376	0.1375	0.1389	0.957	
		$\hat{\beta}_C$	-0.0040	0.1134	0.1134	0.1175	0.964	-0.1269	0.1698	0.1128	0.1182	0.781	
		$\hat{\beta}_A$	-0.0034	0.1160	0.1160	0.1187	0.959	-0.0036	0.1387	0.1388	0.1393	0.955	
200		80	$\hat{\beta}_W$	0.0042	0.0906	0.0905	0.0905	0.944	0.0100	0.1057	0.1053	0.1061	0.962
			$\hat{\beta}_C$	0.0044	0.0903	0.0902	0.0892	0.944	-0.1196	0.1509	0.0921	0.0924	0.696
			$\hat{\beta}_A$	0.0047	0.0904	0.0903	0.0903	0.942	-0.0002	0.1050	0.1050	0.1065	0.956
	85	$\hat{\beta}_W$	-0.0018	0.0950	0.0950	0.0954	0.962	0.0093	0.1089	0.1088	0.1124	0.962	
		$\hat{\beta}_C$	-0.0016	0.0949	0.0950	0.0942	0.952	-0.1258	0.1566	0.0933	0.0960	0.700	
		$\hat{\beta}_A$	-0.0017	0.0944	0.0945	0.0952	0.957	-0.0043	0.1092	0.1092	0.1125	0.958	
	90	$\hat{\beta}_W$	-0.0001	0.1007	0.1008	0.1065	0.972	0.0144	0.1238	0.1231	0.1263	0.965	
		$\hat{\beta}_C$	0.0009	0.0991	0.0991	0.1054	0.968	-0.1211	0.1587	0.1026	0.1055	0.763	
		$\hat{\beta}_A$	0.0000	0.1004	0.1005	0.1064	0.974	0.0026	0.1239	0.1239	0.1264	0.959	
	300	80	$\hat{\beta}_W$	-0.0010	0.0732	0.0732	0.0754	0.959	0.0089	0.0877	0.0873	0.0890	0.965
			$\hat{\beta}_C$	-0.0011	0.0726	0.0726	0.0746	0.962	-0.1216	0.1439	0.0769	0.0757	0.605
			$\hat{\beta}_A$	-0.0011	0.0733	0.0734	0.0754	0.960	-0.0012	0.0879	0.0879	0.0891	0.965
85		$\hat{\beta}_W$	0.0020	0.0810	0.0810	0.0813	0.954	0.0053	0.0963	0.0962	0.0969	0.959	
		$\hat{\beta}_C$	0.0011	0.0800	0.0800	0.0808	0.952	-0.1275	0.1508	0.0807	0.0812	0.628	
		$\hat{\beta}_A$	0.0015	0.0813	0.0813	0.0812	0.952	-0.0053	0.0956	0.0955	0.0967	0.952	
90		$\hat{\beta}_W$	0.0044	0.0926	0.0926	0.0925	0.953	0.0132	0.1092	0.1084	0.1118	0.961	
		$\hat{\beta}_C$	0.0045	0.0923	0.0923	0.0919	0.954	-0.1257	0.1530	0.0872	0.0906	0.687	
		$\hat{\beta}_A$	0.0041	0.0922	0.0922	0.0924	0.949	0.0013	0.1084	0.1085	0.1108	0.960	

Note: $\hat{\beta}_W$ denotes the estimator calculated based on the covariates without distorting factors, $\hat{\beta}_C$ denotes the estimator calculated based on the covariates with distorting factors, and $\hat{\beta}_A$ denotes the estimator calculated based on the adjusted covariates with distorting factors. Algorithm: MM algorithm.

association between these clinical features under smoothed serum sodium and creatinine phosphokinase by ejection fraction and anaemia.

There are 96 deaths in this data, and the censoring rate is nearby 67.9%. Due to the different dimensions of the covariates, we firstly standardize each covariate. Then, we set the standardized age as Z_1 , the standardized creatinine phosphokinase as $\tilde{Z}_2$, the standardized diabetes as Z_3 , the standardized high blood pressure as Z_4 , the standardized platelets as Z_5 , the standardized serum creatinine as Z_6 , the standardized serum sodium as $\tilde{Z}_7$, the standardized sex as Z_8 , the standardized smoking as Z_9 , the standardized ejection fraction as a distorting factor U_1 , and the standardized anaemia as a distorting factor U_2 , where $\tilde{Z}_2$ and $\tilde{Z}_7$ indicate that the covariate is distorted.

The case-cohort sample is composed of the subcohort randomly selected 100 from the whole data and the dead objects in the whole data expect the subcohort. We perform regression analysis on the sample using the proportional hazard function:

$$x(t|Z) = x_0(t) \exp \left\{ \sum_{i=1}^9 \beta_i Z_i \right\}. \quad (28)$$

Because the size of this data is small, to improve the reliability of the sampling method, the sampling process is repeated 100 times. The average value of the 100 distorted estimates is $(2.7224, 14.0226, 0.1603, 0.4295, -0.2667, 3.0520, 11.0269, -0.0742, 0.1933)^\top$.

Then, for the same 100 case-cohort sample, we obtain the adjusted covariates $\tilde{Z}_2$ and $\tilde{Z}_7$ by using the proposed kernel smoothing method on $\tilde{Z}_2$ and $\tilde{Z}_7$ and replace $\tilde{Z}_2$ and $\tilde{Z}_7$ with $\tilde{Z}_2$ and $\tilde{Z}_7$, respectively. On this basis, the average of adjusted estimates is $(2.5077, 1.0344, 0.1136, 0.5474, -0.3552, 2.8577, -2.4303, 0.2115)^\top$.

The kernel and bandwidth selector chosen are the same as those in Section 5. After the covariates were adjusted, the estimator of creatinine phosphokinase is adjusted from 1.0344 to 14.0226, and the estimator of serum sodium is adjusted from -2.4303 to 11.0269 ; the estimators of other covariates changed little after adjustment. The result is reasonable and feasible in medicine and also shows that the proposed method is effective for real data.

In practice, the covariates are measured after the case-cohort sampling to reduce costs and improve efficiency. In this real data, the size of the data is small, and the each covariate has been measured. To reduce the error, we repeat 100 samplings and set the average value as the estimators. For example, based on the 10th case-cohort sample from the original data, the distorted estimator is $(3.5393, -0.5883, 0.1934, 0.3712, 0.4114, 6.5332, -3.9763, 0.1236, 0.3585)^\top$; then, after the adjustment of covariates in the same sample, the adjusted estimator is $(4.5125, 79.7599, 0.1690, 0.2902, 1.0486, 5.9035, 12.7945, 0.1399, 0.1720)^\top$. Therefore, the estimators based on same case-cohort sample are quite different from the average of estimators, but the trends are the

same. So, for real data with a large sample size, we only need to draw a case-cohort sample and measure their covariates.

7. Conclusion

We study the estimation in Cox's model with the adjusted covariates by using the nonparametric method based on the kernel function under case-cohort design. Consistency and asymptotic normality of the proposed estimators are derived. We use MM algorithm for calculating the regression coefficients, where the surrogate function with a diagonal Hessian matrix is established to overcome the operation difficulty in the Newton-Raphson algorithm. Simulation and real data studies suggest that the case-cohort design can be used to reduce the cost for cohort studies and the development of the adjustment for distorted covariates has nice performance.

Here, the adjustment for survival data is discussed in Cox's model. The method can be extended to other models, such as an accelerated failure time model [28], an additive-multiplicative model [29], and an accelerated hazards model [30]. The cost-effective case-cohort design is mainly applied to the data with the high censoring rate. However, in the sampling process, some censoring subjects have not been measured. To improve the efficiency, our future works include developments of inference for estimation with adjusted covariates under a generalized case-cohort design [31] and an outcome-dependent sampling design.

Appendix

A. Proof of Asymptotic Properties

We start with a lemma, which is frequently used in the process of proof. Then, we proceed to the proof of Theorem 1, concerning consistency and asymptotic normality.

Lemma A.1. *Let $\eta(t)$ be a continuous function, satisfying $E[\eta(T)^2] < \infty$. Assume that conditions (C), (D1)-(D2), and (E) hold. The following asymptotic representation holds:*

$$\begin{aligned} & \sum_{i=1}^n (\hat{Z}_{ir} - Z_{ir}) \eta(T_{ir}) \\ &= \frac{1}{2} \sum_{i=1}^n (Z_{ir} - E[Z_r]) \frac{E[Z_r \eta(T_r)]}{E[Z_r]} \\ &+ \sum_{i=1}^n (\tilde{Z}_{ir} - Z_{ir}) \frac{E[Z_r \eta(T_r)]}{E[Z_r]} + O_p(\sqrt{n}). \end{aligned} \quad (A.1)$$

Proof. of Lemma A.1. Recall that $\hat{Z}_{ir} = \tilde{Z}_{ir} / \hat{\phi}_r(U_i)$, which is given by Section 2. Sample calculations give the following expression:

$$\sum_{i=1}^n (\hat{Z}_{ir} - \mathbf{Z}_{ir} \eta)(T_{ir}) = \sum_{i=1}^n \hat{Z}_{ir} \eta(T_{ir}) - \sum_{i=1}^n \mathbf{Z}_{ir} \eta(T_{ir}). \quad (\text{A.2})$$

Thus, it suffices to deal with the equation $\sum_{i=1}^n \mathbf{Z}_{ir} \eta(T_{ir})$. According to the type it holds that

$$\begin{aligned} \sum_{i=1}^n \hat{Z}_{ir} \eta(T_{ir}) &= \sum_{i=1}^n \hat{Z}_{ir} \eta(T_{ir}) \left(\frac{\hat{P}(U_i)}{\hat{g}_r(U_i)} \right) \bar{\bar{Z}}_r \\ &= \sum_{i=1}^n \bar{Z}_{ir} \eta(T_{ir}) \left(\frac{\hat{P}(U_i)}{\hat{g}_r(U_i)} \right) (U_i) E[\mathbf{Z}_r] + \sum_{i=1}^n \bar{Z}_{ir} \eta(T_{ir}) \left(\frac{\hat{P}(U_i)}{\hat{g}_r(U_i)} \right) (\bar{\bar{Z}}_r - E[\mathbf{Z}_r]) \\ &\triangleq L_1 + L_2. \end{aligned} \quad (\text{A.3})$$

First, we analyze L_1 . The proof is divided into three steps. Denote by L_{11} , L_{12} , and L_{13} the quantities:

$$\begin{aligned} L_{11} &= \sum_{i=1}^n \mathbf{Z}_{ir} \eta(T_{ir}), \\ L_{12} &= \sum_{i=1}^n \mathbf{Z}_{ir} \eta(T_{ir}) \frac{\hat{g}_r(U_i)}{g_r(U_i)}, \\ L_{13} &= \sum_{i=1}^n \mathbf{Z}_{ir} \eta(T_{ir}) \frac{\hat{P}(U_i)}{p(U_i)}. \end{aligned} \quad (\text{A.4})$$

Step 1: show that

$$L_1 = L_{11} + L_{12} + L_{13} + o_p(\sqrt{n}). \quad (\text{A.5})$$

Applying the equation $\phi_r(\cdot) = (g_r(\cdot)/p(\cdot)E(\mathbf{Z}_r))$, we have

$$\begin{aligned} L_1 &= \sum_{i=1}^n \mathbf{Z}_{ir} \eta(T_{ir}) \frac{g_r(U_i)}{P(U_i)} \cdot \frac{\hat{P}(U_i)}{\hat{g}_r(U_i)} \\ &= \sum_{i=1}^n \sum_{r=1}^q \mathbf{Z}_{ir} \eta(T_{ir}) \frac{g_r(U_i)}{p(U_i)} \cdot \frac{p(U_i)}{g_r(U_i)} \left(1 - \frac{\hat{g}_r(U_i)}{g_r(U_i)} + \frac{\hat{P}(U_i)}{p(U_i)} \right) + L_1^{k_1} - L_1^{k_2} \\ &= \sum_{i=1}^n \mathbf{Z}_{ir} \eta(T_{ir}) - \sum_{i=1}^n \mathbf{Z}_{ir} \eta(T_{ir}) \frac{\hat{g}_r(U_i)}{g_r(U_i)} + \sum_{i=1}^n \mathbf{Z}_{ir} \eta(T_{ir}) \frac{\hat{P}(U_i)}{p(U_i)} + L_1^{k_1} - L_1^{k_2} \\ &= L_{11} - L_{12} + L_{13} + L_1^{k_1} - L_1^{k_2}, \end{aligned} \quad (\text{A.6})$$

where

$$\begin{aligned} L_1^{k_1} &= \sum_{i=1}^n \mathbf{Z}_{ir} \eta(T_{ir}) \frac{(\hat{g}_r)(U_i) - g_r(U_i)}{g_r(U_i) - \hat{g}_r(U_i)}, \\ L_1^{k_2} &= \sum_{i=1}^n \mathbf{Z}_{ir} \eta(T_{ir}) \frac{(\hat{g}_r)(U_i) - g_r(U_i) (\hat{P}(U_i) - p(U_i))}{p(U_i) \hat{g}_r(U_i)}. \end{aligned} \quad (\text{A.7})$$

Equation (A.5) can be concluded by proven

$$L_1^{k_i} = o_p(n^{1/2}), \quad 1 \leq i \leq 2. \quad (\text{A.8})$$

According to Theorem 1 of [32], Lemma 3 of [23], condition (D2), and the Law of Large Numbers (LLN) for $(1/n) \sum_{i=1}^n \mathbf{Z}_{ir} \eta(T_{ir})$, we then follow the arguments used in [23] to yield the higher order term:

$$\begin{aligned} L_1^{k_1} &= O_p(ch^4) + n^{-1/2} h^{-1} \log n^2 \left\{ \sum_{i=1}^n \mathbf{Z}_{ri} \eta(T_{ri}) \right\} \\ &= o_p(n^{1/2}). \end{aligned} \quad (\text{A.9})$$

Similar argument yield (A.8) for $L_1^{k_2}$.

Step 2: show that

$$\begin{aligned} L_{12} &= \frac{1}{2} \sum_{i=1}^n \mathbf{Z}_{ir} \eta(T_{ir}) + \frac{1}{2} \sum_{i=1}^n \mathbf{Z}_{ri} \frac{E[\mathbf{Z}_r \eta(T_r)]}{E[\mathbf{Z}_r]} + o_p(n^{1/2}), \\ L_{13} &= \frac{1}{2} \sum_{i=1}^n \mathbf{Z}_{ir} \eta(T_{ir}) + \frac{1}{2} \sum_{i=1}^n E[\mathbf{Z}_r \eta(T_r)] + o_p(n^{1/2}). \end{aligned} \quad (\text{A.10})$$

To analyze L_{12} , for ease of understanding, we use $p(U)$ to express the density function of U and define as follows:

$$g_r(U) = E[\tilde{Z}_r | U] p(U), \quad 1 \leq r \leq p. \quad (\text{A.11})$$

We need to work on

$$\sum_{i=1}^n \mathbf{Z}_{ir} \eta(T_{ir}) \frac{\hat{g}_r(U_i)}{g_r(U_i)}. \quad (\text{A.12})$$

We can obtain the above sum by a U-statistic with a varying kernel with the bandwidth h by applying the similar arguments used by Davis and Fang [23] again. Then, we can have the asymptotic representation [4] as

$$\begin{aligned} \sum_{i=1}^n \mathbf{Z}_{ir} \eta(T_{ir}) \frac{\hat{g}_r(U_i)}{g_r(U_i)} &= \frac{1}{2} \sum_{i=1}^n \frac{1}{h} \int \mathbf{Z}_{ir} \eta(T_{ir}) \frac{1}{g_r(U_i)} K\left(\frac{U_i - u}{h}\right) \delta \phi_r(u) P_{r,u}(\delta, u) d\delta du \\ &\quad + \frac{1}{2} \sum_{i=1}^n \frac{1}{h} \int \mathbf{Z}_{ir} \eta(T_{ir}) \frac{1}{g_r(U_i)} K\left(\frac{U_i - u}{h}\right) j P_{J_r,u}(j, u) dj du + o_p(n^{1/2}), \end{aligned} \quad (\text{A.13})$$

with $P_{r,U}(\delta, U)$ being the density function of $(\mathbf{Z}_r, U)$, $J_r = \mathbf{Z}_r \eta(T_r)$ and $P_{J_r,U}(j, U)$ the density of variable (J_r, U) . Note that $\mathbf{Z}_r$ and U are independent and J_r and

U are independent. According to conditions (C), (D1)-(D2), and (E), we obtain

$$\begin{aligned} \sum_{i=1}^n \mathbf{Z}_{ir} \eta(T_{ir}) \frac{\hat{g}_r(U_i)}{g_r(U_i)} &= \frac{1}{2} \sum_{i=1}^n \mathbf{Z}_{ir} \eta(T_{ir}) \frac{E[\mathbf{Z}_r]}{g_r(U_i)} \frac{1}{h} \int K\left(\frac{U_i - u}{h}\right) \phi_r(u) P_U(u) du \\ &\quad + \frac{1}{2} \sum_{i=1}^n \mathbf{Z}_{ir} \phi_r(U_i) E[\mathbf{Z}_r \eta(T_r)] \frac{1}{h} \int \frac{P_U(u)}{g_r(u)} K\left(\frac{U_i - u}{h}\right) du + o_p(n^{1/2}) \\ &= \frac{1}{2} \sum_{i=1}^n \mathbf{Z}_{ir} \eta(T_{ir}) + \frac{1}{2} \sum_{i=1}^n \mathbf{Z}_{ir} \frac{E[\mathbf{Z}_r \eta(T_r)]}{E[\mathbf{Z}_r]} + o_p(n^{1/2}), \end{aligned} \quad (\text{A.14})$$

L_{13} can be similarly proceeded. Combining (A.5) and (A.8), we have (7):

$$\begin{aligned} L_1 &= \sum_{i=1}^n (\mathbf{Z}_{ir} \eta(T_{ir}) - E[\mathbf{Z}_r \eta(T_r)]) - \frac{1}{2} \sum_{i=1}^n (\mathbf{Z}_{ir} - E[\mathbf{Z}_r]) \frac{E[\mathbf{Z}_r \eta(T_r)]}{E[\mathbf{Z}_r]} \\ &\quad + n E[\mathbf{Z}_r \eta(T_r)] + o_p(n^{1/2}). \end{aligned} \quad (\text{A.15})$$

Step 3: show that

$$L_2 = \sum_{i=1}^n (\tilde{Z}_{ir} - E[\mathbf{Z}_r]) \frac{E[\mathbf{Z}_r \eta(T_r)]}{E[T]_r} + o_p(n^{1/2}). \quad (\text{A.16})$$

From (A.16) and the definition of L_2 in (A.3), we derive that

$$\begin{aligned}
 L_2 &= \frac{\bar{\bar{Z}}_r - E[\mathbf{Z}_r]}{E[\mathbf{Z}_r]} L_1 \\
 &= \sum_{i=1}^n [\mathbf{Z}_{ir} \eta(T_{ir}) - E[\mathbf{Z}_r \eta(T_r)]] \frac{\bar{\bar{Z}}_r - E[\mathbf{Z}_r]}{E[\mathbf{Z}_r]} \\
 &\quad - \frac{1}{2} \sum_{i=1}^n (\mathbf{Z}_{ir} - E[\mathbf{Z}_r]) \frac{E[\mathbf{Z}_r \eta(T_r)]}{E[\mathbf{Z}_r]} \frac{\bar{\bar{Z}}_r - E[\mathbf{Z}_r]}{E[\mathbf{Z}_r]} + nE[\mathbf{Z}_r \eta(T_r)] \frac{\bar{\bar{Z}}_r - E[\mathbf{Z}_r]}{E[\mathbf{Z}_r]} + o_p(n^{1/2}),
 \end{aligned} \tag{A.17}$$

where the last equality is obtained by applying LLN to $(1/n) \sum_{i=1}^n (\mathbf{Z}_{ir} \eta(T_{ir}) - E[\mathbf{Z}_r \eta(T_r)])$ and $(1/n) \sum_{i=1}^n (\mathbf{Z}_{ir} - E[\mathbf{Z}_r])$.

Finally, together with the asymptotic representation of L_1 and L_2 in (A.15) and (A.17), the desired result is easy to arrive at. $\square$

Lemma A.2. *Refine*

$$\begin{aligned}
 \tilde{A}(\beta, t) &= \sum_{i \in C} \int_0^t \left[\mathbf{Z}'_i \beta - \log \left\{ \sum_{l \in \tilde{C}} Y_l(U) \exp(\mathbf{Z}'_l \beta) \right\} \right] dN_i(U), \\
 \hat{A}(\beta, t) &= \sum_{i \in C} \int_0^t \left[\hat{\mathbf{Z}}'_i \beta - \log \left\{ \sum_{l \in \tilde{C}} Y_l(U) \exp(\hat{\mathbf{Z}}'_l \beta) \right\} \right] dN_i(U).
 \end{aligned} \tag{A.18}$$

Assume that conditions (C), (D1)-(D2), and (E) hold, then $\tilde{A}(\beta, t)$ converge to $\hat{A}(\beta, t)$ asymptotically.

Proof.

$$\begin{aligned}
 \hat{A}(\beta, t) - \tilde{A}(\beta, t) &= \sum_{i \in C} \int_0^t \left[(\hat{\mathbf{Z}}_i - \mathbf{Z}_i)' \beta - \log \frac{\{\sum_{l \in \tilde{C}} \tilde{Y}_l(U) \exp(\hat{\mathbf{Z}}'_l \beta)\}}{\{\sum_{l \in \tilde{C}} \tilde{Y}_l(U) \exp(\mathbf{Z}'_l \beta)\}} \right] dN_i(U) \\
 &= \int_0^t \left[\sum_{i \in C} (\hat{\mathbf{Z}}_i - \mathbf{Z}_i)' \beta - \sum_{i \in C} \log \frac{\{\sum_{l \in \tilde{C}} \tilde{Y}_l(U) \exp(\hat{\mathbf{Z}}'_l \beta)\}}{\{\sum_{l \in \tilde{C}} \tilde{Y}_l(U) \exp(\mathbf{Z}'_l \beta)\}} \right] dN_i(U).
 \end{aligned} \tag{A.19}$$

For $\sum_{i \in C} (\hat{\mathbf{Z}}_i - \mathbf{Z}_i) \beta$, according to Lemma A.1,

$$\begin{aligned}
 \sum_{i \in C} (\hat{\mathbf{Z}}_i - \mathbf{Z}_i) \beta &= \frac{1}{2} \sum_{i \in C} (\mathbf{Z}_i - E[\mathbf{Z}])' \frac{E[\mathbf{Z}' \beta]}{E[\mathbf{Z}']} \\
 &\quad + \sum_{i \in C} (\tilde{\mathbf{Z}}_i - \mathbf{Z}_i)' \frac{E[\mathbf{Z}' \beta]}{E[\mathbf{Z}']} + o_p(\sqrt{n}).
 \end{aligned} \tag{A.20}$$

According to the Law of Large Numbers, (A.20) converges to 0 in probability.

Next, we prove that

$$\sum_{i \in C} \log \frac{\{\sum_{l \in \tilde{C}} \tilde{Y}_l(U) \exp(\hat{\mathbf{Z}}'_l \beta)\}}{\{\sum_{l \in \tilde{C}} \tilde{Y}_l(U) \exp(\mathbf{Z}'_l \beta)\}} \xrightarrow{P} 0. \tag{A.21}$$

Only

$$\sum_{i \in \tilde{C}} Y_l(U) \exp(\hat{\mathbf{Z}}'_l \beta) - \sum_{i \in \tilde{C}} Y_l(U) \exp(\mathbf{Z}'_l \beta) \longrightarrow 0. \tag{A.22}$$

Let

$$\sum_{i \in \tilde{C}} Y_l(U) [\exp(\hat{\mathbf{Z}}'_l \beta) - \exp(\mathbf{Z}'_l \beta)] \longrightarrow 0. \tag{A.23}$$

Because $\exp(\cdot)$ is differentiable function, the first derivative of $\exp(\cdot)$, and it is obtained from (A.17) that

$$\exp(\tilde{Z}_i'\beta) - \exp(Z_i'\beta) \xrightarrow{L} 0. \quad (\text{A.24})$$

By (16) and (A.21), $\tilde{A}(\beta, t)$ asymptotically converges to $\tilde{A}(\beta, t)$.

Under the case-cohort sampling design, let $\tilde{A}$ denote the sample set of the subcolumn, $\tilde{n}$ denote the number of samples in $\tilde{A}$, A denotes the case-cohort sample, and the set of n represents the number of samples in A . Under the case-cohort sampling design, the covariate is not completely observed; then, its likelihood function cannot be expressed as (2). According to the pseudolikelihood function given in the article of Prentice [1], the pseudolikelihood function expression is as follows:

$$L_F(\beta) = \prod_{i \in A} \left[\frac{\exp(Z_i\beta^\top)}{\sum_{i \in \tilde{R}(t_i)} \exp(Z_i\beta^\top)} \right]^{\Delta_i}, \quad (\text{A.25})$$

where adventure set $\tilde{R}(t) = \{i: T_i \geq t, i \in \tilde{A} \cup B(t)\}$, $\{B(t) = i: N_i(t) \neq N_i(t-1), i = 1, 2, \dots, n\}$. Note that only when the individual fails at time t , the set $B(t)$ is nonempty, and the corresponding log-likelihood function is

$$l_F(\beta) = \sum_{i \in A} \Delta_i \left[Z_i\beta^\top - \log \sum_{i \in \tilde{R}(t_i)} \exp(Z_i\beta^\top) \right]. \quad (\text{A.26})$$

□

Proof of Theorem 1. To derive the asymptotic properties of the proposed estimator $\hat{\beta}$, we first introduce the following working likelihood function:

$$\hat{l}_F(\beta) = \sum_{i \in C} \int_0^\tau [\tilde{Z}_i'\beta - \log\{\sum_{i \in C} Y_i(t) \exp(\tilde{Z}_i'\beta)\}] dN_i(t), \quad (\text{A.27})$$

which differs from the pseudolikelihood function $\hat{l}_F(\beta)$ in (A.16) by using the index set $\tilde{C}$ instead of $\tilde{R}(t)$. The corresponding estimator based on $\hat{l}_F(\beta)$ is defined as

$$\tilde{\beta} = \arg \max_{\beta \in B} \hat{l}_F(\beta). \quad (\text{A.28})$$

Note the argument in [21]; we can prove that the estimator $\tilde{\beta}$ is asymptotically equivalent to the proposed estimator $\hat{\beta}$. That is, the asymptotic properties of $\tilde{\beta}$ can be obtained by proofing the desired asymptotic behaviors of $\hat{\beta}$.

Note that $\tilde{l}_F(\beta) = \tilde{A}(\beta, t)$. Mimicking the discussions in Lemma 1 of [21], we obtain that

$$\begin{aligned} \tilde{D}(\beta, t) &= n^{-1} \{ \tilde{A}(\beta, t) - \tilde{A}(\beta_0, t) \} \\ &= n^{-1} \{ \tilde{A}(\beta, t) - \tilde{A}(\beta, t) \} \\ &\quad + n^{-1} \{ \tilde{A}(\beta, t) - \tilde{A}(\beta_0, t) \}. \end{aligned} \quad (\text{A.29})$$

Define

$$\tilde{D}(\beta, t) = n^{-1} \{ \tilde{A}(\beta, t) - \tilde{A}(\beta_0, t) \}, \quad (\text{A.30})$$

where $\tilde{D}(\beta, t)$ converges in probability to the same limit as

$$D(\beta, t) = \int_0^t [(\beta - \beta_0)' S^{(1)}(\beta_0, U) - \log \left\{ \frac{S^{(0)}(\beta, U)}{S^{(0)}(\beta_0, U)} \right\} S^{(0)}(\beta_0, U)] \cdot \lambda_0(U) du, \quad (\text{A.31})$$

for $\beta \in \beta$. Therefore, $D(\beta, \tau)$ converges uniformly to

$$d(\beta, t) = \int_0^\tau [(\beta - \beta_0)' s^{(1)}(\beta_0, t) - \log \left\{ \frac{s^{(0)}(\beta, t)}{s^{(0)}(\beta_0, t)} \right\} s^{(0)}(\beta_0, t)] \cdot \lambda_0(t) dt, \quad (\text{A.32})$$

which is a continuous and convex function of β and has a unique maximum at β_0 [20, 21]. That is,

$$d(\beta) \leq d(\beta_0), \quad (\text{A.33})$$

with equality if and only if $\beta = \beta_0$.

We assume that $\tilde{\beta}$ does not converge to β_0 by a set of positive probability. So, there exists a subsequence $\{\tilde{\beta}_{in}\}$ of $\{\tilde{\beta}\}$ which converges to $\tilde{\beta}^*$ not equal to β_0 . Since $\{\tilde{\beta}_{in}\}$ is the maximum, we have $D(\tilde{\beta}_{in}, \tau) = D(\beta_0, \tau)$. We obtain the following inequality by the uniform convergency and continuity of limit:

$$d(\beta^*) \geq d(\beta_0), \quad \beta^* \neq \beta_0. \quad (\text{A.34})$$

This inequality is in contradiction with (A.12), so we have the convergency of $\tilde{\beta}$ to β_0 in probability.

From the basic assumptions I, II, and III, we derive that

$$\begin{aligned} E[E(\tilde{Z}_{ij}|U_i)] &= E[E(\phi_j(U_i)Z_{ij}|U_i)] \\ &= E[\phi_j(U_i)E(Z_{ij}|U_i)] \\ &= E[\phi_j(U_i)E(Z_{ij})] \\ &= E(\phi_j(U_i))Z_{ij} \\ &= Z_{ij}, \end{aligned} \quad (\text{A.35})$$

and from the above equation, we have

$$\begin{aligned} \tilde{Z}_i &= (\tilde{Z}_{i1}, \tilde{Z}_{i2}, \dots, \tilde{Z}_{ip})^T = \left(\frac{\tilde{Z}_{i1}}{\hat{\phi}_1(U_i)}, \frac{\tilde{Z}_{i2}}{\hat{\phi}_2(U_i)}, \dots, \frac{\tilde{Z}_{ip}}{\hat{\phi}_p(U_i)} \right)^T \\ &= \begin{pmatrix} \hat{\phi}_1^{-1}(U_i) & 0 & 0 & 0 & 0 & 0 \\ 0 & \hat{\phi}_2^{-1}(U_i) & 0 & 0 & 0 & 0 \\ 0 & 0 & . & 0 & 0 & 0 \\ 0 & 0 & 0 & . & 0 & 0 \\ 0 & 0 & 0 & 0 & . & 0 \\ 0 & 0 & 0 & 0 & 0 & \hat{\phi}_p^{-1}(U_i) \end{pmatrix} \begin{pmatrix} \tilde{Z}_{i1} \\ \tilde{Z}_{i2} \\ . \\ . \\ . \\ \tilde{Z}_{ip} \end{pmatrix}. \end{aligned} \quad (\text{A.36})$$

Under conditions (A1)-(A3), (B1)-(B4), (C), (D1)-(D2), and (E) and similar arguments in Self and Prentice [21], we obtain the following results:

$$-\frac{1}{n} \nabla_{\beta}^2 \hat{I}_F(\beta_0) \xrightarrow{L} \hat{\Sigma}(\beta_0), \frac{1}{\sqrt{n}} \nabla_{\beta} \hat{I}_F(\beta_0) \xrightarrow{L} N(0, \hat{\Sigma}(\beta_0) + \hat{\Sigma}_1(\beta_0)). \quad (\text{A.37})$$

Because $\tilde{\beta}$ converges to β_0 in probability, Slutsky's theorem, and (A.12), we can obtain

$$\sqrt{n}(\tilde{\beta} - t\beta_0) \xrightarrow{L} N(0, \hat{\Omega}(\beta_0)), \quad (\text{A.38})$$

where the asymptotic variance matrix

$$\hat{\Omega}(\beta_0) = \hat{\Sigma}(\beta_0) + \hat{\Sigma}_1(\beta_0), \quad (\text{A.39})$$

where

$$\hat{\Sigma}(\beta_0) = \int_0^\tau \phi^{-1} \gamma(\beta_0, t) s^{(0)}(\beta_0, t) \lambda_0(t) \phi^{-1} dt,$$

$$\hat{\Sigma}_1(\beta_0) = \int_0^\tau \int_0^\tau \hat{\Psi}(\beta_0, t, \omega) s^{(0)}(\beta_0, t) s^{(0)}(\beta_0, \omega) \lambda_0(t) \lambda_0(\omega) dt d\omega,$$

$$\begin{aligned} \hat{\Psi}(\beta_0, t, \omega) &= (1 - \alpha) \alpha^{-1} \{s^{(0)}(\beta_0, t) s^{(0)}(\beta_0, \omega)\}^{-1} \hat{\phi}^{-1} \\ &\quad \cdot \{q^{(2)}(\beta_0, t, \omega) - q^{(1)}(\beta_0, \omega, t) e(\beta_0, \omega)' - e(\beta_0, t) q^{(1)}(\beta_0, \omega)' + q^{(0)}(\beta_0, t, \omega) e(\beta_0, t) e(\beta_0, \omega)'\} \hat{\phi}^{-1}, \\ \hat{\phi}^{-1} &= \sum_{l=1}^{\tilde{n}} \phi_l^{-1}. \end{aligned} \quad (\text{A.40})$$

Data Availability

The simulation data used in this paper are randomly generated by R language software to support the proposed method of this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported in part by the Scientific Research Foundation of Shandong University of Science and Technology for Recruited Talents, Shandong University of Science and Technology (2019RCJJ021 to L. D.), and National Natural Science Foundation of China (71501114 to Q. L.).

References

- [1] R. L. Prentice, "A case-cohort design for epidemiologic cohort studies and disease prevention trials," *Biometrika*, vol. 73, no. 1, pp. 1–11, 1986.
- [2] F. Imamura, M. B. Schulze, S. J. Sharp et al., "Estimated substitution of tea or coffee for sugar-sweetened beverages was associated with lower type 2 Diabetes incidence in case-cohort analysis across 8 European countries in the EPIC-InterAct study," *The Journal of Nutrition*, vol. 149, no. 11, pp. 1985–1993, 2019.
- [3] C. Lassale, I. Tzoulaki, K. G. M. Moons et al., "Separate and combined associations of obesity and metabolic health with coronary heart disease: a pan-European case-cohort analysis," *European Heart Journal*, vol. 39, no. 5, pp. 397–406, 2018.
- [4] R. J. Serfling, "Approximation theorems of mathematical statistics," *Biometrics*, vol. 37, no. 4, 1980.
- [5] S. Yao, M. L. Kwan, I. J. Ergas et al., "Association of serum level of Vitamin D at diagnosis with breast cancer survival," *JAMA Oncology*, vol. 3, no. 3, pp. 351–357, 2017.
- [6] Z. Xu, D. P. Sandler, and J. A. Taylor, "Blood dna methylation and breast cancer: a prospective case-cohort analysis in the sister study," *JNCI: Journal of the National Cancer Institute*, vol. 112, no. 1, pp. 87–94, 2019.
- [7] P. Andronis, L. D. A. F. Amorim, and J. Cai, "Multiplicative rates model for recurrent events in case-cohort studies," *Lifetime Data Analysis*, vol. 26, no. 1, pp. 134–157, 2020.
- [8] E. T. Parner, P. K. Andersen, and M. Overgaard, "Cumulative risk regression in case-cohort studies using pseudo-observations," *Lifetime Data Analysis*, vol. 26, no. 4, pp. 639–658, 2020.
- [9] J. Rosman, P. Ter Wee, S. Meijer Truus et al., "Prospective randomised trial of early dietary protein restriction in chronic

□

- renal failure," *The Lancet*, vol. 324, no. 8415, pp. 1291–1296, 1984.
- [10] S. Klahr, A. S. Levey, G. J. Beck et al., "The effects of dietary protein restriction and blood pressure control on the progression of renal disease," *New England Journal of Medicine*, vol. 330, no. 13, pp. 877–884, 1994.
  - [11] D. Sentürk and H. Müller, "Covariate-adjusted regression," *Biometrika*, vol. 92, no. 1, pp. 75–89, 2005.
  - [12] X. Cui, W. Guo, L. Lin, and L. Zhu, "Covariate-adjusted nonlinear regression," *The Annals of Statistics*, vol. 37, no. 4, pp. 1839–1870, 2009.
  - [13] T. Mantopoulos, P. M. Mitchell, N. J. Welton, R. McManus, and L. Andronis, "Choice of statistical model for cost-effectiveness analysis and covariate adjustment: empirical application of prominent models and assessment of their results," *The European Journal of Health Economics*, vol. 17, no. 8, pp. 927–938, 2016.
  - [14] X. D. Huang, G. P. Wang, and M. M. Li, "Adjusting for high-dimensional covariates in sparse precision matrix estimation by lasso penalized D-trace loss," *Chinese Applied Probability Statistics*, vol. 35, no. 5, pp. 441–452, 2019.
  - [15] S. H. Zhang, J. Y. Zhu, and S. G. Zhang, "A powerful procedure for multiple outcomes comparison with covariate adjustment and its application to genomic data," *Journal of University of Chinese Academy of Sciences*, vol. 36, no. 2, pp. 155–161, 2019.
  - [16] J. Ding, G.-L. Tian, and K. C. Yuen, "A new MM algorithm for constrained estimation in the proportional hazards model," *Computational Statistics & Data Analysis*, vol. 84, pp. 135–151, 2015.
  - [17] L. Deng, J. Ding, Y. Liu, and C. Wei, "Regression analysis for the proportional hazards model with parameter constraints under case-cohort design," *Computational Statistics & Data Analysis*, vol. 117, pp. 194–206, 2018.
  - [18] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.
  - [19] G. K. Eagleson and H. G. Muller, "Transformations for smooth regression models with multiplicative errors," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 59, no. 1, pp. 173–189, 1997.
  - [20] P. K. Andersen and R. D. Gill, "Cox's regression model for counting processes: a large sample study," *The Annals of Statistics*, vol. 10, no. 4, pp. 1100–1120, 1982.
  - [21] S. G. Self and R. L. Prentice, "Asymptotic distribution theory and efficiency results for case-cohort studies," *The Annals of Statistics*, vol. 16, no. 1, pp. 64–81, 1988.
  - [22] D. Sentürk and H. Müller, "Inference for covariate adjusted regression via varying coefficient models," *The Annals of Statistics*, vol. 34, no. 2, pp. 654–679, 2006.
  - [23] L.-X. Davis and K.-T. Fang, "Asymptotics for kernel estimate of sliced inverse regression," *The Annals of Statistics*, vol. 24, no. 3, pp. 1053–1068, 1996.
  - [24] N. L. Hjort, "Bootstrapping cox's regression model," Technical report NSF-241, Department of Statistics, Stanford University, Stanford, CA, USA, 1985.
  - [25] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, New York, NY, USA, 1993.
  - [26] D. Burr, "A comparison of certain bootstrap confidence intervals in the cox model," *Journal of the American Statistical Association*, vol. 89, no. 428, pp. 1290–1302, 1994.
  - [27] R. L. Eubank, *Spline Smoothing and Nonparametric Regression*, Dekker, New York, NY, USA, 1988.
  - [28] D. Y. Lin, L. J. Wei, and Z. Ying, "Accelerated failure time models for counting processes," *Biometrika*, vol. 85, no. 3, pp. 605–618, 1998.
  - [29] D. Y. Lin and Z. Ying, "Semiparametric analysis of general additive-multiplicative hazard models for counting processes," *The Annals of Statistics*, vol. 23, no. 5, pp. 1712–1734, 1995.
  - [30] Y. Q. Chen and M.-C. Wang, "Analysis of accelerated hazards models," *Journal of the American Statistical Association*, vol. 95, no. 450, pp. 608–618, 2000.
  - [31] J. Cai and D. Zeng, "Power calculation for case-cohort studies with nonrare events," *Biometrics*, vol. 63, no. 4, pp. 1288–1295, 2007.
  - [32] C. R. Rao, *Linear Statistical Inference and its Applications*, Wiley, Hoboken, NJ, USA, 2nd edition, 1973.