Open Journal of Statistics, 2016, 6, 133-155
Published Online February 2016 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2016.61013

Transformation Models for Survival Data Analysis with Applications

Yang Liu[1], Qiusheng Chen[2], Xufeng Niu[2]

[1]Senior Biometrician, Merck Research Laboratories, Rahway, NJ, USA
[2]Department of Statistics, Florida State University, Tallahassee, FL, USA
Email: yang.liu2@merck.com, niu@stat.fsu.edu

Received 21 December 2015; accepted 22 February 20146; published 25 February 2016

Abstract

When the event of interest never occurs for a proportion of subjects during the study period, survival models with a cure fraction are more appropriate in analyzing this type of data. Considering the non-linear relationship between response variable and covariates, we propose a class of generalized transformation models motivated by Zeng *et al.* [1] transformed proportional time cure model, in which fractional polynomials are used instead of the simple linear combination of the covariates. Statistical properties of the proposed models are investigated, including identifiability of the parameters, asymptotic consistency, and asymptotic normality of the estimated regression coefficients. A simulation study is carried out to examine the performance of the power selection procedure. The generalized transformation cure rate models are applied to the First National Health and Nutrition Examination Survey Epidemiologic Follow-up Study (NHANES1) for the purpose of examining the relationship between survival time of patients and several risk factors.

Keywords

Link Functions, Mixture Cure Rate Models, Noninformative Improper Priors, Proportional Hazards Models, Proportional Odds Models

1. Introduction

Survival data analysis is an important topic in statistics that focuses on analyzing the expected duration of time until one or more events occur, such as death or cancer in a targeted population. In a standard survival model, it is often assumed that all uncensored subjects will eventually experience the event of interest, which is described by a monotone decreasing survival function $S(t)$. The function $S(t)$ goes to 0 when time t tends to infinity. Survival time T is a continuous nonnegative random variable representing the time of an event. The probability

How to cite this paper: Liu, Y., Chen, Q.S. and Niu, X.F. (2016) Transformation Models for Survival Data Analysis with Applications. *Open Journal of Statistics*, **6**, 133-155. http://dx.doi.org/10.4236/ojs.2016.61013

of a subject's surviving till time t is given by $S(t)=1-F(t)=P(T>t)$, where $F(t)$ is the distribution function with probability density $f(t)$. The hazard function $h(t)$ is defined as the instantaneous failure rate at time t conditional on survival until time t or later. The cumulative hazard $H(t)$ is defined as $H(t)=\int_0^t h(s)\mathrm{d}s$, which represents the total amount of risk up to time t. Usually covariates, such as gender, age, weight, blood pressure, heart rate, stage of surgery, etc., are modeled through survival models. In this paper, we assume that the covariates are independent of time.

Cox [2] brought the idea of separating time t and individual covariate vector $\boldsymbol{x}'=\{x_1,\cdots,x_p\}$ in the hazard function, which led to the popular proportional hazard model with

$$h(t,x)=h_0(t)\exp(\beta' x),$$

where $h_0(t)$ was the baseline hazard function and β was a vector of regression coefficients.

However, in some situations, the event of interest never occurs for a significant proportion of subjects. For example, in a cancer clinical trial, the endpoint of interest is often recurrence. For some patients, the disease will never relapse after being treated. These patients are considered cured. Sometimes, subjects with long-term censored times can be viewed as "cured" as well. Survival models with a cure fraction are very popular in analyzing this type of cancer clinical trials.

Motivated by the transformed proportional time cure model introduced by Zeng *et al*. [1], we propose a class of generalized transformation models to characterize the non-linear relationship between survival function $S(t)$ and relate covariates. Statistical properties of the proposed models are investigated, which include identifiability, asymptotic consistency, and asymptotic normality of the estimated regression coefficients. Powers of fractional polynomials within the proposed models are selected based on the likelihood function. A simulation study is carried out to examine the performance of the power selection procedure. The generalized transformation cure rate models are applied to coronary heart disease and cancer related medical data from both observational cohort studies and clinical trials.

The first cure rate model is the mixture cure rate model proposed by Berkson and Gage [3], which combines the cured and non-cured populations by using a summation function. In their model, the survival function for the entire population, denoted by $S_1(t)$, is given by

$$S_1(t)=\pi+(1-\pi)S_2(t),$$

where π is the proportion in the cured group and $S_2(t)$ is the survival function for the non-cured group in the entire population. Notice that $S_1(t)$ is not a proper survival function since $S_1(\infty)=\pi>0$. This mixture model has been fully discussed by many authors, including Farewell [4], Gary and Tsiatis [5], Sposto *et al*. [6], Laska and Meisner [7], Sy and Taylor [8], and Lu and Ying [9].

Even though the mixture model introduced by Berkson and Gage [3], is attractive and widely used, it has several drawbacks. One of them is that the mixture model cannot have a proportional hazards structure if the covariates are modeled through π. Ibrahim *et al*. [10] also pointed out that a mixture model sometimes yields improper posterior distribution when noninformative improper priors are used from the Bayesian point of view. Yakovlev and Tsodikov [11], Tsodikov [12], Chen *et al*. [13], and Zeng *et al*. [1] proposed and studied promotion time cure model. Instead of dividing the population into two sub-populations so that some subjects are long-term survivors with probability π and others have a proper survival function $S(t)$ with probability $1-\pi$, the promotion time cure model takes long-term survivors into account by putting a restriction on the cumulative hazard function $H(t)$. In general, the population survival function $S(t)$ is represented as $S(t)=\exp(-H(t))$. However, in a cure rate model the function $S(t)$ is improper in the sense of $S(\infty)=\pi>0$, which also implies that $H(t)$ is bounded by some positive number, say θ. When t goes to ∞, we have $\lim_{t\to\infty}H(t)=\theta$. Tsodikov [12] suggested to consider $H(t)=\theta F(t)$, where $F(t)$ is the distribution function of a nonnegative random variable with $F(0)=0$ and covariates can be modeled through $\theta(.)$,

$$S(t)=\exp(-\theta F(t)). \tag{1.1}$$

The promotion time cure model avoids the drawbacks of a mixture model and has a proportional hazards structure through the cure rate parameter. Chen *et al*. [13] also proposed classes of noninformative and informative priors for promotion time cure rate model that lead to proper posterior distributions.

The promotion time cure rate model and the mixture cure rate model are linked by a mathematical relation-

ship, and can be rewritten in a uniform format. Zeng *et al*. [1] proposed a general promotion time cure model with transformation. Their model includes proportional hazards model and proportional odds model as special cases. To take into account the unknown and unobservable risk factor for each individual, they used a subject-specific frailty variable ξ_i, $i=1,\cdots,n$ in model (1.1). The survival function for the time to relapse is given by

$$S\left(t|\boldsymbol{X}_i,\xi_i\right)=\exp\left(-\theta\left(\boldsymbol{X}_i\right)F\left(t\right)\xi_i\right). \tag{1.2}$$

Different parametric distributions may be applied to the frailty ξ_i. The most commonly used one is the gamma distribution $\Gamma\left(1/\gamma,\gamma\right)$, where $\gamma\geq 0$. The mean of the gamma distribution needs to be one due to the model identification issue. Taking expectations with respect to ξ_i on both sides in (1.2), the survival function becomes

$$S\left(t|\boldsymbol{X}_i\right)=E_{\xi_i}\left[\exp\left(-\theta\left(\boldsymbol{X}_i\right)F\left(t\right)\xi_i\right)\right]=\left(1+\gamma\theta\left(\boldsymbol{X}_i\right)F\left(t\right)\right)^{-1/\gamma}. \tag{1.3}$$

As Zeng *et al*. [1] pointed out, (1.3) provides a very wide class of transformation cure models with the form:

$$S\left(t|\boldsymbol{X}_i\right)=G_\gamma\left(\theta\left(\boldsymbol{X}_i\right)F\left(t\right)\right), \tag{1.4}$$

where

$$G_\gamma\left(x\right)=\begin{cases}\left(1+\gamma x\right)^{-1/\gamma}, & \gamma>0,\\ \mathrm{e}^{-x}, & \gamma=0.\end{cases} \tag{1.5}$$

When ξ_i takes other distributions, we may get different transformations. A Box-Cox type transformation is also considered in Zeng *et al*. [1] with

$$G_\gamma\left(x\right)=\begin{cases}\exp\left\{-\dfrac{\left(1+x\right)^\gamma-1}{\gamma}\right\}, & \gamma>0,\\ \dfrac{1}{1+x}, & \gamma=0.\end{cases} \tag{1.6}$$

The proportional hazards model in (1.1) is a special case of the transformation families (1.5) and (1.6) corresponding to $\gamma=0$ and $\gamma=1$, respectively. Another popular survival model, the proportional odds model, is also a special case of (1.5) and (1.6) when $\gamma=1$ and $\gamma=0$, respectively.

From model (1.4) the cure fraction is $S\left(\infty\right)=G_\gamma\left(\theta F\left(\infty\right)\right)=G_\gamma\left(\theta\right)$, and the model can be written as a standard cure rate model,

$$S\left(t\right)=G_\gamma\left(\theta\right)+\left(1-G_\gamma\left(\theta\right)\right)S^*\left(t\right),$$

where $S^*\left(t\right)$ is the survival function for the non-cured population,

$$S^*\left(t\right)=\frac{G_\gamma\left(\theta F\left(t\right)\right)-G_\gamma\left(\theta\right)}{1-G_\gamma\left(\theta\right)}.$$

The covariates can be modeled through a known and strictly positive increasing link function $\theta\left(\boldsymbol{X}_i\right)=\eta\left(\beta'\boldsymbol{X}_i\right)$, where β is the regression vector including an intercept term.

In this paper, we extend the transformed proportional time cure model proposed by Zeng *et al*. [14] to a more general class of transformation models, in which fractional polynomials are used instead of the simple linear combination of the covariates. The statistical properties of our proposed models will be investigated. Estimation and model selection procedures will be discussed. The rest of the paper is organized as follows. In Section 2, we introduce the generalized transformation models and study the identifiability and asymptotic properties of the proposed models. In Section 3, simulation studies are conducted for the purpose of assessing the performance of the power selection procedure. In Section 4, the proposed models will be applied to some real datasets and compared with other models. Conclusions and some discussions are given in Section 5. Proofs of the theorems in Section 2 are provided in **Appendix**.

2. Proposed Models and Their Properties

In survival data analysis, the relationship between hazard rates and covariates is quite often nonlinear. Motivated

by Zeng *et al.* [1], we propose a generalized transformation cure model by using a general additive function of $\boldsymbol{X}_i=\left(\boldsymbol{X}_{i1},\cdots,\boldsymbol{X}_{ip}\right)$ instead of the strictly positive increasing link function $\theta\left(\boldsymbol{X}_i\right)=\eta\left(\beta'\boldsymbol{X}_i\right)$.

The additive models were introduced by Stone [15], which is defined by $E\left[Y\middle|\boldsymbol{X}\right]=\sum_{j=0}^{p}f_j\left(X_j\right)$, where $f_0\left(.\right)$ is a constant term and $f_1\left(.\right),\cdots,f_p\left(.\right)$ are arbitrary univariate functions, one for each covariate. Additive models retain the important additive feature of the linear regression models and are much more flexible to use in practice. Royston and Altman [14] suggested using fractional polynomials for each $f_j\left(X_j\right)$, which is a family of functions of positive covariates. For simplicity, let us consider a single covariate X first. A fractional polynomial with degree m is described as $E\left[Y\middle|X\right]=\beta_0+\sum_{i=1}^{m}\beta_i f_i\left(X\right)$, where

$$f_i\left(X\right)=\begin{cases}X^{p_i}, & p_i\neq p_{i-1}\\ f_{i-1}\left(X\right)\log\left(X\right), & p_i=p_{i-1}\end{cases} \tag{2.1}$$

and $\boldsymbol{p}=\left(p_1,\cdots,p_m\right)$, $p_1\le p_2\le\cdots\le p_m$, is a real-valued vector of powers. If $p_i=0$ for any i, X^{p_i} is defined to be $\log\left(X\right)$ by the Box-Tidwell transformation. For example, $E\left[Y\middle|X\right]=\beta_0+\beta_1X^{-1}+\beta_2X^{-1}\log\left(X\right)+\beta_3X^{-1}\left(\log\left(X\right)\right)^2+\beta_4\log\left(X\right)+\beta_5X^2$ is a fractional polynomial with degree $m=5$ and $\boldsymbol{p}=\left(-1,-1,-1,0,2\right)$. Royston and Altman [14] pointed out that special attention should be paid to low-order fractional polynomials with degrees one and two, since models with degree higher than two are rarely used in practice. They also suggested that the powers could be chosen from the set $\left(-2,-1,-0.5,0,0.5,1,2,\cdots,\max\left(3,m\right)\right)$, since the set is rich enough to cover all conventional polynomials of interest. It is well known that the best estimates of the powers in a transformation model may be determined based on the maximum likelihood method.

For some data sets, especially data from medical studies, fractional polynomials may give a better fit compared to the conventional polynomial. In our proposed models we use a fractional polynomial instead of $\beta'X$ in the link function $\eta\left(\cdot\right)$. Even though in practice fractional polynomials with degree higher than two are not used very often, we consider the following general form for the function $\theta\left(\boldsymbol{X}\right)$ in (1.2),

$$\theta\left(X\right)=\eta\left(\sum_{j=0}^{q}\beta_jX_j+\sum_{i=q+1}^{p}\left(\beta_{i0}\frac{X_i^{\alpha_{i0}}-1}{\alpha_{i0}}+\beta_{i1}\frac{X_i^{\alpha_i}-1}{\alpha_i}+\beta_{i2}\frac{X_i^{\alpha_i}-1}{\alpha_i}\log X_i\right)\right), \tag{2.2}$$

where $\boldsymbol{X}=\left(X_1,\cdots,X_q,X_{q+1},\cdots,X_p\right)$, $\left\{X_1,\cdots,X_q\right\}$ are categorical covariates such as ordinal covariates or dummy variables, and $\left\{X_{q+1},\cdots,X_p\right\}$ are positive continuous covariates. An intercept term β_0 is also considered in (2.2) when we assume that $X_0\equiv1$. Moreover, we assume that $\alpha_{i0}\neq\alpha_i$ for $q+1\le i\le p$, *i.e.*, a degree of three fractional polynomial is used for each continuous covariate X_i. For example, for a given $q+1\le i\le p$ if $\alpha_{i0}=-1,\alpha_i=1$, the powers for predictor X_i are $\boldsymbol{p}_i=\left(-1,1,1\right)$ based on the definition in (2.1).

In a typical survival analysis setting, survival times are often right censored, which means for some subjects we do not know when exactly the failures occurred, but we do know that the survival time is at least beyond some certain time point *C*. Suppose that there are *n* right censored subjects. For the *i*th individual the survival time and the fixed censoring time are denoted by T_i and C_i, respectively. The T_i's are assumed to be independent and identically distributed with a distribution function *F*.

The observed time point for the *i*th subject is $Y_i=\min\left(T_i,C_i\right)$. The exact survival time T_i will be observed only if the failure occurred before being censored, otherwise Y_i is equal to the censoring time. A triple of random variables $\left(Y_i,X_i,\Delta_i\right)$ is used to describe each subject, where X_i is the covariate vector and Δ_i is defined as the following,

$$\Delta_i=\begin{cases}1, & T_i\le C_i\\ 0, & T_i>C_i.\end{cases}$$

In a proportional hazard model, the regression coefficient $\boldsymbol{\beta}$ is estimated by maximizing the partial likelihood function,

$$L\left(\boldsymbol{\beta}\right)\propto\prod_{i=1}^{n}\left(f\left(Y_i\right)^{\Delta_i}\left(1-F\left(Y_i\right)\right)^{1-\Delta_i}\right)$$

$$=\prod_{i=1}^{n}\left(h_0\left(Y_i\right)\exp\left(\beta'X_i\right)\right)^{\Delta_i}\prod_{i=1}^{n}\left[S\left(Y_i\right)\right]^{1-\Delta_i}.$$

In the model (1.4) with link function (2.2), if the parameter γ is given the likelihood function is expressed by

$$\begin{aligned} &L(\boldsymbol{\beta}, F) \\ &= \left[\left\{ -G'(\eta((\boldsymbol{\beta}, \boldsymbol{X}) F(Y)) \eta((\boldsymbol{\beta}, \boldsymbol{X}) f(Y) \right\}^{\Delta} \left\{ G(\eta(\boldsymbol{\beta}, \boldsymbol{X}) F(Y)) \right\}^{1-\Delta} \right]^{I(Y<\infty)} \left[G(\eta(\boldsymbol{\beta}, \boldsymbol{X})) \right]^{I(Y=\infty)}. \end{aligned} \tag{2.3}$$

Given observations $(\{Y_i, \boldsymbol{X}_i, \Delta_i\}, \ i=1, \cdots, n)$ and following the discussion in Zeng *et al.* [1], the maximum likelihood estimates of $(\boldsymbol{\beta}, F)$, denoted by $(\boldsymbol{\beta}_n, \hat{F}_n)$, are derived from the modified semi-parametric version of (2.3),

$$\begin{aligned} &L(\boldsymbol{\beta}, F) \\ &= \prod_{i=1}^{n} \left\{ \left[\left\{ -G'(\eta(\boldsymbol{\beta}, \boldsymbol{X}_i) F(Y_i)) \eta(\boldsymbol{\beta}, \boldsymbol{X}_i) F\{Y_i\} \right\}^{\Delta_i} \times \left\{ G(\eta(\boldsymbol{\beta}, \boldsymbol{X}_i) F(Y_i)) \right\}^{(1-\Delta_i)} \right]^{I(Y_i<\infty)} \times \left[G(\eta(\boldsymbol{\beta}, \boldsymbol{X}_i)) \right]^{I(Y_i=\infty)} \right\}, \end{aligned} \tag{2.4}$$

where $F\{Y_i\}$ is the jump size of F at Y_i and $F(Y_i) = \sum_{\Delta_k=1, Y_k \le Y_i} F\{Y_k\}$.

The three pieces of products in (2.4) are for failures, censored cases, and subjects who never experience failure or censoring, respectively. The estimate of $(\boldsymbol{\beta}, F)$, denoted by $(\boldsymbol{\beta}_n, \hat{F}_n)$, can be obtained by using the nonparametric maximum likelihood estimation approach and Newton-Raphson algorithm iteratively.

2.1. Model Identifiability

For the statistical properties of our proposed models, we first discuss the identifiability of generalized transformation models. Suppose that we use models (1.4) and (1.5) with the link function defined in (2.2). The observed-data likelihood function of parameters (γ, θ, F) is given by

$$\begin{aligned} &L(\gamma, \theta, F) \\ &= \left[\left(-G_\gamma'(\theta(X) F(Y)) \theta(X) f(Y) \right)^{\Delta} \left(G_\gamma(\theta(X) F(Y)) \right)^{(1-\Delta)} \right]^{I(Y<\infty)} \left[G_\gamma(\theta(X)) \right]^{I(Y=\infty)}. \end{aligned} \tag{2.5}$$

The following two lemmas give sufficient conditions of identifiability to a more general class of transformations that include the transformation (1.5) as a special case. Proofs of the lemmas are given in **Appendix**.

Lemma 1. If $G_\gamma(.)$ satisfies the following conditions:

(G1) $G_\gamma(.)$ is strictly monotonic and twice continuously differentiable with $G_\gamma(0) = 1$ and $G_\gamma'(0) \neq 0$.

(G2) If $\gamma \neq \tilde{\gamma}$, then $G_\gamma''(0) / (G_\gamma'(0))^2 \neq G_{\tilde{\gamma}}''(0) / (G_{\tilde{\gamma}}'(0))^2$.

Then $G_\gamma(\theta(x) F(y)) = G_{\tilde{\gamma}}(\tilde{\theta}(x) \tilde{F}(y))$ implies that $\gamma = \tilde{\gamma}, \theta = \tilde{\theta}$, and $F = \tilde{F}$, *i.e.*, (γ, θ, F) is identifiable.

It can be shown that the transformation family given in (1.5) satisfies both conditions (G1) and (G2). Specifically, we have $G_\gamma'(0) = -1$ and $G_\gamma''(0) / [G_\gamma'(0)]^2 = 1 + \gamma$.

Other transformation families can also be considered as long as the conditions (G1) and (G2) hold. For example, the Box-Cox type transformation discussed in Zeng *et al.* [1], also satisfies conditions (G1) and (G2) with $G_\gamma'(0) = -1$ and $G_\gamma''(0) / [G_\gamma'(0)]^2 = 2 - \gamma$.

Next, we consider the following function

$$\theta(X) = \eta(\boldsymbol{\beta}, X) = \eta\left(\sum_{j=0}^{q} \beta_j X_j + \sum_{i=q+1}^{p} f_i(X_i) \right), \tag{2.6}$$

where $\eta(\cdot)$ is strictly monotonic, $f_i(X_i) = \sum_{m,n} \beta_{imn} X_i^{p_{im}} (\log X_i)^{q_{in}}$, p_{im} and q_{in} are not equal to zeros simultaneously, and $\sum_{m,n}$ is used for a finite summation since the number of parameters in our proposed models is finite.

Function in (2.6) is a more general function than that defined in (2.2). The following lemma show that the parameters β_j's and β_{imn}'s in the function are all identifiable.

Lemma 2. For the function $\eta(\boldsymbol{\beta}, X)$ defined in (2.6), if $\eta(\boldsymbol{\beta}, X) = \eta(\boldsymbol{\beta}, X)$, then $\boldsymbol{\beta} = \boldsymbol{\beta}$, *i.e.*, parameters in $\eta(\boldsymbol{\beta}, X)$ are identifiable.

Based on the results in **Lemma 1** and **Lemma 2**, we have the following theorem on the identifiability of the generalized transformation models.

Theorem 1. For the generalized transformation models defined in (1.4) and (1.5) with the link function specified in (2.2), if $L(\gamma,\theta,F(y))=L(\tilde{\gamma},\tilde{\theta},\tilde{F}(y))$, for any $y\in R^+$ and any X, then $\gamma=\tilde{\gamma},\theta=\tilde{\theta}$, and $F=\tilde{F}$. In other words, the generalized transformation models are identifiable.

2.2. Estimation

Zeng *et al*. [1] discussed semiparametric transformation models for survival data with a cure fraction and established theorems describing the asymptotic properties of the maximum likelihood estimation of $(\boldsymbol{\beta},F)$, where $\boldsymbol{\beta}$ is the vector of coefficients and $F(.)$ is the promotion time cumulative distribution function in the model. In our proposed generalized transformation cure models, fractional polynomials are used instead of the simple linear combination of the covariates. Similar to **Theorem 1** and **Theorem 2** in Zeng *et al*. [1], we can prove the asymptotic properties of the maximum likelihood estimation of $(\boldsymbol{\beta},F)$ in the proposed models.

To obtain consistency and asymptotic normality, we make the following assumptions:

(C1) The covariate $\boldsymbol{X}$ belongs to a compact set $\mathcal{X}$.

(C2) The vector of regression coefficients $\boldsymbol{\beta}$ belongs to a compact set $\mathcal{B}_0$. The true value of $\boldsymbol{\beta}$, denoted by $\boldsymbol{\beta}_0$, belongs to the interior of set $\mathcal{B}_0$.

(C3) F is a distribution function with jumps when $\Delta=1$. The true F, denoted by F_0, is differentiable with $F_0'(x)>0$ for all $x\in R^+$.

(C4) Conditional on $\boldsymbol{X}$, the right censoring time C is independent of T, and $S_C(\infty|\boldsymbol{X})>0$.

(C5) The positive link function $\eta(.)$ is a strictly increasing and twice continuously differentiable for $\boldsymbol{X}$.

(C6) The transformation G satisfies $G(0)=1$, $G(x)>0$, $G'(x)<0$ and $G^{(3)}(x)$ exists and is continuous.

Under conditions (C1)-(C6), we can prove the following theorems.

Theorem 2. The maximum likelihood estimates $(\boldsymbol{\beta}_n,\hat{F}_n)$ based on (2.4) are strongly consistent, that is

$$|\boldsymbol{\beta}_n-\boldsymbol{\beta}_0|\to 0,\text{ and }\sup_{y\in[0,\infty)}\left|\hat{F}_n(y)-F_0(y)\right|\to 0\text{ almost surely.}$$

Theorem 3. $\sqrt{n}(\boldsymbol{\beta}_n-\boldsymbol{\beta}_0,\hat{F}_n-F_0)$ converges weakly to a Gaussian process.

Sketched proofs of Theorem 2 and Theorem 3 are provided in **Appendix**.

3. Simulations

In this section, we conduct simulations to study the empirical properties of the generalized transformation models and to examine the performance of the proposed power selection procedure on generalized transformation models. The model used in this simulation was given in Zeng *et al*. [1] and has a survival function of the form:

$$S(t|\boldsymbol{X})=G_\gamma(\theta(\boldsymbol{X})F(t)),\tag{3.1}$$

where $G_\gamma(x)$ is given in (1.5).

For the purpose of illustration, only one continuous variable X_1 and one categorical variable X_2 are considered in the simulation. Specifically, we take γ equal to zero in (1.5) and consider the following link function,

$$\theta(\boldsymbol{X})=\exp(\beta_0+\beta_1X_1^{p_0}+\beta_2X_2),\tag{3.2}$$

where p_0 is a nonzero power varying from −2 to 2. Covariate X_1 is a uniformly distributed random variable in [0.5, 2] and covariate X_2 is a Bernoulli random variable with probability 0.5. The coefficients β_0, β_1, and β_2 are assumed to be constants. When $p_0=0$, we use

$$\theta(\boldsymbol{X})=\exp(\beta_0+\beta_1\log(X_1)+\beta_2X_2).\tag{3.3}$$

$F(t)$ is a proper distribution function. We choose $F(t)=1-\exp(-t)$ in this simulation.

Survival times of subjects with covariates X_1 and X_2 are generated. Each subject has a chance of being cured. We assume the survival life times T equal to ∞ for the cured population. For example, the ith individual in the simulated data set has a cure rate equal to $\exp(-\theta(\boldsymbol{X}_i))$, which means the survival life time T_i equals ∞ with probability $\exp(-\theta(\boldsymbol{X}_i))$. Moreover, with probability $1-\exp(-\theta(\boldsymbol{X}_i))$, the survival time T_i is finite and follows the distribution $1-S(t|\boldsymbol{X}_i)$, where $S(.|\boldsymbol{X}_i)$ is the generalized transformation model given in (3.1) and (1.5). Therefore, the life time T will be generated from

$$T_i=-\log\left(1+\frac{\log\left(1-\left(1-\exp(-\theta(\boldsymbol{X}_i))\right)U\right)}{\theta(\boldsymbol{X}_i)}\right) \tag{3.4}$$

with probability $1-\exp(-\theta(\boldsymbol{X}_i))$, where U has a uniform distribution in [0, 1].

Assume each subject being right-censored with a probability $q<1$, for example q = 80%. So, the censoring time C_i for the ith individual in the data set will equal ∞ with a 20% of chance. For the rest of the population, the censoring time is generated from an exponential distribution with mean one.

The complete data set $\{(Y_i,\boldsymbol{X}_i,\Delta_i),i=1,\cdots,n\}$ is given by

$$\begin{aligned} Y_i&=\min(T_i,C_i),\\ \boldsymbol{X}_i&=(X_{i1},X_{i2})',\\ \Delta_i&=\begin{cases}0, & T_i>C_i,\\ 1, & T_i\le C_i, T_i\neq\infty,\\ 2, & T_i=\infty, C_i=\infty.\end{cases}\end{aligned} \tag{3.5}$$

The whole population is categorized into three groups: right-censoring events when $\Delta=0$, failure events when $\Delta=1$, and cured population when $\Delta=2$.

The coefficients β_0, β_1, and β_2 in model (3.2) are arbitrary constants. We set $\beta_0=-0.5$, $\beta_1=1$, and $\beta_2=0.7$. As p_0 changes from −2 to 2, the cured proportions vary from 5% to 10%. For each simulated date set, we choose a $\hat{p}_0$ from the set A = (−2, −1.5, −1, 0.5, 0, 0.5, 1, 1.5, 2) based on the likelihood function given in (2.3).

Table 1 shows the power selection results under the proposed generalized transformation model based on 200 simulated data sets with q = 80% and sample sizes 2000 and 5000, respectively. The columns labeled "mean" are the average of the selected powers and the columns labeled "freq." are the number of times of selecting the

Table 1. Results of power selection under the proposed generalized transformation model based on 200 simulated data sets with coefficients $\beta_0=-0.5$, $\beta_1=1$, $\beta_2=0.7$, and the probability of each subject being right-censored $q=80\%$.

	n = 2000		n = 5000	
p_0	mean	freq.	mean	freq.
−2	−1.748	127	−1.845	144
−1.5	−1.488	61	−1.448	91
−1	−0.995	75	−1.005	104
−0.5	−0.488	66	−0.528	109
0	−0.045	69	−0.030	110
0.5	0.555	74	0.478	110
1	0.960	90	1.015	115
1.5	1.500	84	1.508	125
2	1.865	151	1.909	163

true power in the 200 simulations. When the sample size is 5000, the power selection procedure work well. The accurate rates of choosing the true power are higher than 50% and the means of the selected powers are very close to the true value for most of the cases. For example, when $p_0 = -1$ the true power is selected for 104 times and the estimated mean is −1.005. When the sample size decreases to 2000, the power selection results are less accurate. For both sample sizes, the accurate rates are higher when $p_0 = -2$ or $p_0 = 2$ than other cases since we select the powers only in the range of −2 to 2. This also explains why the absolute values of means of the selected powers when $p_0 = -2$ or $p_0 = 2$ tends to be smaller. If powers beyond −2 and 2 are allowed to be selected, $p_0 = -2$ or $p_0 = 2$ should have less chance to be underestimated.

Table 2 presents more results on power selection with $n = 5000$ and q = 80% based on 200 simulations. In the table each column represents one scenario. For example, when $p_0 = -1$, the true power −1 is selected 104 times; Powers −1.5 and −0.5 are selected 44 and 42 times, respectively; and powers −1 and 0 are selected 5 times each. These results indicates that the selected powers are all centering around the true power.

In this simulation we assume that the probability of each subject being censored is q = 80%. In fact, the probability q basically does not affect the performance of the power selection procedure. When q takes different values while other factors in the simulation remain the same, the power selection results show a very similar pattern as that when q = 80%.

4. Applications

In this section, we will illustrate the applications of the proposed generalized transformation models and compare the proposed models with the Cox proportional hazards model and the Zeng *et al.* [1] transformation cure model by analyzing data from the First National Health and Nutrition Examination Survey Epidemiologic Follow-up Study (NHANES1). The NHANES1 data set is from the Diverse Populations Collaboration (DPC), which is a pooled database contributed by a group of investigators to examine issues of heterogeneity of results in epidemiological studies. The database includes 21 observational cohorts studies, 3 clinical trials, and 3 national samples. In the dataset NHANES1, information for 14,407 individuals was collected in four cohorts from 1971 to 1992. In this analysis, we use data from two of the four cohorts, the black female cohort and the black male cohort. After dropping all missing observations, a total of 2027 patients remains in these two cohorts, including 1265 black females and 762 black males. Survival times of the 2027 patients are used as the response variable. The endpoint is the overall survival time collected in 1992. In the two cohorts 848 patients, about 40% of the total number of patients, died at the end of followup with a maximum survival life time of 7691 days. There were 1179 patients whose survival times were right censored, among them 115 patients had survival time longer than 7691 days. We consider these 115 patients as cured subjects.

Covariates selected by fitting the Cox model and using the stepwise backward elimination algorithm will be

Table 2. Results of power selection under the proposed generalized transformation model based on 200 simulated data sets with sample size n = 5000 and the probability of each subject being right-censored q = 80%.

	True power								
Selected power	−2	−1.5	−1	−0.5	0	0.5	1	1.5	2
−2	144	46	5						
−1.5	50	91	44	3					
−1	6	59	104	48	4				
−0.5		4	42	109	47	3			
0			5	37	110	48	1		
0.5				3	35	110	38		
1					4	33	115	36	
1.5						6	46	125	37
2								39	163

included to compare different survival models. These covariates are Age, Systolic blood pressure (Sbp), Sex, Body Mass Index (BMI), Diabetes (Diab), and Coronary heart disease (Chd). Summary statistics of continuous covariates are list in **Table 3**. Diab and Chd are categorical and only take the values of 0 and 1 for absence and presence of the corresponding disease. Among the 2027 patients in the two cohorts, there were 121 of them having diabetes and 82 of them having coronary heart disease.

The results of the Cox proportional hazard model are summarized in **Table 4**. All covariates are highly significant at the $\alpha = 0.05$ level. The results show that males have a higher hazard rate than females and older patients have a higher hazard rate than younger patients. People with diabetes or coronary heart disease face a higher hazard rate than people who did not have such disease. The hazard of death increases by 0.4% when the Sbp level of a patient increases 1 mmHg. The results also show that the higher the value of BMI of a patient the lower the hazard rate she/he will face. Particularly, the hazard will decrease about 1.2% when the value of BMI increases by 1 kg/m^2, which is not quite reasonable. The values of BMI often ranges from 15 kg/m^2 to 60 kg/m^2. BMI in the range of 21 kg/m^2 to 25 kg/m^2 is considered as normal weight; 30 kg/m^2 or greater is considered as obesity. It is well known that being obesity will increase the hazard to develop many coronary heart diseases or even death. The relationship between survival time and BMI may not be linear. Therefore, a transformation on the covariate BMI may be needed for the NHANES1 data.

A transformation of $\gamma = 0$ is chosen with maximum likelihood from Zeng *et al*. [1] model with transformation family (1.5). The observed log-likelihood is shown in **Figure 1** with different values of γ. The corresponding estimates of regression coefficients are summarized in **Table 5**. The results are comparable with that in the Cox proportional hazards model.

There are three continuous covariates in our analysis, Age, BMI, and Sbp. The main relationship of interest is between mortality and the factor BMI. In the next step, we will focus on choosing an appropriate power from the set $A = (-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2)$ for BMI within our proposed models. To do so, we fit models

$$S(t|\boldsymbol{X}) = G_\gamma\left(\theta(\boldsymbol{\beta}, \boldsymbol{X})F(t)\right),$$

with link function $\theta(\boldsymbol{\beta}, \boldsymbol{X}) = \exp(\boldsymbol{\beta}, \boldsymbol{X})$. In stead of using the linear terms as in Zeng *et al*. [1] models, we use the following four expressions in the function $\theta(\boldsymbol{\beta}, \boldsymbol{X})$:

$$\theta(\boldsymbol{\beta}, \boldsymbol{X}) = \beta_0 + \beta_1 Age + \beta_2 Sbp + \beta_3 Sex + \beta_4 BMI^{p_{01}} + \beta_5 Diab + \beta_6 Chd, \quad (4.1)$$

$$\theta(\boldsymbol{\beta}, \boldsymbol{X}) = \beta_0 + \beta_1 Age^{p_{02}} + \beta_2 Sbp + \beta_3 Sex + \beta_4 BMI + \beta_5 Diab + \beta_6 Chd, \quad (4.2)$$

Table 3. Summary statistics of continuous covariates in the NHANES1 study.

Variable	Min	Max	Mean	Std.Dev.
Age	25	75	50.12	15.55
BMI	15.07	72.31	26.98	6.11
Sbp	85	266	142.35	28.21

Table 4. Fitted Cox proportional hazards model for the NHANES1 study.

Variable	Coef.	Std. Err.	z	$Prob > \lvert z \rvert$
Age	0.020	0.002	10.31	0.000
Sbp	0.004	0.001	4.31	0.000
Sex	0.275	0.050	5.46	0.000
BMI	−0.012	0.005	−2.65	0.008
Diab	0.299	0.104	2.87	0.004
Chd	0.724	0.126	5.76	0.000

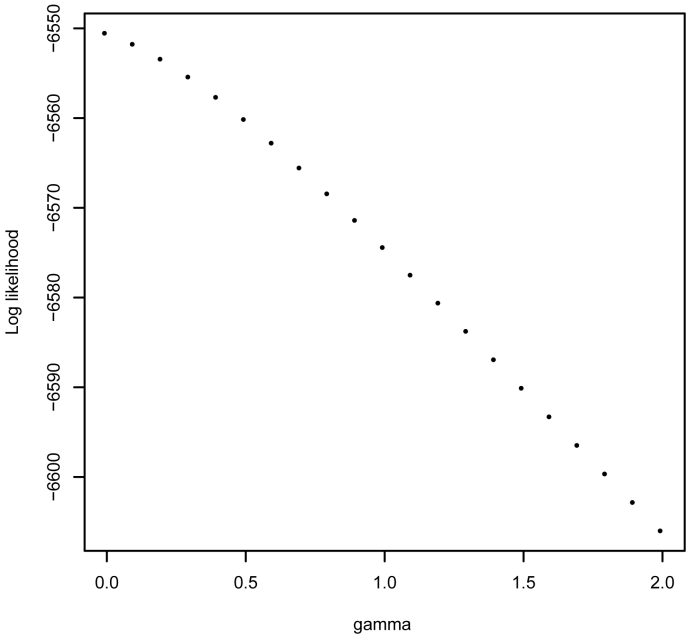

Figure 1. Log-likelihood in Zeng *et al*. [1] model from transformation (1.5) with different γ for the NHANES1 study.

Table 5. Estimates of regression coefficients in Zeng *et al*. [1] model based on transformation class (1.5) with $\gamma = 0$ for the NHANES1 study.

Variable	Coef	Std. Err.	$Prob > \lvert z \rvert$
Intercept	−4.580	0.290	0.000
Age	0.062	0.003	0.000
Sbp	0.008	0.001	0.000
Sex	0.488	0.072	0.000
BMI	−0.020	0.007	0.004
Diab	0.633	0.113	0.000
Chd	0.692	0.129	0.000

$$\theta(\boldsymbol{\beta},\boldsymbol{X})=\beta_0+\beta_1 Age^{p_{03}}+\beta_2 Sbp+\beta_3 Sex+\beta_4 BMI^{-1}+\beta_5 Diab+\beta_6 Chd, \tag{4.3}$$

$$\theta(\boldsymbol{\beta},\boldsymbol{X})=\beta_0+\beta_1 Age^{p_{04}}+\beta_2 Sbp+\beta_3 Sex+\beta_4 BMI^{-2}+\beta_5 Diab+\beta_6 Chd. \tag{4.4}$$

In model (4.1), when we fix Age and Sbp, power $p_{01}=-2$ is selected for BMI. The observed log-likelihood is plotted in **Figure 2(a)**. In the next model (4.2), we fix BMI and Sbp, trying to find a transformation for Age. Power $p_{02}=1$ is selected based on the log-likelihood, which is plotted in **Figure 2(b)**. The selected model corresponds to Zeng *et al*'s model. In many statistical models, the inverse of BMI, BMI^{-1}, lean body mass index is used. So we fit a model (4.3) where BMI^{-1} and Sbp are fixed. In model (4.4), BMI^{-2} and Sbp fixed. Both model (4.3) and model (4.4) select power=1 for Age. The results are plotted in **Figure 2(c)** and **Figure 2(d)**. As a summary, the best transformation based on log-likelihood from model (4.1)-(4.4) is

$$\theta(\boldsymbol{\beta},\boldsymbol{X})=\beta_0+\beta_1 Age+\beta_2 Sbp+\beta_3 Sex+\beta_4 BMI^{-2}+\beta_5 Diab+\beta_6 Chd. \tag{4.5}$$

The corresponding estimates of regression coefficients are listed in **Table 6**.

Now let us compare the Cox model, Zeng *et al*. [1] models, and the proposed models by using the Brier score. The Brier score was originally proposed by Brier [16] to verify the accuracy of weather forecasts and then

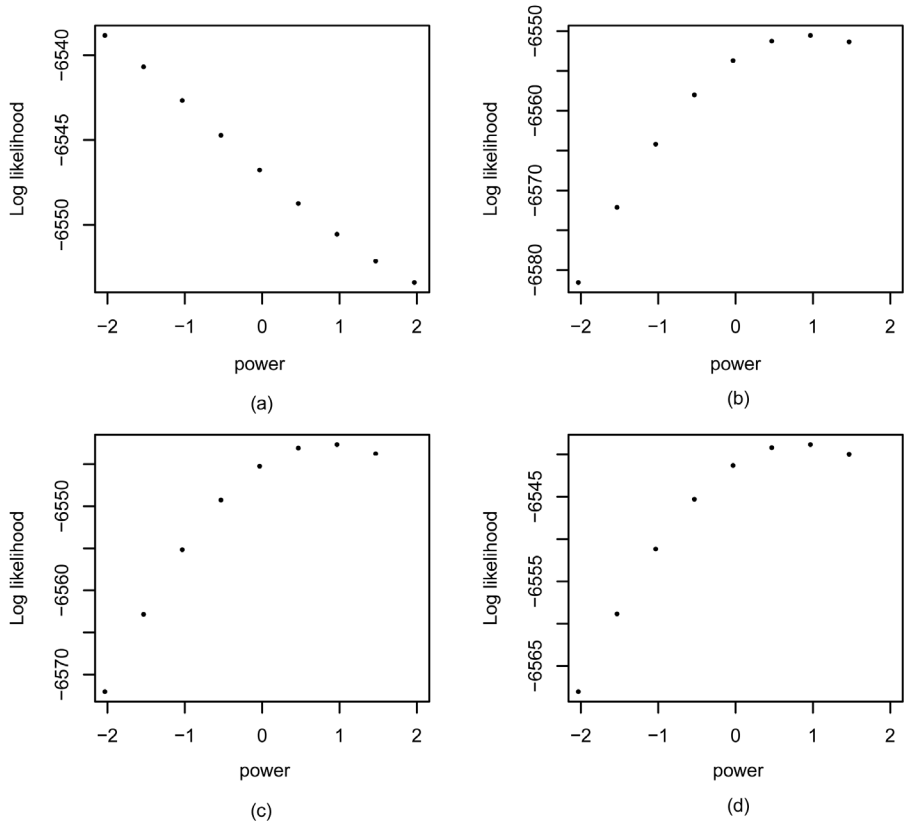

Figure 2. Log-likelihood and selected power in the proposed models from transformation (1.5) for the NHANES1 study. (a) Model (4.1), (b) Model (4.2), (c) Model (4.3), (d) Model (4.4).

Table 6. Estimates of regression coefficients in the proposed models (4.1) based on transformation class (1.5) with $\gamma = 0$ and transformation on BMI ($p_0 = -2$) for the NHANES1 study.

Variable	Coef	Std. Err.	$Prob > \lvert z \rvert$
Intercept	−5.665	0.256	0.000
Age	0.062	0.003	0.000
Sbp	0.008	0.001	0.000
Sex	0.473	0.071	0.000
BMI	328.498	56.203	0.000
Diab	0.646	0.113	0.000
Chd	0.726	0.129	0.000

extended by May *et al*. [17] to survival models. The Brier score (*BS*) at time t^* is given by

$$BS\left(t^*\right) = \frac{\sum_{i=1}^{n}\left(I\left(Y_i > t^*\right) - \hat{S}\left(t^* \middle| \boldsymbol{X}_i\right)\right)^2}{n}, \tag{4.6}$$

where *n* is the total sample size, Y_i is the observed survival time of the *i*th patient, $I\left(Y_i > t^*\right)$ is the indicator function representing the occurrence of the event, and $\hat{S}\left(t^* \middle| \boldsymbol{X}_i\right)$ is the predicted probability of the *i*th patient surviving beyond time t^*. The choices of the time t^* can be arbitrary, such as the quartiles of follow up time, the quartiles of the survival time, or a fixed number of years.

Table 7. Brier scores for different survival models for the NHANES1 study.

t^* in days	Cox Model	Zeng *et al.*'s Model	Proposed Model
Q1 = 2089.5	0.0851	0.0824	0.0815
Q2 = 3894.5	0.1396	0.1308	0.1297
Q3 = 5498.75	0.1890	0.1855	0.1838

It is obvious that the Brier score takes minimum value of 0 for perfect prediction of survival status and its range is from 0 to 1. The lower the value of the Brier score, the better the prediction. To compare the Cox model, Zeng *et al*. [1] models, and proposed models, we calculated the Brier scores at the first quartile Q1, median Q2, and last quartile Q3 of 848 uncensored survival times in the NHANES1 study. The results are summarized in **Table 7**. We can see that the proposed models has the smallest Brier scores at all there time points. For example, at the median uncensored survival time Q2 = 3894.5 days, the Brier score is 0.1396 for the Cox model. It is 0.1308 for Zeng *et al*. [1] model. The value of Brier score drops to 0.1297 for the selected proposed model, which indicates the chosen proposed model can well predict the survival outcome as the other two models, and sometimes better.

5. Conclusions and Discussion

In this paper, we proposed a class of generalized transformation models. Zeng *et al*. [14] introduced semiparametric transformation models for survival data with a cure fraction, which included the commonly used proportional hazards cure rate models and proportional odds models as special cases. Similar to the structure suggested in Zeng *et al*. [1], covariates related to the event of interest were modeled through a link function $\theta(\boldsymbol{X}) = \eta(\boldsymbol{\beta}' X)$, where $\eta(.)$ was a known and strictly positive increasing function, such as exponential functions. In our proposed models, we used generalized additive models instead of $\boldsymbol{\beta}' X$ in the link function $\eta(.)$. Specifically, we considered fractional polynomials proposed by Royston and Altman [14]. We proved that the proposed model was identifiable as long as the transformation families $G_\gamma(.)$ to satisfy some very general conditions. To select transformation powers in fractional polynomials, we proposed choosing powers from set *A* = (−2, −1.5, −1, 0.5, 0, 0.5, 1, 1.5, 2) by comparing likelihood functions. Simulation results showed the power selection procedure works well. An improvement in this direction could consider the power as a parameter and estimate the power by using maximum likelihood methods rather than selecting the power from set *A*.

The proposed generalized transformation models can be applied to a variety of survival data. Even though the cure models are motivated from clinical trials where the end point is not death, such as relapse-free survival time, it can be used to overall survival time as well. In this article, the applications of the proposed models are illustrated by examining the relationship between the survival time of a patient and several risk factors based on two cohorts data from the First National Health and Nutrition Examination Survey Epidemiologic Follow-up Study. In terms of the Brier scores, the selected proposed model provides better fitting compared with the Cox proportional hazards model and the Zeng *et al*. [1] transformation cure model. It should be pointed out that even though the Brier score is commonly used in practice for model comparison, it has its own disadvantages. For instance, although the Brier score can be calculated at any arbitrary time point, but it dose not discriminate competing models over the whole time period. Other model comparison methodologies will be explored in our future study. For example, receiver operating characteristic (ROC) curves may be used to measure the differences of the models over all the relevant time periods.

Acknowledgements

The authors would like to thank Dr. Donglin Zeng from the University of North Carolina at Chapel Hill for sharing his original MATLAB code with us, and to thank the Diverse Populations Collaboration Group for providing data from their studies. The authors would also like to thank the Editor, the Associate Editor, and the referees for their insightful comments and suggestions that provide guidelines for the authors to revise the paper.

References

[1] Zeng, D., Yin, G. and Ibrahim, J.G. (2006) Semiparametric Transformation Models for Survival Data with a Cure

Fraction. *Journal of the American Statistical Association*, **101**, 670-684. http://dx.doi.org/10.1198/016214505000001122

[2] Cox, D.R. (1972) Regression Models and Life-Tables. *Journal of the Royal Statistical Society*, **34**, 187-220. http://www.jstor.org/stable/2985181?seq=1#page_scan_tab_contents

[3] Berkson, J. and Cage, R.P. (1952) Survival Curve for Cancer Patients Following Treatment. *Journal of the American Statistical Association*, **47**, 501-505. http://dx.doi.org/10.1080/01621459.1952.10501187

[4] Farewell, V.T. (1982) The Use of Mixtures Models for the Analysis of Survival Data with Long-Term Survivors. *Biometrics*, **38**, 1041-1046. http://dx.doi.org/10.2307/2529885

[5] Gary, R.J. and Tsiatis, A.A. (1989) A Linear Rank Test for Use When the Main Interest Is in Differences in Cure Rates. *Biometrics*, **45**, 899-904. http://dx.doi.org/10.2307/2531691

[6] Sposto, R., Sather, H.N. and Baker, S.A. (1992) A Comparison of Tests of the Difference in the Proportion of Patients Who Are Cured. *Biometrics*, **48**, 87-99. http://dx.doi.org/10.2307/2532741

[7] Laska, E.M. and Meisner, M.J. (1992) Nonparametric Estimation and Testing in a Cure Model. *Biometrics*, **48**, 1223-1234. http://dx.doi.org/10.2307/2532714

[8] Sy, J.P. and Taylor, J.M.G. (2000) Estimation in a Cox Proportional Hazards Cure Model. *Biometrics*, **56**, 227-236. http://dx.doi.org/10.1111/j.0006-341X.2000.00227.x

[9] Lu, W. and Ying, Z. (2004) On Semiparametric Transformation Cure Models. *Biometrika*, **91**, 331-343. http://dx.doi.org/10.1093/biomet/91.2.331

[10] Ibrahim, J.G., Chen, M.-H. and Sinha, D. (2001) Bayesian Survival Analysis. Springer Series in Statistics, First Edition, Springer, New York. http://dx.doi.org/10.1007/978-1-4757-3447-8

[11] Yakovlev, A.Y. and Tsodikov, A.D. (1996) Stochastic Models of Tumor Latency and Their Biostatistical Application. First Edition, World Scientific, Singapore.

[12] Tsodikov, A. (1998) A Proportional Hazards Model Taking Account of Long-Term Survivors. *Biometrics*, **54**, 1508-1516. http://dx.doi.org/10.2307/2533675

[13] Chen, M.-H., Ibrahim, J.G. and Sinha, D. (1999) A New Bayesian Model for Survival Data with a Surviving Fraction. *Journal of the American Statistical Association*, **94**, 909-919. http://dx.doi.org/10.1080/01621459.1999.10474196

[14] Royston, P. and Altman, D.G. (1994) Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modeling. *Applied Statistics*, **43**, 429-467. http://dx.doi.org/10.2307/2986270

[15] Stone, C.J. (1985) Additive Regression and Other Nonparametric Models. *Annals of Statistics*, **13**, 689-705. http://dx.doi.org/10.1214/aos/1176349548

[16] Brier, G.W. (1950) Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, **78**, 1-3. http://dx.doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

[17] May, M., Roystion, P., Egger, M., Justice, A.C. and Sterne, J.A. (2004) Development and Validation of a Prognostic Model for Survival Time Data: Application to Prognosis of HIV Positive Patients Treated with Antiretroviral Therapy. *Statistics in Medicine*, **23**, 2375-2398. http://dx.doi.org/10.1002/sim.1825

[18] van der Vaart, A.W. and Wellner, J.A. (1996) Weak Convergence and Empirical Processes. Springer Series in Statistics, First Edition, Springer, New York. http://dx.doi.org/10.1007/978-1-4757-2545-2

Appendix: Proofs of the Main Results

In **Appendix**, we first prove the Lemmas on model identifiability listed in Section 2.1. Then we will show the asymptotic properties of the semi-parametric estimates in the proposed models under conditions (C1)-(C6) given in Section 2.2. Proofs of the **Theorems 2** and **3** are similar to those of **Theorems 1** and **2** in Zeng *et al*. [14] with some modifications.

a. Proofs of Model Identifiability

Proof of Lemma 1: Suppose that $\theta(X)$ can take two different non-zero values α_1 and α_2, such that

$$\theta(x_1)=\alpha_1, \theta(x_2)=\alpha_2,$$
$$\tilde{\theta}(x_1)=\beta_1, \tilde{\theta}(x_2)=\beta_2,$$

then we will have the following two equations about $G(\cdot)$,

$$\begin{aligned} G_\gamma\left(\alpha_1 F(y)\right) &= G_{\tilde{\gamma}}\left(\beta_1 \tilde{F}(y)\right), \\ G_\gamma\left(\alpha_2 F(y)\right) &= G_{\tilde{\gamma}}\left(\beta_2 \tilde{F}(y)\right). \end{aligned} \tag{A.1}$$

The inverse function of $G_{\tilde{\gamma}}(\cdot)$ exists because of the monotonicity of $G_{\tilde{\gamma}}(\cdot)$. Applying $G_{\tilde{\gamma}}^{-1}(\cdot)$ to the above we get,

$$\begin{aligned} G_{\tilde{\gamma}}^{-1} \circ G_\gamma\left(\alpha_1 F(y)\right) &= \beta_1 \tilde{F}(y), \\ G_{\tilde{\gamma}}^{-1} \circ G_\gamma\left(\alpha_2 F(y)\right) &= \beta_2 \tilde{F}(y). \end{aligned} \tag{A.2}$$

We want to show that $g(\cdot)=G_{\tilde{\gamma}}^{-1}\circ G_\gamma(\cdot)$ is an identity function. Function $g(\cdot)$ is monotonic since both $G_{\tilde{\gamma}}(\cdot)$ and $G_{\tilde{\gamma}}^{-1}(\cdot)$ are monotonic, which implies that both β_1 and β_2 can not be zero. Otherwise $g(y)\equiv 0$ when y takes different values. Take the ratio of the two equations in (A.2) and let $s=F(y)$. The following equation holds for $s\in[0,1]$,

$$g(\alpha_1 s)=\frac{\beta_1}{\beta_2} g(\alpha_2 s). \tag{A.3}$$

Suppose that $\gamma\neq\tilde{\gamma}$ and the conditions (G1) and (G2) hold. Noting that $G_{\tilde{\gamma}}\left(g(x)\right)=G_\gamma(x)$, we have

$$\begin{aligned} g'(0) &= \left.\frac{G_\gamma'(x)}{G_{\tilde{\gamma}}'\left(G_{\tilde{\gamma}}^{-1}\left(G_\gamma(x)\right)\right)}\right|_{x=0} = \frac{G_\gamma'(0)}{G_{\tilde{\gamma}}'(0)} \neq 0 \\ g''(0) &= \left.\frac{G_\gamma''(x)\left(G_{\tilde{\gamma}}'\left(g(x)\right)\right)^2-\left(G_\gamma'(x)\right)^2 G_{\tilde{\gamma}}''\left(g(x)\right)}{\left(G_{\tilde{\gamma}}'\left(g(x)\right)\right)^3}\right|_{x=0} \\ &= \frac{G_\gamma''(0)(G_{\tilde{\gamma}}'(0))^2-\left(G_\gamma'(0)\right)^2 G_{\tilde{\gamma}}''(0)}{\left(G_{\tilde{\gamma}}'(0)\right)^3} \neq 0. \end{aligned} \tag{A.4}$$

Calculating the first and second order derivatives in both sides of (3), plugging in $s=0$, and taking ratio of the two equations, we will have $\alpha_1=\alpha_2$. This contradiction leads to $\gamma=\tilde{\gamma}$. This concludes that $g(\cdot)$ is an identity function. Therefore, $\theta(X)F(y)=\tilde{\theta}(X)\tilde{F}(y)$. Letting $y\to\infty$, we get $\theta(X)=\tilde{\theta}(X)$ and therefore $F(y)=\tilde{F}(y)$. □

Proof of Lemma 2: Suppose that $\eta(\boldsymbol{\beta},X)=\eta(\boldsymbol{\beta},X)$. Since $\eta(\cdot)$ is a strictly monotonic function, we have $\sum_{j=0}^{q}\beta_j X_j+\sum_{i=q+1}^{p} f_i(X_i)=\sum_{j=0}^{q}\tilde{\beta}_j X_j+\sum_{i=q+1}^{p}\tilde{f}_i(X_i)$. Now, let's fix $X_1,\cdots,X_q$ and $X_{q+2},\cdots,X_p$ for example, and only consider $f_{q+1}\left(X_{q+1}\right)$, where X_{q+1} is a continuous covariate,

$$\begin{aligned}&\sum_{m,n}\beta_{q+1,mn}X_{q+1}^{p_{q+1,m}}\left(\log X_{q+1}\right)^{q_{q+1,n}}+M\\&=\sum_{m,n}\tilde{\beta}_{q+1,mn}X_{q+1}^{\tilde{p}_{q+1,m}}\left(\log X_{q+1}\right)^{\tilde{q}_{q+1,n}}+\tilde{M}.\end{aligned}\tag{A.5}$$

Without loss of generality, assume that $p_{q+1,m}=\tilde{p}_{q+1,m}$ and $q_{q+1,n}=\tilde{q}_{q+1,n}$, since we can always add more terms with coefficients zero to both sides of (A.5).

Let $p_{q+1,0}=q_{q+1,0}=0$ and $\beta_{q+1,00}=M,\tilde{\beta}_{q+1,00}=\tilde{M}$, we have the following equation,

$$\sum_{m,n}\left(\beta_{q+1,mn}-\tilde{\beta}_{q+1,mn}\right)X_{q+1}^{p_{q+1,m}}\left(\log X_{q+1}\right)^{q_{q+1,n}}=0.\tag{A.6}$$

Because the function in the left side on (A.6) is analytic in some interval $I\in R^{+}$, it holds for any $X_{q+1}\in R^{+}$. For different $p_{q+1,m}$ or $q_{q+1,n}$, $X_{q+1}^{p_{q+1,m}}\left(\log X_{q+1}\right)^{q_{q+1,n}}$'s have different orders when $X_{q+1}\to\infty$. But since their summation is always zero, the coefficients for each term must be zero. Therefore, we have $f_{q+1}(.)=\tilde{f}_{q+1}(.)$. Similarly, we can prove that $f_j(\cdot)=\tilde{f}_j(\cdot)$ for $q+1\le j\le p$.

To prove the identifiability of the coefficient of a categorical covariate, for example β_1, fixing $X_2,\cdots,X_p$ we have $\beta_1X_1+M=\tilde{\beta}_1X_1+\tilde{M}$. Coefficient β_1 is identifiable if X_1 can take at least two different values.

Thus all parameters in $\eta(\boldsymbol{\beta},X)$ are identifiable. □

b. Proofs of Strong Consistency of the Maximum Likelihood Estimates

Let $\boldsymbol{E}_n$ be the empirical measure of *n* iid observations and *E* be the expectation, respectively. For any measurable function $g(\Delta,Y,X)$, define $\boldsymbol{E}_n\left(g(\Delta,Y,X)\right)=\frac{1}{n}\sum_{i=1}^{n}g(\Delta_i,Y_i,X_i)$. Suppose that there are *n* independent right censored observations. For the *i*th observation, we have $\{Y_i,\boldsymbol{X}_i,\Delta_i\}$, $i=1,\cdots,n$, where

$$Y_i=\min(T_i,C_i),\tag{A.7}$$

$$\Delta_i=\begin{cases}0, & T_i>C_i,\\1, & T_i\le C_i.\end{cases}\tag{A.8}$$

In applications we may use

$$\Delta_i=\begin{cases}0, & T_i>C_i,\\1, & T_i\le C_i,T_i\ne\infty,\\2, & T_i=\infty,C_i=\infty.\end{cases}$$

to differ the cured and uncured population. which will not affect the proof of consistency and asymptotic normality of the maximum likelihood estimates.

The modified semi-parametric version observed-data likelihood function of parameters $(\boldsymbol{\beta},F)$, denoted by $L(\boldsymbol{\beta},F)$, is given in (2.4). Let $\boldsymbol{\beta}_n$, $\left(\hat{F}_n\{Y_i\}i=1,2,\cdots,n\right)$ be the estimates of $\boldsymbol{\beta}$ and F such that $L(\boldsymbol{\beta},F)$ reaches its maximum. The log likelihood function $l(\boldsymbol{\beta},F)$ is given by

$$\begin{aligned}l(\boldsymbol{\beta},F)=&\sum_{j=1}^{n}\Delta_j\left\{\log\left(-G'\left(\eta\left(\boldsymbol{\beta},X_j\right)F\left(Y_j\right)\right)\right)+\log\eta\left(\boldsymbol{\beta},X_j\right)+\log F\{Y_j\}\right\}I\left(Y_j<\infty\right)\\&+\sum_{j=1}^{n}\left(1-\Delta_j\right)\left\{\log G\left(\eta\left(\boldsymbol{\beta},X_j\right)F\left(Y_j\right)\right)\right\}I\left(Y_j<\infty\right)\\&+\sum_{j=1}^{n}\left\{\log G\left(\eta\left(\boldsymbol{\beta},X_j\right)\right)\right\}I\left(Y_j=\infty\right),\end{aligned}\tag{A.9}$$

where $\left(F\{Y_i\},i=1,\cdots,n\right)$ satisfy the restricted condition $G(F)=\sum_{i=1}^{n}F\{Y_i\}=1$ with $F\{Y_i\}>0$ when $\Delta_i=1$ and $F\{Y_i\}=0$ when $\Delta_i=0$. When $\Delta_i=1$ for $i\in\{1,2,\cdots,n\}$, we have

$$\frac{\partial l\left(\left(\boldsymbol{\beta}_n, \hat{F}_n\right)\right.}{\partial F_n\left\{Y_i\right\}} = n\hat{\lambda}_n \frac{\partial G\left(\hat{F}_n\right)}{\partial F_n\left\{Y_i\right\}}, \tag{A.10}$$

by the method of Lagrange multipliers, where $\hat{\lambda}_n$ is the Lagrange multiplier. That is,

$$\frac{1}{\hat{F}_n\left\{Y_i\right\}} + nH_n\left(Y_i, \boldsymbol{\beta}_n, \hat{F}_n\right) = n\hat{\lambda}_n, \tag{A.11}$$

for $\Delta_i = 1$, where

$$\begin{aligned} H_n\left(y, \hat{\beta}_n, \hat{F}_n\right) = \frac{1}{n}\sum_{Y_j<\infty}\Bigg\{ & \Delta_j \frac{G''\left(\eta\left(\boldsymbol{\beta}_n, X_j\right)\hat{F}_n\left(Y_j\right)\right)\eta\left(\boldsymbol{\beta}_n, X_j\right)}{G'\left(\eta\left(\boldsymbol{\beta}_n, X_j\right)\hat{F}_n\left(Y_j\right)\right)} I\left(Y_j \geq y\right) \\ & + \left(1-\Delta_j\right)\frac{G'\left(\eta\left(\boldsymbol{\beta}_n, X_j\right)\hat{F}_n\left(Y_j\right)\right)\eta\left(\boldsymbol{\beta}_n, X_j\right)}{G\left(\eta\left(\boldsymbol{\beta}_n, X_j\right)\hat{F}_n\left(Y_j\right)\right)} I\left(Y_j \geq y\right)\Bigg\}. \end{aligned} \tag{A.12}$$

Equation (A.11) can be written as

$$\hat{F}_n\left\{Y_i\right\} = \frac{\Delta_i}{n\left(\hat{\lambda}_n - H_n\left(Y_i, \boldsymbol{\beta}_n, \hat{F}_n\right)\right)} \tag{A.13}$$

for i $i \in \{1, 2, \cdots, n\}$ when $\Delta_i = 1$. When $\Delta_i = 0$, we have the same expression for $\hat{F}_n\left\{Y_i\right\}$ considering $\hat{F}_n\left\{Y_i\right\} = 0$. Therefore,

$$\hat{\lambda}_n = \frac{1}{n}\sum_{i=1}^{n}\Delta_i I\left(Y_i < \infty\right) + \int_0^{\infty} H_n\left(y, \boldsymbol{\beta}_n, \hat{F}_n\right)\mathrm{d}\hat{F}_n\left(y\right). \tag{A.14}$$

Since $H_n(.)$ is bounded, the sequence $\left\{\hat{\lambda}_n, n = 1, 2, \cdots\right\}$ is also bounded. Thus we can choose a subsequence from $\hat{\lambda}_n$ such that $\hat{\lambda}_n \to \lambda^*$ almost surely; choose a further subsequence of $\boldsymbol{\beta}_n$ such that $\boldsymbol{\beta}_n \to \beta^*$ almost surely since $\boldsymbol{\beta}_n$ belong to a compact set $\mathcal{B}_0$; choose a subsequence of $\hat{F}_n$ such that $\hat{F}_n \to F^*$ pointwise. Notice that F^* is monotone and $F^*(\infty) = \lim_{x\to\infty}\left(F^*(x)\right) \leq 1$. Later we will prove $F^*(\infty) = 1$ and F^* is a proper distribution function.

The structure of the limit function F^* can be derived from the results of **Lemmas 3** and **4**. In particular, **Lemma 3** shows the convergence of $H_n\left(y, \beta_n, \hat{F}_n\right)$. Proof of the lemma was given in Zeng *et al.* [1].

Lemma 3. Under conditions (C1)-(C6), $H_n\left(y, \boldsymbol{\beta}_n, \hat{F}_n\right) \to^{a.s.} H^*\left(y\right)$ uniformly in *y*, where

$$\begin{aligned} & H^*\left(y\right) \\ & = E\left\{\Delta\frac{G''\left(\eta\left(\boldsymbol{\beta}^*, \boldsymbol{X}\right)F^*\left(Y\right)\right)\eta\left(\boldsymbol{\beta}^*, \boldsymbol{X}\right)I\left(\infty > Y \geq y\right)}{G'\left(\eta\left(\boldsymbol{\beta}^*, \boldsymbol{X}\right)F^*\left(Y\right)\right)} + \left(1-\Delta\right)\frac{G'\left(\eta\left(\boldsymbol{\beta}^*, \boldsymbol{X}\right)F^*\left(Y\right)\right)\eta\left(\boldsymbol{\beta}^*, \boldsymbol{X}\right)I\left(\infty > Y \geq y\right)}{G\left(\eta\left(\boldsymbol{\beta}^*, \boldsymbol{X}\right)F^*\left(Y\right)\right)}\right\}. \end{aligned} \tag{A.15}$$

Actually, the right hand side of Equation (A.14) converges to $\lambda^* = E\left(\Delta I\left(Y < \infty\right)\right) + E\left(I\left(Y < \infty\right)\int_0^Y H^*\left(y\right)\mathrm{d}F^*\left(y\right)\right)$. For the difference $\lambda^* - H^*\left(y\right)$, we have the following result.

Lemma 4. Under conditions (C1)-(C6), $\lambda^* - H^*\left(y\right) > 0$ for $0 \leq y < \infty$, and $E\left(\frac{\Delta I\left(Y<\infty\right)}{\lambda^* - H^*\left(Y\right)}\right) \leq 1$.

Proof: Because $\sum_{i=1}^{n}\hat{F}_n\left\{Y_i\right\} = 1$, we have

$$\begin{aligned} 1 &= \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i I\left(Y_i<\infty\right)}{\hat{\lambda}_n - H_n\left(Y_i,\boldsymbol{\beta}_n,\hat{F}_n\right)} \\ &\geq \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i I\left(Y_i<\infty\right)}{\left|\hat{\lambda}_n - H_n\left(Y_i,\boldsymbol{\beta}_n,\hat{F}_n\right)\right|+\varepsilon} \\ &\geq \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i I\left(Y_i<\infty\right)}{\left|\lambda^* - H^*\left(Y_i\right)\right|+\varepsilon}+\frac{o_n(1)}{\varepsilon^2}. \end{aligned} \tag{A.16}$$

Letting $n\to\infty$, then $\varepsilon\to 0$, we obtain

$$1\geq E\left(\frac{\Delta I\left(Y<\infty\right)}{\left|\lambda^* - H^*\left(Y\right)\right|}\right). \tag{A.17}$$

We then calculate the right hand side of (A.17) by using conditional expectations.

$$\begin{aligned} E\left(\frac{\Delta I\left(Y<\infty\right)}{\left|\lambda^* - H^*\left(Y\right)\right|}\right) &= E_T\left(E_X\left(E_C\left(\frac{\Delta I\left(T<\infty\right)}{\left|\lambda^* - H^*\left(T\right)\right|}\middle| T,X\right)\middle| T\right)\right) \\ &= E_T\left(E_X\left(\frac{S_c\left(T\middle| X\right)I\left(T<\infty\right)}{\left|\lambda^* - H^*\left(T\right)\right|}\middle| T\right)\right) \\ &= \int_0^\infty \frac{E_X\left(-S_c\left(T\middle| X\right)G'\left(\eta\left(\beta_0,X\right)F_0\left(T\right)\right)\eta\left(\beta_0,X\right)\right)}{\left|\lambda^* - H^*\left(T\right)\right|} f_0\left(T\right)\mathrm{d}T \\ &= \int_0^\infty \frac{k\left(T\right)f_0\left(T\right)}{\left|\lambda^* - H^*\left(T\right)\right|}\mathrm{d}T, \end{aligned} \tag{A.18}$$

where

$$k\left(T\right) = E_X\left(-S_c\left(T\middle| X\right)G'\left(\eta\left(\beta_0,X\right)F_0\left(T\right)\right)\eta\left(\beta_0,X\right)\right). \tag{A.19}$$

Function $t\left(T\right) = -S_c\left(T\middle| X\right)G'\left(\eta\left(\beta_0,X\right)F_0\left(T\right)\right)\eta\left(\beta_0,X\right)$ is positive and continuous on $\left[0,\infty\right)$. When $T\to\infty$, $t\left(\infty\right) = -S_c\left(\infty\middle| X\right)G'\left(\eta\left(\beta_0,X\right)\right)\eta\left(\beta_0,X\right)$ exists and is positive. Therefore there exists positive constants c_0, c_1 such that $c_0\leq t\left(T\right)\leq c_1$, for any T. Hence $c_0\leq k\left(T\right)\leq c_1$ for any T. Combining (A.17) and (A.18), we then have $1\geq c_0\int_0^\infty \frac{f_0\left(T\right)}{\left|\lambda^* - H^*\left(T\right)\right|}\mathrm{d}T$.

It can be shown that $H^*\left(y\right)$ is Lipshitz continuous and $\lambda^* - H^*\left(T\right)\neq 0$ for any $T\in\left[0,\infty\right)$. Because of the continuity of $H^*\left(y\right)$ and $\hat{F}_n\left\{Y_i\right\} = \frac{\Delta_i}{n\left(\hat{\lambda}_n - H_n\left(Y_i,\hat{\beta}_n,\hat{F}_n\right)\right)} > 0$ for $\Delta_i = 1$, we have $\hat{\lambda}_n - H_n\left(Y_i,\hat{\beta}_n,\hat{F}_n\right) > 0$. Therefore, for any *i* we have

$$\frac{1}{n}\sum_{i=1}^{n}\Delta_i\left(\hat{\lambda}_n - H_n\left(Y_i,\hat{\beta}_n,\hat{F}_n\right)\right)\geq 0,$$

which implies that $E\left(\Delta\left(\lambda^* - H^*\left(Y\right)\right)\right)\geq 0$ and $\Delta\left(\lambda^* - H^*\left(Y\right)\right)\geq 0$. Taking $T\leq C$, we have $\lambda^* - H^*\left(T\right) > 0$ for any $T\in\left[0,\infty\right)$. □

Based on the results in Lemma 4, for a given M there exists $\delta = \delta\left(M\right) > 0$ such that $\lambda^* - H^*\left(y\right)\geq\delta$ for any $y\in\left[0,M\right]$. Define class $\mathcal{B}_M = \left\{\frac{\Delta I\left(Y\leq y\right)}{\lambda^* - H^*\left(Y\right)} : y\in\left[0,M\right].\right\}$. Class $\mathcal{B}_M$ is a Donsker class (van der Vaart,

A. W. and Wellner, J. A. [18]) because $\lambda^*-H^*(Y)$ is bounded away from zero. Thus, $\hat{F}_n(y)$ converges to $E\left(\frac{\Delta I(Y\le y)}{\lambda^*-H^*(Y)}\right)$ uniformly in $y\in[0,M]$, *i..e*,

$$\hat{F}_n(y)=\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i I(Y_i\le y)}{\left|\hat{\lambda}_n-H_n\left(Y_i,\hat{\beta}_n,\hat{F}_n\right)\right|}\to F^*(y)=E\left(\frac{\Delta I(Y\le y)}{\lambda^*-H^*(Y)}\right). \tag{A.20}$$

Following the calculations in (A.18), we have $F^*(y)=\int_0^y\frac{k(t)}{\lambda^*-H^*(t)}f_0(t)\mathrm{d}t$ with the density function $f^*(y)=\frac{k(y)}{\lambda^*-H^*(y)}f_0(y)$.

Based on the expression of $F^*(y)$, we can construct a sequence of distribution functions $\{\tilde{F}_n\}$ with jumps $\tilde{F}_n\{Y_i\}$ such that $\tilde{F}_n(y)\to F_0(y)$ for any $y\in(0,\infty)$. For $\Delta_i=1$ and $Y_i<\infty$, define

$$\tilde{F}_n\{Y_i\}=\frac{1}{nC_n}\frac{\Delta_i I(Y_i<\infty)}{k(Y_i)}, \tag{A.21}$$

where C_n is a constant such that $\sum_{i=1}^{n}\tilde{F}_n\{Y_i\}=1$ and $k(Y)$ is defined in (A.19). Let

$$\tilde{F}_n(y)=\sum_{i=1}^{n}\tilde{F}_n\{Y_i\}I(Y_i\le y)=\frac{1}{C_n}\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i I(Y_i\le y)}{k(Y_i)}. \tag{A.22}$$

Then it is obviously $\tilde{F}_n(\infty)=1$. Because $k(Y)$ is bounded away from zero, we have

$$C_n=\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i I(Y_i\le\infty)}{k(Y_i)}\to E\left(\frac{\Delta I(Y\le\infty)}{k(Y)}\right)=\int_0^\infty f_0(T)\mathrm{d}T=1. \tag{A.23}$$

The calculation here is similar to that in (A.18) with $k(Y)$ in the denominator instead of $\left|\lambda^*-H^*(Y)\right|$. Therefore, combining (A.22) and (A.23) we have

$$\tilde{F}_n(y)\to E\left(\frac{\Delta I(Y\le y)}{k(Y)}\right)=\int_0^y f_0(T)\mathrm{d}T=F_0(y). \tag{A.24}$$

Because $\boldsymbol{\beta}_n,\hat{F}_n$ are the maximum likelihood estimates, from (A.9) we have

$$\begin{aligned}&\log\frac{L\left(\hat{\beta}_n,\hat{F}_n\right)}{L\left(\beta_0,\tilde{F}_n\right)}\\&=\frac{1}{n}\sum_{i=1}^{n}\Delta_i I(Y_i<\infty)\log\frac{\hat{F}_n\{Y_i\}}{\tilde{F}_n\{Y_i\}}+\frac{1}{n}\sum_{i=1}^{n}I(Y_i=\infty)\log\frac{G\left(\eta\left(\hat{\beta}_n,X_i\right)\right)}{G\left(\eta\left(\beta_0,X_i\right)\right)}\\&+\frac{1}{n}\sum_{i=1}^{n}I(Y_i<\infty)\left(\Delta_i\log\frac{G'\left(\eta\left(\hat{\beta}_n,X_i\right)\hat{F}_n(Y_i)\right)\eta\left(\hat{\beta}_n,X_i\right)}{G'\left(\eta\left(\beta_0,X\right)\tilde{F}_n(Y_i)\right)\eta\left(\beta_0,X_i\right)}+(1-\Delta_i)\frac{G\left(\eta\left(\hat{\beta}_n,X_i\right)\hat{F}_n(Y_i)\right)}{G\left(\eta\left(\beta_0,X_i\right)\tilde{F}_n(Y_i)\right)}\right)\ge 0.\end{aligned} \tag{A.25}$$

For the strong convergency of the maximum likelihood estimates, we need to show that $\beta^*=\beta_0,F^*=F_0$. Letting $n\to\infty$ in (A.25), we have $E\left(\log\frac{L(\beta^*,F^*)}{L(\beta_0,F_0)}\right)\ge 0$. By the Jenson inequality, we have $\log E\left(\frac{L(\beta^*,F^*)}{L(\beta_0,F_0)}\right)\ge E\left(\log\frac{L(\beta^*,F^*)}{L(\beta_0,F_0)}\right)\ge 0$, where "=" holds if and only if $L\left(\beta^*,F^*\right)\equiv L\left(\beta_0,F_0\right)$ which con-

cludes $\beta^* = \beta_0, F^* = F_0$ since the model is identifiable. Therefore, We only need to show $E\left(\frac{L(\beta^*, F^*)}{L(\beta_0, F_0)}\right) \le 1$.

Theorem 2. Under conditions (C1)-(C6), $E\left(\frac{L(\beta^*, F^*)}{L(\beta_0, F_0)}\right) = 1$. The maximum likelihood estimates $(\boldsymbol{\beta}_n, \hat{F}_n)$ based on the modified likelihood function are strongly consistent, that is $|\beta_n - \beta_0| \to 0$, and $\sup_{y\in[0,\infty)} |\hat{F}_n(y) - F_0(y)| \to 0$ a.s., where $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$ and function F_0 is the true promotion time cumulative distribution function.

Proof: First of all, we want to prove $E\left(\frac{L(\beta^*, F^*)}{L(\beta_0, F_0)}\right) \le 1$. In fact, from (A.9) we have

$$\begin{aligned} E\left(\frac{L(\beta^*, F^*)}{L(\beta_0, F_0)}\right) &= E\left(\Delta I(Y<\infty)\frac{G'\left(\eta(\beta^*, X)F^*(Y)\right)\eta(\beta^*, X) f^*(Y)}{G'\left(\eta(\beta_0, X)F_0(Y)\right)\eta(\beta_0, X) f_0(Y)}\right) \\ &\quad + E\left((1-\Delta) I(Y<\infty)\frac{G\left(\eta(\beta^*, X)F^*(Y)\right)}{G\left(\eta(\beta_0, X)F_0(Y)\right)}\right) \\ &\quad + E\left(I(Y=\infty)\frac{G\left(\eta(\beta^*, X)\right)}{G\left(\eta(\beta_0, X)\right)}\right) \\ &= 1 - E_X\left(S_C(\infty|X)G\left[\eta(\beta^*, X)F^*(\infty)\right]\right) \\ &\quad + E_X\left(S_C(\infty|X)G\left[\eta(\beta^*, X)\right]\right) \end{aligned} \tag{A.26}$$

by a direct calculation.

Because $F^*(\infty) \le 1$ and $G(\cdot)$ is a monotone decreasing function, which implies that

$$G\left[\eta(\beta^*, X)F^*(\infty)\right] \ge G\left[\eta(\beta^*, X)\right]$$

and

$$E_X\left(S_C(\infty|X)G\left[\eta(\beta^*, X)F^*(\infty)\right]\right) \ge E_X\left(S_C(\infty|X)G\left[\eta(\beta^*, X)\right]\right).$$

Thus from (A.26) we have $E\left(\frac{L(\beta^*, F^*)}{L(\beta_0, F_0)}\right) \le 1$. Thus $L(\beta^*, F^*) \equiv L(\beta_0, F_0)$, which concludes $\beta^* = \beta_0$, $F^* = F_0$ and also concludes $F^*(\infty) = F_0(\infty) = 1$.

We have proved that any subsequence of $\boldsymbol{\beta}_n$, which is also denoted by $\boldsymbol{\beta}_n$, converges to $\boldsymbol{\beta}_0$ almost surely. Therefore, we conclude that the whole sequence $\boldsymbol{\beta}_n$ converges to $\boldsymbol{\beta}_0$ with probability 1.

We also proved that $\hat{F}_n(y)$ converges to $F_0(y)$ uniformly in y on $[0, M]$ for any fixed M and $\hat{F}_n(y)$ converges to $F_0(y)$ pointwise on $[0,\infty)$ since $F_0(\infty) = 1$. Therefore, $\hat{F}_n(y)$ converges to $F_0(y)$ uniformly in y on $[0,\infty)$ because of the continuity of F_0. □

c. Proofs of Asymptotic Normality of the Maximum Likelihood Estimates

We consider the likelihood function

$$\begin{aligned} &L(\boldsymbol{\beta}, F) \\ &= \left[\left\{-G'\left(\eta(\boldsymbol{\beta}, \boldsymbol{X})F(Y)\right)\eta(\boldsymbol{\beta}, \boldsymbol{X}) f(Y)\right\}^{\Delta}\left\{G\left(\eta(\boldsymbol{\beta}, \boldsymbol{X})F(Y)\right)\right\}^{1-\Delta}\right]^{I(Y<\infty)}\left[G\left(\eta(\boldsymbol{\beta}, \boldsymbol{X})\right)\right]^{I(Y=\infty)}, \end{aligned} \tag{A.27}$$

and write $L(\boldsymbol{\beta},F)$ as $L(\boldsymbol{\beta},F)=f(Y)^{\Delta I(Y<\infty)}K(\boldsymbol{\beta},F)$, where

$$K(\boldsymbol{\beta},F)$$
$$=\left[\left\{-G'(\eta(\boldsymbol{\beta},\boldsymbol{X})F(Y))\eta(\boldsymbol{\beta},\boldsymbol{X})\right\}^{\Delta}\left\{G(\eta(\boldsymbol{\beta},\boldsymbol{X})F(Y))\right\}^{1-\Delta}\right]^{I(Y<\infty)}\left[G(\eta(\boldsymbol{\beta},\boldsymbol{X}))\right]^{I(Y=\infty)}.$$

Then the log likelihood function, denoted by $l(\boldsymbol{\beta},F)$, can be written as

$$l(\boldsymbol{\beta},F)=\Delta I(Y<\infty)\log f(Y)+\log K(\boldsymbol{\beta},F). \tag{A.28}$$

Lemma 5. For any $\boldsymbol{\beta}$ and any distribution function F with a density, we have

$$\boldsymbol{E}\left(\log L(\boldsymbol{\beta}_0,F_0)\right)\geq \boldsymbol{E}\left(\log L(\boldsymbol{\beta},F)\right), \tag{A.29}$$

where $L(\boldsymbol{\beta},F)$ is the likelihood function given in (A.27), $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$ and F_0 is the true promotion time cumulative distribution function.

Proof: By Jensen inequality $\boldsymbol{E}\left(\log\dfrac{L(\beta,F)}{L(\beta_0,F_0)}\right)\leq\log\boldsymbol{E}\left(\dfrac{L(\beta,F)}{L(\beta_0,F_0)}\right)$. Thus, it suffices to show that

$\boldsymbol{E}\left(\dfrac{L(\beta,F)}{L(\beta_0,F_0)}\right)=1$. The proof is similar to that of **Theorem 2** thus omitted. □

From (A.29) we can derive a differential equation with $(\boldsymbol{\beta}_0,F_0)$. Let us consider function $H(t)$ such that:

(1) $H(t)$ is continuously differentiable with $H'(t)=h(t)$.

(2) $H(0)=0$, $H(\infty)=\lim_{t\to\infty}H(t)=0$.

(3) For $\nu\in\mathrm{IR}$ and $|\nu|$ is small enough, $f_0(t)+\nu h(t)\geq 0$.

Under conditions (1)-(3), $f_0(t)+\nu h(t)$ is a density function with corresponding distribution $F_0(t)+\nu H(t)$. For any $\boldsymbol{\alpha}\in\mathrm{IR}^d$, where d is the dimension of $\boldsymbol{\beta}$, we have $\boldsymbol{\beta}_0+\nu\boldsymbol{\alpha}\in\mathrm{IR}^d$, and $\boldsymbol{E}\left(\log L(\boldsymbol{\beta}_0,F_0)\right)\geq\boldsymbol{E}\left(\log L(\boldsymbol{\beta}_0+\nu\,\alpha,F_0+\nu H)\right)$ when $|\nu|$ is small. Therefore,

$$\boldsymbol{E}\left(\Delta I(Y<\infty)\frac{h}{f_0}+\frac{\partial}{\partial F}\log K(\boldsymbol{\beta}_0,F_0)H+\frac{\partial}{\partial\boldsymbol{\beta}}\log K(\boldsymbol{\beta}_0,F_0)\boldsymbol{\alpha}\right)=0. \tag{A.30}$$

Particularly, we can construct $H(t)$ satisfying conditions (1)-(3) through a function $h(t)$, which is defined on $[0,\infty)$ with bounded total variation. The total variation of $h(t)$ is given by $\|h(t)\|_V=\sup\sum_{i=1}^{m}\left|h(t_{i+1}-h(t_i))\right|$, where the supreme is taken over all finite partitions $0=t_1<t_2<\cdots<t_{m+1}=\infty$.

Define

$$\begin{aligned}Q_{F_0}h(y)&=h(y)-\int_0^{\infty}h(y)\mathrm{d}F_0(y),\\ H(y)&=\int_{[0,y]}Q_{F_0}h(s)\mathrm{d}F_0(s).\end{aligned} \tag{A.31}$$

We can show $H(y)$ in (A.31) satisfies conditions (1)-(3). Equation (A.30) can be written as

$$\boldsymbol{E}\left(\Delta I(Y<\infty)Q_{F_0}h+\frac{\partial}{\partial F}\log K(\boldsymbol{\beta}_0,F_0)\int_{[0,Y]}Q_{F_0}h\mathrm{d}F_0+\frac{\partial}{\partial\,\beta}\log K(\boldsymbol{\beta}_0,F_0)\boldsymbol{\alpha}\right)=0. \tag{A.32}$$

Let us consider a modified semiparametric version likelihood function,

$$\begin{aligned}&L_1(\boldsymbol{\beta},F_n)\\ &=\left[\left\{-G'(\eta(\boldsymbol{\beta},\boldsymbol{X})F_n(Y))\eta(\boldsymbol{\beta},\boldsymbol{X})F_n\{Y\}\right\}^{\Delta}\left\{G(\eta(\boldsymbol{\beta},\boldsymbol{X})F_n(Y))\right\}^{1-\Delta}\right]^{I(Y<\infty)}\left[G(\eta(\boldsymbol{\beta},\boldsymbol{X}))\right]^{I(Y=\infty)}\\ &=F_n\{Y\}^{\Delta I(Y<\infty)}K(\boldsymbol{\beta},F_n),\end{aligned} \tag{A.33}$$

where

$$K(\boldsymbol{\beta}, F_n) = \left[\left\{-G'\left(\eta(\boldsymbol{\beta},\boldsymbol{X})F_n(Y)\right)\eta(\boldsymbol{\beta},\boldsymbol{X})\right\}^{\Delta}\left\{G\left(\eta(\boldsymbol{\beta},\boldsymbol{X})F_n(Y)\right)\right\}^{1-\Delta}\right]^{I(Y<\infty)}\left[G\left(\eta(\boldsymbol{\beta},\boldsymbol{X})\right)\right]^{I(Y=\infty)}. \tag{A.34}$$

For any $\boldsymbol{\beta}$ and any step function $F_n(.) \in \mathcal{F}_n$, where

$$\mathcal{F}_n = \left\{F_n(.): F_n\{Y_i\} > 0 \text{ at } Y_i \text{ when } \Delta_i = 1, F_n\{Y_i\} = 0 \text{ o.w.}, \sum_{i=1}^{n} F_n\{Y_i\} = 1\right\}.$$

we have

$$\boldsymbol{E}_n\left(\log L_1\left(\boldsymbol{\beta}_n, \hat{F}_n\right)\right) \geq \boldsymbol{E}_n\left(\log L_1(\boldsymbol{\beta}, F_n)\right), \tag{A.35}$$

where $\left(\boldsymbol{\beta}_n, \hat{F}_n\right)$ is the maximum likelihood estimate of $(\boldsymbol{\beta}, F_n)$ based on (A.33).

Similar to the continuous case, now we can derive a differential equation with $\left(\boldsymbol{\beta}_n, \hat{F}_n\right)$. Consider function $H_n(.)$ satisfying the following conditions:

(1)′ $H_n(.)$ has a jump of size $H_n\{Y_i\}$ at Y_i when $\Delta_i = 1$ and a value of zero elsewhere.

(2)′ The summation of $H_n(.)$ over all Y_i's is zero, that is $\sum_{i=1}^{n} H_n\{Y_i\} = 0$.

(3)′ When $|\nu|$ is small enough, $\hat{F}_n\{Y_i\} + \nu H_n\{Y_i\} \geq 0$ for any Y_i.

Under conditions (1)′ -(3)′, $\hat{F}_n + \nu H_n$ is a qualified distribution function for likelihood (A.33). Take $\boldsymbol{\alpha} \in \mathrm{IR}^d$ such that $\boldsymbol{\beta}_n + \nu\boldsymbol{\alpha} \in \mathrm{IR}^d$. Therefore, because of (A.35) we have $\boldsymbol{E}_n\left(\log L_1\left(\boldsymbol{\beta}_n, \hat{F}_n\right)\right) \geq \boldsymbol{E}_n\left(\log L_1\left(\boldsymbol{\beta}_n + \nu\,\alpha, \hat{F}_n + \nu H_n\right)\right)$, when $|\nu|$ is small enough. After some algebra, we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\Delta_i I(Y_i < \infty)\frac{H_n\{Y_i\}}{\hat{F}_n\{Y_i\}} + E_n\left(\frac{\partial}{\partial F}\log K\left(\boldsymbol{\beta}_n, \hat{F}_n\right)H_n + \frac{\partial}{\partial \boldsymbol{\beta}}\log K\left(\boldsymbol{\beta}_n, \hat{F}_n\right)\boldsymbol{\alpha}\right) = 0. \tag{A.36}$$

Define

$$\begin{aligned} Q_{\hat{F}_n} h(y) &= h(y) - \int_0^{\infty} h(s)\mathrm{d}\hat{F}_n(s) \\ &= h(y) - \sum_{i=1}^{\infty} h(Y_i)\hat{F}_n\{Y_i\} \end{aligned}$$

$$H_n(y) = \int_{[0,y]} Q_{\hat{F}_n} h(s)\mathrm{d}\hat{F}_n(s) = \sum_{Y_i \leq y} Q_{\hat{F}_n} h(Y_i)\hat{F}_n\{Y_i\}. \tag{A.37}$$

With such a function H_n in (A.37), we have

$$\boldsymbol{E}_n\left(\Delta I(Y < \infty)Q_{\hat{F}_n} h + \frac{\partial}{\partial F}\log K\left(\boldsymbol{\beta}_n, \hat{F}_n\right)\int_{[0,Y]} Q_{\hat{F}_n}\mathrm{d}\hat{F}_n + \frac{\partial}{\partial \boldsymbol{\beta}}\log K\left(\boldsymbol{\beta}_n, \hat{F}_n\right)\boldsymbol{\alpha}\right) = 0. \tag{A.38}$$

Now, let us consider functions h with bounded total variation such that $\int_0^{\infty} h(y)\mathrm{d}F_0(y) = 0$ and define a set of such functions as $V_0 = \left\{h \in V \middle| \int_0^{\infty} h\mathrm{d}F_0 = 0\right\}$. For any $(\boldsymbol{\alpha}, h) \in \mathrm{IR}^d \times V_0$, define

$$\Omega_{\beta}(\boldsymbol{\alpha}, h) = \boldsymbol{E}\left(\frac{\partial^2}{\partial \boldsymbol{\beta}^2}\log K(\boldsymbol{\beta}_0, F_0)\right)\alpha + \boldsymbol{E}\left(\frac{\partial^2}{\partial F \partial \boldsymbol{\beta}}\log K(\boldsymbol{\beta}_0, F_0)\int_{[0,Y]} h\mathrm{d}F_0\right), \tag{A.39}$$

$$\Omega_F(\boldsymbol{\alpha}, h) = \omega - \int_0^{\infty}\omega\mathrm{d}F_0,$$

where

$$\omega = -\boldsymbol{E}\left(\Delta I\left(Y<\infty\right)+\frac{\partial}{\partial F}\log K\left(\boldsymbol{\beta}_0,F_0\right)\left(F_0\left(Y\right)-I\left(Y\geq s\right)\right)\right)h$$
$$+\boldsymbol{E}\left(\frac{\partial^2}{\partial F^2}\log K\left(\boldsymbol{\beta}_0,F_0\right)I\left(Y\geq s\right)\int_{[0,Y]}h\mathrm{d}F_0\right)$$
$$+\boldsymbol{E}\left(\frac{\partial^2}{\partial F\partial\boldsymbol{\beta}}\log K\left(\boldsymbol{\beta}_0,F_0\right)I\left(Y\geq s\right)\right)\alpha.$$

Lemma 6. With the notations defined in (A.39), we have

$$\begin{aligned}&\sqrt{n}\left(\Omega_\beta\left(\boldsymbol{\alpha},h\right)\left(\boldsymbol{\beta}_n-\beta_0\right)+\int_0^\infty\Omega_F\left(\boldsymbol{\alpha},h\right)\mathrm{d}\left(\hat{F}_n-F_0\right)\right)\\&=-\sqrt{n}\left(\boldsymbol{E}_n-\boldsymbol{E}\right)\left(\frac{\partial}{\partial\boldsymbol{\beta}}\log L\left(\boldsymbol{\beta}_0,F_0\right)\boldsymbol{\alpha}+\frac{\partial}{\partial F}\log L\left(\boldsymbol{\beta}_0,F_0\right)\int_{[0,Y]}h\mathrm{d}F_0\right)\\&\quad+o_p\left(\sqrt{n}\left\|\beta_n-\beta_0\right\|+\sqrt{n}\left\|\hat{F}_n-F_0\right\|_{L^\infty}\right)+o_p\left(1\right),\end{aligned}\tag{A.40}$$

where $h\in V_0$ and $L\left(\cdot,\cdot\right)$ is the likelihood function given in (A.27).

Proof: It follows from (A.32) and (A.38) that

$$\begin{aligned}&-\sqrt{n}\left(\boldsymbol{E}_n-\boldsymbol{E}\right)\left(\Delta I\left(Y<\infty\right)Q_{\hat{F}_n}h+\frac{\partial}{\partial F}\log K\left(\boldsymbol{\beta}_n,\hat{F}_n\right)\int_{[0,Y]}Q_{\hat{F}_n}h\mathrm{d}\hat{F}_n+\frac{\partial}{\partial\boldsymbol{\beta}}\log K\left(\boldsymbol{\beta}_n,\hat{F}_n\right)\boldsymbol{\alpha}\right)\\&=\sqrt{n}\boldsymbol{E}\left(\Delta I\left(Y<\infty\right)Q_{\hat{F}_n}h+\frac{\partial}{\partial F}\log K\left(\boldsymbol{\beta}_n,\hat{F}_n\right)\int_{[0,Y]}Q_{\hat{F}_n}h\mathrm{d}\hat{F}_n+\frac{\partial}{\partial\boldsymbol{\beta}}\log K\left(\boldsymbol{\beta}_n,\hat{F}_n\right)\boldsymbol{\alpha}\right)\\&\quad-\sqrt{n}\boldsymbol{E}\left(\Delta I\left(Y<\infty\right)Q_{F_0}h+\frac{\partial}{\partial F}\log K\left(\boldsymbol{\beta}_0,F_0\right)\int_{[0,Y]}Q_{F_0}h\mathrm{d}F_0+\frac{\partial}{\partial\boldsymbol{\beta}}\log K\left(\boldsymbol{\beta}_0,F_0\right)\boldsymbol{\alpha}\right).\end{aligned}\tag{A.41}$$

We consider a class $\mathcal{A}$,

$$\begin{aligned}\mathcal{A}=\Bigg\{&\Delta I\left(Y<\infty\right)Q_Fh+\frac{\partial}{\partial F}\log K\left(\boldsymbol{\beta},F\right)\int_{[0,Y]}Q_Fh\mathrm{d}F\\&+\frac{\partial}{\partial\boldsymbol{\beta}}\log K\left(\boldsymbol{\beta},F\right)\boldsymbol{\alpha}:\left\|\boldsymbol{\alpha}\right\|\leq1,\left\|h\right\|_V\leq1,\left\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\right\|+\left\|F-F_0\right\|_{L^\infty}\leq1\Bigg\}.\end{aligned}\tag{A.42}$$

$\mathcal{A}$ is a Donsker class. By Theorem 3.3.1 in van der Vaart and Wellner [18], the left hand side of (A.41) equals

$$\begin{aligned}&-\sqrt{n}\left(\boldsymbol{E}_n-\boldsymbol{E}\right)\left(\Delta I\left(Y<\infty\right)Q_{F_0}h+\frac{\partial}{\partial F}\log K\left(\boldsymbol{\beta}_0,F_0\right)\int_{[0,Y]}Q_{F_0}h\mathrm{d}F_0+\frac{\partial}{\partial\boldsymbol{\beta}}\log K\left(\boldsymbol{\beta}_0,F_0\right)\boldsymbol{\alpha}\right)+o_p\left(1\right)\\&=-\sqrt{n}\left(\boldsymbol{E}_n-\boldsymbol{E}\right)\left(\frac{\partial}{\partial F}\log L\left(\boldsymbol{\beta}_0,F_0\right)\int_{[0,Y]}Q_{F_0}h\mathrm{d}F_0+\frac{\partial}{\partial\boldsymbol{\beta}}\log L\left(\boldsymbol{\beta}_0,F_0\right)\boldsymbol{\alpha}\right)+o_p\left(1\right).\end{aligned}\tag{A.43}$$

Using the Taylor expansion at $\left(\boldsymbol{\beta}_0,F_0\right)$ for the right hand side of (A.41) and notations (A.39), Equation (A.41) can be simplified as

$$\begin{aligned}&\sqrt{n}\left(\Omega_\beta\left(\boldsymbol{\alpha},h\right)\left(\boldsymbol{\beta}_n-\beta_0\right)+\int_0^\infty\Omega_F\left(\boldsymbol{\alpha},h\right)\mathrm{d}\left(\hat{F}_n-F_0\right)\right)\\&=-\sqrt{n}\left(\boldsymbol{E}_n-\boldsymbol{E}\right)\left(\frac{\partial}{\partial\beta}\log L\left(\boldsymbol{\beta}_0,F_0\right)\boldsymbol{\alpha}+\frac{\partial}{\partial F}\log L\left(\boldsymbol{\beta}_0,F_0\right)\int_{[0,Y]}h\mathrm{d}F_0\right)\\&\quad+o_p\left(\sqrt{n}\left\|\beta_n-\beta_0\right\|+\sqrt{n}\left\|\hat{F}_n-F_0\right\|_{L^\infty}\right)+o_p\left(1\right).\end{aligned}\tag{A.44}$$

□

Lemma 7. The linear operator $\left(\boldsymbol{\alpha},h\right)\to\left(\Omega_\beta\left(\boldsymbol{\alpha},h\right),\Omega_F\left(\boldsymbol{\alpha},h\right)\right)$ is invertible from $\mathrm{IR}^d\times V_0$ to itself.

Proof: Decompose $\Omega_F(\boldsymbol{\alpha},h)$ as a summation of $\Omega_1(h)$ and $\Omega_2(\boldsymbol{\alpha},h)$, and

$$\begin{aligned}&\omega_1 = gh,\\ &\Omega_1(h) = \omega_1 - \int_0^\infty \omega_1 \mathrm{d}F_0 = -gh + \int_0^\infty gh\mathrm{d}F_0,\\ &\omega = \omega_1 + \omega_2,\\ &\Omega_2(\boldsymbol{\alpha},h) = \omega_2 - \int_0^\infty \omega_2 \mathrm{d}F_0.\end{aligned} \tag{A.45}$$

We can show $\Omega_1(h)$ is an invertible operator from V_0 to V_0, and $\Omega_2(\boldsymbol{\alpha},h)$ is a compact operator. Rewrite Ω as $\Omega = (\Omega_\beta, \Omega_F) = (I_d, \Omega_1) + (\Omega_\beta - I_d, \Omega_2)$. To prove Ω is invertible, we only need to show $Ker(\Omega) = \{0\}$. Suppose that $\Omega_\beta(\boldsymbol{\alpha},h)\boldsymbol{\alpha} + \int_0^\infty \Omega_F(\boldsymbol{\alpha},h)h\mathrm{d}F_0 = 0$. Because

$$\Omega_\beta(\boldsymbol{\alpha},h)\boldsymbol{\alpha} + \int_0^\infty \Omega_F(\boldsymbol{\alpha},h)h\mathrm{d}F_0 = -E\left(\frac{\partial}{\partial \boldsymbol{\beta}}\log(\boldsymbol{\beta}_0,F_0)\boldsymbol{\alpha} + \frac{\partial}{\partial F}\log(\boldsymbol{\beta}_0,F_0)\int_{[0,Y]} h\mathrm{d}F_0\right)^2 \tag{A.46}$$

with probability one, we have

$$\frac{\partial}{\partial \boldsymbol{\beta}}\log(\boldsymbol{\beta}_0,F_0)\boldsymbol{\alpha} + \frac{\partial}{\partial F}\log(\boldsymbol{\beta}_0,F_0)\int_{[0,Y]} h\mathrm{d}F_0 = 0, \tag{A.47}$$

which implies $\boldsymbol{\alpha} = 0$ and $h = 0$. □

Theorem 3. Under condition (C1)-(C6), $\sqrt{n}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0, \hat{F}_n - F_0)$ converges weakly to a Gaussian process in $l^\infty(\mathrm{IR}^d \times V_0)$.

Proof: Because $(\boldsymbol{\alpha},h) \to (\Omega_\beta(\boldsymbol{\alpha},h), \Omega_F(\boldsymbol{\alpha},h))$ has an inverse, denoted by $(\boldsymbol{\alpha},h) \to (\tilde{\Omega}_\beta(\boldsymbol{\alpha},h), \tilde{\Omega}_F(\boldsymbol{\alpha},h))$, Equation (A.44) can be written as

$$\begin{aligned}&\sqrt{n}\left(\boldsymbol{\alpha}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0) + \int_0^\infty h\mathrm{d}(\hat{F}_n - F_0)\right)\\ &= -\sqrt{n}(\boldsymbol{P}_n - \boldsymbol{P})\left(\frac{\partial}{\partial \boldsymbol{\beta}}\log L(\boldsymbol{\beta}_0,F_0)\tilde{\Omega}_\beta(\boldsymbol{\alpha},h) + \frac{\partial}{\partial F}\log L(\boldsymbol{\beta}_0,F_0)\int_{[0,Y]}\tilde{\Omega}_F(\boldsymbol{\alpha},h)\mathrm{d}F_0\right)\\ &\quad + o_p\left(\sqrt{n}\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_0\| + \sqrt{n}\|\hat{F}_n - F_0\|_{L^\infty}\right) + o_p(1).\end{aligned} \tag{A.48}$$

Immediately from (A.44) and (A.48), we have $\sqrt{n}\left(\|\beta_n - \beta_0\| + \|\hat{F}_n - F_0\|\right) = o_p(1)$. Back to (A.48), we obtain

$$\begin{aligned}&\sqrt{n}\left(\boldsymbol{\alpha}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0) + \int_0^\infty h\mathrm{d}(\hat{F}_n - F_0)\right)\\ &= -\sqrt{n}(\boldsymbol{P}_n - \boldsymbol{P})\left(\frac{\partial}{\partial \boldsymbol{\beta}}\log L(\boldsymbol{\beta}_0,F_0)\tilde{\Omega}_\beta(\boldsymbol{\alpha},h) + \frac{\partial}{\partial F}\log L(\boldsymbol{\beta}_0,F_0)\int_{[0,Y]}\tilde{\Omega}_F(\boldsymbol{\alpha},h)\mathrm{d}F_0\right) + o_p(1).\end{aligned} \tag{A.49}$$

Equation (A.49) holds uniformly for any $\|\boldsymbol{\alpha}\| \le 1$ and $\|h\|_V \le 1$. By using **Theorem 3.3.1** in van der Vaart and Wellner [18], $\sqrt{n}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0, \hat{F}_n - F_0)$ converges weakly to a Gaussian process in $l^\infty(\mathrm{IR}^d \times V_0)$. □