MAXIMUM LIKELIHOOD METHOD FOR LINEAR TRANSFORMATION MODELS WITH COHORT SAMPLING DATA

Yuan Yao

Hong Kong Baptist University

Abstract: 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 the Stanford heart transplant data are presented.

Key words and phrases: 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. 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 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)).

Thomas (1977) proposed a partial likelihood approach to estimating the regression parameter of the 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 the 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.

Samuelsen (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. The resulting estimator and its inference are also very easy to compute. Empirical evidence shows that Samuelsen's estimator is sometimes better than Thomas'. As Samuelsen (1997) has pointed out, however, the weighting techniques used to construct the pseudo-likelihood can be inefficient.

Chen (2001) proposed an estimation method based on local averaging that is essentially an alternative sample reuse method other than those of Thomas and Samuelsen. 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 Samuelsen, but the efficient estimator can be quite difficult to obtain, as it involves estimating and computing an integral operator.

There are also some other methods in the literature. Almost all use Cox model. This puts serious limitations 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. 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, ϵ is the unobserved random variable with known distribution and β is the unknown regression parameter of main interest. When ϵ follows the extreme value distribution or the standard logistic distribution, the model reduces to the Cox's proportional hazards model or the proportional odds model, respectively. Cheng, Wei, and Ying (1995) contains a very simple and elegant idea based on pairwise comparison of the survival times. This type of rank based method relies on the assumption that the censoring variables are independent of the covariates.

In this paper, we propose an efficient estimation method for linear transformation models with cohort sampling designs. 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 application to data are given in Sections 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 ith individual respectively. We assume that C_i is independent of (T_i, Z_i). We set

(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 ith individual being sampled for covariate ascertainment: Δ_i, with conditional probability $\pi_i = P(\Delta_i = 1|(Y_j, \delta_j), 1 \leq j \leq n)$.

It is necessary in the inference procedure that $(\Delta_1, \ldots, \Delta_n)$ are conditionally independent of $(Z_1, \ldots, Z_n)$, given (Y_j, δ_j), $1 \leq j \leq n$.

In the linear transformation model

$$\log H(T) = -\beta Z + \epsilon, \tag{2.1}$$

it is supposed that β is the unknown parameter of main interest, H is an unknown increasing function, and ϵ is a continuous random variable whose hazard function λ_ϵ is known. We take ϵ to be independent of Z and C.

The likelihood function from (2.1) can then be written as

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

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

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

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

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

The last term in the likelihood function 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 λ and Λ, and the probability density and distribution of covariate Z as f and F, respectively. The log-likelihood then takes a final form

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

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

$$+ \delta_i \log h(Y_i) + \Delta_i \log f(Z_i)$$

$$+ \log(\lambda_C(Y_i)^{1-\delta_i} e^{-\Lambda_C(Y_i)}) + \log f_\Delta(\Delta_1, \ldots, \Delta_n | (Y_j, \delta_j), 1 \le j \le n),$$

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 jth 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

$$H(t) = \sum_{j=1}^{n_1} q_j I(s_j \le t), \quad \text{and} \quad h(t) = \sum_{j=1}^{n_1} q_j I(t = s_j).$$

To estimate F, put probability mass on all known covariates, with

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

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

$$\int [e^{\beta Z} \lambda(H(Y_i)e^{\beta Z})]^{\delta_i} e^{-\Lambda(H(Y_i)e^{\beta Z})} f(Z) dZ$$

$$= \sum_{j=1}^{n_2} [e^{\beta Z_j} \lambda(H(Y_i)e^{\beta Z_j})]^{\delta_i} e^{-\Lambda(H(Y_i)e^{\beta Z_j})} p_j.$$

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

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

To describe the variance estimation, let τ denote the duration of the study and suppose Z lies in a bounded set M. Let $\mathcal{Q}_1 = \{p \in BV[0, \tau] : |p| \le 1\}$ and $\mathcal{Q}_2 = \{q \in BV[M] : |q| \le 1\}$, 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 $L^\infty(\mathcal{Q}_1)$ as

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

and similarly, $\hat{F}_n$ can be treated as a bounded linear functional in $L^\infty(\mathcal{Q}_2)$.

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* $\mathbb{R}^d \times L^\infty(\mathcal{Q}_1) \times L^\infty(\mathcal{Q}_2)$. *The limiting covariance matrix of* $\sqrt{n}(\hat{\beta}_n - \beta_0)$ *attains the semiparametric efficiency bound.*

Theorem 3. *For any* $(b, p, q) \in \mathcal{V} \times \mathcal{Q}_1 \times \mathcal{Q}_2$, *where* $\mathcal{V} = \{v \in \mathbb{R}^d : |v| \leq 1\}$, *the asymptotic variance for*

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

can be consistently estimated by $(v^T, \vec{p}^T, \vec{q}^T)I_n^{-1}(v^T, \vec{p}^T, \vec{q}^T)^T$, *where* nI_n *is the negative Hessian matrix of the log-likelihood function* $l_n(\beta, H, F)$ *with respect to* $(\beta, q_1, \ldots, q_{n_1}, p_1, \ldots, p_{n_2})$, *and the vectors* $\vec{p} = (p(s_1), \ldots, p(s_{n_1}))^T$, $\vec{q} = (q(Z_1), \ldots, q(Z_{n_2}))^T$. *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 does well when handling moderate sample size like 200. For the initial value, one can use estimates from the Cox model or try different initial values to make the maximization guaranteed.

3. Simulation Study

We took independent covariates Z_1 uniform distributed over $[0, 1]$ and Z_2 Bernoulli with success probability 0.5. The hazard function of the error term ϵ was chosen as

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

with $r = 0, 1$ and 2 (Dabrowska and Doksum (1988); Chen, Jin, and Ying (2002)). Here 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 was uniform on $[0, c]$, where c was chosen such that the expected proportion of censoring was 0.7 and 0.8, respectively. The sample size n was set at 200 and all simulations were based on 500 replications. For the case-cohort design, we selected all failures and a subcohort of size 85 and 50 when the censoring rate was 0.7 and 0.8, respectively. For the classical case-control design, we selected all failures and a group of non-failures of the same size as the failures. For the nested case-control design, we selected all failures and two non-failures in each risk set of failure times.

Table 1 summarizes the simulation results estimating β_1 and β_2. The true values were $\beta_1 = 1$ 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

Table 1. Summary of simulation results.

Censoring rate=0.7

	Case-cohort design											
r	Bias		SSD		ESE		CP		RE		RE*	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
0	0.006	-0.027	0.542	0.333	0.584	0.375	0.962	0.976	0.84	0.77	0.73	0.64
1	0.014	-0.027	0.676	0.405	0.695	0.424	0.954	0.960	0.65	0.60	0.62	0.62
2	-0.003	0.021	0.733	0.433	0.750	0.477	0.958	0.962	0.85	0.82	0.72	0.67
	Classical case-control design											
r	Bias		SSD		ESE		CP		RE		RE*	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
0	0.018	-0.012	0.550	0.343	0.580	0.364	0.964	0.968	0.82	0.72	0.60	0.59
1	0.016	-0.040	0.663	0.402	0.674	0.415	0.966	0.956	0.68	0.61	0.62	0.60
2	-0.022	-0.018	0.779	0.460	0.823	0.469	0.944	0.946	0.76	0.73	0.67	0.70
	Nested case-control design											
r	Bias		SSD		ESE		CP		RE		RE*	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
0	-0.013	-0.006	0.531	0.304	0.544	0.341	0.960	0.974	0.88	0.92	0.79	0.83
1	0.039	-0.035	0.642	0.374	0.653	0.398	0.954	0.970	0.72	0.70	0.78	0.74
2	0.016	-0.002	0.741	0.443	0.781	0.451	0.950	0.950	0.83	0.78	0.78	0.82

Censoring rate=0.8

	Case-cohort design											
r	Bias		SSD		ESE		CP		RE		RE*	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
0	-0.007	-0.033	0.623	0.393	0.642	0.411	0.962	0.972	0.83	0.92	0.62	0.76
1	0.032	0.019	0.760	0.467	0.736	0.458	0.950	0.954	0.74	0.74	0.55	0.63
2	0.018	-0.018	0.928	0.535	1.070	0.560	0.958	0.956	0.65	0.72	0.58	0.62
	Classical case-control design											
r	Bias		SSD		ESE		CP		RE		RE*	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
0	0.036	-0.002	0.732	0.436	0.725	0.449	0.940	0.960	0.60	0.74	0.51	0.59
1	0.068	-0.004	0.829	0.480	0.917	0.516	0.958	0.950	0.62	0.70	0.45	0.61
2	-0.007	0.003	0.933	0.540	0.920	0.547	0.956	0.958	0.64	0.71	0.59	0.66
	Nested case-control design											
r	Bias		SSD		ESE		CP		RE		RE*	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
0	0.012	-0.031	0.612	0.380	0.649	0.415	0.970	0.962	0.86	0.98	0.75	0.88
1	0.051	-0.036	0.768	0.424	0.843	0.485	0.960	0.968	0.72	0.90	0.70	0.87
2	0.024	0.020	0.792	0.486	0.855	0.508	0.966	0.962	0.89	0.87	0.74	0.76

of the estimated standard errors (ESE) of the estimates, and the 95% empirical coverage probabilities (CP) for β_1 and β_2 based on the asymptotic normality in Theorem 2. The proposed estimation procedures perform well in all cases.

We compared our approach with that of Chen, Sun, and Tong (2012), 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 was conducted and it found that the Chen et al. estimators less efficient in most of the senarios, especially when the censoring rate was large. This is mainly because the Chen et al. estimator does not involve the event times whose covariates are not sampled. 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 was 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 incompatability between the initial donor and the recipient with respect to HLA antigens. The goal of this study was to analyze the relationship among the survival time, T5 mismatch score and the age of the patients.

We compared the results using the full cohort data with those from different sampling schemes. As in Miller and Halpern (1982) and Jin, Lin, and Ying (2006), only the 157 patients with complete records were used. Following Miller and Halpern (1982), in Model 1 we regressed the logarithm of the survival time against the ages and T5 mismatch scores for the 157 patients. Then the T5 mismatch score was deleted due to its nonsignificance in Model 1 and a quadratic age model was tried to achieve better fit. In Model 2 we regressed 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 were set as follows: For the case-cohort design we selected all failures and a subcohort of size 75 among the whole set of subjects. For the classical case-control design we selected all failures and 25 non-failures. For the nested case-control design we selected 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 led to the same conclusion as using the full data set. For comparison, we also present the estimates from Chen, Sun, and Tong (2012) in the table, from which one can draw similar conclusions in general. It can also be seen from the comparison of standard error that the proposed method has the advantage of being asymptotically efficient.

Table 2. Comparison of parameter estimation (estimated standard error, p-value) for the Stanford heart transplant data.

Full data

Model 1	Age	T5
	0.0295 (0.0115, 0.0056)	0.1692 (0.2110,0.2120)
Model 2	Age	Age^2
	-0.1457 (0.0528, 0.0033)	0.0023 (0.0007,0.0004)

Model 1

Design	Proposed		Chen	
	Age	T5	Age	T5
C-C	0.0281 (0.0179, 0.0400)	0.1175 (0.2817, 0.3368)	0.0292 (0.0182, 0.0677)	0.1621 (0.2756, 0.2960)
C-C-C	0.0275 (0.0186, 0.0459)	0.1145 (0.2882, 0.3406)	0.0295 (0.0178, 0.0638)	0.1636 (0.2757, 0.2947)
N-C-C	0.0263 (0.0146, 0.0301)	0.1825 (0.2426, 0.2263)	0.0291 (0.0174, 0.0542)	0.1691 (0.2565, 0.2693)

Model 2

Design	Proposed		Chen	
	Age	Age^2	Age	Age^2
C-C	-0.1416 (0.0593, 0.0294)	0.0023 (0.0008, 0.0108)	-0.1478 (0.0830, 0.0599)	0.0024 (0.0010, 0.0317)
C-C-C	-0.1429 (0.0652, 0.0320)	0.0023 (0.0008, 0.0125)	-0.1498 (0.0850, 0.0614)	0.0024 (0.0011, 0.0337)
N-C-C	-0.1474 (0.0552, 0.0150)	0.0023 (0.0007, 0.0065)	-0.1489 (0.0697, 0.0355)	0.0024 (0.0009, 0.0136)

Note: Proposed represents the proposed method; Chen represents the Chen, Sun, and Tong (2012) method; C-C represents case-cohort sampling; C-C-C represents classical case-control sampling; N-C-C represents nested case-control sampling.

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. The variance estimation is obtained from the inverse of the negative Hessian matrix. We prove that such estimation method attains the semiparametric efficiency bound. The performance of the estimator is illustrated with simulation studies and a Stanford heart transplant data study. Further work will consider extending this method to partially linear transformation models, where kernel smoothing could be applied.

Acknowledgement

Appendix: Regularity conditions and proofs of theorems

Some regularity conditions are needed:

(C1) The function $H_0(t)$ and the distribution $F_0(t)$ are strictly increasing and differentiable, with derivatives $h(t)$ and $f(t)$ that are absolutely continuous, and β_0 lies in the interior of a known compact set in the domain of $\mathcal{B}$.

(C2) The covariate Z is bounded, and $P(C \geq \tau) > \delta > 0$ for some constant δ.

(C3) $\lambda(t) > 0$, and λ is twice continuously differentiable.

(C4) $\limsup\limits_{x \to \infty} \Lambda(C_0 x)^{-1} \log(x \sup_{y \leq x} \lambda(y)) = 0$ holds for every $C_0 > 0$.

(C5) (First Identifiability) Let

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

If

$$\Psi_i(\beta^*, H^*, F^*) h^*(Y_i)^{\delta_i} f^*(Z_i)^{\Delta_i} = \Psi_i(\beta_0, H_0, F_0) h_0(Y_i)^{\delta_i} f_0(Z_i)^{\Delta_i},$$

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

(C6) (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. The jump size of $\hat{H}_n$ must be finite, otherwise the log-likelihood would diverge to $-\infty$ by (C4). $\hat{H}_n$ is also bounded almost surely, otherwise if a new estimator $\bar{H}_n = \hat{H}_n/\hat{H}_n(\tau)$ 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 is 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 β^* for the same subsequence. Then consistency is 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

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

Hence

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

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

Now define

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

The Glivenko-Cantelli Theorem leads to $\lim_{n\to\infty} \tilde{H}_n(t) = H_0(t)$. Here $\hat{H}_n(t)$ can be written as

$$\hat{H}_n(t) = \int_0^t \frac{\dfrac{1}{n}\displaystyle\sum_{j=1}^{n} \frac{\Psi_{jH}(\beta_0, H_0, F_0)[I(\cdot \geq u)]}{\Psi_j(\beta_0, H_0, F_0)}}{\dfrac{1}{n}\displaystyle\sum_{j=1}^{n} \frac{\Psi_{jH}(\hat{\beta}_n, \hat{H}_n, \hat{F}_n)[I(\cdot \geq u)]}{\Psi_j(\hat{\beta}_n, \hat{H}_n, \hat{F}_n)}} d\tilde{H}_n(u). \qquad (A.1)$$

It is now possible to show that H^* is continuously differentiable. For simplicity, define $E\left(\Psi_{jH}(\beta_0, H_0, F_0)[I(\cdot \geq u)]/\Psi_j(\beta_0, H_0, F_0)\right)$ as $g_{1j}(u)$ and $E\left(\Psi_{jH}(\beta^*, H^*, F^*)[I(\cdot \geq u)]/\Psi_j(\beta^*, H^*, F^*)\right)$ as $g_{2j}(u)$. It can be shown that they are both bounded away from zero.

Taking the limit of (A.1), we get

$$H^*(t) = \int_0^t \frac{g_{1j}(u)}{g_{2j}(u)} dH_0(u).$$

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

$$\lim_{n\to\infty} \frac{d\hat{H}_n(t)}{d\tilde{H}_n(t)} = \frac{h^*(t)}{h_0(t)}$$

uniformly in $t \in [0, \tau]$, where h^* is the derivative of H^*.

Repeating the same process with F 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

$$\frac{1}{n} \sum_{j=1}^{n} \log \frac{\Psi_j(\hat{\beta}_n, \hat{H}_n, \hat{F}_n)}{\Psi_j(\beta_0, \tilde{H}_n, \tilde{F}_n)} + \frac{1}{n} \sum_{j=1}^{n} \delta_j \log \frac{d\hat{H}_n(Y_j)}{d\tilde{H}_n(Y_j)} + \frac{1}{n} \sum_{j=1}^{n} \Delta_j \log \frac{d\hat{F}_n(Z_j)}{d\tilde{F}_n(Z_j)} \geq 0.$$

Let $n \to \infty$, then

$$E\left(\log \frac{\Psi_j(\beta^*, H^*, F^*) h^*(Y_j)^{\delta_j} f^*(Z_j)^{\Delta_j}}{\Psi_j(\beta_0, H_0, F_0) h_0(Y_j)^{\delta_j} f_0(Z_j)^{\Delta_j}} \right) \geq 0.$$

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 of Van der Vaart (1998, pp.419-424). Let

$$\mathcal{L}(\beta, H, F) = \log \Psi_j + \delta_j \log h(Y_j) + \Delta_j \log f(Z_j),$$

$$\Phi_n(\beta, H, F) = \mathcal{P}_n \left\{ v^T \mathcal{L}_\beta + \mathcal{L}_H \left[\int p \, dH \right] + \mathcal{L}_F \left[\int q \, dF \right] \right\},$$

$$\Phi(\beta, H, F) = \mathcal{P} \left\{ v^T \mathcal{L}_\beta + \mathcal{L}_H \left[\int p \, dH \right] + \mathcal{L}_F \left[\int q \, dF \right] \right\},$$

where we use $v^T \mathcal{L}_\beta$, $\mathcal{L}_H[g_1]$, and $\mathcal{L}_F[g_2]$ to 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 with $\mathcal{P}$ as its expectation.

For any $\delta_0 > 0$, when n large enough,

$$(\hat{\beta}_n, \hat{H}_n, \hat{F}_n) \in \mathcal{N}_0 = \left\{ (\beta, H, F) : |\beta - \beta_0| + |H - H_0| + |F - F_0| < \delta_0 \right\}$$

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

$$\sqrt{n}(\mathcal{P}_n - \mathcal{P}) \left\{ v^T \mathcal{L}_\beta(\beta_0, H_0, F_0) + \mathcal{L}_H(\beta_0, H_0, F_0) \left[\int p \, dH_0 \right] \right.$$

$$+ \mathcal{L}_F(\beta_0, H_0, F_0) \Big[\int q dF_0 \Big] \Big\}$$

$$= \sqrt{n}(\mathcal{P}_n - \mathcal{P}) \Big\{ v^T \mathcal{L}_\beta(\hat{\beta}_n, \hat{H}_n, \hat{F}_n) + \mathcal{L}_H(\hat{\beta}_n, \hat{H}_n, \hat{F}_n) \Big[\int p d\hat{H}_n \Big]$$

$$+ \mathcal{L}_F(\hat{\beta}_n, \hat{H}_n, \hat{F}_n) \Big[\int q d\hat{F}_n \Big] \Big\} + o_p(1)$$

$$= -\sqrt{n}\mathcal{P} \Big\{ v^T \mathcal{L}_\beta(\hat{\beta}_n, \hat{H}_n, \hat{F}_n) - v^T \mathcal{L}_\beta(\beta_0, H_0, F_0) + \mathcal{L}_H(\hat{\beta}_n, \hat{H}_n, \hat{F}_n) \Big[\int p d\hat{H}_n \Big]$$

$$- \mathcal{L}_H(\beta_0, H_0, F_0) \Big[\int p dH_0 \Big] + \mathcal{L}_F(\hat{\beta}_n, \hat{H}_n, \hat{F}_n) \Big[\int q d\hat{F}_n \Big]$$

$$- \mathcal{L}_F(\beta_0, H_0, F_0) \Big[\int q dF_0 \Big] \Big\} + o_p(1)$$

$$= -\sqrt{n}\mathcal{P} \Big\{ \frac{v^T \Psi_{j\beta}(\hat{\beta}_n, \hat{H}_n, \hat{F}_n)}{\Psi_j(\hat{\beta}_n, \hat{H}_n, \hat{F}_n)} - \frac{v^T \Psi_{j\beta}(\beta_0, H_0, F_0)}{\Psi_j(\beta_0, H_0, F_0)} + \frac{\Psi_{jH}(\hat{\beta}_n, \hat{H}_n, \hat{F}_n)[\int p d\hat{H}]}{\Psi_j(\hat{\beta}_n, \hat{H}_n, \hat{F}_n)}$$

$$- \frac{\Psi_{jH}(\beta_0, H_0, F_0)[\int p dH_0]}{\Psi_j(\beta_0, H_0, F_0)} + \frac{\Psi_{jF}(\hat{\beta}_n, \hat{H}_n, \hat{F}_n)[\int q d\hat{F}]}{\Psi_j(\hat{\beta}_n, \hat{H}_n, \hat{F}_n)}$$

$$- \frac{\Psi_{jF}(\beta_0, H_0, F_0)[\int q dF_0]}{\Psi_j(\beta_0, H_0, F_0)} \Big\} + o_p(1). \tag{A.2}$$

For the continuous linear functional from $BV[0, \tau]$ to $\mathbb{R}$:

$$p \mapsto \mathcal{P} \Big(\frac{\Psi_{jH}[p]}{\Psi_j}(\beta, H, F) \Big),$$

by Theorem 4.2 in Edwards and Wayment (1971), there exists a bounded function η_H such that

$$\mathcal{P} \Big(\frac{\Psi_{jH}[p]}{\Psi_j}(\beta, H, F) \Big) = \int_0^\tau \eta_H(t; \beta, H, F) dp(t).$$

To be specific, let $p(s) = I(s \geq t)$, so that

$$\eta_H(t; \beta, H, F) = \mathcal{P} \Big(\frac{\Psi_{jH}[I(\cdot \geq t)]}{\Psi_j}(\beta, H, F) \Big).$$

Analogously, we define

$$\eta_F(z; \beta, H, F) = \mathcal{P} \Big(\frac{\Psi_{jF}[I(\cdot \geq z)]}{\Psi_j}(\beta, H, F) \Big),$$

and second order partial derivatives functions:

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

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

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

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

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

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

We also set

$$\zeta_\beta(\beta, H, F) = \frac{\partial}{\partial \beta} \mathcal{P}\left(\frac{\Psi_{j\beta}}{\Psi_j}\right)(\beta, H, F),$$

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

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

The right-hand side of (A.2) is now

$$= -\sqrt{n}\Big(B_1[v, p, q]^T(\hat{\beta}_n - \beta_0) + \int B_{21}[v, p, q]d(\hat{H}_n - H_0) + \int B_{22}[v, p, q]d(\hat{F}_n - F_0)\Big)$$

$$+ o\Big(\sqrt{n}|\hat{\beta}_n - \beta_0| + \sqrt{n}|\hat{H}_n - H_0| + \sqrt{n}|\hat{F}_n - F_0|\Big),$$

with linear operators B_1, B_{21}, and B_{22} defined as

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

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

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

$$+ \int_0^\tau \eta_{HH}(s, t; \beta_0, H_0, F_0)p(s)dH_0(s)$$

$$+ \int_M \eta_{FH}(x, t; \beta_0, H_0, F_0)q(x)dF_0(x);$$

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

$$+ \int_0^\tau \eta_{HF}(s, z; \beta_0, H_0, F_0)p(s)dH_0(s)$$

$$+ \int_M \eta_{FF}(x, z; \beta_0, H_0, F_0)q(x)dF_0(x).$$

The operator $[B_1, B_{21}, B_{22}]$ can then be written as

$$B_1[v, p, q]^T \tilde{v} + \int B_{21}[v, p, q]\tilde{p}dH_0 + \int B_{22}[v, p, q]\tilde{q}dF_0$$

$$= \frac{d}{d\epsilon}\Big|_{\epsilon=0} \mathcal{P}\Big\{v^T \mathcal{L}_\beta(\beta_0 + \epsilon\tilde{v}, H_0 + \epsilon\int \tilde{p}dH_0, F_0 + \epsilon\int \tilde{q}dF_0)$$

$$+ \mathcal{L}_H(\beta_0 + \epsilon\tilde{v}, H_0 + \epsilon\int \tilde{p}dH_0, F_0 + \epsilon\int \tilde{q}dF_0)\Big[\int pdH_0\Big]$$

$$+ \mathcal{L}_F(\beta_0 + \epsilon\tilde{v}, H_0 + \epsilon\int \tilde{p}dH_0, F_0 + \epsilon\int \tilde{q}dF_0)\Big[\int qdF_0\Big]\Big\}.$$

We now show that (B_1, B_{21}, B_{22}) is continuously invertible. By the Open Mapping Theorem, we need only 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

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

where the right-hand side is the derivative of

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

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

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

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

$$\begin{pmatrix} v \\ \eta_H(t; \beta_0, H_0, F_0)p(t) \\ \eta_F(z; \beta_0, H_0, F_0)q(z) \end{pmatrix}$$

and $I_2[v, p, q]$ is defined as

$$\begin{pmatrix} v^T \zeta_\beta(\beta_0, H_0, F_0) + \int_0^\tau \eta_{H\beta}(t; \beta_0, H_0, F_0)p(t)dH_0(t) \\ \qquad + \int_M \eta_{F\beta}(z; \beta_0, H_0, F_0)q(z)dF_0(z) - v \\ v^T \zeta_H(t; \beta_0, H_0, F_0) + \int_0^\tau \eta_{HH}(s, t; \beta_0, H_0, F_0)p(s)dH_0(s) \\ \qquad + \int_M \eta_{FH}(x, t; \beta_0, H_0, F_0)q(x)dF_0(x) \\ v^T \zeta_F(z; \beta_0, H_0, F_0) + \int_0^\tau \eta_{HF}(s, z; \beta_0, H_0, F_0)p(s)dH_0(s) \\ \qquad + \int_M \eta_{FF}(x, z; \beta_0, H_0, F_0)q(x)dF_0(x) \end{pmatrix}.$$

Since $\eta_H(t; \beta_0, H_0, F_0) = \mathcal{P}(\mathcal{L}_H(\beta_0, H_0, F_0)[I(\cdot \geq t)])$, the score function along the direction of $H_0 + \epsilon I(\cdot \geq t)$, $\eta_H(t; \beta_0, H_0, F_0)$ is negative and continuous for all t. Similarly, $\eta_F(z; \beta_0, H_0, F_0)$ is the score function along the direction $F_0 + \epsilon I(\cdot \geq z)$, 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, and is continuously invertible.

By the 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 the Arzelá-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, and so is $I_1 + I_2$. 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 (A.2) we have

$$\sqrt{n}\left\{ v^T(\hat{\beta}_n - \beta_0) + \int pd(\hat{H}_n - H_0) + \int qd(\hat{F}_n - F_0) \right\}$$

$$= -\sqrt{n}(\mathcal{P}_n - \mathcal{P})\left\{ \tilde{v}^T \mathcal{L}_\beta(\beta_0, H_0, F_0) + \mathcal{L}_H(\beta_0, H_0, F_0)\left[\int \tilde{p}dH_0 \right] \right. \tag{A.3}$$

$$\left. + \mathcal{L}_F(\beta_0, H_0, F_0)\left[\int \tilde{q}dF_0 \right] \right\} + o\left(\sqrt{n}|\hat{\beta}_n - \beta_0| + \sqrt{n}|\hat{H}_n - H_0| + \sqrt{n}|\hat{F}_n - F_0| \right).$$

The first term of the right side of (A.3) converges in distribution to a zero-mean Gaussian process in the metric space $\mathbb{R}^d \times L^\infty(\mathcal{Q}_1) \times L^\infty(\mathcal{Q}_2)$. By Slutsky's theorem, we now need only show that

$$\sqrt{n}|\hat{\beta}_n - \beta_0| + \sqrt{n}|\hat{H}_n - H_0| + \sqrt{n}|\hat{F}_n - F_0| = O_p(1). \tag{A.4}$$

By definition,

$$|\hat{\beta}_n - \beta_0| + |\hat{H}_n - H_0| + |\hat{F}_n - F_0|$$
$$= \sup_{(v,p,q)\in\mathcal{V}\times\mathcal{Q}_1\times\mathcal{Q}_2} \left| v^T(\hat{\beta}_n - \beta_0) + \int_0^\tau pd(\hat{H}_n - H_0) + \int_Z qd(\hat{F}_n - F_0) \right|,$$

so (A.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 (A.4) holds immediately.

Now we have

$$\sqrt{n}\left\{ v^T(\hat{\beta}_n - \beta_0) + \int pd(\hat{H}_n - H_0) + \int qd(\hat{F}_n - F_0) \right\}$$

$$= -\sqrt{n}(\mathcal{P}_n - \mathcal{P})\Big\{\tilde{v}^T \mathcal{L}_\beta(\beta_0, H_0, F_0) + \mathcal{L}_H(\beta_0, H_0, F_0)\Big[\int \tilde{p}dH_0\Big]$$

$$+\mathcal{L}_F(\beta_0, H_0, F_0)\Big[\int \tilde{q}dF_0\Big]\Big\} + o_p(1). \tag{A.5}$$

Since the right side of (A.5) converges to normal 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. If $p = q = 0$ in equality (A.5), then the estimate $\hat{\beta}_n$ is an asymptotically linear estimator with influence function $\tilde{v}^T \mathcal{L}_\beta(\beta_0, H_0, F_0) + \mathcal{L}_H(\beta_0, H_0, F_0)[\int \tilde{p}dH_0] + \mathcal{L}_F(\beta_0, H_0, F_0)[\int \tilde{q}dF_0]$, which lies in the linear space spanned by the score functions: $\{\tilde{v}^T \mathcal{L}_\beta + \mathcal{L}_H[\int \tilde{p}dH] + \mathcal{L}_F[\int \tilde{q}dF] : \tilde{v} \in \mathbb{R}^d, \tilde{p} \in \mathcal{Q}_1, \tilde{q} \in \mathcal{Q}_2\}$. By Proposition 1 in Bickel et al. (1993), the estimate $\hat{\beta}_n$ is semi-parametrically efficient.

Proof of Theorem 3. Let $\tilde{H}_n(t)$ be a step function with a jump size $\tilde{h}_n(s_i) = H_0(s_i) - \max_{s_j < s_i} H_0(s_j)$ at each failure time s_i, and $\tilde{F}_n(z)$ be a step function with a jump size $\tilde{f}_n(Z_i) = F_0(Z_i) - \max_{Z_{j_1} \leq Z_i, \dots, Z_{j_d} \leq Z_i} F_0(Z_{j_1}(1), \dots, Z_{j_d}(d))$ at each observed covariate Z_i, where $Z_k(m)$ is the value on the mth dimension of the covariate Z_k, and the order $Z_k < Z_i$ is taken as $Z_k(m) < Z_i(m)$ for all $1 \leq m \leq d$. Then $\tilde{H}_n(s_i) = H_0(s_i)$ and $\tilde{F}_n(Z_i) = F_0(Z_i)$ holds for every s_i and Z_i.

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

$$\begin{pmatrix} v_1 \\ \int p_1 d\hat{H}_n \\ \int q_1 d\hat{F}_n \end{pmatrix} = I_n^{-1} \begin{pmatrix} v \\ \vec{p} \\ \vec{q} \end{pmatrix},$$

where $I_n, \vec{p}$, and $\vec{q}$ are as defined in Theorem 3 and

$$\int p_1 d\hat{H}_n = (p_1(s_1)\hat{h}_n(s_1), \dots, p_1(s_{n_1})\hat{h}_n(s_{n_1}))^T,$$

$$\int q_1 d\hat{F}_n = (q_1(Z_1)\hat{f}_n(Z_1), \dots, q_1(Z_{n_2})\hat{f}_n(Z_{n_2}))^T.$$

Let

$$\hat{h}_n - \tilde{h} = (\hat{h}_n(s_1) - \tilde{h}(s_1), \dots, \hat{h}_n(s_{n_1}) - \tilde{h}(s_{n_1}))^T$$

and

$$\hat{f}_n - \tilde{f} = (\hat{f}_n(Z_1) - \tilde{f}(Z_1), \dots, \hat{f}_n(Z_{n_2}) - \tilde{f}(Z_{n_2}))^T,$$

then we have

$$\sqrt{n}v^T(\hat{\beta}_n - \beta_0) + \sqrt{n}\sum_{i=0}^{n_1} p(s_i)(\hat{h}_n(s_i) - \tilde{h}(s_i)) + \sqrt{n}\sum_{i=0}^{n_2} q(Z_i)(\hat{f}_n(Z_i) - \tilde{f}(Z_i))$$

$$= \sqrt{n} \begin{pmatrix} \hat{\beta}_n - \beta_0 \\ \hat{h}_n - \tilde{h} \\ \hat{f}_n - \tilde{f} \end{pmatrix}^T \begin{pmatrix} v \\ \vec{p} \\ \vec{q} \end{pmatrix}$$

$$= \sqrt{n} \begin{pmatrix} \hat{\beta}_n - \beta_0 \\ \hat{h}_n - \tilde{h} \\ \hat{f}_n - \tilde{f} \end{pmatrix}^T I_n \begin{pmatrix} v_1 \\ \int p_1 d\hat{H}_n \\ \int q_1 d\hat{F}_n \end{pmatrix}$$

$$= \sqrt{n} \mathcal{P}_n \begin{pmatrix} \mathcal{L}_{\beta\beta} & \mathcal{L}_{\beta H} & \mathcal{L}_{\beta F} \\ \mathcal{L}_{H\beta} & \mathcal{L}_{HH} & \mathcal{L}_{HF} \\ \mathcal{L}_{F\beta} & \mathcal{L}_{FH} & \mathcal{L}_{FF} \end{pmatrix} (\hat{\beta}_n, \hat{H}_n, \hat{F}_n) \left[\begin{pmatrix} v_1 \\ \int p_1 d\hat{H}_n \\ \int q_1 d\hat{F}_n \end{pmatrix}, \begin{pmatrix} \hat{\beta}_n - \beta_0 \\ \hat{H}_n - \tilde{H}_n \\ \hat{F}_n - \tilde{F}_n \end{pmatrix} \right]$$

$$= \sqrt{n} \mathcal{P} \begin{pmatrix} \mathcal{L}_{\beta\beta} & \mathcal{L}_{\beta H} & \mathcal{L}_{\beta F} \\ \mathcal{L}_{H\beta} & \mathcal{L}_{HH} & \mathcal{L}_{HF} \\ \mathcal{L}_{F\beta} & \mathcal{L}_{FH} & \mathcal{L}_{FF} \end{pmatrix} (\beta_0, H_0, F_0) \left[\begin{pmatrix} v_1 \\ \int p_1 dH_0 \\ \int q_1 dF_0 \end{pmatrix}, \begin{pmatrix} \hat{\beta}_n - \beta_0 \\ \hat{H}_n - H_0 \\ \hat{F}_n - F_0 \end{pmatrix} \right] + o_p(1)$$

$$= -\sqrt{n}(\mathcal{P}_n - \mathcal{P}) \left\{ v_1{}^T \mathcal{L}_\beta(\beta_0, H_0, F_0) + \mathcal{L}_H(\beta_0, H_0, F_0) \left[\int p_1 dH_0 \right] \right.$$

$$\left. + \mathcal{L}_F(\beta_0, H_0, F_0) \left[\int q_1 dF_0 \right] \right\} + o_p(1),$$

where the last equality is from the proof of Theorem 2.

Note that

$$= -\sqrt{n}(\mathcal{P}_n - \mathcal{P}) \left\{ v_1{}^T \mathcal{L}_\beta(\beta_0, H_0, F_0) + \mathcal{L}_H(\beta_0, H_0, F_0) \left[\int p_1 dH_0 \right] \right.$$

$$\left. + \mathcal{L}_F(\beta_0, H_0, F_0) \left[\int q_1 dF_0 \right] \right\}$$

$$= -\sqrt{n}(\mathcal{P}_n - \mathcal{P}) \left\{ v_1{}^T \mathcal{L}_\beta(\beta_0, H_0, F_0) + \mathcal{L}_H(\beta_0, H_0, F_0) \left[\int p_1 d\hat{H} \right] \right.$$

$$\left. + \mathcal{L}_F(\beta_0, H_0, F_0) \left[\int q_1 d\hat{F} \right] \right\} + o_p(1),$$

and the variance of the right side is consistently estimated by

$$- \mathcal{P} \begin{pmatrix} \mathcal{L}_{\beta\beta} & \mathcal{L}_{\beta H} & \mathcal{L}_{\beta F} \\ \mathcal{L}_{H\beta} & \mathcal{L}_{HH} & \mathcal{L}_{HF} \\ \mathcal{L}_{F\beta} & \mathcal{L}_{FH} & \mathcal{L}_{FF} \end{pmatrix} (\beta_0, H_0, F_0) \left[\begin{pmatrix} v_1 \\ \int p_1 d\hat{H}_n \\ \int q_1 d\hat{F}_n \end{pmatrix}, \begin{pmatrix} v_1 \\ \int p_1 d\hat{H}_n \\ \int q_1 d\hat{F}_n \end{pmatrix} \right].$$

This can be further estimated by

$$- \mathcal{P}_n \begin{pmatrix} \mathcal{L}_{\beta\beta} & \mathcal{L}_{\beta H} & \mathcal{L}_{\beta F} \\ \mathcal{L}_{H\beta} & \mathcal{L}_{HH} & \mathcal{L}_{HF} \\ \mathcal{L}_{F\beta} & \mathcal{L}_{FH} & \mathcal{L}_{FF} \end{pmatrix} (\hat{\beta}_n, \hat{H}_n, \hat{F}_n) \left[\begin{pmatrix} v_1 \\ \int p_1 d\hat{H}_n \\ \int q_1 d\hat{F}_n \end{pmatrix}, \begin{pmatrix} v_1 \\ \int p_1 d\hat{H}_n \\ \int q_1 d\hat{F}_n \end{pmatrix} \right]$$

$$= \begin{pmatrix} v_1 \\ \int p_1 d\hat{H}_n \\ \int q_1 d\hat{F}_n \end{pmatrix}^T I_n \begin{pmatrix} v_1 \\ \int p_1 d\hat{H}_n \\ \int q_1 d\hat{F}_n \end{pmatrix}$$

$$= (v^T, \vec{p}^T, \vec{q}^T) I_n^{-1} I_n I_n^{-1} (v^T, \vec{p}^T, \vec{q}^T)^T$$

$$= (v^T, \vec{p}^T, \vec{q}^T) I_n^{-1} (v^T, \vec{p}^T, \vec{q}^T)^T.$$

The proof of Theorem 3 is complete.

References

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* John Hopkins University Press, Baltimore.

Breslow, N. E. (1996). Statistics in epidemiology: The case-control studies. *J. Amer. Statist. Assoc.* **91**, 14-28.

Breslow, N. E. and Day, N. E. (1980). Statistical methods in cancer research. *Vol.I, The Design and Analysis of Case-Control Studies.* International Agency for Research on Cancer, Lyon.

Chen, K. (2001). Generalized case-cohort sampling. *J. Roy. Statist. Soc. Ser. B* **63**, 791-809.

Chen, K., Jin, Z. and Ying, Z. (2002). Semiparametric analysis of transformation models with censored data. *Biometrika* **89**, 659-668.

Chen, K., Sun, L. and Tong, X. (2012). Analysis of cohort survival data with transformation model. *Statist. Sinica* **22**, 489-508.

Cheng, S. C., Wei, L. J. and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82**, 835-45.

Dabrowska, D. M. and Doksum, K. A. (1988). Estimation and testing in the two-sample generalized odds rate model. *J. Amer. Statist. Assoc.* **83**, 744-749.

Edwards, J. R. and Wayment, S. G. (1971). Representations for transformations continuous in the *BV* norm. *Trans. Amer. Math. Soc.* **154**, 251-265.

Goldstein, L. and Langholz, B. (1992). Asymptotic theory for nested case-control sampling in the Cox regression model. *Ann. Statist.* **20**, 190-1928.

Jin, Z., Lin, D. Y. and Ying, Z. (2006). On least-squares regression with censored data. *Biometrika* **93**, 147-161.

Langholz, B. and Goldstein, L. (1996). Risk set sampling in epidemiologic cohort studies. *Statist. Sci.* **11**, 35-53.

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

Samuelsen, S. (1997). A pseudo-likelihood approach to analysis of nested case-control data. *Biometrika* **84**, 379-394.

Thomas, D. C. (1977). Appendum to "Methods of cohort analysis: appraisal by application to asbestos mining," by Liddell, F. D. K., McDonald, J. C., and Thomas, D. C. *J. Roy. Statist. Soc. Ser. A* **140**, 469-490.

Van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press, Cambridge, U.K.

Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Kowloon, Hong Kong.

E-mail: yaoyuan@hkbu.edu.hk