<table>
<thead>
<tr>
<th><strong>Statistica Sinica Preprint No:</strong> SS-11-194R2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
</tr>
<tr>
<td><strong>Manuscript ID</strong></td>
</tr>
<tr>
<td><strong>URL</strong></td>
</tr>
<tr>
<td><strong>DOI</strong></td>
</tr>
<tr>
<td><strong>Complete List of Authors</strong></td>
</tr>
<tr>
<td><strong>Corresponding Author</strong></td>
</tr>
<tr>
<td><strong>E-mail</strong></td>
</tr>
<tr>
<td>Notice: Accepted version subject to English editing.</td>
</tr>
</tbody>
</table>
Maximum likelihood method for linear transformation models with cohort sampling data

BY YUAN YAO

Department of Mathematics, Hong Kong Baptist University,
Kowloon Tong, Kowloon, Hong Kong
yaoyuan@hkbu.edu.hk

Three widely used sampling designs—the nested case-control, case-cohort and classical case-control designs—can be categorized as generalized case-cohort designs. Maximum likelihood methods are used to perform regression analysis of linear transformation models with these sampling designs, and the resulting estimator is proved to be consistent, asymptotically normal and semiparametrically efficient. Simulation studies and an application to Stanford heart transplant data are presented.

Some key words: Linear transformation models; Maximum likelihood estimation; Missing at random; Nested case-control sampling.
1. Introduction

Cohort sampling is a popular methodology in epidemiological studies and clinical trials. When a time-to-failure response subject to censoring is involved, nested case-control design is a sampling method that is widely used. The main advantage of such a design, compared with prospective studies, is that collecting covariate information is relatively quick and inexpensive. A number of publications have addressed the analysis of nested case-control designs using the well-known Cox model. On the other hand, transformation models are among the most popular statistical models, and are particularly useful in analyzing time-to-event data. They include the Cox proportional hazards model and the proportional odds model as special cases. In the recent decades, considerable effort has been devoted to the statistical analysis of censored data using transformation models, leading to some recent breakthroughs. It is now timely to consider analyzing cohort sampling data using transformation models.

There is a vast literature on cohort designs, and in particular on nested case-control (n-c-c) designs (see, e.g., Breslow and Day (1980), Langholz and Goldstein (1996) and Breslow (1996)). The case-control studies of aircraft manufacturing employee data and the Colorado Plateau uranium miners are two real examples illustrating the popularity and importance of n-c-c studies. The regression model commonly used in the analysis of these data is the Cox proportional hazards model. There are typically two types of n-c-c data. Extended n-c-c data contain also the censoring times for the non-failures while time-restricted n-c-c data do not. The additional information in extended n-c-c data are less useful and may be less reliable, as Chen (2004) has discussed.

Thomas (1977) proposed a partial likelihood approach to estimate the regression parameter of Cox model. Because of its partial likelihood nature, his estimation method possesses useful properties similar to those of Cox’s partial likelihood estimation based on full cohort data. The estimator is easy to compute, and its variance estimator is simply negative of the derivative of the log-partial likelihood. The asymptotic properties of Thomas’ estimator can be formally established using the counting process martingale theory of Goldstein and Langholz (1992). However, Thomas’ estimator is not semiparametrically efficient, and it is likely that other more efficient and easy to compute estimates exist.
Samuelson (1997) presented an entirely new estimator based on maximizing pseudo-likelihood. His key observation is that the conditional probability of a censored subject’s ever being selected as a control can be explicitly computed with an expression similar to that of the Kaplan–Meier estimator. This observation leads to the construction of the pseudo-likelihood. The resulting estimator and its inference are also rather easy to compute. Empirical evidence shows that Samuelson’s estimator is sometimes better than Thomas’. As Samuelson (1997) has pointed out, however, the weighting techniques used to construct the pseudo-likelihood could be inefficient.

Chen (2001) has proposed an estimation method based on local averaging which is essentially an alternative sample reuse method other than those of Thomas and Samuelson. This method is applicable to so-called generalized case-cohort designs, which includes not only n-c-c, but case-control and case-cohort designs as well. Chen (2001) discusses how the accuracy of this estimator is comparable to those of Thomas and Samuelson, but the efficient estimator can be quite difficult to obtain, as it involves estimating and computing an integral operator. The estimators of Samuelson (1997) and Chen (2001) use extended n-c-c data, while Thomas (1977) and Chen’s later formulation (2004) use only time-restricted n-c-c data. The estimator of Chen’s study is based on certain optimal combinations of martingale residuals, and it has an asymptotic variance as small as or smaller than Thomas’ estimator.

There are also some other methods reported in the literature but not reviewed here. Almost all use Cox model. This poses serious limitation on practitioners because of the lack of statistical tools to analyze n-c-c or more general cohort designs. More general statistical methodologies for this purpose are badly needed. The transformation model, as a natural generalization of the Cox model, is an ideal choice. This summary will cover some recent developments on transformation models with full cohort data.

Typically a linear transformation model can be written as

$$\log H(T) = -\beta Z + \epsilon,$$

where \((T, Z)\) is the response-covariates pair, \(H\) is an unknown monotone function, \(\epsilon\) is the unobserved random variable whose distribution is known and \(\beta\) is the unknown regression parameter of main interest. When \(\epsilon\) follows the extreme value distribution or the standard
logistic distribution, the model reduces to the proportional hazards model or the proportional
odds model, respectively. The study from Cheng’s group (Cheng et al., 1995) contains a very
simple and elegant idea based on pairwise comparison of the survival times. The resulting
estimator is rather easy to compute and has a variance widely used (see Cheng et al., 1997
among many others). However, this type of rank based estimation method relies on the nec-
essary assumption that the censoring variables are independent of the covariates. Although
this assumption can be overcome or relaxed to some extent by stratification, stratification
may cause poor estimation of the distribution of the censoring variable within each stratum.
Nevertheless, this approach is still outstanding in terms of its simplicity.

The rest of the paper is organized as follows. We describe the maximum likelihood based
estimation method and asymptotic properties in Section 2. Simulation studies and an ap-
plication to real data are given in Section 3 and 4. Concluding remarks are given in Section
5. Technical details are provided in the Appendix.

2. Estimation and Inference

Consider a cohort of $n$ i.i.d. individuals. Let $T_i$ and $C_i$ denote the failure time and censoring
time of the $i$-th individual respectively, and assume that given covariate $Z_i$, $T_i$ and $C_i$ are
conditionally independent. Some notations can now be set as follows:

(i) the event time: $Y_i = \min(T_i, C_i)$;
(ii) the failure/censoring index: $\delta_i = I(T_i \leq C_i)$;
(iii) the indicator of the $i$-th individual being sampled for covariate ascertainment: $\Delta_i$, with
    conditional probability $\pi_i = P(\Delta_i = 1|(Y_j, \delta_j), 1 \leq j \leq n)$.

Furthermore, it is necessary in the inference procedure that $(\Delta_1, \cdots, \Delta_n)$ are condition-
ally independent of $(Z_1, \cdots, Z_n)$, given $(Y_j, \delta_j), 1 \leq j \leq n$.

Using the linear transformation model

$$\log H(T) = -\beta Z + \epsilon,$$  \hspace{1cm} (1)

where it is supposed that $\beta$ is the unknown parameter of main interest, $H$ is an unknown in-
creasing function, and $\epsilon$ is a continuous random variable whose hazard function $\lambda_\epsilon$ is known.
Also, it is assumed that \( \epsilon \) is independent of \( Z \) and \( C \).

Denote the distribution and probability density of covariate \( Z \) as \( F \) and \( f \). The likelihood function due to linear transformation model (1) can then be written as

\[
L(\beta, H, F) = \prod_{i=1}^{n} f(Y_i, \delta_i) \times f(\Delta_1, \cdots, \Delta_n|(Y_j, \delta_j), 1 \leq j \leq n)
\]

\[
\times \prod_{i=1}^{n} f(Z_i|(Y_j, \delta_j, \Delta_j), 1 \leq j \leq n)\Delta_i
\]

\[
= \prod_{i=1}^{n} f(Y_i, \delta_i) \times f(\Delta_1, \cdots, \Delta_n|(Y_j, \delta_j), 1 \leq j \leq n)
\]

\[
\times \prod_{i=1}^{n} f(Z_i|Y_i, \delta_i)\Delta_i,
\]

where the second equation comes from the conditional independence of \((\Delta_1, \cdots, \Delta_n)\) and \((Z_1, \cdots, Z_n)\), and independence between individuals.

Note that the last term in the likelihood function above can be calculated as

\[
f(Z_i|Y_i, \delta_i)\Delta_i = \left( \frac{f(Y_i, \delta_i|Z_i)f(Z_i)}{f(Y_i, \delta_i)} \right)^{\Delta_i}.
\]

For simplicity, denote the hazard and cumulative hazard functions of \( \exp(\epsilon) \) as \( \lambda \) and \( \Lambda \), respectively. The log-likelihood then takes the final form

\[
l_n(\beta, H, F) = \sum_{i=1}^{n} \Delta_i \left\{ \delta_i [\beta Z_i + \log \lambda(H(Y_i)e^{\beta Z_i})] - \Lambda(H(Y_i)e^{\beta Z_i}) \right\}
\]

\[
+ (1 - \delta_i) \log \left\{ \int [e^{\beta Z}\lambda(H(Y_i)e^{\beta Z})]^\delta e^{-\Lambda(H(Y_i)e^{\beta Z})} f(Z) dZ \right\}
\]

\[
+ \delta_i \log h(Y_i) + \Delta_i \log f(Z_i)
\]

\[
+ \Delta_i \log(\lambda C(Y_i)^{1-\delta_i} e^{-\Lambda C(Y_i)}) + \log f(\Delta_1, \cdots, \Delta_n|(Y_j, \delta_j), 1 \leq j \leq n),
\]

where \( h(\cdot) \) is the derivative function of \( H(\cdot) \).

Using the method of discretization of \( H \) and \( F \), let \( q_j \) represent the size of the increment of \( H \) at the \( j \)-th smallest observed failure times, say \( s_j, j = 1, \ldots, n_1 \), where \( n_1 = \sum_{i=1}^{n} \delta_i \) is the number of failures. Set

\[
H(t) = \sum_{j=1}^{n_1} q_j I(s_j \leq t), \quad \text{and} \quad h(t) = \sum_{j=1}^{n_1} q_j I(t = s_j).
\]
To estimate $F$, put probability mass on all known covariates, i.e., set

$$f(Z_j) = P(Z = Z_j) = p_j \quad \text{satisfying } \sum_{j=1}^{n_2} p_j = 1,$$

where $n_2 = \sum_{i=1}^{n} \Delta_i$ is the number of individuals with known covariates. Then the integral over $Z$ in the likelihood can be estimated as

$$\int [e^{\beta Z} \lambda(H(Y_i)e^{\beta Z})]^{1/2} e^{-\Lambda(H(Y_i)e^{\beta Z})} f(Z) dZ = \sum_{j=1}^{n_2} [e^{\beta Z_j} \lambda(H(Y_i)e^{\beta Z_j})]^{1/2} e^{-\Lambda(H(Y_i)e^{\beta Z_j})} p_j.$$

We show that maximizing the log-likelihood function over $(\beta, q_1, \cdots, q_{n_1}, p_1, \cdots, p_{n_2})$ leads to its being consistent, asymptotically normal and semiparametrically efficient for $\beta$ under certain regularity conditions in the following theorems. Regularity conditions and the proof of the theorems are shown in the Appendix.

**Theorem 1.** Under the regularity conditions (C1)-(C5), $|\hat{\beta}_n - \beta_0| \rightarrow 0$, $\sup_{t \in [0,\tau]} |\hat{H}_n(t) - H_0(t)| \rightarrow 0$ and $\sup_{Z \in M} |\hat{F}_n(Z) - F_0(Z)| \rightarrow 0$ almost surely.

To describe the variance estimation, let $\tau$ denote the duration of the study and suppose $Z$ lies in a bounded set $M$. Denote the set of functions $Q_1 = \{ p \in BV[0,\tau] : |p| \leq 1 \}$ and $Q_2 = \{ q \in BV[M] : |q| \leq 1 \}$, where $BV[D]$ is the set of functions on $D$ with bounded total variation. Then $\hat{H}_n$ can be treated as a bounded linear functional in $L^\infty(Q_1)$ by definition

$$\hat{H}_n(p) = \int_0^\tau p(t)d\hat{H}_n(t),$$

and similarly, $\hat{F}_n$ can be treated as a bounded linear functional in $L^\infty(Q_2)$.

**Theorem 2.** Under the conditions (C1)-(C6), $\sqrt{n}(\hat{\beta}_n - \beta_0, \hat{H}_n - H_0, \hat{F}_n - F_0)$ converges weakly to a zero-mean Gaussian process in the metric space $\mathbb{R}^d \times L^\infty(Q_1) \times L^\infty(Q_2)$. And the limiting covariance matrix of $\sqrt{n}(\hat{\beta}_n - \beta_0)$ attains the semiparametric efficiency bound.

**Theorem 3.** For any $(b, p, q) \in \mathcal{V} \times Q_1 \times Q_2$, where $\mathcal{V} = \{ v \in \mathbb{R}^d : |v| \leq 1 \}$. The asymptotic variance for

$$\sqrt{n}v^T(\hat{\beta}_n - \beta_0) + \sqrt{n} \int_0^\tau p(t)d[\hat{H}_n(t) - H_0(t)] + \sqrt{n} \int Z q(Z)d[\hat{F}_n(Z) - F_0(Z)]$$

can be consistently estimated by $(v^T, p^T, q^T)I_n^{-1}(v^T, p^T, q^T)^T$, where we denote the negative Hessian matrix of the log-likelihood function $l_n(\beta, H, F)$ with respect to $(\beta, q_1, \cdots, q_{n_1}, p_1, \cdots, p_{n_2})$.
as $nI_n$, and the vectors $\tilde{p} = (p(s_1), \cdots, p(s_n))^T, \tilde{q} = (q(Z_1), \cdots, q(Z_n))^T$. Therefore, by taking $p = 0$ and $q = 0$, the variance matrix of $\sqrt{n}(\hat{\beta}_n - \beta_0)$ can be estimated by the upper left $d \times d$ matrix of $I_n^{-1}$.

**Remark.** The computation can be carried out in many scientific computing packages. For example, the algorithm of `fmincon` in the optimization toolbox of Matlab can be used to find a minimizer and calculate the Hessian matrix as well. Our simulation shows that this algorithm performs good when handling moderate sample size like 200. For the initial value, one may use the estimates from the Cox model or try different initial values to make the maximization guaranteed.

### 3. Simulation study

Let us generate two independent covariates: $Z_1$ is uniform distributed over $[0, 1]$ and $Z_2$ follows Bernoulli distribution with success probability 0.5. The hazard function of the error term $\epsilon$ is chosen as

$$
\lambda(t) = \frac{\exp(t)}{1 + r \exp(t)},
$$

with $r = 0, 1$ and 2 (Dabrowska and Doksum, 1988; Chen et al., 2002). Note that the proportional hazards and proportional odds models correspond to $r = 0$ and $r = 1$, respectively. For the transformation function in the model

$$
\log H(t) = -\beta Z + \epsilon,
$$

we take $H(t) = t$ for $r = 0$, $H(t) = \exp(t) - 1$ for $r = 1$ and $H(t) = 0.5 \exp(2t) - 0.5$ for $r = 2$. The censoring time $C$ is generated from a uniform distribution on $[0, c]$, where $c$ is chosen such that the expected proportion of censoring is 0.7 and 0.8, respectively. The sample size $n$ is set to be 200 and all simulations are based on 500 replications. For the case-cohort design, we select all failures and a subcohort of size 85 and 50 when censoring rate is 0.7 and 0.8, respectively. For the classical case-control design, we select all failures and a group of non-failures the same amount as the failures. For the nested case-control design, we select all failures and 2 non-failures in each risk set of failure times.

Table 1 summarizes the simulation results estimating $\beta_1$ and $\beta_2$. The true values are set as $\beta_1 = 1$ and $\beta_2 = -1$. Results include the mean of the bias (Bias) of the estimates,
the sample standard deviations (SSD) of the estimates, the mean of the estimated standard errors (ESE) of the estimates, and the 95% empirical coverage probabilities (CP) for \(\beta_1\) and \(\beta_2\) based on the asymptotic normality in Theorem 2. The proposed estimation procedures perform well in all cases.

We compare our approach with that of Chen et al. (2012), which is a likelihood based method involving inverse probability that can apply to different cohort sampling schemes. A series of simulation with the same settings as above is conducted and it shows that the Chen et al. estimators are less efficient among most of the scenarios, especially when censoring rate is large. This is mainly because the Chen et al. estimator does not involve the event times whose covariates are not sampled. Therefore it usually cannot produce efficient estimation. This can be seen in Table 1 by comparing the relative efficiency of our estimators (RE) with that of the Chen et al. estimators (RE*), where the relative efficiency is computed by comparing the sample variance of an estimator to the sample variance of the full-cohort estimator.

4. Application

The Stanford heart transplant data, consisting of censored or uncensored survival times in February 1980 of 184 patients who had received heart transplants, was reported in Miller and Halpern (1982). In the data set, the patients’ T5 mismatch score is a measure of the degree of tissue incompatibility between the initial donor and the recipient with respect to HLA antigens. The goal of this study is to analyze the relationship among the survival time, T5 mismatch score and the age of the patients.

In this example, we compare the results using the full cohort data with those from different sampling schemes. Like Miller and Halpern (1982) and Jin et al. (2006), only the 157 patients with complete records are used. Following Miller and Halpern (1982), in model 1 we regress the logarithm of the survival time against the ages and T5 mismatch scores for the 157 patients. Then the T5 mismatch score is deleted due to its nonsignificance in model 1 and then a quadratic age model is tried to achieve better fit. In model 2 we regress the logarithm of the survival time against the age and age\(^2\) for the 152 patients whose survival times were not less than 10 days.
The sampling designs for both models are set as follows: For the case-cohort design we select all failures and a subcohort of size 75 among the whole set of subjects. For the classical case-control design we select all failures and 25 non-failures. For the nested case-control design we select all failures and one non-failure in each risk set.

Table 2 shows the average of the estimates of parameters with standard errors and p-values based on 400 replications. The proposed method under all the sampling designs leads to the same conclusion as using the full data set. For comparison, we also present the estimates from Chen et al. (2012) in the table, from which one can draw similar conclusion in general. It can also be seen from the comparison of standard error that the proposed method has the advantage of being asymptotically efficient.

5. Concluding remarks

This paper presents an efficient estimation technique for a broad class of sampling designs using linear transformation models. The computation procedure is based on the maximization of the discretized likelihood function and the variance estimation is obtained from the inverse of the negative Hessian matrix. We prove that the proposed estimator attains the semiparametric efficiency bound. The performance of the estimator is illustrated with simulation studies and Stanford heart transplant data study. Further work will consider extending this method to partially linear transformation models, where kernel smoothing could be applied.

ACKNOWLEDGEMENT

The author is grateful to the referee and associate editor for their constructive comments which greatly improved this manuscript. The author also deeply thanks Professor Kani Chen for his guidance during the study.

APPENDIX: Regularity conditions and proofs of theorems

Assume that the following regularity conditions hold:
The function $H_0(t)$ and the distribution $F_0(t)$ are strictly increasing and differentiable, with derivatives $h(t)$ and $f(t)$ which are absolutely continuous, and $\beta_0$ lies in the interior of a known compact set in the domain of $B$.

$\lambda(t) > 0$, and $P(C \geq \tau|Z) > \delta > 0$ for some constant $\delta$ with probability one.

The covariate $z$ is bounded, and $\lambda$ is twice continuously differentiable.

$$\limsup_{x \to \infty} \Lambda(C_0x)^{-1}\log(x \sup_{y \leq x} \lambda(y)) = 0$$ holds for every $C_0 > 0$.

(First Identifiability) Denote

$$\Psi_j(\beta, H, F) = \left\{ \left[ e^{\beta Z_i} \lambda(H(Y_i))e^{\beta Z_i} \right]^{\delta_i} e^{-\Lambda(h(Y_i)e^{\beta Z_i})} \right\}^{\Delta_i} \times \left\{ \int [e^{\beta Z} \lambda(H(Y_i)e^{\beta Z})]^{\delta_i} e^{-\Lambda(h(Y_i)e^{\beta Z})} f(Z) dZ \right\}^{(1-\Delta_i)}.$$

If

$$\Psi_j(\beta^*, H^*, F^*)h^*(Y_j)^{\delta_j}f^*(Z_j)^{\Delta_j} = \Psi_j(\beta_0, H_0, F_0)h_0(Y_j)^{\delta_j}f_0(Z_j)^{\Delta_j},$$

almost surely, then $\beta^* = \beta_0$, $H^* = H_0$ and $F^* = F_0$.

(Second Identifiability) If

$$v^T l_\beta(\beta_0, H_0, F_0) + l_H(\beta_0, H_0, F_0) \left[ \int p dH_0 \right] + l_F(\beta_0, H_0, F_0) \left[ \int q dF_0 \right] = 0$$

almost surely for some $v \in \mathbb{R}^d$, $p \in BV[0, \tau]$ and $q \in BV[M]$, then $(v, p, q) = 0$, where $v^T l_\beta$, $l_H[g_1]$ and $l_F[g_2]$ denote the partial derivatives of $l$ along direction of $v$, $g_1$ and $g_2$, respectively.

Remark. Conditions (C1)-(C3) assume certain smoothness and identifiability of the model, which are standard requirements in censored data analysis. (C4) is a technical condition on the structure of the model that is used in the proof of consistency. (C5) is the usual parameter identifiability condition. (C6) ensures that the Fisher information is non-singular.

Proof of Theorem 1. Firstly, the jump size of $\hat{H}_n$ must be finite, otherwise the log-likelihood would diverge to $-\infty$ by condition (C4). Secondly, $\hat{H}_n$ is also bounded almost surely, otherwise if a new estimator $\hat{H}_n = \hat{H}_n/\hat{H}_n(\tau)$ were considered, it would contradict the maximum property of $\hat{H}_n$. Since $\hat{H}_n$ is uniformly bounded and monotone, Helly’s selection theorem requires that for any subsequence of $\{\hat{H}_n\}$, there will always be a further...
subsequence which converges to some monotone function $H^*$ pointwise. Without loss of generality, assume that $\hat{F}_n$ converges to $F^*$ and $\hat{\beta}_n$ converges to $\beta^*$ for the same subsequence. Then consistency will be proved if we can show that $H^* = H_0$, $F^* = F_0$ and $\beta^* = \beta_0$ with probability one. Furthermore, the continuity of $H_0$ and $F_0$ ensures that the convergence is uniform in $t$ and $Z$.

Take the derivative of the log-likelihood with respect to $H$ along $H + \epsilon I(\cdot \geq Y_i)$, and let it be zero. Then

$$\hat{h}(Y_i) = -\frac{\delta_i}{\sum_{j=1}^{n} \frac{\Psi_j H(\hat{\beta}_n, H_n, \hat{F}_n)[I(\cdot \geq Y_i)]}{\Psi_j(\beta_0, H_0, F_0)}}.$$ 

Hence

$$\hat{H}_n(t) = \sum_{i=1}^{n} \int_0^t h(u) dN_i(u) = -\sum_{i=1}^{n} \int_0^t \sum_{j=1}^{n} \frac{dN_i(u)}{\Psi_j H(\hat{\beta}_n, H_n, \hat{F}_n)[I(\cdot \geq u)]},$$

where we use $N_i(u)$ to denote the counting process of $Y_i$.

Now define

$$\tilde{H}_n(t) = -\sum_{i=1}^{n} \int_0^t \frac{dN_i(u)}{\sum_{j=1}^{n} \frac{\Psi_j H(\beta_0, H_0, F_0)[I(\cdot \geq u)]}{\Psi_j(\beta_0, H_0, F_0)}}.$$ 

The Glivenko-Cantelli Theorem leads to

$$\lim_{n \to \infty} \tilde{H}_n(t) = H_0(t).$$

Note that $\tilde{H}_n(t)$ can be written as

$$\tilde{H}_n(t) = \int_0^t \frac{1}{n} \sum_{j=1}^{n} \frac{\Psi_j H(\beta_0, H_0, F_0)[I(\cdot \geq u)]}{\Psi_j(\beta_0, H_0, F_0)} d\tilde{H}_n(u).$$

(1)

It is now possible to show that $H^*$ is continuously differentiable. For simplicity, define

$$E\left(\frac{\Psi_j H(\beta_0, H_0, F_0)[I(\cdot \geq u)]}{\Psi_j(\beta_0, H_0, F_0)}\right) \text{ as } g_{1j}(u) \text{ and } E\left(\frac{\Psi_j H^*(\beta^*, H^*, F^*)[I(\cdot \geq u)]}{\Psi_j(\beta^*, H^*, F^*)}\right) \text{ as } g_{2j}(u).$$

It can be shown that they are both bounded from zero.
Taking the limit of (1), we get
\[ H^*(t) = \int_0^t g_{1j}(u) g_{2j}(u) dH_0(u). \]
We have now shown that \( H^*(t) \) is absolutely continuous with respect to \( H_0(t) \). By assumption, \( H_0(t) \) is continuously differentiable, and so is \( H^*(t) \). Then
\[ \lim_{n \to \infty} \frac{d\hat{H}_n(t)}{dH_n(t)} = \frac{h^*(t)}{h_0(t)} \]
uniformly in \( t \in [0, \tau] \), where \( h^* \) is the derivative of \( H^* \).

Repeating the same process with \( F(t) \) similarly yields
\[ \lim_{n \to \infty} \frac{d\hat{F}_n(Z)}{d\tilde{F}_n(Z)} = \frac{f^*(Z)}{f_0(Z)} \]
uniformly.

It follows from the inequality \( l_n(\hat{\beta}_n, \hat{H}_n, \hat{F}_n) \geq l_n(\beta_0, \tilde{H}_n, \tilde{F}_n) \) that
\[ \frac{1}{n} \sum_{j=1}^n \log \frac{\Psi_j(\hat{\beta}_n, \hat{H}_n, \hat{F}_n)}{\Psi_j(\beta_0, \tilde{H}_n, \tilde{F}_n)} + \frac{1}{n} \sum_{j=1}^n \delta_j \log \frac{d\hat{H}_n(Y_j)}{dH_n(Y_j)} + \frac{1}{n} \sum_{j=1}^n \Delta_j \log \frac{d\hat{F}_n(Z_j)}{d\tilde{F}_n(Z_j)} \geq 0. \]
Let \( n \to \infty \), then
\[ E \left( \log \frac{\Psi_j(\beta^*, H^*, F^*)h^*(Y_j)\delta_j f^*(Z_j)\Delta_j}{\Psi_j(\beta_0, H_0, F_0)h_0(Y_j)\delta_j f_0(Z_j)\Delta_j} \right) \geq 0. \]
Note that the left-hand side is the negative Kullback-Leibler distance, therefore condition C5 requires that \( \beta^* = \beta_0, H^* = H_0 \) and \( F^* = F_0 \) with probability one.

**Proof of Theorem 2.** This proof is based on the argument on maximum likelihood estimators by Van der Vaart (1998, pp. 419-424). Let
\[ \mathcal{L}(\beta, H, F) = \log \Psi_j + \delta_j \log h(Y_j) + \Delta_j \log f(Z_j), \]
\[ \Phi_n(\beta, H, F) = \mathcal{P}_n \left\{ v^T \mathcal{L}_\beta + \mathcal{L}_H \left[ \int p dH \right] + \mathcal{L}_F \left[ \int q dF \right] \right\} \]
and
\[
\Phi(\beta, H, F) = \mathcal{P}\left\{ v^T \mathcal{L}_\beta + \mathcal{L}_H\left[ \int pdH \right] + \mathcal{L}_F\left[ \int qdF \right] \right\},
\]
where we use \( v^T \mathcal{L}_\beta \), \( \mathcal{L}_H[g_1] \) and \( \mathcal{L}_F[g_2] \) denote the partial derivatives of \( \mathcal{L} \) along direction of \( \beta + \epsilon v \), \( H + \epsilon g_1 \) and \( F + \epsilon g_2 \) respectively as above, and let \( \mathcal{P}_n \) denote the empirical measure based on \( n \) i.i.d. observations and \( \mathcal{P} \) be its expectation.

Note that for any \( \delta > 0 \), when \( n \) large enough,
\[
(\hat{\beta}_n, \hat{H}_n, \hat{F}_n) \in N_0 = \left\{ (\beta, H, F_2) : |\beta - \beta_0| + |H - H_0| + |F_2 - F_{20}| < \delta_0 \right\}
\]
almost surely. By the Donsker theorem,
\[
\sqrt{n}(\Phi_n - \Phi)(\hat{\beta}_n, \hat{H}_n, \hat{F}_n) - \sqrt{n}(\Phi_n - \Phi)(\beta_0, H_0, F_0) = o_p(1).
\]

Direct calculations show
\[
\sqrt{n}(\mathcal{P}_n - \mathcal{P})\left\{ v^T \mathcal{L}_\beta(\beta_0, H_0, F_0) + \mathcal{L}_H(\beta_0, H_0, F_0)\left[ \int pdH_0 \right] + \mathcal{L}_F(\beta_0, H_0, F_0)\left[ \int qdF_0 \right] \right\} = \sqrt{n}(\mathcal{P}_n - \mathcal{P})\left\{ v^T \mathcal{L}_\beta(\hat{\beta}_n, \hat{H}_n, \hat{F}_n) + \mathcal{L}_H(\hat{\beta}_n, \hat{H}_n, \hat{F}_n)\left[ \int pd\hat{H} \right] + \mathcal{L}_F(\hat{\beta}_n, \hat{H}_n, \hat{F}_n)\left[ \int qd\hat{F} \right] \right\} + o_p(1)
\]
\[
= -\sqrt{n}\mathcal{P}\left\{ v^T \mathcal{L}_\beta(\beta_0, H_0, F_0) - v^T \mathcal{L}_\beta(\hat{\beta}_n, \hat{H}_n, \hat{F}_n) + \mathcal{L}_H(\beta_0, H_0, F_0)\left[ \int pdH_0 \right] + \mathcal{L}_H(\hat{\beta}_n, \hat{H}_n, \hat{F}_n)\left[ \int pd\hat{H} \right] - \mathcal{L}_F(\beta_0, H_0, F_0)\left[ \int qdF_0 \right] + \mathcal{L}_F(\hat{\beta}_n, \hat{H}_n, \hat{F}_n)\left[ \int qd\hat{F} \right] \right\}
\]
\[
= -\sqrt{n}\mathcal{P}\left\{ v^T \mathcal{L}_\beta(\hat{\beta}_n, \hat{H}_n, \hat{F}_n) - v^T \mathcal{L}_\beta(\beta_0, H_0, F_0) \right\} + \frac{\Psi_jH(\beta_0, H_0, F_0)[\int pdH_0]}{\Psi_j(\beta_0, H_0, F_0)} - \frac{\Psi_jH(\hat{\beta}_n, \hat{H}_n, \hat{F}_n)[\int pd\hat{H}]}{\Psi_j(\hat{\beta}_n, \hat{H}_n, \hat{F}_n)} + \frac{\Psi_jF(\beta_0, H_0, F_0)[\int qdF_0]}{\Psi_j(\beta_0, H_0, F_0)} - \frac{\Psi_jF(\hat{\beta}_n, \hat{H}_n, \hat{F}_n)[\int qd\hat{F}]}{\Psi_j(\hat{\beta}_n, \hat{H}_n, \hat{F}_n)}. \hspace{1cm} (2)
\]

For the continuous linear functional from \( BV[0, \tau] \) to \( \mathbb{R} \):
\[
p \mapsto \mathcal{P}\left( \frac{\Psi_jH[p]}{\Psi_j}(\beta, H, F) \right),
\]
by theorem 4.2 in Edwards and Wayment (1971), there exists a bounded function \( \eta_H \) such that
\[
\mathcal{P}\left( \frac{\Psi_jH[p]}{\Psi_j}(\beta, H, F) \right) = \int_0^\tau \eta_H(t; \beta, H, F)dp(t).
\]
To be specific, let \( p(s) = I(s \geq t) \), then we have

\[
\eta_H(t; \beta, H, F) = \mathcal{P}\left(\frac{\Psi_j H[I(\cdot \geq t)](\beta, H, F)}{\Psi_j}\right).
\]

Follow the same procedure, we define

\[
\eta_F(z; \beta, H, F) = \mathcal{P}\left(\frac{\Psi_j F[I(\cdot \geq z)](\beta, H, F)}{\Psi_j}\right),
\]

and second order partial derivatives functions:

\[
\eta_{H\beta}(t; \beta, H, F) = \frac{\partial}{\partial \beta} \eta_H(t; \beta, H, F),
\]

\[
\eta_{F\beta}(t; \beta, H, F) = \frac{\partial}{\partial \beta} \eta_F(t; \beta, H, F),
\]

\[
\eta_{HH}(s, t; \beta, H, F) = \frac{\partial}{\partial H} \eta_H(s; \beta, H, F)[I(\cdot \geq t)],
\]

\[
\eta_{HF}(s, z; \beta, H, F) = \frac{\partial}{\partial F} \eta_H(s; \beta, H, F)[I(\cdot \geq z)],
\]

\[
\eta_{FH}(x, t; \beta, H, F) = \frac{\partial}{\partial H} \eta_F(x; \beta, H, F)[I(\cdot \geq t)]
\]

and

\[
\eta_{FF}(x, z; \beta, H, F) = \frac{\partial}{\partial F} \eta_F(x; \beta, H, F)[I(\cdot \geq z)].
\]

We also denote the following metrics and functions:

\[
\zeta_{\beta}(\beta, H, F) = \frac{\partial}{\partial \beta} \mathcal{P}\left(\frac{\Psi_j \beta}{\Psi_j}\right)(\beta, H, F),
\]

\[
\zeta_H(t; \beta, H, F) = \frac{\partial}{\partial H} \mathcal{P}\left(\frac{\Psi_j \beta}{\Psi_j}\right)(\beta, H, F)[I(\cdot \geq t)]
\]

and

\[
\zeta_F(z; \beta, H, F) = \frac{\partial}{\partial F} \mathcal{P}\left(\frac{\Psi_j \beta}{\Psi_j}\right)(\beta, H, F)[I(\cdot \geq z)].
\]

With these notations, the right-hand side of (2) is equal to

\[
= -\sqrt{n}\left(B_1[v, p, q]^T(\hat{\beta}_n - \beta_0) + \int B_{21}[v, p, q]d(\hat{H}_n - H_0) + \int B_{22}[v, p, q]d(\hat{F}_n - F_0)\right) + o\left(\sqrt{n}|\hat{\beta}_n - \beta_0| + \sqrt{n}|\hat{H}_n - H_0| + \sqrt{n}|\hat{F}_n - F_0|\right),
\]
where linear operators $B_1$, $B_{21}$, $B_{22}$ are defined as

\[ B_1[v, p, q] = v^T \zeta_\beta(\beta_0, H_0, F_0) + \int_0^T \eta_{H\beta}(t; \beta_0, H_0, F_0)p(t)dH_0(t) \]

\[ + \int_M \eta_{F\beta}(z; \beta_0, H_0, F_0)q(z)dF_0(z); \]

\[ B_{21}[v, p, q] = v^T \zeta_H(t; \beta_0, H_0, F_0) + \eta_H(t; \beta_0, H_0, F_0)p(t) \]

\[ + \int_0^T \eta_{HH}(s,t; \beta_0, H_0, F_0)p(s)dH_0(s) + \int_M \eta_{FH}(x,t; \beta_0, H_0, F_0)q(x)dF_0(x); \]

\[ B_{22}[v, p, q] = v^T \zeta_F(z; \beta_0, H_0, F_0) + \eta_F(z; \beta_0, H_0, F_0)q(z) \]

\[ + \int_0^T \eta_{HF}(s,z; \beta_0, H_0, F_0)p(s)dH_0(s) + \int_M \eta_{FF}(x,z; \beta_0, H_0, F_0)q(x)dF_0(x). \]

In other word, the operator $[B_1, B_{21}, B_{22}]$ can be written as

\[ B_1[v, p, q] = \frac{d}{d\epsilon} \bigg|_{\epsilon=0} \mathcal{P} \left\{ v^T \mathcal{L}_\beta(\beta_0 + \epsilon \hat{v}, H_0 + \epsilon \int \hat{p}dH_0, F_0 + \epsilon \int \hat{q}dF_0) \right. \]

\[ + \mathcal{L}_H(\beta_0 + \epsilon \hat{v}, H_0 + \epsilon \int \hat{p}dH_0, F_0 + \epsilon \int \hat{q}dF_0) \left[ \int \hat{p}dH_0 \right] \]

\[ + \mathcal{L}_F(\beta_0 + \epsilon \hat{v}, H_0 + \epsilon \int \hat{p}dH_0, F_0 + \epsilon \int \hat{q}dF_0) \left[ \int \hat{q}dF_0 \right] \bigg\}. \]

We now show that $(B_1, B_{21}, B_{22})$ is continuously invertible. By the open mapping theorem, we need only to prove that the linear operator is one-to-one.

To show $(B_1, B_{21}, B_{22})$ is injective, suppose that $B_1[v, p, q] = 0$, $B_{21}[v, p, q] = 0$ and $B_{22}[v, p, q] = 0$. Then

\[ 0 = B_1[v, p, q]^Tv + \int B_{21}[v, p, q]pdH_0 + \int B_{22}[v, p, q]qdF_0, \]

where the right-hand side is the derivative of

\[ \mathcal{P} \left\{ v^T \mathcal{L}_\beta(\beta_0, H_0, F_0) + \mathcal{L}_H(\beta_0, H_0, F_0) \left[ \int pdH_0 \right] + \mathcal{L}_F(\beta_0, H_0, F_0) \left[ \int qdF_0 \right] \right\} \]

along the path $(\beta_0 + \epsilon v, H_0 + \epsilon \int pdH_0, F_0 + \epsilon \int qdF_0)$. This implies that the information along this path is zero, hence

\[ v^T \mathcal{L}_\beta(\beta_0, H_0, F_0) + \mathcal{L}_H(\beta_0, H_0, F_0) \left[ \int pdH_0 \right] + \mathcal{L}_F(\beta_0, H_0, F_0) \left[ \int qdF_0 \right] = 0 \]

almost surely. By the identifiability condition (C6), we get $(v, p, q) = 0$. 
To show \((B_1, B_{21}, B_{22})\) is surjective, write
\[
(B_1, B_{21}, B_{22})[v, p, q] = I_1[v, p, q] + I_2[v, p, q],
\]
where \(I_1[v, p, q]\) is defined as
\[
\left( \begin{array}{c} v \\ \eta_H(t; \beta_0, H_0, F_0)p(t) \\ \eta_F(z; \beta_0, H_0, F_0)q(z) \end{array} \right)
\]
and \(I_2[v, p, q]\) is defined as
\[
\left( \begin{array}{c} v^T \zeta_\beta(\beta_0, H_0, F_0) + \int_0^t \eta_{H\beta}(t; \beta_0, H_0, F_0)p(t)dH_0(t) + \int_M \eta_{F\beta}(z; \beta_0, H_0, F_0)q(z)dF_0(z) - v \\ v^T \zeta_H(t; \beta_0, H_0, F_0) + \int_0^t \eta_{HH}(s, t; \beta_0, H_0, F_0)p(s)dH_0(s) + \int_M \eta_{HH}(x, t; \beta_0, H_0, F_0)q(x)dH_0(x) \\ v^T \zeta_F(z; \beta_0, H_0, F_0) + \int_0^t \eta_{HF}(s, z; \beta_0, H_0, F_0)p(s)dH_0(s) + \int_M \eta_{HF}(x, z; \beta_0, H_0, F_0)q(x)dF_0(x) \end{array} \right).
\]
Since \(\eta_H(t; \beta_0, H_0, F_0) = \mathcal{P}(\mathcal{L}_H(\beta_0, H_0, F_0)[I(\cdot \geq t)])\), which is the score function along the direction of \(H_0 + \epsilon I(\cdot \geq 0)\), hence \(\eta_H(t; \beta_0, H_0, F_0)\) is negative and continuous for all \(t\). Similarly, \(\eta_F(z; \beta_0, H_0, F_0)\) is the score function along the direction of \(F_0 + \epsilon I(\cdot \geq 0)\), hence \(\eta_F(z; \beta_0, H_0, F_0)\) is negative and continuous for all \(z\). Thus \(I_1[v, p, q]\) is a bijective continuous operator, hence it is continuously invertible.

Furthermore, by smoothness conditions (C1) and (C3), the image of \(I_2\) is a set of uniformly bounded and equi-continuous functions, hence \(I_2\) is a compact operator by Arzelà-Ascoli theorem. Now we can write \((B_1, B_{21}, B_{22})\) as
\[
I_1 + I_2 = I_1(I + I_1^{-1}I_2),
\]
where \(I\) is identity mapping. It is clear that \(I_1^{-1}I_2\) is a compact operator by the continuity of \(I_1^{-1}\). Then \(I + I_1^{-1}I_2\) is a Fredholm operator. Now since \(\ker(I + I_1^{-1}I_2) = \ker(I_1 + I_2) = 0\), the Fredholm operator theory gives that \(I + I_1^{-1}I_2\) is surjective, so is \(I_1 + I_2\). So far the continuous invertibility of \((B_1, B_{21}, B_{22})\) has been proved.

Now with
\[
(\tilde{v}, \tilde{p}, \tilde{q}) = (B_1, B_{21}, B_{22})^{-1}(v, p, q),
\]
from (2) we have
\[
\sqrt{n} \left\{ v^T(\hat{\beta}_n - \beta_0) + \int pd(\hat{H}_n - H_0) + \int qd(\hat{F}_n - F_0) \right\} = -\sqrt{n} \left( \mathcal{P}_n - \mathcal{P} \right) \left\{ \tilde{v}^T \mathcal{L}_H(\beta_0, H_0, F_0) + \mathcal{L}_H(\beta_0, H_0, F_0) \left[ \int \tilde{p}dH_0 \right] + \mathcal{L}_F(\beta_0, H_0, F_0) \left[ \int \tilde{q}dF_0 \right] \right\} + o \left( \sqrt{n}|\hat{\beta}_n - \beta_0| + \sqrt{n} |\hat{H}_n - H_0| + \sqrt{n} |\hat{F}_n - F_0| \right).
\]
The first term of the right side of (3) converges in distribution to a zero-mean Gaussian process in the metric space $\mathbb{R}^d \times L^\infty(Q_1) \times L^\infty(Q_2)$. By the Slutsky’s theorem, we now need only to show that
\[
\sqrt{n}|\hat{\beta}_n - \beta_0| + \sqrt{n}|\hat{H}_n - H_0| + \sqrt{n}|\hat{F}_n - F_0| = O_p(1). \tag{4}
\]
By definition,
\[
|\hat{\beta}_n - \beta_0| + |\hat{H}_n - H_0| + |\hat{F}_n - F_0| = \sup_{(v,p,q) \in \mathcal{V} \times Q_1 \times Q_2} \left| v^T(\hat{\beta}_n - \beta_0) + \int_0^T pd(\hat{H}_n - H_0) + \int qd(\hat{F}_n - F_0) \right|,
\]
so (3) gives
\[
\sqrt{n}|\hat{\beta}_n - \beta_0| + \sqrt{n}|\hat{H}_n - H_0| + \sqrt{n}|\hat{F}_n - F_0| = O_p(1) + o\left(\sqrt{n}|\hat{\beta}_n - \beta_0| + \sqrt{n}|\hat{H}_n - H_0| + \sqrt{n}|\hat{F}_n - F_0|\right),
\]
therefore (4) holds immediately.

Now we have
\[
\sqrt{n}\left\{v^T(\hat{\beta}_n - \beta_0) + \int pd(\hat{H}_n - H_0) + \int qd(\hat{F}_n - F_0)\right\} \\
= -\sqrt{n}(\mathcal{P}_n - \mathcal{P})\left\{\hat{v}^T\mathcal{L}_\beta(\beta_0, H_0, F_0) + \mathcal{L}_H(\beta_0, H_0, F_0)\left[\int (\hat{p}dH_0)\right] + \mathcal{L}_F(\beta_0, H_0, F_0)\left[\int (\hat{q}dF_0)\right]\right\} \\
+ o_p(1). \tag{5}
\]
Since the right side of (5) converges to a normal distribution by the central limit theorem, we have proved that $\sqrt{n}(\hat{\beta}_n - \beta_0, \hat{H}_n - H_0, \hat{F}_n - F_0)$ converges weakly to a zero-mean Gaussian process. Let $p = q = 0$ in equality (5), then the estimate $\hat{\beta}_n$ is an asymptotically linear estimator with influence function $\hat{v}^T\mathcal{L}_\beta(\beta_0, H_0, F_0) + \mathcal{L}_H(\beta_0, H_0, F_0)[\int (\hat{p}dH_0)] + \mathcal{L}_F(\beta_0, H_0, F_0)[\int (\hat{q}dF_0)]$, which lies in the linear space spanned by the score functions: $\{\hat{v}^T\mathcal{L}_\beta + \mathcal{L}_H[\int (\hat{p}dH)] + \mathcal{L}_F[\int (\hat{q}dF)]: \hat{v} \in \mathbb{R}^d, \hat{p} \in Q_1, \hat{q} \in Q_2\}$. By proposition 1 in Bickel et al. (1993, PP. 65), the estimate $\hat{\beta}_n$ is semiparametrically efficient.

**Proof of Theorem 3.** Let $\hat{H}_n(t)$ be a step function with a jump size $\hat{h}_n(s_i) = H_0(s_i) - \max_{s_{i-1} < s_i} H_0(s_i)$ at each failure time $s_i$, and $\hat{F}_n(z)$ is a step function with a jump size $\hat{f}_n(Z_i) = F_0(Z_i) - \max_{Z_{i_1} \leq Z_i \cdots \leq Z_{i_d}} F_0(Z_{j_1}(1), \cdots, Z_{j_d}(d))$ at each observed covariate $Z_i$, where $Z_k(m)$ is the value on the $m$-th dimension of the covariate $Z_k$, and the order $Z_k < Z_i$ is defined as $Z_k(m) < Z_i(m)$ for all $1 \leq m \leq d$. With such definitions, $\hat{H}_n(s_i) = H_0(s_i)$ and $\hat{F}_n(Z_i) = F_0(Z_i)$ holds for every $s_i$ and $Z_i$ clearly.

17
Choose $v_1 \in \mathbb{R}^d$ and bounded variation functions $p_1 \in \mathcal{Q}_1, q_1 \in \mathcal{Q}_2$ such that
\[
\begin{pmatrix}
v_1 \\
\int p_1 d\hat{H}_n \\
\int q_1 d\hat{F}_n
\end{pmatrix} = I_n^{-1} \begin{pmatrix} v \\
\tilde{p} \\
\tilde{q}
\end{pmatrix}
\]
where $I_n, \tilde{p}, \tilde{q}$ are as defined in Theorem 3 and
\[
\int p_1 d\hat{H}_n = (p_1(s_1)\hat{h}_n(s_1), \cdots, p_1(s_{n_1})\hat{h}_n(s_{n_1}))^T, \\
\int q_1 d\hat{F}_n = (q_1(Z_1)\hat{f}_n(Z_1), \cdots, q_1(Z_{n_2})\hat{f}_n(Z_{n_2}))^T.
\]
Denote
\[
\hat{h}_n - \hat{h} = (\hat{h}_n(s_1) - \hat{h}(s_1), \cdots, \hat{h}_n(s_{n_1}) - \hat{h}(s_{n_1}))^T
\]
and
\[
\hat{f}_n - \hat{f} = (\hat{f}_n(Z_1) - \hat{f}(Z_1), \cdots, \hat{f}_n(Z_{n_2}) - \hat{f}(Z_{n_2}))^T,
\]
then we have
\[
\sqrt{n}v^T(\hat{\beta}_n - \beta_0) + \sqrt{n} \sum_{i=0}^{n_1} p_i(s_i)(\hat{h}_n(s_i) - \hat{h}(s_i)) + \sqrt{n} \sum_{i=0}^{n_2} q_i(Z_i)(\hat{f}_n(Z_i) - \hat{f}(Z_i))
\]
\[
= \sqrt{n} \begin{pmatrix}
\hat{\beta}_n - \beta_0 \\
\hat{h}_n - \hat{h} \\
\hat{f}_n - \hat{f}
\end{pmatrix}^T \begin{pmatrix}
v \\
\int p_1 d\hat{H}_n \\
\int q_1 d\hat{F}_n
\end{pmatrix}
\]
\[
= \sqrt{n} \begin{pmatrix}
\hat{\beta}_n - \beta_0 \\
\hat{h}_n - \hat{h} \\
\hat{f}_n - \hat{f}
\end{pmatrix}^T I_n \begin{pmatrix} v_1 \\
\int p_1 d\hat{H}_n \\
\int q_1 d\hat{F}_n
\end{pmatrix}
\]
\[
= \sqrt{n} \mathbb{P}_n \begin{pmatrix}
\mathcal{L}\beta & \mathcal{L}\beta & \mathcal{L}\beta \\
\mathcal{L}H & \mathcal{L}H & \mathcal{L}H \\
\mathcal{L}F & \mathcal{L}F & \mathcal{L}F
\end{pmatrix} \begin{pmatrix}
\hat{\beta}_n, \hat{H}_n, \hat{F}_n
\end{pmatrix} \begin{pmatrix}
v_1 \\
\int p_1 d\hat{H}_n \\
\int q_1 d\hat{F}_n
\end{pmatrix}, \begin{pmatrix}
\hat{\beta}_n - \beta_0 \\
\hat{H}_n - \hat{H}_n \\
\hat{F}_n - \hat{F}_n
\end{pmatrix}
\]
\[
= \sqrt{n} \mathbb{P} \begin{pmatrix}
\mathcal{L}\beta & \mathcal{L}\beta & \mathcal{L}\beta \\
\mathcal{L}H & \mathcal{L}H & \mathcal{L}H \\
\mathcal{L}F & \mathcal{L}F & \mathcal{L}F
\end{pmatrix} \begin{pmatrix}
\beta_0, H_0, F_0
\end{pmatrix} \begin{pmatrix}
v_1 \\
\int p_1 dH_0 \\
\int q_1 dF_0
\end{pmatrix}, \begin{pmatrix}
\hat{\beta}_n - \beta_0 \\
\hat{H}_n - H_0 \\
\hat{F}_n - F_0
\end{pmatrix}
\]
\[
+ o_p(1)
\]
\[
= -\sqrt{n}(\mathbb{P}_n - \mathbb{P}) \left\{ v_1^T \mathcal{L}\beta(\beta_0, H_0, F_0) + \mathcal{L}H(\beta_0, H_0, F_0) \left[ \int p_1 dH_0 \right] + \mathcal{L}F(\beta_0, H_0, F_0) \left[ \int q_1 dF_0 \right] \right\}
\]
\[
+ o_p(1),
\]
where the last equality is proved in the proof of Theorem 2.
Note that
\[
-\sqrt{n}(\mathcal{P}_n - \mathcal{P})\left\{ v_1^T L_\beta(\beta_0, H_0, F_0) + L_H(\beta_0, H_0, F_0) \left[ \int p_1 dH_0 \right] + L_F(\beta_0, H_0, F_0) \left[ \int q_1 dF_0 \right] \right\} = -\sqrt{n}(\mathcal{P}_n - \mathcal{P})\left\{ v_1^T L_\beta(\beta_0, H_0, F_0) + L_H(\beta_0, H_0, F_0) \left[ \int p_1 d\hat{H}_n \right] + L_F(\beta_0, H_0, F_0) \left[ \int q_1 d\hat{F}_n \right] \right\} + o_p(1),
\]
and variance of the right side is consistently estimated by
\[
-\mathcal{P} \begin{pmatrix} L_{\beta\beta} & L_{\beta H} & L_{\beta F} \\ L_{H\beta} & L_{HH} & L_{HF} \\ L_{F\beta} & L_{FH} & L_{FF} \end{pmatrix} (\beta_0, H_0, F_0) \left[ \begin{pmatrix} v_1 \\ \int p_1 d\hat{H}_n \\ \int q_1 d\hat{F}_n \end{pmatrix}, \begin{pmatrix} v_1 \\ \int p_1 d\hat{H}_n \\ \int q_1 d\hat{F}_n \end{pmatrix} \right],
\]
which can be further estimated by
\[
-\mathcal{P}_n \begin{pmatrix} L_{\beta\beta} & L_{\beta H} & L_{\beta F} \\ L_{H\beta} & L_{HH} & L_{HF} \\ L_{F\beta} & L_{FH} & L_{FF} \end{pmatrix} (\hat{\beta}_n, \hat{H}_n, \hat{F}_n) \left[ \begin{pmatrix} v_1 \\ \int p_1 d\hat{H}_n \\ \int q_1 d\hat{F}_n \end{pmatrix}, \begin{pmatrix} v_1 \\ \int p_1 d\hat{H}_n \\ \int q_1 d\hat{F}_n \end{pmatrix} \right] = (v^T, p^T, q^T) I_n^{-1} I_n^{-1} (v^T, p^T, q^T)^T
\]
\[
= (v^T, p^T, q^T) I_n^{-1} (v^T, p^T, q^T)^T.
\]

The proof of Theorem 3 is complete.

References


Table 1: Summary of simulation results.

Censoring rate=0.7

<table>
<thead>
<tr>
<th>r</th>
<th>Bias</th>
<th>SSD</th>
<th>ESE</th>
<th>CP</th>
<th>RE</th>
<th>RE*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
</tr>
<tr>
<td>0</td>
<td>0.006</td>
<td>-0.027</td>
<td>0.542</td>
<td>0.333</td>
<td>0.584</td>
<td>0.375</td>
</tr>
<tr>
<td>1</td>
<td>0.014</td>
<td>-0.027</td>
<td>0.676</td>
<td>0.405</td>
<td>0.695</td>
<td>0.424</td>
</tr>
<tr>
<td>2</td>
<td>-0.003</td>
<td>0.021</td>
<td>0.733</td>
<td>0.433</td>
<td>0.750</td>
<td>0.477</td>
</tr>
</tbody>
</table>

Censoring rate=0.8

<table>
<thead>
<tr>
<th>r</th>
<th>Bias</th>
<th>SSD</th>
<th>ESE</th>
<th>CP</th>
<th>RE</th>
<th>RE*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
</tr>
<tr>
<td>0</td>
<td>-0.007</td>
<td>-0.033</td>
<td>0.623</td>
<td>0.393</td>
<td>0.642</td>
<td>0.411</td>
</tr>
<tr>
<td>1</td>
<td>0.032</td>
<td>0.019</td>
<td>0.760</td>
<td>0.467</td>
<td>0.736</td>
<td>0.458</td>
</tr>
<tr>
<td>2</td>
<td>0.018</td>
<td>-0.018</td>
<td>0.928</td>
<td>0.535</td>
<td>1.070</td>
<td>0.560</td>
</tr>
</tbody>
</table>

Statistica Sinica: Newly accepted Paper (accepted version subject to English editing)
Table 2: Comparison of parameter estimation (estimated standard error, p-value) for the Stanford heart transplant data.

<table>
<thead>
<tr>
<th>Design</th>
<th>Model 1</th>
<th></th>
<th></th>
<th>Model 2</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Age</td>
<td>T5</td>
<td></td>
<td>Age</td>
<td>T5</td>
<td></td>
</tr>
<tr>
<td>Full data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age T5</td>
<td>0.0295 (0.0115, 0.0056)</td>
<td>0.1692 (0.2110, 0.2120)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age Age²</td>
<td>-0.1457 (0.0528, 0.0033)</td>
<td>0.0023 (0.0007, 0.0004)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Design</th>
<th>Model 1</th>
<th></th>
<th></th>
<th>Model 2</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Age</td>
<td>T5</td>
<td></td>
<td>Age</td>
<td>T5</td>
<td></td>
</tr>
<tr>
<td>Full data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age T5</td>
<td>0.0281 (0.0179, 0.0400)</td>
<td>0.1175 (0.2817, 0.3368)</td>
<td></td>
<td>0.0292 (0.0182, 0.0677)</td>
<td>0.1621 (0.2756, 0.2960)</td>
<td></td>
</tr>
<tr>
<td>C-C-C</td>
<td>0.0275 (0.0186, 0.0459)</td>
<td>0.1145 (0.2882, 0.3406)</td>
<td></td>
<td>0.0295 (0.0178, 0.0638)</td>
<td>0.1636 (0.2757, 0.2947)</td>
<td></td>
</tr>
<tr>
<td>N-C-C</td>
<td>0.0263 (0.0146, 0.0301)</td>
<td>0.1825 (0.2426, 0.2263)</td>
<td></td>
<td>0.0291 (0.0174; 0.0542)</td>
<td>0.1691 (0.2565, 0.2693)</td>
<td></td>
</tr>
<tr>
<td>Model 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>Age²</td>
<td></td>
<td>Age</td>
<td>Age²</td>
<td></td>
</tr>
<tr>
<td>Full data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age Age²</td>
<td>-0.1416 (0.0593, 0.0294)</td>
<td>0.0023 (0.0008, 0.0108)</td>
<td></td>
<td>-0.1478 (0.0830, 0.0599)</td>
<td>0.0024 (0.0010, 0.0317)</td>
<td></td>
</tr>
<tr>
<td>C-C-C</td>
<td>-0.1429 (0.0652, 0.0320)</td>
<td>0.0023 (0.0008, 0.0125)</td>
<td></td>
<td>-0.1498 (0.0850, 0.0614)</td>
<td>0.0024 (0.0011, 0.0337)</td>
<td></td>
</tr>
<tr>
<td>N-C-C</td>
<td>-0.1474 (0.0552, 0.0150)</td>
<td>0.0023 (0.0007, 0.0065)</td>
<td></td>
<td>-0.1489 (0.0697, 0.0355)</td>
<td>0.0024 (0.0009, 0.0136)</td>
<td></td>
</tr>
</tbody>
</table>

Note: Proposed represents the proposed method; Chen represents the Chen et al. (2012) method; C-C represents case-cohort sampling; C-C-C represents classical case-control sampling; N-C-C represents nested case-control sampling.