2009

Transformation Models for Survival Data Analysis and Applications

Yang Liu

FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

TRANSFORMATION MODELS FOR SURVIVAL DATA ANALYSIS AND APPLICATIONS

By

YANG LIU

A Dissertation submitted to the
Department of Statistics
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Degree Awarded:
Spring Semester, 2009

The members of the Committee approve the Dissertation of Yang Liu defended on March 23, 2009.

Xu-Feng Niu
Professor Directing Dissertation

Donald Lloyd
Outside Committee Member

Dan McGee
Committee Member

Debajyoti Sinha
Committee Member

Approved:

Dan McGee, Chair, Department of Statistics

Joseph Travis, Dean, College of Arts and Sciences

The Graduate School has verified and approved the above named committee members.

This thesis is dedicated to my parents, Jiaquan Liu and Xiaobo Yang.

ACKNOWLEDGEMENTS

I would first like to thank my advisor, Dr. Xu-Feng Niu, for his encouragement and support throughout my academic program. This dissertation could not have been written without his enthusiastic guidance.

I would also like to acknowledge my gratitude to my committee members, Dr. Dan McGee, Dr. Debajyoti Sinha, and Dr. Donald Lloyd for their insightful comments.

Thanks to many professors in the Department of Statistics at the Florida State University, Dr. Flori Bunea, Dr. Miles Hollander, Dr. Eric Chicken, Dr. Marten Wegkamp, and Dr. Kai-Sheng Song who taught excellent statistics classes and helped me to build a solid statistical background. Special thanks go to Dr. Fred Huffer. He always opens his door and is willing to help.

I would also like to thank the staff members, Pam McGhee, Megan Trautman, Chauncey Richburg, Evangelous Robinson, James Stricherz and Jennifer Rivera for their kind help.

Thanks to Dr. Donglin Zeng in the Department of Biostatistics at the University of North Carolina, Chapel Hill. The code he provided were extremely helpful.

Most important, thanks to my family. Thanks to my parents for their constant love. They are the greatest parents in the world. Love you, baba and mama! Thanks to my husband for his unconditional support.

— Yang Liu

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

It is often assumed that all uncensored subjects will eventually experience the event of interest in standard survival models. However, in some situations when the event considered is not death, it will never occur for a proportion of subjects. Survival models with a cure fraction are becoming popular in analyzing this type of study. We propose a generalized transformation model motivated by Zeng et al's (2006) transformed proportional time cure model. In our proposed model, fractional polynomials are used instead of the simple linear combination of the covariates. The proposed models give us more flexibility without loosing any good properties of the original model, such as asymptotic consistency and asymptotic normality of the regression coefficients. The proposed model will better fit the data where the relationship between a response variable and covariates is non-linear. We also provide a power selection procedure based on the likelihood function. A simulation study is carried out to show the accuracy of the proposed power selection procedure. The proposed models are applied to coronary heart disease and cancer related medical data from both observational cohort studies and clinical trials.

CHAPTER 1

INTRODUCTION

Survival analysis arises from many fields of study, such as medicine, public health, engineering, economics and others. It is concerned with analyzing the time to the occurrence of an event. For example, the event can be the time until the air conditioning of an aircraft dies, the time until a stock goes down, the time until a man commits further crime after being released from prison, the time until a person finds a decent job, etc. Especially, survival analysis is very popular in medicine and public health study. It involves modeling the medical or related data, such as the time until a cancer patient relapses or dies, the time until some severe side effect occurs after radiation therapy, etc.

In standard survival models, it is often assumed that all uncensored subjects will eventually experience the event of interest, which is described by a monotone decreasing function $S(t) = 1 - F(t)$, where $F(.)$ is a distribution function. The survival function goes to 0 when time t tends to infinity. However, in some situations when the considered event is not death, it will never occur for a significant proportion of subjects. For example, in a cancer clinical trial, the endpoint of interest is often recurrence. For some patients, they will never relapse after being treated. These patients are considered cured. Survival models with a cure fraction are becoming very popular in analyzing this type of cancer clinical trials. The focus of cure rate models is on the estimation of the proportion of cured population and of the failure time distribution of the uncured population.

The first cure rate model was proposed by Berkson and Gage [1],which is often referred as the standard cure rate model. It combines the cured and non-cured populations by using a summation function. This model has been extensively discussed and used by many researchers. Some alternative models including parametric cure rate models and semiparametric cure rate models are discussed later in several articles. Yakovlev and

Tsodikov [2], Chen et al. [3] explored the parametric cure rate models from the frequentist and the Bayesian point of view respectively. The book by Ibrahim et al. [4] gives an extensive discussion for the semiparametric cure rate models. Cure rate models are useful when patients have long-term censored times. A clinical study of melanoma cancer patients, analyzed by Zeng et al. [5] is a typical example. This trial consisted of 427 patients on two treatment arms. The response variable was relapse-free survival time. They considered patients without any recurrence beyond some certain point as being cured and used 5.5 years or greater as the threshold value to identify the "cured" patients.

In a survival model, T is often used as a continuous nonnegative random variable representing the time of an event. The probability of a subject's surviving till time t is given by the survival function $S(t) = 1 - F(t) = P(T > t)$, where $F(t)$ is the distribution function with probability density $f(t)$. The hazard function $h(t)$ is defined as the limit:

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t < T < t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)}.$$

It is the instantaneous failure rate at time t conditional on survival until time t or later. The cumulative hazard $H(t)$ is defined as $H(t) = \int_0^t h(s)ds$. It represents the total amount of risk up to time t. These four functions, $f(t), F(t), h(t)$, and $H(t)$ describe the distribution of the event time. Given one of them, the other three are completely determined. Here we list some relationships of them as well as some properties of the functions:

(1) $f(t) = h(t)\exp(-H(t))$,

(2) $S(t) = \exp(-H(t)) = \exp(-\int_0^t h(s)ds)$,

(3) $S(0) = 1, S(\infty) = \lim_{t\to\infty} S(t) = 0$ and $H(\infty) = \lim_{t\to\infty} H(t) = \infty$.

Usually, there are more than one covariate for each subject in a survival model. The covariates may be categorical, such as gender, location, level of education, stage of surgery, presence or absence of any behavior, exposure or unexposure to any particular treatment, etc. The variables may also be numerical, treated as either discrete or continuous data, such as age, weight, blood pressure, heart rate, etc. Some covariates may be time-dependent, but in this study, we only consider the covariates that are independent of time.

The hazard function now depends on both time and a set of covariates. Cox [6] brought the idea of separating time t and individual covariate vector $\boldsymbol{x} = \{x_1, ..., x_p\}$, which led to the popular proportional hazard model with the hazard function

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

where $h_0(t)$ is the baseline hazard function and $\boldsymbol{\beta}$ is a vector of regression coefficients. $\eta = \boldsymbol{\beta'x}$ is called the linear predictor, and it can be a more complicated function of $\boldsymbol{\beta}$ and $\boldsymbol{x}$. This is a semiparametric model because a parametric form is only assumed for the covariate effect. The base line function is being treated nonparametrically. Proportional hazards is a desirable property when comparing two subjects with different covariate values. Their hazard ratio is constant over time, which makes the interpretation very easy.

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. Right censored cases are quite common in clinical trials. Patients may also be left censored or interval censored, but here we only consider right censored data. Suppose there are n right censored subjects. For the ith 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 ith subject is Y_i, $Y_i = \min(T_i, C_i)$. 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. Now we can use a triple of random variables (Y_i, X_i, Δ_i) to describe each subject, where X_i is the covariate vector and Y_i and Δ_i are defined as the following,

$$\begin{aligned} Y_i &= \min(T_i, C_i) \\ \Delta_i &= \begin{cases} 1, & T_i \leq C_i \\ 0, & T_i > C_i. \end{cases} \end{aligned}$$

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

$$\begin{aligned} L(\boldsymbol{\beta}) &\propto \prod_{i=1}^{n} f(Y_i)^{\Delta_i}(1 - F(Y_i))^{1-\Delta_i} \\ &= \prod_{i=1}^{n}(h_0(Y_i)\exp(\beta' X_i))^{\Delta_i} \prod_{i=1}^{n} S(Y_i). \end{aligned}$$

The Newton-Raphson algorithm is often used here to obtain the maximum likelihood estimates of β.

In the following chapter an overview of the cure rate model will be given. Both the mixture cure rate model and the promotion time cure model will be introduced. Emphasis

is on discussing the transformation models introduced by Zeng et al. [5]. The fractional polynomial regression model is discussed in Section 2.2. The proposed transformation model is given in Section 2.3, in which we use fractional polynomials instead of the simple linear combination of the covariates. The proposed models give us more flexibility without loosing any good properties of the transformation cure rate model. A discussion about identifiability of the proposed model is given at the end of Chapter 2.

In Chapter 3, we conduct a series of simulations and show how to select powers for the continuous covariates when the fractional polynomials are used in different models. Different model selection criteria are used, such as the residual sum of squares and the partial likelihood function. The power selection procedure appears to work well in the linear regression model as long as the variability of the random error term is not too large. In logistic regression, the power selection procedure may not perform well if the sample size is small. We also consider survival models with monotone increasing, decreasing or constant hazard rate. The power selection procedure performs well in the survival models if there are not too many outliers for the life time T. A simulation study on the proposed model will be carried out later.

Some examples based on the proposed model will be given.

CHAPTER 2

MODELS FOR SURVIVAL DATA

Two commonly used cure models are introduced in this chapter. One is called the mixture cure rate model. It combines the cured group and the non-cured group by using a summation function. Another one is the promotion time cure model, which takes long-term survivors into account by applying a restriction on the cumulative hazard function. The promotion time cure model can avoid some drawbacks of a mixture model. A fractional polynomial regression is also introduced in this chapter. A proposed generalized transformation model and a discussion about identifiability of the proposed model is given at the end.

2.1 Cure Rate Model

Survival models with a cure fraction are often called as cure rate models, where the population is a mixture of cured and non-cured individuals. The objective is usually to study the cure rate and the effect of any covariates.

2.1.1 Mixture Cure Rate Model

One type of commonly used cure rate model is the mixture model proposed by Berkson and Gage [1]. 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. $S_2(t)$ is a proper survival function, while $S_1(t)$ is not since $S_1(\infty) = \pi > 0$. The mixture model has been fully discussed by many authors,

including Farewell [7], Gray and Tsiatis [8], Sposto et al. [9], Laska and Meisner [10], Sy and Taylor [11], and Lu and Ying [12].

Farewell [7] considered the data from toxicological experiments with laboratory animal and discussed the effect of toxicant levels. In the data set Farewell [7] used, a substantial proportion of the animals do not die by the end of the study at some toxicant levels. The probability that animals will die in the experiment is represented by a logistic regression model. The time to death among these animals is modelled by a Weibull distribution and a series of conditional probabilities. Gray and Tsiatis [8] derived a linear rank test to compare the cure rates in two treatment groups by using a mixture model, which is more efficient than the Mantel and Haenszel [13]'s log-rank test. Sy and Taylor [11] used a nonparametric form of the likelihood function and EM algorithm to estimate the probability of event occurrence and the regression parameters jointly. Lu and Ying [12] proposed a unified approach for a class of semiparametric transformation models with the cure fraction modeled by logistic regression, including the proportional hazards cure model and the proportional odds cure model as special cases.

The mixture model is attractive and commonly used, however, it has some drawbacks. One of them is that it cannot have a proportional hazards structure if the covariates are modeled through π. Ibrahim et al. [4] also pointed out that a mixture model sometimes yields improper posterior distribution when noninformative improper priors are used from the Bayesian point of view.

2.1.2 Promotion Time Cure Model

Yakovlev and Tsodikov [2], Tsodikov [14], Chen et al. [3], and Zeng et al. [5]proposed and studied another type of cure rate model called promotion time cure model. Instead of dividing the population into two subpopulations 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. The population survival function $S(t)$ is represented in the form $S(t) = \exp(-H(t))$, where $H(t)$ is the cumulative hazard function. In a cure rate model, the function $S(t)$ is improper, which 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 [14] suggested to consider $H(t) = \theta F(t)$, where $F(t)$ is the distribution function of a nonnegative random

variable,

$$S(t) = \exp(-\theta F(t)). \tag{2.1}$$

We get a proportional hazard model from (2.1) if the covariates are modeled through $\theta(.)$.

A biological motivation of this model was provided by Chen et al. [3], and later extended by Zeng et al. [5]. For the ith subject in a population, we let N_i denote the number of tumor cells left active after the initial treatment. Assume N_i is Poisson distributed with mean $\theta(\boldsymbol{X_i})$, where $\boldsymbol{X_i} = (X_{i1}, ...X_{ip})$ is the covariate vector for the ith individual. Let $T_k(k = 1, ..., N_i)$ denote the random time for the kth tumor cell to produce a detectable cancer mass. Conditional on $N_i = k$, the T_k's are identically and independently distributed with a common distribution $F(t)$. The time to relapse of cancer T is defined as the minimum of the T_k's, $T = \min(T_1, ..., T_{N_i})$. The survival function for the population is given by

$$\begin{aligned} S(t|\boldsymbol{X_i}) &= P(T > t) \\ &= P(N_i = 0) + \sum_{k\geq 1} P(T_1 > t, ..., T_{N_i} > t, N_i = k) \\ &= \exp(-\theta(\boldsymbol{X_i})) + \sum_{k=1}^{\infty} S(t)^k \frac{(\theta(\boldsymbol{X_i}))^k}{k!} \exp(-\theta(\boldsymbol{X_i})) \\ &= \exp(-\theta(\boldsymbol{X_i})F(t)), \end{aligned} \tag{2.2}$$

where $F(t)$ is a proper distribution function. When $t \to \infty$, $S(t|\boldsymbol{X_i}) = \exp(-\theta(\boldsymbol{X_i})) > 0$ is the cure rate for the ith subject. The first derivative of $1 - S(t|\boldsymbol{X_i})$ is given by

$$g(t|\boldsymbol{X_i}) = \frac{d}{dt}(1 - S(t|\boldsymbol{X_i})) = \theta(\boldsymbol{X_i}) f(t) \exp(-\theta(\boldsymbol{X_i})F(t)), \tag{2.3}$$

where $f(t)$ is the density corresponding to $F(t)$. Since $S(t|\boldsymbol{X_i})$ is not a proper survival function, function $g(t|\boldsymbol{X_i})$ in (2.3) is also not a proper density, but it is needed when derive the likelihood.

The promotion time cure model avoids the drawbacks of a mixture model. It has a proportional hazards structure through the cure rate parameter. Chen et al. [3] 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 relationship. For simplicity, let us ignore the covariates for now. The

relationship is given by

$$S(t) = \exp(-\theta F(t)) = \exp(-\theta) + (1 - \exp(-\theta))S^*(t),$$

where

$$S^*(t) = P(T > t|N \geq 1) = \frac{\exp(-\theta F(t)) - \exp(-\theta)}{1 - \exp(-\theta)}.$$

$S^*(t)$ can be viewed as the proper survival function for the subpopulation who will eventually experience the event. $\pi = \exp(-\theta)$ is the cure rate in the mixture model.

Zeng et al. [5] proposed a generalized 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, $i = 1, ..., n$ in model (2.1). The number of tumor cells N_i is assumed to be Poisson distributed with mean $\theta(\boldsymbol{X_i})\xi_i$. Conditional on both $N_i = k$ and ξ_i, the random promotion times $T_k(k = 1, ..., N_i)$ are identically and independently distributed with distribution function $F(t)$. Following the derivation of (2.2), the survival function for the time to relapse, $T = \min(T_1, ..., T_{N_i})$, is given by

$$S(t|\boldsymbol{X_i}, \xi_i) = \exp(-\theta(\boldsymbol{X_i})F(t)\xi_i). \tag{2.4}$$

Different parametric distributions can be used for the frailty ξ_i. The most commonly used one is $\Gamma(1/\gamma, \gamma)$, where $\gamma \geq 0$. The mean in the gamma distribution need to be one due to the model identification issue. We take expectations with respect to ξ_i for both sides in (2.4). Since both $\theta(\boldsymbol{X_i})$ and $F(t)$ are non-negative, we can apply the first order moment generating function of a gamma distribution. The survival function becomes

$$S(t|\boldsymbol{X_i}) \quad = \quad E_{\xi_i}[\exp(-\theta(\boldsymbol{X_i})F(t)\xi_i)] = (1 + \gamma\theta(\boldsymbol{X_i})F(t))^{-1/\gamma}. \tag{2.5}$$

As Zeng et al. [5] pointed out, (2.5) provides a very wide class of transformation cure models,

$$S(t|\boldsymbol{X_i}) \quad = \quad G_\gamma(\theta(\boldsymbol{X_i})F(t)), \tag{2.6}$$

where

$$G_\gamma(x) \quad = \quad \begin{cases} (1 + \gamma x)^{-1/\gamma}, & \gamma > 0, \\ e^{-x}, & \gamma = 0. \end{cases} \tag{2.7}$$

When ξ_i takes other distributions, we may get different transformations. A Box-Cox type transformation is also considered in Zeng et al. [5] with

$$G_\gamma(x) = \begin{cases} \exp\{-\frac{(1+x)^\gamma - 1}{\gamma}\}, & \gamma > 0, \\ \frac{1}{1+x}, & \gamma = 0. \end{cases} \tag{2.8}$$

The proportional hazards model in (2.1) is a special case of the transformation families (2.7) and (2.8) corresponding to $\gamma = 0$ and $\gamma = 1$ respectively. Another popular survival model, the proportional odds model, is also a special case of (2.7) and (2.8) when $\gamma = 1$ and $\gamma = 0$ respectively.

We also see from model (2.6) that the cure fraction is $S(\infty) = G_\gamma(\theta F(\infty)) = G_\gamma(\theta)$, and the model can be written as a standard cure rate model,

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

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

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

The covariates can be modeled through a known and strictly positive increasing link function $\theta(\boldsymbol{X_i}) = \eta(\boldsymbol{\beta' X_i})$, where $\boldsymbol{\beta}$ is the regression vector including an intercept term. Following model (2.6) and the notations defined in the previous section, the observed data likelihood of parameters $(\boldsymbol{\beta}, F)$ is given by

$$\begin{aligned} &\prod_{i=1}^{n} \{[\{-G'(\eta(\boldsymbol{\beta' X_i})F(Y_i))\eta(\boldsymbol{\beta' X_i})f(Y_i)\}^{\Delta_i} \\ &\times \{G(\eta(\boldsymbol{\beta' X_i})F(Y_i))\}^{(1-\Delta_i)}]^{I(Y_i<\infty)} \times [G(\eta(\boldsymbol{\beta' X_i}))]^{I(Y_i=\infty)}\} \end{aligned} \tag{2.9}$$

The three pieces in the product are for failures, censoring and subjects who never experience failure or censoring respectively. Referring to (2.3), the derivation of the likelihood function is quite straightforward. They applied a nonparametric maximum likelihood estimation approach and Newton-Raphson algorithm to get the estimate of $(\boldsymbol{\beta}, F)$ iteratively. They also showed asymptotic properties of the estimate $(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)$, including consistency and asymptotic normality.

2.2 Fractional Polynomial Regression

Motivated by Zeng et al. [5] and consideration of data where there is a nonlinear relationship between the covariates and the hazard rates, we propose a generalized transformation cure model by using a general additive function of $\boldsymbol{X_i} = (X_{i1}, ..., X_{ip})$' instead of the simple linear combination $\boldsymbol{\beta' X_i}$, such as a fractional polynomial for each continuous covariate.

The multiple linear regression model $E[Y|\boldsymbol{X}] = \boldsymbol{\beta' X}, \boldsymbol{X} = (X_1, ...X_p)$', has been widely used by researchers in many fields to study the dependence of the response on the covariates, and to make predictions. It has many attractive properties but also few limitations. One of them is that the response has to be linearly dependent on each covariate. But it has been recognized that simple linear relationships do not always fit survival data well, especially in medicine and public health. For example, curves are often needed when describe human growth. One way to generalize the linear regression model is the additive models by Stone [15], which is defined by $E[Y|\boldsymbol{X}] = \sum_{j=0}^{p} f_j(X_j)$, where $f_0(.)$ is the constant term and $f_1(.), ..., f_p(.)$ are arbitrary univariate functions, one for each covariate. Additive models are much more flexible to use and it retains the important additive feature of the linear regression models.

One important issue in applying additive models is to identify or estimate the function $f_j(X_j)'s$. Royston and Altman [16] suggested using fractional polynomials for each $f_j(X_j)$, which is a family of functions of positive covariates. When non-positive values occur, a shift of the covariates is needed to ensure positivity. For simplicity, let us consider a single covariate X for now. A fractional polynomial with degree m is described as $E[Y|X] = \beta_0 + \sum_{i=1}^{m} \beta_i f_i(X)$, where

$$f_i(X) \quad = \quad \begin{cases} X^{p_i}, & p_i \neq p_{i-1} \\ f_{i-1}(X)\log(X), & p_i = p_{i-1} \end{cases} \tag{2.10}$$

and $\boldsymbol{p} = (p_1, ..., p_m), p_1 \leq p_2 \leq ... \leq p_m$, is a real-valued vector of powers. If $p_i = 0$ for any i, X^{p_i} is defined to be $\log(X)$ by the Box-Tidwell transformation. For example, $E[Y|X] = \beta_0 + \beta_1 X^{-1} + \beta_2 X^{-1}\log(X) + \beta_3 X^{-1}(\log(X))^2 + \beta_4 \log(X) + \beta_5 X^2$ is a fractional polynomial with degree $m = 5$ and $\boldsymbol{p} = (-1, -1, -1, 0, 2)$. Royston and Altman [16] pointed out that it is worth considering fractional polynomials with degrees one and two specially, since models with degree higher than two are rarely used in practice. A fractional polynomial

with degree one is simply

$$E[Y|X] = \begin{cases} \beta_0 + \beta_1 X^{p_1}, & p \neq 0, \\ \beta_0 + \beta_1 \log(X), & p = 0. \end{cases} \tag{2.11}$$

A second degree fractional polynomial is given by

$$E[Y|X] = \begin{cases} \beta_0 + \beta_1 X^{p_1} + \beta_2 X^{p_2}, & p_1 < p_2, \\ \beta_0 + \beta_1 X^{p_1} + \beta_2 X^{p_1} \log(X), & p_1 = p_2. \end{cases} \tag{2.12}$$

Royston and Altman [16] also suggested that the powers in (2.11) and (2.12) can be chosen from the set $(-2, -1, -0.5, 0, 0.5, 1, 2, ..., \max(3, m))$, since it is difficult to estimate the powers precisely and the set is rich enough to cover all conventional polynomials of interest. The best estimates of the powers are determined based on the maximum likelihood method.

Suppose there is more than one covariate, the power selection procedure can be done iteratively. Let us take two covariates $X_i, i = 1, 2$, for example to illustrate the procedure. The goal is to fit the model $E[Y|X_1, X_2] = \beta_0 + \sum_{j=1}^{m_1} \beta_{1j} f_{1j}(X_1) + \sum_{j=1}^{m_2} \beta_{2j} f_{1j}(X_2)$. First we fit a fractional polynomial for X_1 and keep X_2 as a linear term in the model, which is $E[Y|X_1, X_2] = \beta_0 + \sum_{j=1}^{m_1} \beta_{1j} f_{1j}(X_1) + \beta_2 X_2$. Then we fix the functions $f_{1j}(X_1)$ and fit the model $E[Y|X_1, X_2] = \beta_0 + \sum_{j=1}^{m_1} \beta_{1j} f_{1j}(X_1) + \sum_{j=1}^{m_2} \beta_{2j} f_{1j}(X_2)$. The coefficients β_{1j}'s are re-estimated and we get a fractional polynomial for X_2. Then fix the functions $f_{2j}(X_2)$ and determine a new sequence of $f_{1j}(X_1)$. Repeat the procedure iteratively until the fractional polynomial function $f_{1j}(X_1), j = 1, ..., m_1, f_{2j}(X_2), j = 1, ..., m_2$, do not change from one iteration to the next.

For some data set, especially some medical related data, fractional polynomial can give a better fit compared to the conventional polynomial. Our proposed model is trying to adopt this idea and use it in the transformation cure rate model.

2.3 Generalized Transformation Models

Cure rate models have been widely used for modeling data from cancer and AIDS clinical trials where a proportion of patients is cured. One type of commonly used model is the promotion time cure model, which has been discussed by many authors. For example, Zeng et al. [5] provided a semiparametric transformation model with a cure fraction. For an individual with covariate vector $\boldsymbol{X}$ in the study, the survival function is given by

$$S(t|\boldsymbol{X}) = G_\gamma(\theta(\boldsymbol{X})F(t)), \tag{2.13}$$

where $\theta(\boldsymbol{X}) = \eta(\boldsymbol{\beta'X})$ and $\eta(.)$ is a known and strictly positive increasing function that is often referred as the link function. Many functions have these properties, such as $\eta(x) = e^x$ and $\eta(x) = e^x/(1+e^x)$. $G_\gamma(.)$ is a transformation family given by

$$G_\gamma(x) = \begin{cases} (1+\gamma x)^{-1/\gamma}, & \gamma > 0, \\ e^{-x}, & \gamma = 0. \end{cases} \tag{2.14}$$

In our proposed model, we use a fractional polynomial instead of $\boldsymbol{\beta'X}$ in the link function $\eta(.)$. Although in practice, fractional polynomials with degree higher than two are not used very often, but to make the proposed model more flexible, we consider $\theta(\boldsymbol{X})$ as the following,

$$\theta(\boldsymbol{X}) = \eta\left(\sum_{j=0}^{q}\beta_j X_j + \sum_{i=q+1}^{p}\left(\beta_{i0}\frac{X_i^{\alpha_{i0}}-1}{\alpha_{i0}} + \beta_{i1}\frac{X_i^{\alpha_i}-1}{\alpha_i} + \beta_{i2}\frac{X_i^{\alpha_i}-1}{\alpha_i}\log X_i\right)\right), \tag{2.15}$$

where $\boldsymbol{X} = (X_1, ..., X_q, X_{q+1}, ..., X_p)$, $X_1, ..., X_q$ are categorical covariates and $X_{q+1}, ..., X_p$ are positive continuous covariates. An intercept term β_0 is also considered when we assume that $X_0 \equiv 1$. As shown in (2.15), when $\alpha_{i0} \neq \alpha_i$ for $\forall\ q+1 \leq i \leq p$, a degree of three fractional polynomial is used for each of the continuous covariates. A proof of identifiability of this model will be given in Section 2.4. The proof can be easily adapted to a simplified model with lower degree of freedom by letting one or two coefficients among β_{i0}, β_{i1} and β_{i2} equal to 0.

The powers α_{i0} and α_i can be chosen from a special set, for example A=(-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2), by comparing the likelihood. Considering the continuity of the transformation family when α approaching to 0, we use transformation $\frac{X^\alpha-1}{\alpha}$ not X^α in (2.15) such that the model is appropriate for any $\alpha \in [-2, 2]$. But actually in practice, since the powers are chosen only from a certain set of numbers, continuity of the transformation family will not be an issue. In the simulation study and examples we will show in the following chapters, transformation X^α is used sometimes instead of $\frac{X^\alpha-1}{\alpha}$, but it will not affect the identifiability of the model.

2.4 Identifiability of the Proposed Models

Suppose we use model (2.13) and (2.14) with link function (2.15), the observed-data likelihood function of parameters (γ, θ, F) is given by

$$L(\gamma, \theta, F) = \tag{2.16}$$
$$[(-G'_\gamma(\theta(X)F(Y))\theta(X)f(Y))^{\triangle}(G_\gamma(\theta(X)F(Y)))^{(1-\triangle)}]^{I(Y<\infty)}[G_\gamma(\theta(X))]^{I(Y=\infty)}.$$

We first discuss the identifiability of this likelihood function. Suppose two sets of parameters, (γ, θ, F) and $(\tilde{\gamma}, \tilde{\theta}, \tilde{F})$, give the same likelihood function, that is $L(\gamma, \theta, F) = L(\tilde{\gamma}, \tilde{\theta}, \tilde{F})$, we want to show that $\gamma = \tilde{\gamma}, \theta = \tilde{\theta}$, and $F = \tilde{F}$.

Claim 1. Suppose $L(\gamma, \theta, F(y)) = L(\tilde{\gamma}, \tilde{\theta}, \tilde{F}(y))$, for any $y \in R^+$ and any X, then $\gamma = \tilde{\gamma}, \theta = \tilde{\theta}$, and $F = \tilde{F}$.

Proof: We choose $\triangle = 1, Y = y < \infty, X = x$ and obtain

$$G'_\gamma(\theta(x)F(y))\theta(x)f(y) = G'_{\tilde{\gamma}}(\tilde{\theta}(x)\tilde{F}(y))\tilde{\theta}(x)\tilde{f}(y). \tag{2.17}$$

We can also take $\triangle = 0, Y = y < \infty, X = x$ and get

$$G_\gamma(\theta(x)F(y)) = G_{\tilde{\gamma}}(\tilde{\theta}(x)\tilde{F}(y)). \tag{2.18}$$

Notice that in the transformation (2.14), we have

$$G'_\gamma(t) = -\frac{G_\gamma(t)}{1+\gamma t}. \tag{2.19}$$

Combine (2.17), (2.18) and (2.19), we get the following equation

$$\frac{\theta(x)f(y)}{1+\gamma\theta(x)F(y)} = \frac{\tilde{\theta}(x)\tilde{f}(y)}{1+\tilde{\gamma}\tilde{\theta}(x)\tilde{F}(y)}. \tag{2.20}$$

Now (2.20) can be written as

$$\frac{f(y)}{\tilde{\theta}(x)} - \frac{\tilde{f}(y)}{\theta(x)} = k(y), \tag{2.21}$$

where $k(y) = \gamma F(y)\tilde{f}(y) - \tilde{\gamma}\tilde{F}(y)f(y)$.

Choose $y_0 < \infty$, such that $\tilde{f}(y_0) \neq 0$. Since (2.21) holds for any $y < \infty$, we obtain a system of linear equations in the two variables $1/\theta(x)$ and $1/\tilde{\theta}(x)$

$$\begin{cases} f(y)/\tilde{\theta}(x) - \tilde{f}(y)/\theta(x) = k(y), \\ f(y_0)/\tilde{\theta}(x) - \tilde{f}(y_0)/\theta(x) = k(y_0). \end{cases}$$

Since $\theta(x)$ is not a constant function, there exist x_1 and x_2 such that $\theta(x_1) \neq \theta(x_2)$. Now the system has two different solutions $(1/\theta(x_1), 1/\tilde{\theta}(x_1))$ and $(1/\theta(x_2), 1/\tilde{\theta}(x_2))$. Therefore, its determinant must be zero, that is $f(y)\tilde{f}(y_0) = f(y_0)\tilde{f}(y)$. We get $f(y_0) = \tilde{f}(y_0)$ after integrating with respect to y on both sides since both $f(y)$ and $\tilde{f}(y)$ are density functions. Thus, we get $f(y) = \tilde{f}(y)$ and $F(y) = \tilde{F}(y)$ for any y.

Now immediately from (2.21), we get

$$\frac{1}{\tilde{\theta}(x)} - \frac{1}{\theta(x)} = (\gamma - \tilde{\gamma})F(y). \tag{2.22}$$

Letting $y = 0$, we have $F(y) = 0$ and $\theta(x) = \tilde{\theta}(x)$ for any x. Therefore, $(\gamma - \tilde{\gamma})F(y) = 0$ and $\gamma = \tilde{\gamma}$. $\square$

Then, we can prove that for $\theta(X)$ defined in (2.15), $\beta_j, \beta_{i0}, \beta_{i1}, \beta_{i2}, \alpha_{i0}$ and α_i are identifiable for $\forall\ i$ and j.

Claim 2. Suppose $\sum_{j=0}^{q} \beta_j X_j + \sum_{i=q+1}^{p} \left(\beta_{i0} \frac{X_i^{\alpha_{i0}} - 1}{\alpha_{i0}} + \beta_{i1} \frac{X_i^{\alpha_i} - 1}{\alpha_i} + \beta_{i2} \frac{X_i^{\alpha_i} - 1}{\alpha_i} \log X_i \right) = \sum_{j=0}^{q} \tilde{\beta}_j X_j + \sum_{i=q+1}^{p} \left(\tilde{\beta}_{i0} \frac{X_i^{\tilde{\alpha}_{i0}} - 1}{\tilde{\alpha}_{i0}} + \tilde{\beta}_{i1} \frac{X_i^{\tilde{\alpha}_i} - 1}{\tilde{\alpha}_i} + \tilde{\beta}_{i2} \frac{X_i^{\tilde{\alpha}_i} - 1}{\tilde{\alpha}_i} \log X_i \right)$, for $\forall\ \boldsymbol{X} = (X_1, ..., X_p)$, then $\beta_j = \tilde{\beta}_j, \beta_{i0} = \tilde{\beta}_{i0}, \beta_{i1} = \tilde{\beta}_{i1}, \beta_{i2} = \tilde{\beta}_{i2}, \alpha_{i0} = \tilde{\alpha}_{i0}$, and $\alpha_i = \tilde{\alpha}_i$.

Proof: To prove the identifiability of the coefficient of a categorical covariate, for example β_1, fix $X_2, ..., X_p$, then we have $\beta_1 X_1 + M = \tilde{\beta}_1 X_1 + \tilde{M}$. Coefficient β_1 is identifiable if X_1 can take at least two different values. The proof for a continuous covariate is different. Let's take X_p for example. Fix $X_1, ..., X_{p-1}$, then we have

$$\begin{aligned} & M + \beta_{p0} \frac{X_p^{\alpha_{p0}} - 1}{\alpha_{p0}} + \beta_{p1} \frac{X_p^{\alpha_p} - 1}{\alpha_p} + \beta_{p2} \frac{X_p^{\alpha_p} - 1}{\alpha_p} \log X_p \\ = \ & \tilde{M} + \tilde{\beta}_{p0} \frac{X_p^{\tilde{\alpha}_{p0}} - 1}{\tilde{\alpha}_{p0}} + \tilde{\beta}_{p1} \frac{X_p^{\tilde{\alpha}_p} - 1}{\tilde{\alpha}_p} + \tilde{\beta}_{p2} \frac{X_p^{\tilde{\alpha}_p} - 1}{\tilde{\alpha}_p} \log X_p. \end{aligned} \tag{2.23}$$

For simplicity of calculation, we use $\xi_{p0} = \beta_{p0}/\alpha_{p0}, \xi_{p1} = \beta_{p1}/\alpha_p, \xi_{p2} = \beta_{p2}/\alpha_p$, and write all constant terms together. Proving the β's and the α's are identifiable is equivalent to prove the ξ's and the α's are identifiable. Now we can use the following equation instead of (2.23) ,

$$\begin{aligned} & M + \xi_{p0} X_p^{\alpha_{p0}} + \xi_{p1} X_p^{\alpha_p} + \xi_{p2} X_p^{\alpha_p} \log X_p \\ = \ & \tilde{M} + \tilde{\xi}_{p0} X_p^{\tilde{\alpha}_{p0}} + \tilde{\xi}_{p1} X_p^{\tilde{\alpha}_p} + \tilde{\xi}_{p2} X_p^{\tilde{\alpha}_p} \log X_p. \end{aligned} \tag{2.24}$$

Let's assume that none of the coefficients equals zero. Suppose $\alpha_p = \max(\alpha_{p0}, \alpha_p, \tilde{\alpha}_{p0}, \tilde{\alpha}_p)$. Divide both sides by $X_p^{\alpha_p}$ and let $X_p \to \infty$. If $\alpha_p > \tilde{\alpha}_p, \alpha_p > \tilde{\alpha}_{p0}$, then $\xi_{p1} + \xi_{p2} \log X_p = 0$, for any $X_p \in R_+$. This cannot be true unless both ξ_{p1} and ξ_{p2} are equal to 0. If $\alpha_p > \tilde{\alpha}_p, \alpha_p = \tilde{\alpha}_{p0}$, then $\xi_{p1} + \xi_{p2} \log X_p = \tilde{\xi}_{p0}$ for any X_p, which is also impossible. Therefore, we have $\alpha_p = \tilde{\alpha}_p, \alpha_p > \tilde{\alpha}_{p0}$, and $\xi_{p1} + \xi_{p2} \log X_p = \tilde{\xi}_{p1} + \tilde{\xi}_{p2} \log X_p$ for any X_p. Thus, $\xi_{p1} = \tilde{\xi}_{p1}, \xi_{p2} = \tilde{\xi}_{p2}$.

Using (2.24) and the above results we can get $\xi_{p0} = \tilde{\xi}_{p0}, \alpha_{p0} = \tilde{\alpha}_{p0}$ and $M = \tilde{M}$. Similarly, we can prove identifiability in (2.24) by assuming $\alpha_{p0} = \max(\alpha_{p0}, \alpha_p, \tilde{\alpha}_{p0}, \tilde{\alpha}_p)$. $\square$

In fact, we can have a more general result. Let $\theta(X)$ equal $\eta(\sum_{j=0}^{q} \beta_j X_j + \sum_{i=q+1}^{p} f_i(X_i))$, where $\eta(.)$ is strictly monotonic and $f_i(X_i) = \sum_{m,n} \beta_{imn} X_i^{p_{im}} (\log X_i)^{q_{in}}$. 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 is finite. We want to show that all parameters in $\theta(X)$ are identifiable.

Suppose $\theta(X) = \tilde{\theta}(X)$, since $\eta(.)$ 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 $X_1, ..., X_q, X_{q+2}, ..., X_p$ for example, and only consider $f_{q+1}(X_{q+1})$, where X_{q+1} is a continuous covariate,

$$\sum_{m,n} \beta_{q+1,mn} X_{q+1}^{p_{q+1,m}} (\log X_{q+1})^{q_{q+1,n}} + M = \sum_{m,n} \tilde{\beta}_{q+1,mn} X_{q+1}^{\tilde{p}_{q+1,m}} (\log X_{q+1})^{\tilde{q}_{q+1,n}} + \tilde{M} \quad (2.25)$$

Without loss of generality, assume that $p_{q+1,m} = \tilde{p}_{q+1,m}$ and $q_{q+1,n} = \tilde{q}_{q+1,n}$, because we can always add more terms with coefficients zero to both sides of (2.25). 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,

$$\sum_{m,n} (\beta_{q+1,mn} - \tilde{\beta}_{q+1,mn}) X_{q+1}^{p_{q+1,m}} (\log X_{q+1})^{q_{q+1,n}} = 0 \quad (2.26)$$

Because (2.26) is analytic in some interval $I \in R^+$, it holds for any $X_{q+1} \in R^+$. For different $p_{q+1,m}$ or $q_{q+1,n}$, $X_{q+1}^{p_{q+1,m}} (\log X_{q+1})^{q_{q+1,n}}$'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 $f_j(.) = \tilde{f}_j(.)$ for $q+1 \leq j \leq p$. The proof for the coefficient of a categorical covariate is same as before. Now we conclude that all parameters in $\theta(.)$ are identifiable.

Claims one and two proved the identifiability of (γ, θ, F) for transformation (2.14). The following lemma applies to a more general class of transformations.

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

(C1) $G_\gamma(.)$ is strictly monotonic and twice continuously differentiable with $G_\gamma(0) = 1$ and $G'_\gamma(0) \neq 0$.

(C2) If $\gamma \neq \tilde{\gamma}$, then $G''_\gamma(0)/(G'_\gamma(0))^2 \neq G''_{\tilde{\gamma}}(0)/(G'_{\tilde{\gamma}}(0))^2$.

Then (γ, θ, F) is identifiable.

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

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

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

$$\begin{aligned}G_\gamma(\alpha_1 F(y)) = G_{\tilde{\gamma}}(\beta_1 \tilde{F}(y)),\\ G_\gamma(\alpha_2 F(y)) = G_{\tilde{\gamma}}(\beta_2 \tilde{F}(y)).\end{aligned} \tag{2.27}$$

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

$$\begin{aligned}G^{-1}_{\tilde{\gamma}} \circ G_\gamma(\alpha_1 F(y)) = \beta_1 \tilde{F}(y),\\ G^{-1}_{\tilde{\gamma}} \circ G_\gamma(\alpha_2 F(y)) = \beta_2 \tilde{F}(y).\end{aligned} \tag{2.28}$$

We want to show that $g(.) = G^{-1}_{\tilde{\gamma}} \circ G_\gamma(.)$ is an identity function. Function $g(.)$ is monotonic since both $G_\gamma(.)$ and $G^{-1}_{\tilde{\gamma}}(.)$ are monotonic, which implies that β_1 and β_2 can not be zero. Otherwise $g \equiv 0$ when y takes different values. Take the ratio of the two equations in (2.28) and let $s = F(y)$. Equation (2.29) holds for $s \in [0, 1]$,

$$g(\alpha_1 s) = \frac{\beta_1}{\beta_2} g(\alpha_2 s). \tag{2.29}$$

Suppose that $\gamma \neq \tilde{\gamma}$ and the conditions (C1) and (C2) hold, we have

$$\begin{aligned} g'(0) &= \frac{G'_\gamma(x)}{G'_{\tilde{\gamma}}(G^{-1}_{\tilde{\gamma}}(G_\gamma(x)))} \mid_{x=0} = \frac{G'_\gamma(0)}{G'_{\tilde{\gamma}}(0)} \neq 0 \\ g''(0) &= \frac{G''_\gamma(x)(G'_{\tilde{\gamma}}(g(g)))^2 - (G'_\gamma(x))^2 G''_{\tilde{\gamma}}(g(x))}{(G'_{\tilde{\gamma}}(g(x)))^3} \mid_{x=0} \\ &= \frac{G''_\gamma(0)(G'_{\tilde{\gamma}}(0))^2 - (G'_\gamma(0))^2 G''_{\tilde{\gamma}}(0)}{(G'_{\tilde{\gamma}}(0))^3} \neq 0. \end{aligned} \tag{2.30}$$

Calculating the first and second order derivatives in (2.29) and plugging in $s = 0$, we will have $\alpha_1 = \alpha_2$. This contradiction leads to $\gamma = \tilde{\gamma}$. This concludes that $g(.)$ 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 $F(y) = \tilde{F}(y)$. □

It can be shown that the transformation in (2.14)

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

satisfies both conditions (C1) and (C2). We will have $G'_\gamma(0) = -1$ and $G''_\gamma(0)/[G'_\gamma(0)]^2 = 1+\gamma$ after some calculations. We may consider other transformation families as long as the conditions hold. For example, the Box-Cox type transformation mentioned in Zeng et al. [5],

$$G_\gamma(x) = \begin{cases} \exp\{-\frac{(1+x)^\gamma - 1}{\gamma}\}, & \gamma > 0, \\ \frac{1}{1+x}, & \gamma = 0. \end{cases} \tag{2.31}$$

also satisfies conditions (C1) and (C2) with $G'_\gamma(0) = -1$ and $G''_\gamma(0)/[G'_\gamma(0)]^2 = 2-\gamma$.

From the mathematics point of view, we can create a new transformation family, such as a mixture of transformations (2.14) and (2.31), although it may not be popular in practice. For example, the new transformation can be

$$G_\gamma(x) = \begin{cases} (1+\gamma x)^{-1/\gamma}, & \gamma > 0, \\ \frac{1}{1+x}, & \gamma = 0. \end{cases} \tag{2.32}$$

Now, we still have $G'_\gamma(0) = -1$ for any γ, but condition (C2) no longer holds since $G''_\gamma(0)/[G'_\gamma(0)]^2 = 2$ for both $\gamma = 0$ and 1. Therefore, we are unable to use Lemma 1 to show transformation family in (2.32) is identifiable or not. However, condition (C2) is not a necessary condition for a transformation family to be identifiable. Lemma 2 gives another set of sufficient conditions, which can show the identifiability of transformation family (2.32).

Lemma 2 . Suppose a transformation family $G_\gamma(.)$ satisfies $G_\gamma(0) = 1, G'_\gamma(0) = -1$, and $G_\gamma(.)$ is analytic for any γ, then (γ, θ, F) is identifiable.

***Proof*:** We use the same notations as in the proof of Lemma 1. Specifically, we want to show that $g(.) = G^{-1}_{\tilde{\gamma}} \circ G_\gamma(.)$ is an identity function. Since $G_\gamma(.)$ is analytic for any γ, $g(.)$ is analytic around zero too. We can use Taylor expansion for g(.) in a small neighborhood of zero. Then, we will have $g(0) = G^{-1}_{\tilde{\gamma}} \circ G_\gamma(0) = 0$ and $g'(0) = \frac{G'_\gamma(0)}{G'_{\tilde{\gamma}}(G^{-1}_{\tilde{\gamma}}(G_\gamma(0)))} = 1$. Follow (2.27), (2.28) and (2.29) in the proof of Lemma 1. Suppose there exists $k > 1$, such that the k*th* derivative of g(.) evaluated at zero is not zero, then we obtain $\alpha_1 = \alpha_2$, which is a contradiction. Therefore, $g^{(k)}(0) = 0$ for any $k > 1$. Since $g(.)$ is analytic around zero, $g(t) = t$. Thus, (γ, θ, F) is identifiable. □

2.5 Asymptotic Properties of the Semi-parametric Estimates

Zeng et al. [5] 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. We proposed a generalized transformation cure model (2.13) with link function (2.15). In our proposed model, fractional polynomials are used instead of the simple linear combination of the covariates. Following the Theorem 1 and Theorem 2 in Zeng et al. [5], we can prove the same properties of the maximum likelihood estimation of $(\boldsymbol{\beta}, F)$ in the proposed model.

In the model (2.13) with link function (2.15), the likelihood function is given by

$$\begin{aligned} L(\boldsymbol{\beta}, F) &= [\{-G'(\eta(\boldsymbol{\beta}, \boldsymbol{X})F(Y))\eta(\boldsymbol{\beta}, \boldsymbol{X})f(Y)\}^{\Delta} \\ &\quad \{G(\eta(\boldsymbol{\beta}, \boldsymbol{X})F(Y))\}^{1-\Delta}]^{I(Y<\infty)}[G(\eta(\boldsymbol{\beta}, \boldsymbol{X}))]^{I(Y=\infty)}. \end{aligned} \tag{2.33}$$

Given observations $(\{Y_i, \boldsymbol{X}_i, \Delta_i\}, i = 1, ..., n)$, the maximum likelihood estimates of $(\boldsymbol{\beta}, F)$, denoted by $(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)$, are derived from the modified semi-parametric version of (2.33),

$$\begin{aligned} L(\boldsymbol{\beta}, F) &= \prod_{i=1}^{n}\{[\{-G'(\eta(\boldsymbol{\beta}, \boldsymbol{X_i})F(Y_i))\eta(\boldsymbol{\beta}, \boldsymbol{X_i})F\{Y_i\}\}^{\Delta_i} \\ &\quad \times\{G(\eta(\boldsymbol{\beta}, \boldsymbol{X_i})F(Y_i))\}^{(1-\Delta_i)}]^{I(Y_i<\infty)} \times [G(\eta(\boldsymbol{\beta}, \boldsymbol{X_i}))]^{I(Y_i=\infty)}\}, \end{aligned} \tag{2.34}$$

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

To obtain consistency and asymptotic normality, we make the following assumptions through out this section:

(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 $\in \mathbb{R}^+$. The density function of F_0, denoted by f_0, is bounded from above and below in any compact sets.

(C4). Conditional on $\boldsymbol{X}$, the right censoring time C is independent of T, and $S_C(\infty|\boldsymbol{X}) > 0$. The density functions of T and C are bounded from below and above in any compact sets, respectively.

(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

Theorem 1. The maximum likelihood estimates $(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)$ based on (2.34) are strongly consistent, that is

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

Theorem 2. $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0, \hat{F}_n - F_0)$ converges weakly to a Gaussian process.

Proofs and more discussions are provided in Appendix A.

CHAPTER 3

SIMULATION

Many studies have shown that for a continuous covariate, say X_1, $X_1^{p_0}$ sometimes is a better predictor than X_1 itself. The power p_0 can be any real numbers, positive, negative, integer or fraction. When the power p_0 is zero, a logarithm transformation is often used for X_1. This has been seen in various models, including linear regression model, logistic regression, survival model, etc. How to select the power p_0 to fit the data better is something worth considering. In this chapter, we carried out a series of simulation studies to show how to select the power transformation in different models. The sum of squared residuals and the likelihood function are used in the model selection procedure.

3.1 Linear Regression Model

In the first simulation, we consider the linear regression model

$$Y = \beta_0 + \beta_1 X_1^{p_0} + \beta_2 X_2 + \varepsilon, \tag{3.1}$$

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]; covariate X_2 is a Bernoulli random variable with probability 0.5; and the random error ε is normally distributed with mean 0 and standard deviation σ. When $p_0 = 0$, we use the model

$$Y = \beta_0 + \beta_1 \log(X_1) + \beta_2 X_2 + \varepsilon. \tag{3.2}$$

Actually, the range of X_1 can be arbitrary in the linear regression model. Here, we choose the lower bound and the upper bound of X_1 to be reciprocals of each other. The benefit of doing this is to control the values of $X_1^{p_0}$ in a similar range when p_0 varies from -2 to 2, which will result in the response variables in a similar range for different p_0 values. This restriction

of X_1 values can improve the accuracy of p_0 estimation, especially in survival models where $X_1^{p_0}$ will greatly affect the length of life time and the censoring rate. In practice, if X_1 is not in such a range, we can always apply a transformation on X_1, such as divided by the geometric mean of the values of X_1. Then the standardized data will be in a range, such that the product of the lower and upper bounds is roughly one. In this chapter, to make the simulations comparable we choose the same range for X_1, from 0.5 to 2, in different models.

We choose σ to be 0.25, 1, and 2. The coefficients in model (3.1) and (3.2) are unknown constants. In this simulation, we set $\beta_0 = 0.5, \beta_1 = 1.5, \beta_2 = 0.7$. The sample size n in each simulated data set is set to be 500. For each combination of p_0 and σ, we fit a series of regression models and calculate the residual sum of squares. As Royston and Altman [16] suggested, the selected power, denoted by $\hat{p}_0$, is chosen from a set of numbers $A = (-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2)$. Suppose the generated data set is $\{(X_{1i}, X_{2i}, Y_i), i = 1, ..., n\}$, where n is the number of subjects, we calculate

$$SS(\boldsymbol{\beta}, p_0) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{1i}^{p_0} - \beta_2 X_{2i})^2 \tag{3.3}$$

for each fitted model when $p_0 \in (-2, -1.5, -1, -0.5, 0.5, 1, 1.5, 2)$. When $p_0 = 0$,

$$SS(\boldsymbol{\beta}, p_0) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 \log(X_{1i}) - \beta_2 X_{2i})^2 \tag{3.4}$$

is calculated instead of (3.3). The power p_0 is chosen such that the residual sum of squares $SS(\boldsymbol{\beta}, p_0)$ reaches its minimum value. We simulated 200 data sets for each combination of p_0 and σ based on model (3.1) and (3.2), and choose p_0 for each data set. Then we calculate the mean of $\hat{p}_0$'s and count the number of times of choosing the true value. The results are summarized in Table 3.1.

In Table 3.1, the columns labeled "mean" show the average of the 200 selected powers and the columns " freq." present the number of times that the true power was chosen among the 200 replications. When $\sigma = 0.25$, the power selection procedure performs well. For example, when $p_0 = -1$, the mean of selected power is -0.993, which is very close to -1. In this case the true power is chosen 195 times among the 200 simulations. When σ increases, the value of "mean" begins to move away from the true power and the value of "freq." drops. When $\sigma = 1$ and $p_0 = -1$, the mean of selected power is -0.948. The value of "freq." is 90, which decreases more than 50% compared with "freq." =195 when $\sigma = 0.25$. When the standard

Table 3.1: Results of power selection under the linear regression model based on 200 simulated data sets with sample size n=500 and coefficients $\beta_0 = 0.5, \beta_1 = 1.5, \beta_2 = 0.7$

	σ=0.25		σ=1		σ=2	
p_0	mean	freq.	mean	freq.	mean	freq.
-2	-2.000	200	-1.945	178	-1.845	145
-1.5	-1.500	200	-1.500	119	-1.448	68
-1	-0.993	195	-0.948	90	-1.088	50
-0.5	-0.480	155	-0.493	34	-0.390	20
0	-0.010	194	-0.018	106	-0.045	48
0.5	0.488	143	0.508	54	0.508	25
1	1.003	197	1.003	84	0.968	46
1.5	1.500	200	1.488	127	1.465	73
2	2.000	200	1.963	185	1.858	148

deviation is increased to $\sigma = 2$, the results are worse than other two cases. The simulation results indicate that large variability of the random error term may cause problems when selecting the true power value p_0.

For a fixed σ, the power selection procedure performs better when p_0 is equal to -2 or 2 than other powers. The number of times of choosing the true power decreases gradually when p_0 is towards 0.5 and -0.5. For example, when $\sigma = 1$, "freq." is 178 for $p_0 = -2$, and it decreases to 34 when $p_0 = 0.5$. The "freq." of choosing the correct power also decreases when p_0 is positive and changes from 2 to 0.5. For $\sigma = 1$, "freq." is 185 for $p_0 = 2$ and drops to 54 for $p_0 = 0.5$.

Figure 3.1 shows the histograms of the estimated powers for different p_0 when $\beta_0 = 0.5, \beta_1 = 1.5, \beta_2 = 0.7$, and $\sigma = 1$, which indicates that as long as the variance of the random errors is not too large, even if the estimated powers are not as same as the true one, they are still very close to p_0 and all centered around it. For example, when $p_0 = 0$, there are 106 times that $\hat{p}_0$ equals 0, 42 times and 32 times $\hat{p}_0$ is selected to be -0.5 and 0.5, respectively. There are only 8 times and 11 times that $\hat{p}_0$ equals -1 and 1, respectively.

Another simulation is conducted with coefficients $\beta_0 = 5, \beta_1 = 15$, and $\beta_2 = 7$. In the previous simulation, the mean of Y is between 1 and 3, but now it is ten times larger. The results are shown in Tables 3.2. When $\sigma = 2$, the power selection procedure appears to perform well. For example, there are 199 times we select the true power when

$p_0 = -1$. The average of $\hat{p}_0$ is -0.998. The power selection procedure performs perfectly when $\sigma = 0.25$. When $\sigma = 10$, both "mean" and "freq." are comparable to the results when $\beta_0 = 0.5, \beta_1 = 1.5$, $\beta_2 = 0.7$, and $\sigma = 1$. All of these show that the power selection procedure performs well when the variance of random errors is relatively small compared to the mean of Y. If the variance of the random error ϵ is small, the generated data will be centered closed around its mean and the power selection procedure may perform perfectly.

Table 3.2: Results of power selection under the linear regression model based on 200 simulated data sets with sample size n=500 and coefficients $\beta_0 = 5, \beta_1 = 15, \beta_2 = 7$

	σ=0.25		σ=2		σ=10	
p_0	mean	freq.	mean	freq.	mean	freq.
-2	-2.000	200	-2.000	200	-1.935	174
-1.5	-1.500	200	-1.500	200	-1.490	116
-1	-1.000	200	-0.998	199	-0.983	88
-0.5	-0.500	200	-0.520	166	-0.483	38
0	0.000	200	0.003	199	0.025	76
0.5	0.500	200	0.505	168	0.410	52
1	1.000	200	1.000	200	1.033	81
1.5	1.500	200	1.500	200	1.528	131
2	2.000	200	2.000	200	1.948	179

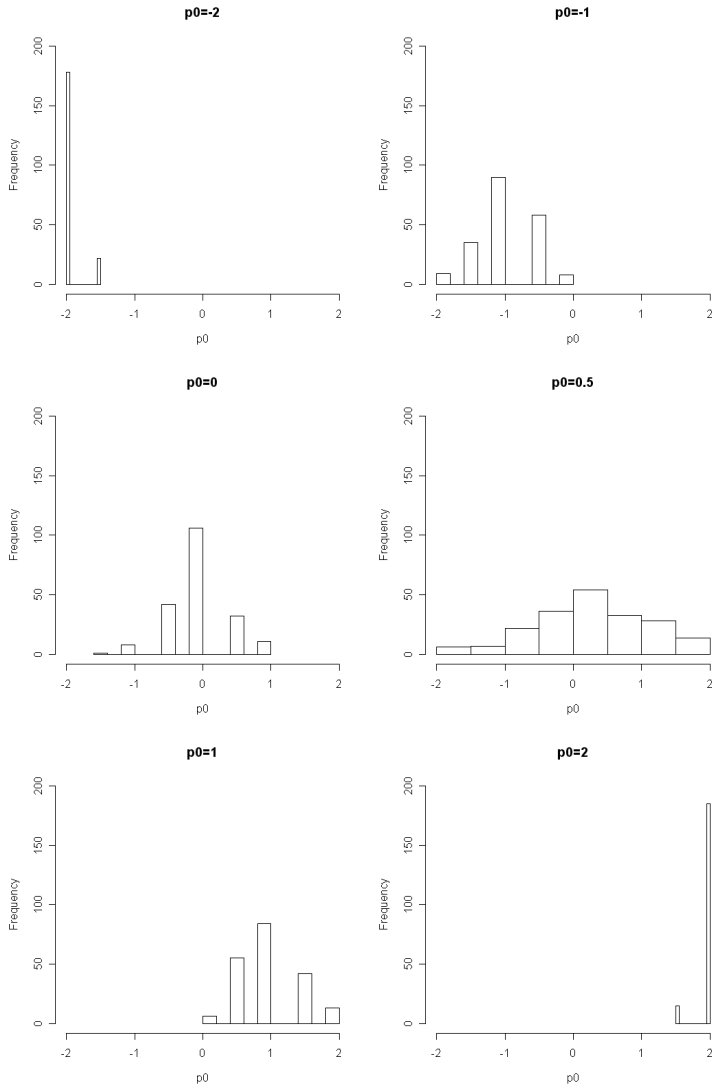

Figure 3.1: Frequencies of true power selected for different p_0's based on 200 replications under the linear regression model with sample size $n = 500$, and coefficients $\beta_0 = 0.5, \beta_1 = 1.5, \beta_2 = 0.7$, and $\sigma = 1$.

3.2 Logistic Regression Model

Logistic regression is very commonly used in many fields, especially in medical or social sciences when people want to predict the probability of occurrence of an event. For example, the probability that a person satisfies his/her job might be predicted from the person's age, salary, education, and location, etc.

Logistic regression model fits the log odds by a linear function of the covariates. What we considered in this simulation is the following model,

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1^{p_0} + \beta_2 X_2, \tag{3.5}$$

where p_0 is nonzero. The binary response variable, denoted by Y, takes value 0 or 1 only for non-occurrence or occurrence of the event respectively. The variable Y equals 1 with probability π and is equal to 0 with probability $1-\pi$. As in the linear regression model defined in (3.1), covariate X_1 is a uniform random variable in [0.5, 2] and covariate X_2 is Bernoulli distributed with probability 0.5. The coefficients β_0, β_1, and β_2 are assumed to be constants. When power $p_0 = 0$, we fit the model

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \log(X_1) + \beta_2 X_2, \tag{3.6}$$

For a given data set, we are interested in selecting a transformation power p_0 from a set of numbers A=(-2, -1.5, -1, 0.5, 0, 0.5, 1, 1.5, 2), such that the model fits the data best. We generate data from either model (3.5) or (3.6), and then fit models with different powers in set A. The optimal power is chosen such that the corresponding model has the highest likelihood.

In logistic regression, taking different values for coefficients β_0, β_1, and β_2 yields different probabilities π's, which are the probabilities of Y being equal to 1. We want to choose appropriate β_0, β_1, and β_2 such that the probabilities are not smaller than 0.1 and also not greater than 0.9 when p_0 varies from -2 to 2. Otherwise, there is a high chance of generating some data sets where Y equals 0 or 1 for all subjects. We conduct two simulation with different sample sizes and set $\beta_0 = 1, \beta_1 = -3$, and $\beta_2 = 0.7$ for both of the simulations. The results are summarized in Table 3.3. The columns labeled "prob" show the probabilities of Y being equal to 1 for each p_0 on average. For example, when p_0 equals -1, the probability is 0.249 for both sample sizes 500 and 2000.

The columns labeled "mean" and "freq." are defined same as in the Section 3.1. When the sample size is 500, the power selection procedure does not appear to work well. When p_0 equals -2 and 2, the mean of selected power $\hat{p}_0$'s are -1.713 and 1.780 respectively, which are not very close to the true powers. When p_0 takes other values, the mean of $\hat{p}_0$ appears to be fine. For example, when $p_0 = -1$ the average of $\hat{p}_0$'s is -1.105. But if we look at the number of times the true power is selected, the accurate rates are lower than 30% at most of the time. For example, when $p_0 = -1$, the real power has been chosen for only 44 times among the 200 simulations. But if we increase the sample size to 2000, both "mean" and "freq." can be greatly improved. The mean of $\hat{p}_0$ is very close to the true power and the bias is less than 0.1 for every p_0. The true power is selected for about or more than 50% of the time when $p_0 = -1.5, -1, 0, 1, 1.5$, and for more than 70% of the time when $p_0 = -2$ or 2. For example, when p_0 is -1, the mean of selected power is -1.038 on average, and there are 94 times the true power is selected among the 200 simulations.

Figure 3.2 shows the histograms of the selected $\hat{p}_0$ based on 200 replications under the logistic regression model with sample size n=2000. By using the proposed power selection procedure, we can choose the true p_0 most of the time and the selcted $\hat{p}_0$'s are all center around the true power. For example, when $p_0 = 0$, the true power 0 is selected 121 times; -0.5 is selected 37 times; and 0.5 is selected 36 times.

Table 3.3: Results of power selection under the logistic regression model based on 200 simulated data sets with coefficients $\beta_0 = -1, \beta_1 = -3, \beta_2 = 0.7$

	n=500			n=2000		
p_0	prob.	mean	freq.	prob.	mean	freq.
-2	.327	-1.713	122	.327	-1.843	142
-1.5	.291	-1.580	44	.291	-1.553	104
-1	.249	-1.105	44	.249	-1.038	94
-0.5	.205	-0.428	33	.205	-0.483	71
0	.658	-0.050	54	.658	0.013	121
0.5	.144	0.588	18	.144	0.448	72
1	.139	1.008	42	.139	1.013	98
1.5	.146	1.478	56	.145	1.503	105
2	.156	1.780	143	.156	1.858	148

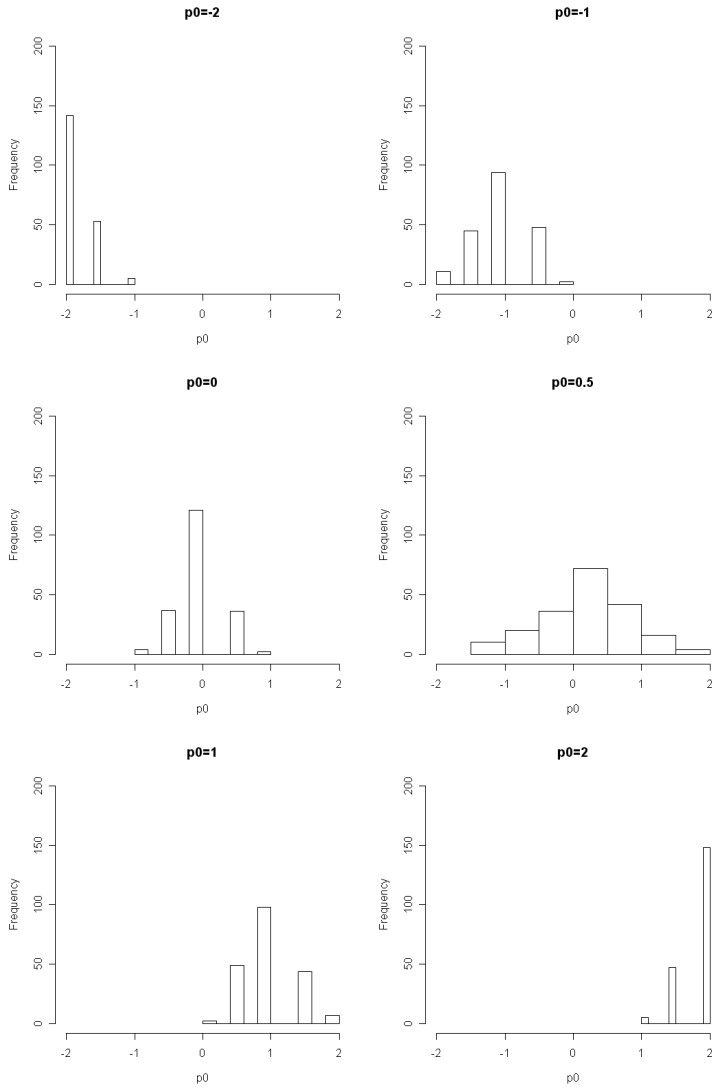

Figure 3.2: Frequencies of true power selected for different p_0's based on 200 replications under the logistic regression model with sample size $n = 2000$,and coefficients $\beta_0 = 1, \beta_1 = -3, \beta_2 = 0.7$.

3.3 Cox Model

In this section, we consider the Cox proportional hazards regression model, which is probably the most popular survival model. This model was first presented by Cox [6], which separates time and covariates with the hazard rate

$$h(t|\boldsymbol{X}) = h_0(t)\exp(\boldsymbol{\beta'X}), \tag{3.7}$$

where $h_0(t)$ is the baseline hazard function. A linear term $\boldsymbol{\beta'X}$ without intercept is used in model (3.7), where $\boldsymbol{X}$ is the vector of covariates, and $\boldsymbol{\beta}$ is the vector of coefficients.

As in the previous sections we use fractional polynomial instead of $\boldsymbol{\beta'X}$, and take two covariates for example. The model is given by

$$h(t|\boldsymbol{X}) = h_0(t)\exp\left(\beta_1 X_1^{(p_0)} + \beta_2 X_2\right). \tag{3.8}$$

where X_1 is uniformly distributed in [0.5, 2] and X_2 is bernoulli distributed with probability 0.5. The coefficients β_1 and β_2 are constants. The Box-Cox transformation is used for X_1 in (3.8),

$$X_1^{(p_0)} = \begin{cases} \frac{X_1^{p_0}-1}{p_0}, & p_0 \neq 0, \\ \log(X_1), & p_0 = 0. \end{cases} \tag{3.9}$$

The baseline hazard function $h_0(t)$ is given by

$$h_0(t) = \lambda t^{\lambda-1}\exp(\beta_0), \tag{3.10}$$

where λ is a shape parameter and β_0 is a constant. When $\lambda = 1$, the model (3.8) reduces to an exponential model with constant baseline hazard $h_0(t) = \exp(\beta_0)$. For other values of λ, (3.8) is a Weibull distribution model, where the baseline hazard is not constant. It is monotone decreasing when $\lambda < 1$ and monotone increasing when $\lambda > 1$.

In this simulation, we generate the survival life time based on (3.8), (3.9) and (3.10). The survival function is given by

$$\begin{aligned} S(t|X) &= \exp\left(-\int_0^t h_0(s)ds\right) \\ &= \exp\left(-t^\lambda \exp(\beta_0)\cdot\exp(\beta_1 X_1^{(p_0)} + \beta_2 X_2)\right). \end{aligned} \tag{3.11}$$

Thus, the life time T will be generated by

$$T = \left(-\frac{\log(U)}{\exp(\beta_0) \cdot \exp(\beta_1 X_1^{(p_0)} + \beta_2 X_2)} \right)^{1/\lambda}, \tag{3.12}$$

where U has a uniform distribution in [0, 1]. The censoring time, C, is generated from an exponential distribution with mean 5, which can be viewed as 5 years. Now, the complete data set $(Y, \boldsymbol{X}, \Delta)$ is given by

$$\begin{aligned} Y &= \min(T, C), \\ \boldsymbol{X} &= (X_1, X_2)', \\ \Delta &= \begin{cases} 1, & T \leq C \\ 0, & T > C. \end{cases} \end{aligned} \tag{3.13}$$

Choosing different values for $\beta_0, \beta_1, \beta_2$, and λ will greatly affect the life time T. The mean of T is expected to be greater than one 1 year and the censoring rate is expected to be around 15%. We conduct three simulations by taking λ equals 0.5, 1, and 2, respectively. To meet the conditions of T and the censoring rate, the values for β_0, β_1, and β_2 in the simulations are also set to be different. Specifically, we generate data from the following 3 cases:

(1) $\lambda = 0.5, \beta_0 = -0.5, \beta_1 = 4, \beta_2 = 3$;

(2) $\lambda = 1, \beta_0 = -1, \beta_1 = 4, \beta_2 = 3$;

(3) $\lambda = 2, \beta_0 = -1, \beta_1 = 5, \beta_2 = 3$.

Using the coefficients in (1)-(3), about 90% of the generated life times T are less than 10, and the means are around or greater than 1. There are always some outliers in the generated life time T, especially when p_0 takes negative values. The outliers may be quite large and sometimes even greater than 1000. Since we set the mean of censoring time to be 5, most of the outliers will be censored.

We fix the sample size to be 500 and conduct 200 simulations for each combination of p_0 and λ. Then, we choose a $\hat{p}_0$ from the set A for each simulated data set, such that the corresponding power transformation model give us the maximum likelihood.

The results are summarized in Table 3.4. The columns labeled "mean" show the mean of the selected $\hat{p}_0$; the columns labeled "freq." present the number of times of choosing the true power among the 200 simulations. The power selection procedure appear to work well under

Table 3.4: Results of power selection under the Cox model based on 200 simulated data sets with sample size n=500

	λ=0.5		λ=1		λ=2	
p_0	mean	freq.	mean	freq.	mean	freq.
-2	-1.920	168	-1.893	158	-1.943	177
-1.5	-1.515	142	-1.520	148	-1.513	163
-1	-1.038	145	-1.000	150	-1.010	170
-0.5	-0.495	152	-0.538	155	-0.500	172
0	0.007	149	0.008	151	-0.003	179
0.5	0.503	165	0.478	167	0.510	174
1	1.018	165	1.035	162	1.020	178
1.5	1.495	174	1.500	168	1.515	180
2	1.975	190	1.965	186	1.985	194

the Cox model. At least 70% of the time we choose $\hat{p}_0$ as same as the true power. When p_0 equals -2 or 2, there are more 80% of the time of choosing the true power. The value of "freq." decreases from both sides towards $p_0 = 0$, which is comparable to the results we get under the linear regression model. See Table 3.1. When $\lambda = 2$, the hazard rate increases linearly over time. Therefore, comparing with $\lambda = 0.5$ and $\lambda = 1$ the generates life times are relatively smaller, which results in fewer outliers in this case. When $\lambda = 2$, the values of "freq." are higher than the ones when $\lambda = 0.5$ or 1 for every p_0. For example, when $p_0 = -1$, we select p_0 correctly 170 times when $\lambda = 2$. It is higher than both "freq." when $\lambda = 0.5$ or 1, which are 145 times and 150 times, respectively.

Figure 3.3 to Figure 3.5 show the histograms of the selected $\hat{p}_0$ for $\lambda = 0.5, 1$ and 2, respectively, when $p_0 = -2, -2, 0, 0.5, 1$ or 2. By using the proposed power selection procedure, we can choose the true p_0 most of the time for different λ's, and the selcted $\hat{p}_0$ are all center around the true power. For example, when $p_0 = 0$ and $\lambda = 2$, the true power 0 is selected 179 times; -0.5 is selected 11 times; and 0.5 is selected 10 times. None of the other powers is selected.

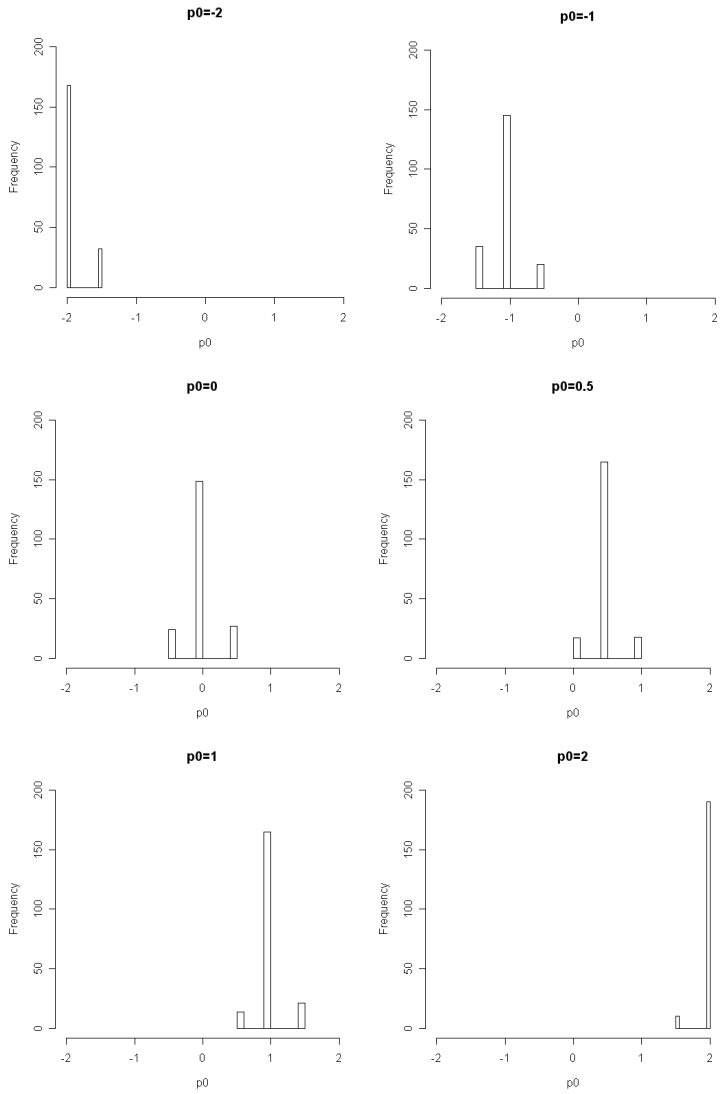

Figure 3.3: Frequencies of true power selected for different p_0's based on 200 replications under Cox model with sample size $n = 500$ and shape parameter $\lambda = 0.5$.

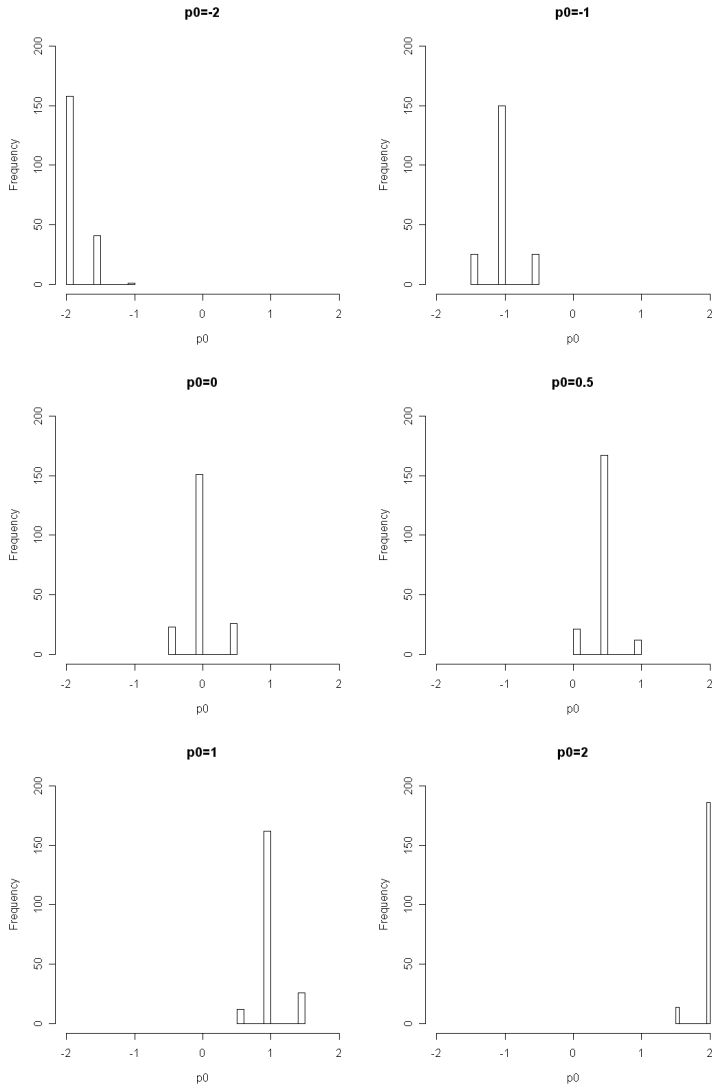

Figure 3.4: Frequencies of true power selected for different p_0's based on 200 replications under Cox model with sample size $n = 500$ and shape parameter $\lambda = 1$.

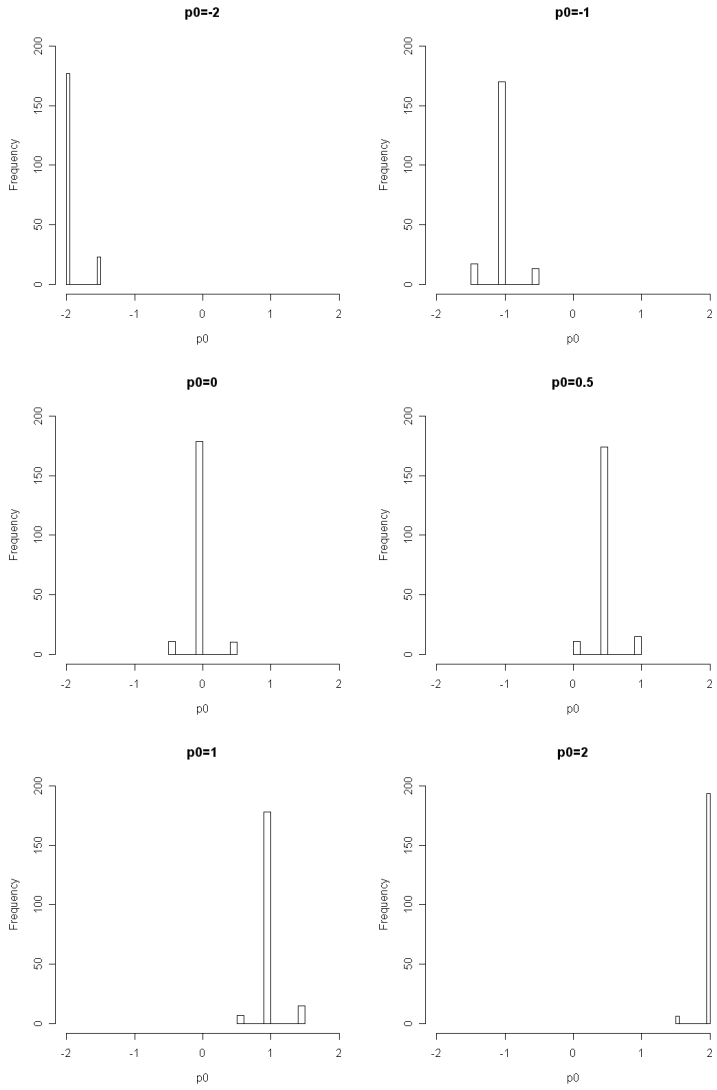

Figure 3.5: Frequencies of true power selected for different p_0's based on 200 replications under Cox model with sample size $n = 500$ and shape parameter $\lambda = 2$.

3.4 Generalized Transformation Model

In this section, we conduct simulations to study generalized transformation models and to examine the performance of our proposed power selection procedure on generalized transformation models. The model has a survival function of the form

$$S(t|\boldsymbol{X}) = G_\gamma(\theta(\boldsymbol{X})F(t)), \tag{3.14}$$

where

$$G_\gamma(x) \;=\; \begin{cases} (1+\gamma x)^{-1/\gamma}, & \gamma > 0, \\ e^{-x}, & \gamma = 0. \end{cases} \tag{3.15}$$

In this simulation study, we take γ equal to 0 and consider the following link function,

$$\theta(\boldsymbol{X}) = \exp(\beta_0 + \beta_1 X_1^{p_0} + \beta_2 X_2), \tag{3.16}$$

where p_0 is a nonzero power varying from -2 to 2. Same as in the previous sections, 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

$$\theta(\boldsymbol{X}) = \exp(\beta_0 + \beta_1 \log(X_1) + \beta_2 X_2). \tag{3.17}$$

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

Survival times of subjects with covariate 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 (3.14) and (3.15). Therefore, the life time T will be generated from

$$T_i = -\log\left(1 + \frac{\log(1-(1-\exp(-\theta(\boldsymbol{X}_i)))U)}{\theta(\boldsymbol{X}_i)}\right) \tag{3.18}$$

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

We assume each subject can be right-censored with a probability $q < 1$, which will guarantee that at least some cured subjects will not be right censored. We will carry out two simulation studies with $q = 80\%$ and $q = 40\%$, respectively, to illustrate the difference when the proportion of censored population differs. So, the censoring time C_i for the ith individual in the data set will equal ∞ with a 20% or 60% of chance in the two simulations, respectively. For the rest of the population, the censoring time is generated from an exponential distribution with mean 1.

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

$$\begin{aligned} Y_i &= \min(T_i, C_i), \\ \boldsymbol{X}_i &= (X_{i1}, X_{i2})', \\ \Delta_i &= \begin{cases} 0, & T_i > C_i, \\ 1, & T_i \leq C_i,\ T_i \neq \infty, \\ 2, & T_i = \infty,\ C_i = \infty. \end{cases} \end{aligned} \tag{3.19}$$

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

We conduct two simulations with different sample sizes and set $\beta_0 = -0.5$, $\beta_1 = 1$, and $\beta_2 = 0.7$. Actually, $\boldsymbol{\beta}$ can take any values, we choose $(-0.5, 1, 0.7)$ such that the cured proportions vary from 5% to 10% as p_0 changes from -2 to 2. 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), such that the corresponding generalized transformation model provide us the maximum likelihood based on the likelihood function given in (2.16).

The results summarized in Table 3.5 show 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 $q = 80\%$ for sample sizes 2000 and 5000, respectively. Same as in the previous sections, the columns labeled "mean" are the average of the selected powers and the columns labeled "freq." are the number of times of selecting the true power in the 200 simulations. When the sample size is 2000, the power selection procedure does not work very well. The true powers are selected for less than 45% of the time except for $p_0 = -2$ and $p_0 = 2$. The reason why the accurate rates are higher for these two cases is we only select the powers in the range of -2 to 2. This also explains why the absolute values of means of the selected powers when $p_0 = -2$ and 2 tends to be smaller. If powers beyond -2 and 2 are allowed to be selected, $p =_0= -2$ and 2

Table 3.5: 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

should have less chance to be underestimated. When p_0 takes other values, the means of the selected powers are very close to the true value. For example, when $p_0 = -1$ the estimated mean is -0.995. The power selection procedure performs much better when the sample size increases to 5000. The accurate rates of choosing the true power climb up to higher than 50% for most of the cases. For example, when $p_0 = -1$ the true power is selected for 104 times comparing with 75 times when $n = 2000$.

Table 3.6 shows more results of power selection with sample size $n = 5000$ and the probability of each subject being right-censored $q = 80\%$ based on 200 simulations. Each column represents one scenario. For example, when $p_0 = -1$, the true power -1 is selected 104 times; -1.5 and -0.5 are selected 44 and 42 times, respectively; and -1 and 0 are selected 5 times each. These results indicates that the selected powers are all center around the true power. Figure 3.6 also shows frequencies of power selected for several different p_0's based on 200 replications under the proposed model with sample size $n = 5000$ and the probability of each subject being right-censored $q = 80\%$.

Table 3.6: 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

When the probability of each subject being censored drops to $q = 40\%$, the proportion of censored population decreases. If $q = 80\%$, there are $35-40\%$ of the population will be right censored when the power p varies from -2 to 2. When $q = 40\%$, this percentage will drop to $30 - 35\%$ and occasionally lower than 30%. The results of power selection with sample size $n = 5000$ and the probability of each subject being right-censored $q = 40\%$ based on 200 simulations are summarized in Table 3.7. Results in Table 3.7 have the same pattern as the ones in Table 3.6 and are slightly better most of the times than Table 3.6 where $q = 80\%$. For example, when p is -1.5, the true power is selected 91 times and 116 times when $q = 80\%$ and 40%, respectively. The mean of selected power based on the 200 simulations is -1.488 when $q = 80\%$ and is -1.498 when $q = 40\%$. This indicates that the probability of being right censored for each subject will not effect the power selection much when sample size $n = 5000$. We use two different probabilities $q = 80\%$ and $q = 40\%$ as an example and the power selection procedure works equally well.

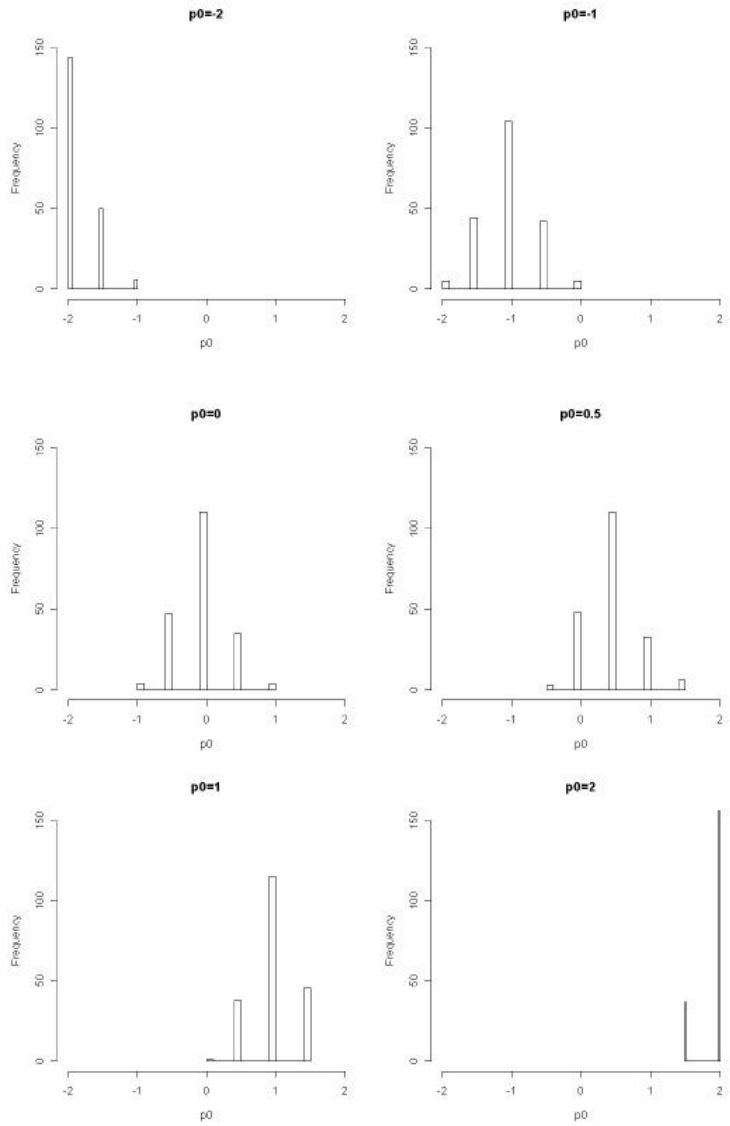

Figure 3.6: Frequencies of true power selected for different p_0's based on 200 replications under the proposed model with sample size $n = 5000$ and the probability of each subject being right-censored $q = 80\%$.

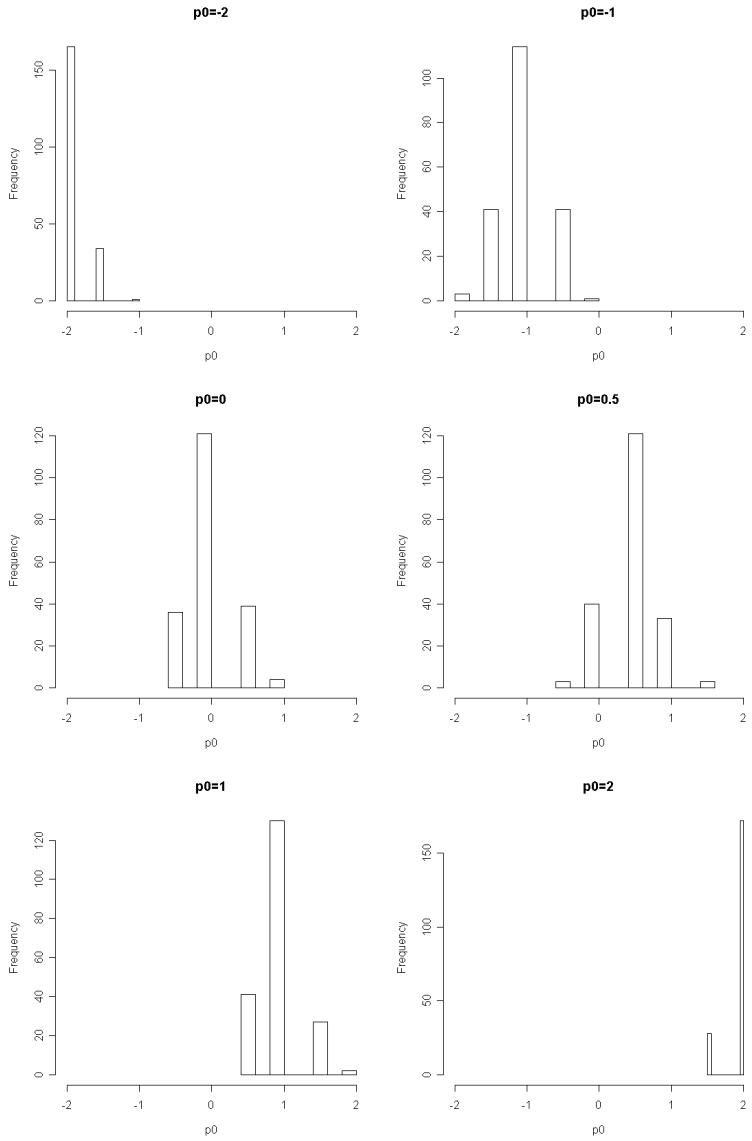

Figure 3.7: Frequencies of true power selected for different p_0's based on 200 replications under the proposed model with sample size $n = 5000$ and the probability of each subject being right-censored $q = 40\%$.

Table 3.7: 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 $q = 40\%$

	True power								
Selected power	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2
-2	165	43	3						
-1.5	34	116	41						
-1	1	38	114	50					
-0.5		3	41	107	36	3			
0			1	41	121	40			
0.5				1	39	121	41		
1				1	4	33	130	30	
1.5						6	27	134	28
2							2	36	172
mean	-1.910	-1.498	-1.010	-0.510	0.028	0.483	0.975	1.515	1.930

CHAPTER 4

APPLICATION

In this chapter, we will apply the proposed transformation models to some cancer and coronary heart disease related medical data from both clinical trials and observational cohort studies.

4.1 Melanoma Data E1690

In the first example, we will illustrate our proposed models by using melanoma data labeled E1690 [17]. Melanoma is one type of cancer with high risk of disease recurrence after definitive surgery. The incidence rate of melanoma depends on the number of reginal lymph nodes and the deepness of the primary tumor. Several adjuvant chemotherapies have been used for melanoma patients, including interferon alpha-2b (IFN) that has shown significant efficacy on both relapse free survival and overall survival. The phase III clinical trial E1690 conducted by Eastern Cooperative Oncology Group (ECOG) was one of the clinical trials to investigate the impact of interferon on melanoma patients.

In this clinical trial, there were two arms comparing high-dose interferon and observation. The study accrued a total number of 427 patients from year 1991 to year 1995, of which 215 patients were in the treatment group and 212 patients were in the control group. Several other covariates were included in the E1690 study, such as age, gender, and nodes. Specifically, age is a continuous covariate varying from 19.13 years to 78.05 years. The mean and the standard deviation of age are 47.93 years and 13.15 years, respectively. There were 268 males and 159 females in the study. The number of positive nodes was recorded in four categories, 0, 1, 2-3, and more than 4 nodes. But in this application we will simply group nodes into two categories. We use 0 for zero nodes and use 1 for one or more than one nodes.

The response variable in the E1690 study is relapse-free survival time. Among a total of

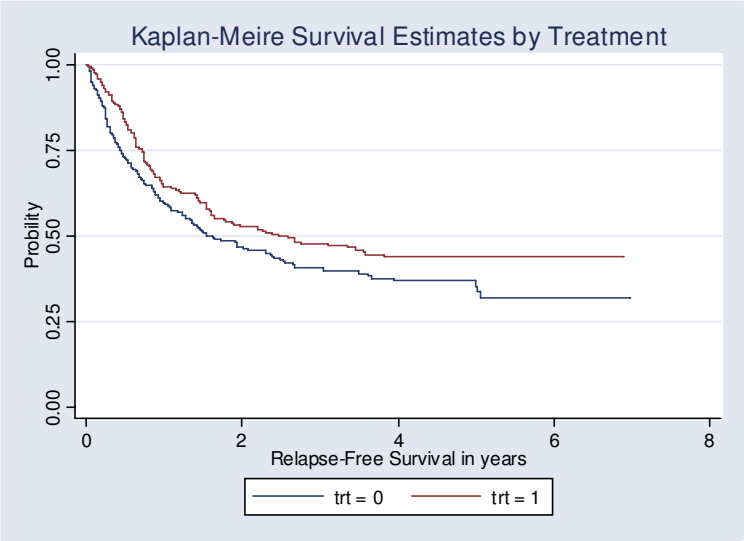

Figure 4.1: Kaplan-Meier estimates of relapse-free survival by treatment arm based on the E1690S study. The red line is treatment arm and the blue line is control arm.

427 patients, there were 241 patients experienced the event. Figure 4.1 shows the Kaplan-Meier estimates of survival time by treatment arm. From the estimated curves, we can clearly see the impact of IFN on the probability of surviving over time. For example, the 4-year survival rate for patients in the treatment arm is around 45%, compared with 40% in the control arm. From Figure 4.1, we also see that a shape of plateau appears at the tail of the survival curves. Such curves are often analyzed by using cure rate models. In this section, besides the Cox proportional hazards model we will apply the transformation cure model by Zeng et al. [5], as well as our proposed generalized transformation model to the E1690 data.

4.1.1 Cox Proportional Hazards Model

First of all, we fit a Cox Proportional hazards model including treatment, age, nodal category, and gender. The results are summarized in Table 4.1. There is a significant (at 0.1 level) impact of treatment on melanoma patients and the corresponding hazard ratio is exp(-0.229)=0.795. This implies that the hazard rate for relapse faced by a patient treated by using IFN is only 79.5% of the hazard rate faced by a patient who did not have the IFN treatment. Cavitate age is a very significant factor in the fitted model. The hazard of melanoma relapse increases by 1.2% when the age of a patient increases one year. The results in Table 4.1 also show that males faced a higher hazard rate than females, and of course with or without primary nodes would make a huge difference. A test of checking the proportional hazards assumption based on Schoenfeld residuals [18] is performed and shown in Table 4.2. We find no evidence against the assumption.

Table 4.1: Fitted Cox proportional hazards model to E1690 study.

Variable	Coef.	Std. Err.	z	Prob> $\|z\|$
Treatment	-0.229	0.130	-1.76	0.079
Age	0.011	0.005	2.27	0.023
Nodal category	0.550	0.160	3.43	0.001
Gender	-0.233	0.138	-1.69	0.091

Table 4.2: Test of proportional hazards assumption based on the Schoenfeld residuals.

Variable	Rho	Chi2	Df.	Prob>chi2
Treatment	0.005	0.01	1	0.9410
Age	0.070	1.21	1	0.2723
Nodal category	-0.069	1.20	1	0.2725
Gender	-0.008	0.01	1	0.9041
Global test		2.94	4	0.5674

4.1.2 Transformation Cure Model by Zeng et al. [5]

In the E1690 data, there were 241 patients experienced the event with a mean of 1.04 years and a maximum of 5.07 years, and 186 patients were right censored in this study. We choose 5.5 years as a threshold value to identify the cured population. There were 30 patients who were right censored and had censoring time longer than 5.5 years. These patients are considered cured subjects. In fact, the estimates of regression coefficients do not depend on the choice of such a threshold value as long as it is greater than 5.07 years, the maximum event time.

We fit the transformation cure model in (2.6)

$$S(t|\boldsymbol{X_i}) \quad = \quad G_\gamma(\theta(\boldsymbol{X_i})F(t)), \tag{4.1}$$

with transformation family (2.7)

$$G_\gamma(x) \quad = \quad \begin{cases} (1+\gamma x)^{-1/\gamma}, & \gamma > 0, \\ e^{-x}, & \gamma = 0. \end{cases} \tag{4.2}$$

Let γ vary from 0 to 2 and we choose the one that maximizes the log-likelihood function

$$\begin{aligned} &\prod_{i=1}^{n}\{[\{-G'(\eta(\boldsymbol{\beta' X_i})F(Y_i))\eta(\boldsymbol{\beta' X_i})F\{Y_i\}\}^{\Delta_i} \\ &\times\{G(\eta(\boldsymbol{\beta' X_i})F(Y_i))\}^{(1-\Delta_i)}]^{I(Y_i<\infty)} \times [G(\eta(\boldsymbol{\beta' X_i}))]^{I(Y_i=\infty)}\}. \end{aligned} \tag{4.3}$$

The observed log-likelihood is plotted in Figure 4.2 with different values of γ. In the E1690 data, $\gamma = 0$ is chosen with log-likelihood being equal to -1579.36. The corresponding estimates of regression coefficients are summarized in Table 4.3. The results are comparable with that in the Cox proportional hazards model. P-values for treatment INF and gender are 0.0906 and 0.1076, respectively. Covariates age and nodal category are both significant at the 0.05 level. Besides estimating the effect of each covariate, the cure model can be used to estimate the cure rate for each subject, which is given by

$$S(\infty|\boldsymbol{X}) = G_\gamma(\theta(\boldsymbol{X})). \tag{4.4}$$

When $\gamma = 0$, the cure rate becomes $S(\infty|\boldsymbol{X}) = \exp(-\theta(\boldsymbol{X}))$ in the transformation family (2.7). For example, the cure rate for a 60 years old male with more than one positive node

in the IFN arm is

$$\begin{aligned} & S(\infty|(1,60,1,0)) \\ = \ & \exp(-\exp(-0.8-0.2198+60\times 0.0115+0.5519)) \\ = \ & 28.69\%. \end{aligned}$$

We also fitted a transformation cure model with transformation family (2.8), where $\gamma = 2$ is selected when γ varies from 0 to 2 based on the log-likelihood function. The selected model has log-likelihood equal to -1579.29, which is slightly improved compared with the model selected from transformation family (2.7). The estimates of regression coefficients are summarized in Table 4.4.

Table 4.3: Estimates of regression coefficients in Zeng et al.'s model based on transformation class (2.7) with $\gamma = 0$ for the E1690 study.

Variable	Coef	Std. Err.	Prob> $\|z\|$
Intercept	-0.8000	0.3270	0.0144
Treatment	-0.2198	0.1299	0.0906
Age	0.0115	0.0050	0.0223
Nodal category	0.5519	0.1601	0.0006
Gender	-0.2209	0.1373	0.1076
Log likelihood	-1579.36		

Table 4.4: Estimates of regression coefficients in Zeng et al.'s model based on transformation class (2.8) with $\gamma = 2$ for the E1690 study.

Variable	Coef	Std. Err.	Prob> $\|z\|$
Intercept	-0.9826	0.2069	0.0000
Treatment	-0.1660	0.1047	0.1129
Age	0.0098	0.0040	0.0153
Nodal category	0.4416	0.1318	0.0008
Gender	-0.1933	0.1123	0.0851
Log likelihood	-1579.29		

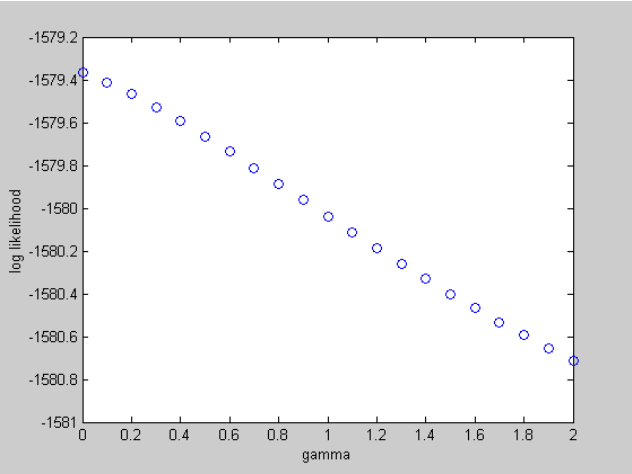

Figure 4.2: Log-likelihood in Zeng et al.'s model from transformation (2.7) with different γ for the E1690 study.

4.1.3 Generalized Transformation Model

In general, younger patients will have a longer relapse-free survival time and a higher cure rate than older patients. But sometimes age and the response variable may not be linear related. We fit a proposed generalized transformation model and try to apply fractional polynomial on covariate age in the E1690 data. Instead of the linear term, we will use

$$\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \text{Treatment} + \boldsymbol{\beta}_2 \text{Age}^{p_0} + \boldsymbol{\beta}_3 \text{Nodal} + \boldsymbol{\beta}_4 \text{Gender} \tag{4.5}$$

in the link function, where p_0 is a nonzero power varying from -2 to 2. We $p_0 = 0$, we will use

$$\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \text{Treatment} + \boldsymbol{\beta}_2 \log(\text{Age}) + \boldsymbol{\beta}_3 \text{Nodal} + \boldsymbol{\beta}_4 \text{Gender}. \tag{4.6}$$

As Royston and Altman [16] suggested, we select p_0 from a set of numbers $A = (-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2)$ based on the likelihood function.

In the transformation family (2.7) with $\gamma = 0$, power $p_0 = -0.5$ is selected, where the log-likelihood reaches its maximum at -1579.26. The log-likelihood with different powers is plotted in Figure 4.3(a). The estimates of regression coefficients of the selected model are summarized in Table 4.5. After transformation, the covariate age is still significant with a P-value equal to 0.0288. Based on the selected model, the cure rate for a 60 years old male with more than one positive node in the IFN arm is

$$\begin{aligned} & S(\infty|(1,60,1,0)) \\ = \; & \exp(-\exp(0.8074 - 0.2177 - 7.0960 \times \frac{1}{\sqrt{60}} + 0.5507)) \\ = \; & 28.61\%, \end{aligned}$$

which is comparable with that obtained by using Zeng et al.'s model in section 4.1.2.

One interesting finding is how the cure rate changes when age varies from 20 years old to 80 years old. We plotted the cure rate for males with more than one positive node in the IFN arm in Figure 4.4 to compare the results obtained from Zeng et al.'s model and our proposed model. In both models, the cure rate decreases when age increases, which means given other factors fixed, younger patients will have a higher chance to survive and cure compared to older patients. The difference is that in Zeng et al.'s model the cure rate changes almost linearly with age; while in the proposed model we see a nonlinear curve with decreasing slope, which means for younger patients the effect of change in age on the cure rate is stronger than that for older patients. For example, for males with more than one positive nodes in the IFN arm, when age changes from 20 years old to 30 years old, the cure rate drops rapidly from 52.73% to 42.47%; while when age increases from 70 years old to 80 years old, the cure rate only changes from 26.20% to 24.30%.

We also fitted proposed models in the transformation family (2.7) when γ takes other values. The results are plotted in Figure 4.3 (b)-(d). For example, when $\gamma = 0.5$, a logarithm transformation is selected for covariate age; when $\gamma = 1$, $p_0 = 0.5$ is selected.

Table 4.5: Estimates of regression coefficients in the proposed model based on transformation class (2.7) with $\gamma = 0$ and selected power=-0.5 for the E1690 study.

Variable	Coef	Std. Err.	Prob> $\|z\|$
Intercept	0.8074	0.4754	0.0895
Treatment	-0.2177	0.1300	0.0939
Age	-7.0960	3.1156	0.0228
Nodal category	0.5507	0.1599	0.0006
Gender	-0.2224	0.1372	0.1050
Log likelihood	-1579.26		

4.1.4 Model Comparison

In this section, we will compare the Cox model, Zeng et al's models, and the proposed models by using the Brier score. The Brier score was originally proposed by Brier [19] to verify the accuracy of weather forecasts. May et al. [20] developed the method to survival models in 2004.

For a sample of size n, the Brier score (BS) at time t^* is given by

$$BS(t^*) = \frac{\sum_{i=1}^{n} \left(I(Y_i > t^*) - \hat{S}(t^*|\boldsymbol{X}_i) \right)^2}{n}, \tag{4.7}$$

where Y_i is the observed survival time of the ith patient, $I(Y_i > t^*)$ is an indicator function representing the occurrence of the event, and $\hat{S}(t^*|\boldsymbol{X}_i)$ is the predicted probability of the ith patient surviving beyond time t^*. Under the Cox model, the surviving probability $\hat{S}(t^*|\boldsymbol{X}_i)$ is obtained by $\hat{S}(t^*|\boldsymbol{X}_i) = \hat{S}_0(t^*)^{\exp(\hat{\beta}X_i)}$, where $\hat{S}_0(t^*)$ is the estimated baseline probability of survival at time t^*. Under Zeng et al's models and the proposed models, $\hat{S}(t^*|\boldsymbol{X}_i)$ is given by $\hat{S}(t^*|\boldsymbol{X_i}) = G_\gamma(\theta(\boldsymbol{X_i})\hat{F}(t^*))$, where $\hat{F}(t^*)$ is the estimated value of promotion time cumulative distribution function at time t^*.

The choices of the time t^* can be arbitrary as long as the survival status for each patient can be ascertained at that time point. Some researchers use a fixed number of years. For example, May et al. [20] evaluated the Brier score at six monthly interval up to 2.5 years. Some other common choices for t^* include the quartiles of follow up time and the quartiles of the survival time.

From the definition in (4.7), the Brier score will have a minimum value of 0 for perfect

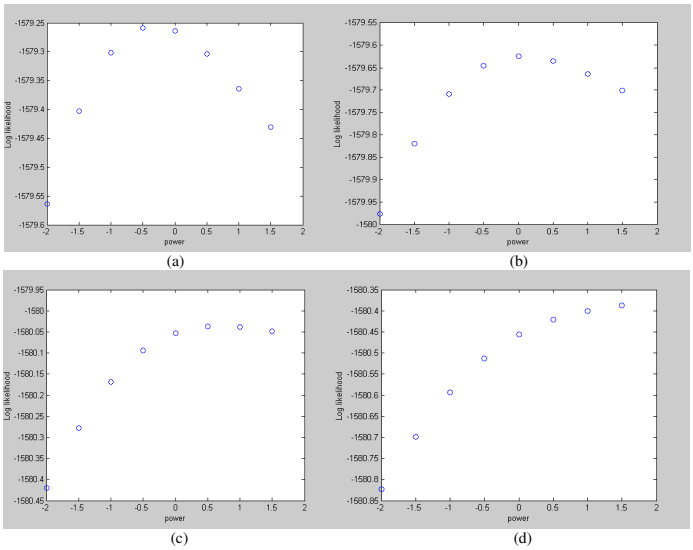

Figure 4.3: Log-likelihood and selected power in proposed model from transformation (2.7) with different γ for the E1690 study. (a) $\gamma = 0$, (b)$\gamma = 0.5$, (c)$\gamma = 1$, (d)$\gamma = 1.5$.

prediction of survival status and will range from 0 to 1. The lower the value of the Brier score, the better the prediction. In the E1690 study, we calculated the Brier scores for the Cox model, Zeng et al's models, and proposed models at the first quartile Q1, median Q2, and last quartile Q3 of 241 uncensored survival times. The results are summarized in Table 4.6. We can see that the Brier scores for different models are very close. They differ only in the third decimal point. We can also see that the Brier scores show an increasing trend in inaccuracy with time. For example, under the proposed models, when t^* equals 0.334 years the Brier score is 0.1451. It increases to 0.2127 at the median uncensored survival time $t^* = 0.723$. When $t^* = 1.460$ the Brier score is 0.2448.

We also want to assess the performance of different survival models on data other than those from which they are derived, i.e. training-validation methods. To do so, we divide the sample into four quarters, use three quarters of the data to fit the models, and use one

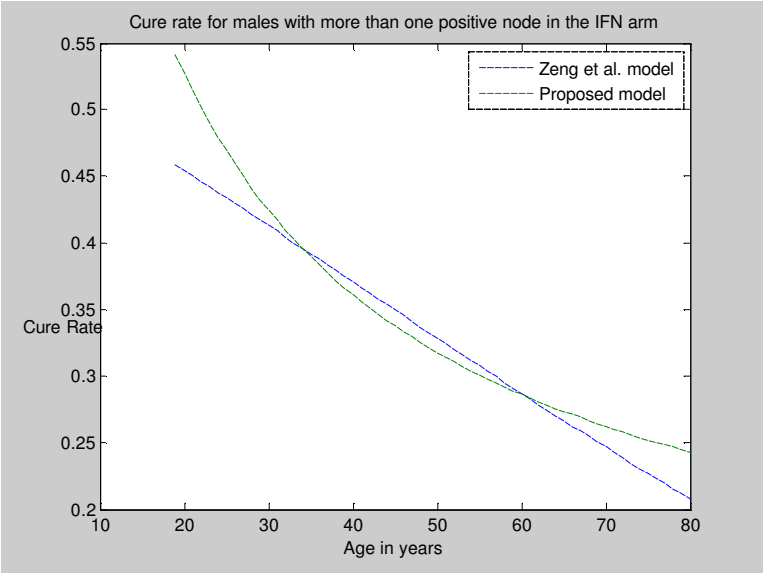

Figure 4.4: Comparison of cure rates in proposed model and in Zeng et al. model for males in the IFN arm with more than one positive node.

quarter to predict the outcome event. We evaluated the Brier scores at the first quartile, median, and last quartile of uncensored survival times of the three quarters of the sample. The results are listed in Table 4.7 to Table 4.10. We can see that sometimes Zeng et al.'s models and the proposed models can predict the outcome event more accurate than the Cox model. All three models predict well at the first quartile and median of survival times. At the third quartile of uncensored survival times, the value of the Brier scores are close to, sometimes greater than, 0.25.

The Brier score is a commonly used method, but it may not be the best way to compare survival models. This approach has some disadvantages. The Brier score can be measured at any arbitrary time point, but it can not measure the discrimination of the models over all the relevant time period. If there are patients who are censored before time t^*, the value of Brier scores tend to be greater than expected.

Table 4.6: Brier scores for different survival models in the E1690 study.

t^* in years	Cox Model	Zeng et al.'s Model	Proposed Model
Q1=0.334	0.1452	0.1450	0.1451
Q2=0.723	0.2124	0.2122	0.2127
Q3=1.460	0.2450	0.2446	0.2448

Table 4.7: Brier scores for different survival models in E1690 study when the first quarter of the data set is used to predict the event.

t^* in years	Cox Model	Zeng et al.'s Model	Proposed Model
Q1=0.333	0.1385	0.1363	0.1366
Q2=0.665	0.1878	0.1851	0.1851
Q3=1.425	0.2299	0.2263	0.2260

Table 4.8: Brier scores for different survival models in E1690 study when the second quarter of the data set is used to predict the event.

t^* in years	Cox Model	Zeng et al.'s Model	Proposed Model
Q1=0.331	0.1158	0.1077	0.1073
Q2=0.750	0.2226	0.2289	0.2289
Q3=1.536	0.2637	0.2595	0.2594

4.2 More Examples

Although the cure rate models are motivated from clinical trials where the endpoint is not death, it can be used in a variety of survival data. In this section, we will apply the proposed generalized transformation models to several data sets from the Diverse Populations Collaboration [21], where the endpoint is overall survival. The Diverse Populations Collaboration is a pooled database contributed by a group is of investigators to to examine issues of heterogeneity of results in epidemiological studies. There are 27 studies in the database, including 21 observational cohorts studies, 3 clinical trials, and 3 national samples.

One of the examples we will show from this database is the First National Health

Table 4.9: Brier scores for different survival models in E1690 study when the third quarter of the data set is used to predict the event.

t^* in years	Cox Model	Zeng et al.'s Model	Proposed Model
Q1=0.346	0.1741	0.1820	0.1823
Q2=0.735	0.2223	0.2288	0.2301
Q3=1.526	0.2461	0.2571	0.2579

Table 4.10: Brier scores for different survival models in E1690 study when the fourth quarter of the data set is used to predict the event.

t^* in years	Cox Model	Zeng et al.'s Model	Proposed Model
Q1=0.322	0.1516	0.1563	0.1560
Q2=0.668	0.2197	0.2241	0.2243
Q3=1.426	0.2445	0.2562	0.2570

and Nutrition Examination Survey Epidemiologic Follow-up Study (NHANES1) [22]. This study collected information for 14,407 individuals from 1971 to 1992. The first medical examinations were conducted from 1971-1975, and the next from 1982-1984, and then in 1986, 1987, and 1992. There were 4 cohorts in the NHANES1 study. In our analysis, we will use two of them, the black female cohort and the black male cohort. After all missing observations dropped, there are 2027 patients in this two cohorts, including 1265 black females and 762 black males. Some covariates we will use in this analysis are Age, Systolic blood pressure (Sbp), Sex, Body Mass Index (BMI), Diabetes (Diab), and Coronary heart disease (Chd), which are selected by fitting the Cox model and using the stepwise backward elimination algorithm. Some summary statistics of Age, Sbp, and BMI, the continuous covariates in the analysis, are list in Table 4.11. Diab and Chd are categorical and only take the value of 0 and 1 for absence and presence of the corresponding disease. Among the 2027 patients, there were 121 of them having diabetes and 82 of them having coronary heart disease. The response variable we will use is the overall survival time collected in 1992. There were 848 deaths at the end of followup, which was about 40% of the total number of patients.

We first fit the Cox proportional hazard model for the NHANES1 study. The results are

Table 4.11: 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

summarized in Table 4.12. The covariates are all highly significant at $\alpha = 0.05$. 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. This result is not quite reasonable. The vaules 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 and we all know that it will increase the hazard to develop many coronary heart diseases or even death. Therefore, a transformation on the covariate BMI may needed for the NHANES1 study.

Table 4.12: Fitted Cox proportional hazards model for the NHANES1 study.

Variable	Coef.	Std. Err.	z	Prob$>|z|$
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

Secondly, we fit Zeng et al's model with transformation family (2.7) for the NHANES1 study. In the Black male and female cohorts of the NHANES1 study, there were 848 patients dead at the end of the followup with a maximum survival life time of 7691 days. 1179 patients

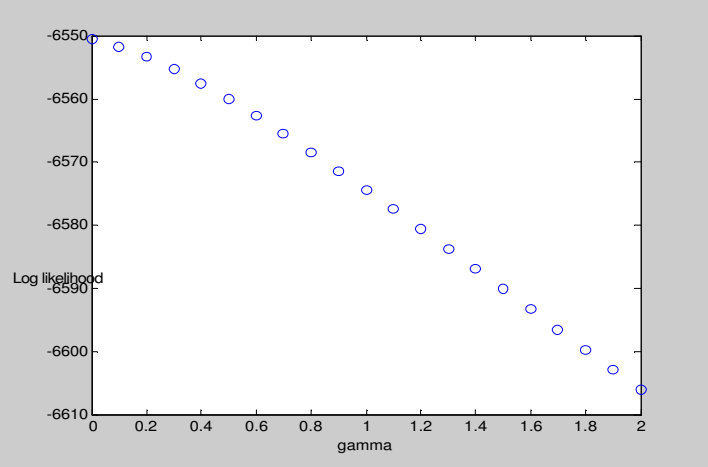

Figure 4.5: Log-likelihood in Zeng et al.'s model from transformation (2.7) with different γ for the NHANES1 study.

were right censored, among which 115 patients had survival time longer than 7691 days. We consider these 115 patients are cured subjects in Zeng et al's model. The observed log-likelihood is plotted in Figure 4.5 with different value of γ. A transformation of $\gamma = 0$ is chosen with maximum likelihood. The corresponding estimates of regression coefficients are summarized in Table 4.13. 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 proposed models. 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. To do so, we fit models (2.13)

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

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's

Table 4.13: Estimates of regression coefficients in Zeng et al.'s model based on transformation class (2.7) with $\gamma = 0$ for the NHANES1 study.

Variable	Coef	Std. Err.	Prob> $\|z\|$
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

models, we use the following four expressions in the function $\theta(\boldsymbol{\beta}, \boldsymbol{X})$, respectively,

$$\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \text{Age} + \boldsymbol{\beta}_2 \text{Sbp} + \boldsymbol{\beta}_3 \text{Sex} + \boldsymbol{\beta}_4 \text{BMI}^{p_{01}} + \boldsymbol{\beta}_5 \text{Diab} + \boldsymbol{\beta}_6 \text{Chd} \tag{4.8}$$

$$\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \text{Age}^{p_{02}} + \boldsymbol{\beta}_2 \text{Sbp} + \boldsymbol{\beta}_3 \text{Sex} + \boldsymbol{\beta}_4 \text{BMI} + \boldsymbol{\beta}_5 \text{Diab} + \boldsymbol{\beta}_6 \text{Chd} \tag{4.9}$$

$$\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \text{Age}^{p_{03}} + \boldsymbol{\beta}_2 \text{Sbp} + \boldsymbol{\beta}_3 \text{Sex} + \boldsymbol{\beta}_4 \text{BMI}^{-1} + \boldsymbol{\beta}_5 \text{Diab} + \boldsymbol{\beta}_6 \text{Chd} \tag{4.10}$$

$$\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \text{Age}^{p_{04}} + \boldsymbol{\beta}_2 \text{Sbp} + \boldsymbol{\beta}_3 \text{Sex} + \boldsymbol{\beta}_4 \text{BMI}^{-2} + \boldsymbol{\beta}_5 \text{Diab} + \boldsymbol{\beta}_6 \text{Chd} \tag{4.11}$$

In model (4.8), when we fix Age and Sbp, power $p_{01} = -2$ is selected for BMI. The observed log-likelihood is plotted in Figure 4.6 (a). In the next model (4.9), 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 4.6 (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 (4.10) where BMI^{-1} and Sbp are fixed. We also tried model (4.11) with BMI^{-2} and Sbp fixed. Both model (4.10) and model (4.11) select power=1 for Age. The results are plotted in Figure 4.6 (c) and (d). As a summary, the best transformation based on log-likelihood from model (4.8)-(4.11) is

$$\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \text{Age} + \boldsymbol{\beta}_2 \text{Sbp} + \boldsymbol{\beta}_3 \text{Sex} + \boldsymbol{\beta}_4 \text{BMI}^{-2} + \boldsymbol{\beta}_5 \text{Diab} + \boldsymbol{\beta}_6 \text{Chd}. \tag{4.12}$$

The corresponding estimates of regression coefficients are listed in Table 4.14.

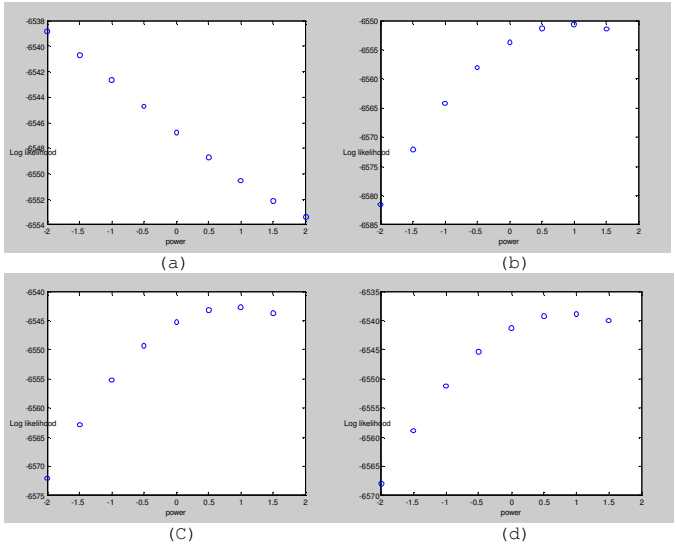

Figure 4.6: Log-likelihood and selected power in proposed models from transformation (2.7) for the NHANES1 study. (a)Model (4.8), (b)Model (4.9), (c)Model (4.10), (d)Model (4.11).

To compare the Cox model, Zeng et al's model, and the proposed model (4.12), we calculate the Brier scores at the first quartile Q1, median Q2, and the last quartile Q3 of 848 uncensored survival times. The results are summarized in Table 4.15. We can see that the proposed model 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's model. The value of Brier score drops to 0.1297 for the proposed model, which indicates the proposed model can predict the survival outcome better than the other two models.

Table 4.14: Estimates of regression coefficients in the proposed model (4.8) based on transformation class (2.7) with $\gamma = 0$ and transformation on BMI ($p_0 = -2$) for the NHANES1 study.

Variable	Coef	Std. Err.	Prob> $\|z\|$
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

Table 4.15: 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

CHAPTER 5

FUTURE WORK

In this dissertation, we proposed a class of generalized transformation models, which is motivated by the work of Zeng et al. [5]. They introduced semiparametric transformation models that can be used for survival data with a cure fraction. Two transformation families $G_\gamma(.)$ were discussed. The commonly used proportional hazards cure rate models and proportional odds models can be included as special cases of the transformation models. Covariates are modeled through a link function $\theta(\boldsymbol{X}) = \eta(\boldsymbol{\beta'X})$, where $\eta(.)$ is a known and strictly positive increasing function, such as exponential functions. In our proposed model, we want to use generalized additive models instead of $\boldsymbol{\beta'X}$ in the link function $\eta(.)$. Specifically, we consider fractional polynomials proposed by Royston and Altman [16]. We proved that the proposed model is identifiable as long as the transformation families $G_\gamma(.)$ satisfy some very general conditions.

How to select transformation powers when fractional polynomials are used, such that the model fits the data best? We proposed a power selection procedure, choosing a power from set A=(-2, -1.5, -1, 0.5, 0, 0.5, 1, 1.5, 2), for different models by comparing residual sum of squares or likelihood functions. We conducted some simulation studies for linear regression models, logistic regression models, Cox models, and the proposed models. The results show the power selection procedure works well in different models when the response variables are continuous. When the response variables are categorical relatively large sample size is needed. Instead of selecting the power from set A, an improvement in this direction could be considering the power as a parameter and estimate it by using maximum likelihood methods.

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 also can be used to overall survival time.

We showed several examples where the data are collect from both observational cohort studies and clinical trials to illustrate idea. We used Brier scores to compare the Cox models, Zeng et al's models, and the proposed models, but it may not be the best way to discriminate between models. We may consider other model comparison methodologies. Receiver operating characteristic (ROC) curves can be used to measure the differences of the models over all the relevant time period. We can also consider using deviance differences [20], which are more sensitive in measuring the predictive performance between models.

APPENDIX A

ASYMPTOTIC PROPERTIES OF THE SEMI-PARAMETRIC ESTIMATES

Zeng et al. [5] 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 this appendix, we will show the asymptotic properties of the semi-parametric estimates in our proposed model. Proofs of the theorems are similar to those of theorems 1 and 2 in Zeng et al. [5] with some modifications.

Before we prove the theorems, we will give a brief introduction of Glivenko-Cantelli theorem and Donsker theorem. These two theorems are main results describing the empirical process of a sample of observations by using empirical measure. Glivenko-Cantelli theorem is uniform law of large number and Donsker theorem is uniform central limit theorem. More discussion of the theorems and their applications will be shown later in this section. Classes that satisfy the conditions of the theorems are called Glivenko-Cantelli class and Donsker class, respectively.

Suppose $\mathbf{E}_n$ is the empirical measure of a sample of random elements $X_1, ..., X_n$ in a measurable space $\mathcal{X}$ and $\mathbf{E}$ is the common distribution of the X_i for $i = 1, ..., n$.

Definition 1. Glivenko-Cantelli Class [23]: A class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is a measurable space, is called a Glivenko-Cantelli class if

$$\sup_{f\in\mathcal{F}} \{|\mathbf{E}_n f - \mathbf{E}f|\} = \sup_{f\in\mathcal{F}} \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - \mathbf{E}f) \right| \right\} \to^{a.s.} 0.$$

Definition 2. Donsker Class [23]: Assuming that $\sup_{f\in\mathcal{F}} \{|f(x) - \mathbf{E}f|\} < \infty$ for every $x \in \mathcal{X}$, a class $\mathcal{F}$ is called Donsker class if $\sqrt{n}(\mathbf{E}_n - \mathbf{E})$ converges to a tight Borel

measurable element $\mathbf{G}$ in $l^\infty(\mathcal{F})$ in distribution, i.e.,

$$\sqrt{n}(\mathbf{E}_n - \mathbf{E}) \to \mathbf{G}, \text{ in } l^\infty(\mathcal{F}).$$

A.1 Strong Consistency

We use the notations introduced in Zeng et al. [5] and van der Vaart and Wellner [23]. Let $\mathbf{E}_n$ denote the empirical measure of n iid observations and let $\mathbf{E}$ be the expectation. For any measurable function $g(\Delta, Y, X)$, define

$$\mathbf{E}_n(g(\Delta, Y, X)) = \frac{1}{n}\sum_{i=1}^{n} g(\Delta_i, Y_i, X_i). \tag{A.1}$$

Recall the definition of (Δ, Y, X) in (3.13) and (3.16). Δ is an indicator function; Y is the minimum of survival life time and right censoring time; X is a vector of covariates. $\{(\Delta_i, Y_i, X_i), i = 1, ..., n\}$ are iid observations. More detailed description of this data structure will be shown in proofs of the theorems.

Following from the law of large number, for a given function g, $\mathbf{E}_n g$ converges to $\mathbf{E}g$ almost surely [24]. Given a collection $\mathcal{F}$ of measurable functions $g : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is a measurable space, the uniform version of the law of large number becomes

$$\sup_{g\in\mathcal{F}}\{|\mathbf{E}_n g - \mathbf{E}g|\} \to 0, \text{ almost surely.} \tag{A.2}$$

A class $\mathcal{F}$ for which (A.2) is true is called a Glivenko-Cantelli class.

For a given function g, it follows from the central limit theorem that,

$$\sqrt{n}(\mathbf{E}_n - \mathbf{E})g \to N(0, \mathbf{E}(g - \mathbf{E}g)^2), \tag{A.3}$$

provided $\mathbf{E}g$ and $\mathbf{E}g^2$ exist. We assume

$$\sup_{g\in\mathcal{F}} |g(x) - \mathbf{E}g| < \infty, \quad \text{for every } x, \tag{A.4}$$

and view the empirical process $\{\sqrt{n}(\mathbf{E}_n - \mathbf{E})g : g \in \mathcal{F}\}$ as a map into $l^\infty(\mathcal{F})$, which is a set of all uniformly bounded real functions on $\mathcal{F}$. Then we can consider a uniform version of the central limit theorem. If the empirical process $\sqrt{n}(\mathbf{E}_n - \mathbf{E})$ converges to an element in $l^\infty(\mathcal{F})$, i.e,

$$\sqrt{n}(\mathbf{E}_n - \mathbf{E}) \to \mathbf{G}, \quad \text{in } \ l^\infty(\mathcal{F}), \tag{A.5}$$

where $\mathbf{G}$ is a tight Borel measurable element in $l^\infty(\mathcal{F})$, then the class $\mathcal{F}$ is called a Donsker class. It can be shown that the limit process $\{\mathbf{G}g : g \in \mathcal{F}\}$ must be a zero-mean Gaussian process. If a class $\mathcal{F}$ is a Donsker class, then it is also a Glivenko-Cantelli class.

Some examples of Donsker class and Glivenko-Cantelli class are provided in van der Vaart and Wellner [23]. For example, the class $\mathcal{F}$ of all nondecreasing functions $f : \mathbb{R} \to \mathbb{R}$, such that $0 \leq f \leq F$, for a given nondecreasing function F, is a Donsker class provided the $\|F\|_{L^2} < \infty$. The class of indicator functions of sets $(-\infty, c]$ is also Donsker, where $c \in \mathbb{R}$. These examples are shown as Example 2.10.27 and Example 2.4.2 in van der Vaart and Wellner [23], respectively. The Donsker property is preserved under the summation, product, and quotient given some very general conditions, which is shown as Example 2.10.7-2.10.9 in van der Vaart and Wellner [23].

We will study the asymptotic properties of the maximum likelihood estimates of $(\boldsymbol{\beta}, F)$ based on the proposed model. Suppose there are n independent right censored observations, we use the following notations to describe the data structure. For the ith observation, we have $\{Y_i, \boldsymbol{X}_i, \Delta_i\}$, $i = 1, ..., n$, where

$$\begin{aligned} Y_i &= \min(T_i, C_i), \\ \Delta_i &= \begin{cases} 0, & T_i > C_i, \\ 1, & T_i \leq C_i. \end{cases} \end{aligned} \tag{A.6}$$

Something worth mentioning is that in application we may use

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

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

We assume $P(Y = \infty|\boldsymbol{X}) > 0$, which means a proportion of subjects never experience failure or right-censoring. The modified semi-parametric version observed-data likelihood function of parameters $(\boldsymbol{\beta}, F)$, denoted by $L(\boldsymbol{\beta}, F)$, is given by

$$\begin{aligned} L(\boldsymbol{\beta}, F) = \prod_{i=1}^{n} \{ & [\{-G'(\eta(\boldsymbol{\beta}, \boldsymbol{X_i})F(Y_i))\eta(\boldsymbol{\beta}, \boldsymbol{X_i})F\{Y_i\}\}^{\Delta_i} \\ & \times \{G(\eta(\boldsymbol{\beta}, \boldsymbol{X_i})F(Y_i))\}^{(1-\Delta_i)}]^{I(Y_i<\infty)} \times [G(\eta(\boldsymbol{\beta}, \boldsymbol{X_i}))]^{I(Y_i=\infty)}\}, \end{aligned} \tag{A.7}$$

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

To obtain consistency and asymptotic normality, we make the following assumptions through out this section:

(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 $\in \mathbb{R}^+$. The density function of F_0, denoted by f_0, is bounded from above and below in any compact sets.

(C4). Conditional on $\boldsymbol{X}$, the right censoring time C is independent of T, and $S_C(\infty|\boldsymbol{X}) > 0$. The density functions of T and C are bounded from below and above in any compact sets, respectively.

(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.

Suppose $\hat{\boldsymbol{\beta}}_n$, $(\hat{F}_n\{Y_i\},\ i = 1, 2, ..., n)$, are the estimates of $\boldsymbol{\beta}$ and F such that $L(\boldsymbol{\beta}, F)$ reaches its maximum. The log likelihood function, denoted by $l(\boldsymbol{\beta}, F)$, is given by

$$\begin{aligned} l(\boldsymbol{\beta}, F) \;=\; & \sum_{j=1}^{n} \Delta_j \{\log(-G'(\eta(\boldsymbol{\beta}, X_j)F(Y_j))) \\ & + \log \eta(\boldsymbol{\beta}, X_j) + \log F\{Y_j\}\} I(Y_j < \infty) \\ & + \sum_{j=1}^{n} (1 - \Delta_j \{\log G(\eta(\theta(\boldsymbol{\beta}, X_j))F(Y_j))\} I(Y_j < \infty) \\ & + \sum_{j=1}^{n} \{\log G(\eta(\boldsymbol{\beta}, X_j))\} I(Y_j = \infty), \end{aligned} \tag{A.8}$$

where $(F\{Y_i\},\ i = 1, ..., n)$ 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 $\Delta_i = 0$. Introducing the Lagrange multiplier λ and maximizing the unconstrained function $l(\boldsymbol{\beta}, F) - \lambda(G(F) - 1)$ will give us the maximum likelihood estimates based on the constrained likelihood function $l(\boldsymbol{\beta}, F)$.

For any i such that $\Delta_i = 1$, the maximum likelihood estimates are the set of solutions for equation $\frac{\partial l(\boldsymbol{\beta},F)}{\partial F\{Y_i\}} = \lambda \frac{\partial G(F)}{\partial F\{Y_i\}} = \lambda$. Therefore we obtain a sequence of equations about $\hat{\boldsymbol{\beta}}_n$ and $\hat{F}_n\{Y_i\}$. When $\Delta_i = 1$ for $i = 1, 2, ..., n$, we have

$$\frac{\partial l(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)}{\partial F_n\{Y_i\}} = n\hat{\lambda}_n \frac{\partial G(\hat{F}_n)}{\partial F_n\{Y_i\}} = n\hat{\lambda}_n, \tag{A.9}$$

where $\hat{\lambda}_n$ denotes the estimation of the Lagrange multiplier λ based on the sample size n.

The left hand side of (A.9) is

$$\begin{aligned} \frac{\partial l(\boldsymbol{\beta}, F)}{\partial F\{Y_i\}} &= \sum_{j=1}^{n} \Delta_j \frac{G''(\eta(\boldsymbol{\beta}, X_j))F(Y_j)\eta(\boldsymbol{\beta}, X_j)I(Y_j \geq Y_i)}{G'(\eta(\boldsymbol{\beta}, X_j)F(Y_j))} I(Y_j < \infty) \\ &\quad + \frac{1}{F\{Y_i\}} + \sum_{j=1}^{n} (1 - \Delta_j) \frac{G'(\eta(\boldsymbol{\beta}, X_j)F(Y_j))\eta(\boldsymbol{\beta}, X_j)I(Y_j \geq Y_i)}{G(\eta(\boldsymbol{\beta}, X_j)F(Y_j))} I(Y_j < \infty) \end{aligned} \tag{A.10}$$

evaluated at $(\hat{\beta}_n, \hat{F}_n)$. Combining (A.9) and (A.10), we obtain

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

where

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

Function $H_n(.)$ is bounded since both $G'(\eta(\hat{\boldsymbol{\beta}}_n, X)\hat{F}_n(Y))$ and $G(\eta(\hat{\boldsymbol{\beta}}_n, X)\hat{F}_n(Y))$ are bounded away from zero under conditions (C1),(C2) and (C6).

Equation (A.11) can be written as

$$\hat{F}_n\{Y_i\} = \frac{1}{n(\hat{\lambda}_n - H_n(Y_i, \hat{\boldsymbol{\beta}}_n, \hat{F}_n))} \tag{A.13}$$

for $i = 1, 2, ..., n$ when $\Delta_i = 1$. Considering $\hat{F}_n\{Y_i\} = 0$ when $\Delta_i = 0$, we also have $\hat{F}_n\{Y_i\} = \frac{\Delta_i}{n(\hat{\lambda}_n - H_n(Y_i, \hat{\boldsymbol{\beta}}_n, \hat{F}_n))}$, $i = 1, ..., n$. From (A.9) and (A.10), $\hat{\lambda}_n$ can be written as

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

If we can prove the sequence $\{\hat{\lambda}_n, n = 1, 2, ...\}$ is bounded from above and below, then we can choose a subsequence from $\hat{\lambda}_n$ such that $\hat{\lambda}_n \to \lambda^*$ almost surely. In fact, $\hat{F}_n\{Y_i\} > 0$ when $\Delta_i = 1$, therefore from (A.13)$\hat{\lambda}_n - H_n(Y_i, \hat{\boldsymbol{\beta}}_n, \hat{F}_n) > 0$. $\hat{\lambda}_n$ is then bounded from below because $H_n(.)$ is bounded. On the other hand, there at least exists one i such that $\Delta_i = 1$ and $\hat{\lambda}_n - H_n(Y_i, \hat{\boldsymbol{\beta}}_n, \hat{F}_n) \leq 1$ that implies that $\{\hat{\lambda}_n, n = 1, 2, ...\}$ is bounded from above. Otherwise, if $\hat{\lambda}_n - H_n(Y_i, \hat{\boldsymbol{\beta}}_n, \hat{F}_n) > 1$ for all the i's with $\Delta_i = 1$, then

$$1 = \sum_{i=1}^{n} \Delta_i \hat{F}_n\{Y_i\} < \sum_{i=1}^{n} \frac{\Delta_i}{n} \leq 1,$$

which leads to a contradiction. We can then choose a further subsequence of $\hat{\boldsymbol{\beta}}_n$ such that $\hat{\boldsymbol{\beta}}_n \to \boldsymbol{\beta}^*$ almost surely since $\hat{\boldsymbol{\beta}}_n$ belong to a compact set $\mathcal{B}_0$. Subsequence of $\hat{F}_n$ is also chosen such that $\hat{F}_n \to F^*$ pointwise. Here we used the same notations for subsequences and the original sequences.

Following the proof in Zeng et al. [5], we consider the class

$$\begin{aligned} \mathcal{A} \;=\; & \{\Delta \frac{G''(\eta(\boldsymbol{\beta}, \boldsymbol{X})F(Y))\eta(\boldsymbol{\beta}, \boldsymbol{X})I(\infty > Y \geq y)}{G'(\eta(\boldsymbol{\beta}, \boldsymbol{X})F(Y))} \\ & + (1 - \Delta) \frac{G'(\eta(\boldsymbol{\beta}, \boldsymbol{X})F(Y))\eta(\boldsymbol{\beta}, \boldsymbol{X})I(\infty > Y \geq y)}{G(\eta(\boldsymbol{\beta}, \boldsymbol{X})F(Y))} : \\ & y \in [0, \infty), \boldsymbol{\beta} \in \mathcal{B}_0, F \text{ is a distribution function}\}. \end{aligned} \tag{A.15}$$

The class $\mathcal{A}$ is generated by

$$\begin{aligned} \mathcal{F}_1 &= \{F(Y) : \text{F is a distribution function.,}\} \\ \mathcal{F}_2 &= \{I(\infty > Y \geq y) : y \in [0, \infty)\}, \\ \mathcal{F}_3 &= \{\eta(\boldsymbol{\beta}, \boldsymbol{X}) : \boldsymbol{\beta} \in \mathcal{B}_0\}. \end{aligned}$$

Based on the results from theorem 2.7.5, theorem 2.4.1, and theorem 2.7.11 of van der Vaart and Wellner [23], $\mathcal{F}_1$, $\mathcal{F}_2$, and $\mathcal{F}_3$ are Donsker classes. Moreover, by assumptions (C5) and (C6), function G, G', G'' and η are continuous. Since $\boldsymbol{\beta}$ and $\boldsymbol{X}$ both belong to compact sets by conditions (C1) and (C2), the values of $\eta(\boldsymbol{\beta}, \boldsymbol{X})$ belong to a compact set and the denominators $G'(\eta(\boldsymbol{\beta}, \boldsymbol{X})F(Y))$ and $G(\eta(\boldsymbol{\beta}, \boldsymbol{X})F(Y))$ are both bounded away from zero. Therefore, the class $\mathcal{A}$ is also a Donsker class by theorem 2.10.6 of van der Vaart and Wellner [23].

Lemma 1. Under conditions (C1)-(C6), $H_n(y, \hat{\boldsymbol{\beta}}_n, \hat{F}_n) \rightarrow^{a.s.} H^*(Y_i)$ uniformly in y, where

$$\begin{aligned} H^*(y) &= E\{\Delta \frac{G''(\eta(\boldsymbol{\beta}^*, \boldsymbol{X})F^*(Y))\eta(\boldsymbol{\beta}^*, \boldsymbol{X})I(\infty > Y \geq y)}{G'(\eta(\boldsymbol{\beta}^*, \boldsymbol{X})F^*(Y))} \\ &\quad +(1-\Delta)\frac{G'(\eta(\boldsymbol{\beta}^*, \boldsymbol{X})F^*(Y))\eta(\boldsymbol{\beta}^*, \boldsymbol{X})I(\infty > Y \geq y)}{G(\eta(\boldsymbol{\beta}^*, \boldsymbol{X})F^*(Y))}\}. \end{aligned} \tag{A.16}$$

Proof: Simply denote the function in class $\mathcal{A}$ as $K(\Delta, Y, \boldsymbol{X}; F, \boldsymbol{\beta}, y)$. Class $\mathcal{A}$ is a Donsker class, we have

$$\sup_{F,\beta,y} |\mathbf{E}_n K(\Delta, Y, \boldsymbol{X}; F, \boldsymbol{\beta}, y) - \mathbf{E}K(\Delta, Y, \boldsymbol{X}; F, \boldsymbol{\beta}, y)| = o_n(1), \text{a.s.}. \tag{A.17}$$

Particularly, we have

$$\sup_{y} |\mathbf{E}_n K(\Delta, Y, \boldsymbol{X}; \hat{F}_n, \hat{\boldsymbol{\beta}}_n, y) - \mathbf{E}K(\Delta, Y, \boldsymbol{X}; \hat{F}_n, \hat{\boldsymbol{\beta}}_n, y)| = o_n(1), \text{a.s.}. \tag{A.18}$$

On the other hand by the bounded convergence theorem, since function K is bounded and $\hat{\boldsymbol{\beta}}_n \rightarrow \boldsymbol{\beta}^*$ almost surely, $\hat{F}_n \rightarrow F^*$ pointwise, we have

$$\begin{aligned} & \sup_y |\mathbf{E}K(\Delta, Y, \boldsymbol{X}; \hat{F}_n, \hat{\boldsymbol{\beta}}_n, y) - \mathbf{E}K(\Delta, Y, \boldsymbol{X}; F^*, \boldsymbol{\beta}^*, y)| \\ \leq & \sup_y \mathbf{E}|K(\Delta, Y, \boldsymbol{X}; \hat{F}_n, \hat{\boldsymbol{\beta}}_n, y) - K(\Delta, Y, \boldsymbol{X}; F^*, \boldsymbol{\beta}^*, y)| \\ \leq & \mathbf{E}|K(\Delta, Y, \boldsymbol{X}; \hat{F}_n, \hat{\boldsymbol{\beta}}_n, 0) - K(\Delta, Y, \boldsymbol{X}; F^*, \boldsymbol{\beta}^*, 0)| \\ = & o_n(1), \text{a.s.}. \end{aligned} \tag{A.19}$$

Combining (A.17) and (A.18), we proved $H_n(y, \hat{\boldsymbol{\beta}}_n, \hat{F}_n)$ converges to $H^*(y)$ uniformly in y. ■

Moreover, the right hand side of (A.14) converges to

$$\lambda^* = E(\Delta I(Y < \infty)) + E\left(I(Y < \infty)\int_0^Y H^*(y)dF^*(y)\right).$$

Lemma 2. Under conditions (C1)-(C6), $\lambda^* - H^*(y) > 0$ for $0 \leq y < \infty$, and $E(\frac{\Delta I(Y<\infty)}{\lambda^* - H^*(Y)}) \leq 1$.

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

$$\begin{aligned} 1 &= \frac{1}{n}\sum_{i=1}^n \frac{\Delta_i I(Y_i < \infty)}{\hat{\lambda}_n - H_n(Y_i, \hat{\boldsymbol{\beta}}_n, \hat{F}_n)} \\ &\geq \frac{1}{n}\sum_{i=1}^n \frac{\Delta_i I(Y_i < \infty)}{|\hat{\lambda}_n - H_n(Y_i, \hat{\boldsymbol{\beta}}_n, \hat{F}_n)| + \epsilon}. \end{aligned} \tag{A.20}$$

" $\geq$ " is true in (A.20) because $\hat{\lambda}_n - H_n(Y_i, \hat{\boldsymbol{\beta}}_n, \hat{F}_n) > 0$ when $\Delta_i = 1$. When n is large enough, by Lemma 1 we have $|\hat{\lambda}_n - \lambda^*| = o_n(1)$ and $|H_n(Y_i, \hat{\boldsymbol{\beta}}_n, \hat{F}_n) - H^*(Y_i)| = o_n(1)$ almost surely. Hence,

$$\begin{aligned}
& \left| \frac{\Delta_i I(Y_i < \infty)}{|\hat{\lambda}_n - H_n(Y_i, \hat{\beta}_n, \hat{F}_n)| + \epsilon} - \frac{\Delta_i I(Y_i < \infty)}{|\lambda^* - H^*(Y_i)| + \epsilon} \right| \\
\leq & \left| \frac{|\lambda^* - H^*(Y_i)| - |\hat{\lambda}_n - H_n(Y_i, \hat{\beta}_n, \hat{F}_n)|}{(|\hat{\lambda}_n - H_n(Y_i, \hat{\beta}_n, \hat{F}_n)| + \epsilon)(|\lambda^* - H^*(Y_i)| + \epsilon)} \right| \\
\leq & \frac{|\lambda^* - H^*(Y_i) - \hat{\lambda}_n + H_n(Y_i, \hat{\beta}_n, \hat{F}_n)|}{(|\hat{\lambda}_n - H_n(Y_i, \hat{\beta}_n, \hat{F}_n)| + \epsilon)(|\lambda^* - H^*(Y_i)| + \epsilon)} \\
\leq & \frac{|\hat{\lambda}_n - \lambda^*| + |H_n(Y_i, \hat{\beta}_n, \hat{F}_n) - H^*(Y_i)|}{(|\hat{\lambda}_n - H_n(Y_i, \hat{\beta}_n, \hat{F}_n)| + \epsilon)(|\lambda^* - H^*(Y_i)| + \epsilon)} \\
\leq & \frac{o_n(1)}{\epsilon^2}, a.s..
\end{aligned} \tag{A.21}$$

Therefore,

$$1 \geq \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta_i I(Y_i < \infty)}{|\lambda^* - H^*(Y_i)| + \epsilon} + \frac{o_n(1)}{\epsilon^2}. \tag{A.22}$$

When $n \to \infty$, we have $1 \geq E\left(\frac{\Delta I(Y<\infty)}{|\lambda^* - H^*(Y)| + \epsilon}\right)$. Letting $\epsilon \to 0$, we obtain

$$1 \geq E\left(\frac{\Delta I(Y < \infty)}{|\lambda^* - H^*(Y)|}\right) \tag{A.23}$$

by Fatou lemma.

We then calculate the right hand side of (A.23) by using conditional expectations.

$$\begin{aligned}
& E\left(\frac{\Delta I(Y < \infty)}{|\lambda^* - H^*(Y)|}\right) \\
= & E_T\left(E_X\left(E_C \frac{\Delta I(T < \infty)}{|\lambda^* - H^*(T)|} \middle| T, X\right) \middle| T\right) \\
= & E_T\left(E_X\left(\frac{S_c(T|X) I(T < \infty)}{|\lambda^* - H^*(T)|} \middle| T\right)\right) \\
= & \int_0^\infty \frac{E_X(-S_c(T|X) G'(\eta(\beta_0, X) F_0(T)) \eta(\beta_0, X))}{|\lambda^* - H^*(T)|} f_0(T) dT \\
= & \int_0^\infty \frac{k(T) f_0(T)}{|\lambda^* - H^*(T)|} dT
\end{aligned} \tag{A.24}$$

where

$$k(T) = E_X(-S_c(T|X) G'(\eta(\beta_0, X) F_0(T)) \eta(\beta_0, X)). \tag{A.25}$$

Function $t(T) = -S_c(T|X)G'(\eta(\beta_0, X)F_0(T))\eta(\beta_0, X)$ is positive and continuous on $[0, \infty)$. When $T \to \infty$, $t(\infty) = -S_c(\infty|X)G'(\eta(\beta_0, X))\eta(\beta_0, X)$ exists and is positive. Therefore there exists c_0, c_1 positive such that $c_0 \leq t(T) \leq c_1$, for $\forall T$. Hence, $c_0 \leq k(T) \leq c_1$, for $\forall T$. Combining (A.23) and (A.24), we then have $1 \geq c_0 \int_0^\infty \frac{f_0(T)}{|\lambda^* - H^*(T)|} dT$.

It can be shown that $H^*(y)$ is Lipshitz continuous. In fact, suppose $0 \leq y_1 < y_2 < \infty$ and notice $H^*(y)$ in (A.16) where y only appears in $I[\infty > Y \geq y]$, we have

$$\begin{aligned} & |H^*(y_1) - H^*(y_2)| \\ \leq \; & cE|I[\infty > Y \geq y_1] - I[\infty > Y \geq y_2]| \\ \leq \; & cE(I[y_1 \leq Y < y_2]) \\ \leq \; & cE(I[y_1 \leq T < y_2]) + cE(I[y_1 \leq C < y_2]). \end{aligned} \tag{A.26}$$

By condition (C4), the density functions of T and C are both bounded from below and above in any compact sets, then $E(I[y_1 \leq T < y_2]) \leq c_1|y_2 - y_1|$ and $E(I[y_1 \leq C < y_2]) \leq c_2|y_2 - y_1|$. Therefore, we proved $H^*(y)$ is Lipshitz continuous.

We claim $\lambda^* - H^*(T) \neq 0$ for $\forall\, T \in [0, \infty)$. Otherwise, suppose there exists T_0 such that $\lambda^* - H^*(T_0) = 0$, then we consider a small neighborhood of T_0, $[-\delta + T_0, \delta + T_0]$. Since $H^*(y)$ is Lipshitz continuous, there exists a constant c_1 such that $|H^*(T) - H^*(T_0)| \leq c_1|T - T_0|$. By assumptions, $f_0(.)$ is bounded from below in $[-\delta + T_0, \delta + T_0]$. Suppose we have $f_0(T) \geq c_2$, where c_2 is a constant.

Then,

$$1 \geq c_0 \int_{T_0 - \delta}^{T_0 + \delta} \frac{c_2 dT}{c_1|T - T_0|} = \infty, \tag{A.27}$$

which leads to a contradiction. Therefore, $\lambda^* - H^*(T) \neq 0$ for any $T \in [0, \infty)$. Because of the continuity of $H^*(y)$, $\lambda^* - H^*(T)$ is either positive or negative for any $T \in [0, \infty)$. When $\Delta_i = 1$, we have $\hat{F}_n\{Y_i\} = \frac{\Delta_i}{n(\hat{\lambda}_n - H_n(Y_i, \hat{\beta}_n, \hat{F}_n))} > 0$. Hence, $\hat{\lambda}_n - H_n(Y_i, \hat{\beta}_n, \hat{F}_n) > 0$. Therefore, for any i we have

$$\Delta_i(\hat{\lambda}_n - H_n(Y_i, \hat{\beta}_n, \hat{F}_n)) \geq 0 \tag{A.28}$$

and

$$\frac{1}{n} \sum_{i=1}^{n} \Delta_i(\hat{\lambda}_n - H_n(Y_i, \hat{\beta}_n, \hat{F}_n)) \geq 0. \tag{A.29}$$

The left hand side of (A.29) converges to $E(\Delta(\lambda^* - H^*(Y)))$. So far, we have proved $E(\Delta(\lambda^* - H^*(Y))) \geq 0$ and $\Delta(\lambda^* - H^*(Y))$ will not change sign for all Y. Therefore, $\Delta(\lambda^* - H^*(Y)) \geq 0$. Taking $T \leq C$, we have $\lambda^* - H^*(T) > 0$ for any $T \in [0, \infty)$.■

Based on the results in Lemma 2, fix M, there exists $\delta = \delta(M) > 0$ such that $\lambda^* - H^*(y) \geq \delta$ for any $y \in [0, M]$. Define class $\mathcal{B}_M$ as

$$\mathcal{B}_M = \left\{ \frac{\Delta I(Y \leq y)}{\lambda^* - H^*(Y)} : y \in [0, M]. \right\}.$$

Class $\mathcal{B}_M$ is a Donsker class because $\lambda^* - H^*(Y)$ is bounded away from zero. Thus, $\hat{F}_n(y)$ converges to $E\left(\frac{\Delta I(Y \leq y)}{\lambda^* - H^*(Y)}\right)$ uniformly in $y \in [0, M]$, i.e,

$$\begin{aligned} \hat{F}_n(y) &= \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta_i I(Y_i \leq y)}{|\hat{\lambda}_n - H_n(Y_i, \hat{\beta}_n, \hat{F}_n)|} \\ &\to F^*(y) = E\left(\frac{\Delta I(Y \leq y)}{\lambda^* - H^*(Y)}\right). \end{aligned} \tag{A.30}$$

Following the calculation in (A.24), we will have $F^*(y) = \int_0^y \frac{k(t)}{\lambda^* - H^*(t)} f_0(t)dt$, therefore function,

$$f^*(y) = \frac{k(y)}{\lambda^* - H^*(y)} f_0(y). \tag{A.31}$$

is the density function of $F^*(y)$. We will prove $F^*(\infty) = 1$ at the end of this section, which also concludes $F^*(y)$ is a proper distribution function.

We consider distribution functions $\tilde{F}_n$ with jumps $\tilde{F}_n\{Y_i\}$ when $\Delta_i = 1$ and $Y_i < \infty$,

$$\tilde{F}_n\{Y_i\} = \frac{1}{nC_n} \frac{\Delta_i I(Y_i < \infty)}{k(Y_i)}, \tag{A.32}$$

where C_n is a constant such that $\sum_{i=1}^{n} \tilde{F}_n\{Y_i\} = 1$. Let

$$\begin{aligned} \tilde{F}_n(y) &= \sum_{i=1}^{n} \tilde{F}_n\{Y_i\} I(Y_i \leq y) \\ &= \frac{1}{C_n} \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta_i I(Y_i \leq y)}{k(Y_i)}, \end{aligned} \tag{A.33}$$

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

$$\begin{aligned} C_n &= \frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i I(Y_i \le \infty)}{k(Y_i)} \\ &\to E\left(\frac{\Delta I(Y \le \infty)}{k(Y)}\right) \\ &= \int_0^\infty f_0(T)dT \\ &= 1. \end{aligned} \tag{A.34}$$

The calculation here is similar to that in (A.24) with $k(Y)$ in the denominator instead of $|\lambda^* - H^*(Y)|$. Therefore, combining (A.33) and (A.34) we have

$$\begin{aligned} \tilde{F}_n(y) &\to E\left(\frac{\Delta I(Y \le y)}{k(Y)}\right) \\ &= \int_0^y f_0(T)dT \\ &= F_0(y). \end{aligned} \tag{A.35}$$

Because $\hat{\boldsymbol{\beta}}_n, \hat{F}_n$ are maximum likelihood estimates, we have $\log\frac{L(\hat{\beta}_n,\hat{F}_n)}{L(\beta_0,\tilde{F}_n)} \ge 0$. Therefore

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} I(Y_i < \infty)\Delta_i \log\frac{\hat{F}_n\{Y_i\}}{\tilde{F}_n\{Y_i\}} + \frac{1}{n}\sum_{i=1}^{n} I(Y_i = \infty)\log\frac{G(\eta(\hat{\beta}_n, X_i))}{G(\eta(\beta_0, X_i))} \\ +\ &\frac{1}{n}\sum_{i=1}^{n} I(Y_i < \infty)(\Delta_i \log\frac{G'(\eta(\hat{\beta}_n, X_i)\hat{F}_n(Y_i))\eta(\hat{\beta}_n, X_i)}{G'(\eta(\beta_0, X)\tilde{F}_n(Y_i))\eta(\beta_0, X_i)} \\ +\ &(1-\Delta_i)\frac{G(\eta(\hat{\beta}_n, X_i)\hat{F}_n(Y_i))}{G(\eta(\beta_0, X_i)\tilde{F}_n(Y_i))}) \ge 0 \end{aligned} \tag{A.36}$$

Lemma 3. Under conditions (C1)-(C6), $E(\Delta I(Y < \infty)|\log\frac{f^*(Y)}{f_0(Y)}|) < \infty$, and

$$\frac{1}{n}\sum_{i=1}^{n}\Delta_i I(Y_i < \infty)\log\frac{\hat{F}_n\{Y_i\}}{\tilde{F}_n\{Y_i\}} \to E\left(\Delta I(Y < \infty)\log\frac{f^*(Y)}{f_0(Y)}\right).$$

Proof: For $t > 0$, we have $|\log t| \le t + \frac{1}{t}$. Therefore,

$$\begin{aligned} \left|\log\frac{f^*(y)}{f_0(y)}\right| &\le \frac{f^*(y)}{f_0(y)} + \frac{f_0(y)}{f^*(y)} \\ &= \frac{k(y)}{\lambda^* - H^*(y)} + \frac{\lambda^* - H^*(y)}{k(y)} \\ &\le \frac{c_1}{\lambda^* - H^*(y)} + c_2, \end{aligned} \tag{A.37}$$

where c_1 and c_2 are constants. " $\leq$ " is true in (A.37) because $\lambda^* - H^*$ is bounded and $k(.)$ is bounded from above and below. Therefore,

$$E\left(\Delta I(Y<\infty)|\log\frac{f^*(Y)}{f_0(Y)}|\right) \leq c_1 E\left(\frac{\Delta I(Y<\infty)}{\lambda^* - H^*(y)}\right) + c_2 < \infty \tag{A.38}$$

When $\Delta_i = 1$ and $Y_i < \infty$, we have $\hat{F}_n\{Y_i\} = \frac{1}{n(\hat{\lambda}_n - H_n(Y_i,\hat{\beta}_n,\hat{F}_n))}$ and $\tilde{F}_n\{Y_i\} = \frac{1}{nC_n k(Y_i)}$. Therefore,

$$\begin{aligned}\frac{\hat{F}_n\{Y_i\}}{\tilde{F}_n\{Y_i\}} &= C_n\frac{k(Y_i)}{\hat{\lambda}_n - H_n(Y_i,\hat{\beta}_n,\hat{F}_n)} \\ &= C_n\frac{\lambda^* - H^*(Y_i)}{\hat{\lambda}_n - H_n(Y_i,\hat{\beta}_n,\hat{F}_n)}\frac{f^*(Y_i)}{f_0(Y_i)}.\end{aligned} \tag{A.39}$$

Then, the first term in (A.36) becomes

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n I(Y_i<\infty)\Delta_i \log\frac{\hat{F}_n\{Y_i\}}{\tilde{F}_n\{Y_i\}} \\ =\ &\frac{1}{n}\sum_{i=1}^n I(Y_i<\infty)\Delta_i \log C_n \\ &+\frac{1}{n}\sum_{i=1}^n I(Y_i<\infty)\Delta_i \log\frac{\lambda^* - H^*(Y_i)}{\hat{\lambda}_n - H_n(Y_i,\hat{\beta}_n,\hat{F}_n)} \\ &+\frac{1}{n}\sum_{i=1}^n I(Y_i<\infty)\Delta_i \log\frac{f^*(Y_i)}{f_0(Y_i)} \\ =\ &\text{I+II+III.}\end{aligned} \tag{A.40}$$

We will then calculate I, II and III separately. For the first term, $|\text{I}| \leq |\log C_n| \to 0$ because $C_n \to 1$ when $n \to \infty$. By the law of large number, we have

$$|\text{III}| \to E\left(\Delta I(Y<\infty)\log\frac{f^*(Y)}{f_0(Y)}\right).$$

Because $\log(1+x) \leq x$ when $x \geq 0$, if $\lambda^* - H^*(Y_i) \geq \hat{\lambda}_n - H_n(Y_i,\hat{\beta}_n,\hat{F}_n)$, then

$$\begin{aligned}0 &\leq \log\frac{\lambda^* - H^*(Y_i)}{\hat{\lambda}_n - H_n(Y_i,\hat{\beta}_n,\hat{F}_n)} \\ &= \log\left(1 + \frac{(\lambda^* - H^*(Y_i)) - (\hat{\lambda}_n - H_n(Y_i,\hat{\beta}_n,\hat{F}_n))}{\hat{\lambda}_n - H_n(Y_i,\hat{\beta}_n,\hat{F}_n)}\right) \\ &\leq \frac{o_n(1)}{\hat{\lambda}_n - H(Y_i,\hat{\beta}_n,\hat{F})}, a.s.\end{aligned} \tag{A.41}$$

when $n \to \infty$. When $\lambda^* - H^*(Y_i) \leq \hat{\lambda}_n - H_n(Y_i, \hat{\beta}_n, \hat{F}_n)$, we have

$$\left| \log \frac{\lambda^* - H^*(Y_i)}{\hat{\lambda}_n - H_n(Y_i, \hat{\beta}_n, \hat{F})} \right| = \log \frac{\hat{\lambda}_n - H_n(Y_i, \hat{\beta}_n, \hat{F})}{\lambda^* - H^*(Y_i)} \leq \frac{o_n(1)}{\lambda^* - H^*(Y_i)}, a.s. \tag{A.42}$$

when $n \to \infty$.

Now, let us calculate the second term in (A.40). Combining (A.41) and (A.42), we have

$$\begin{aligned} \text{II} &\leq \frac{1}{n} \sum_{i=1}^{n} I(Y_i < \infty) \Delta_i \left| \log \frac{\lambda^* - H^*(Y_i)}{\hat{\lambda}_n - H_n(Y_i, \hat{\beta}_n, \hat{F}_n)} \right| \\ &\leq o_n(1) \frac{1}{n} \sum_{i=1}^{n} \frac{I(Y_i < \infty) \Delta_i}{\hat{\lambda}_n - H_n(Y_i, \hat{\beta}_n, \hat{F}_n)} + o_n(1) \frac{1}{n} \sum_{i=1}^{n} \frac{I(Y_i < \infty) \Delta_i}{\lambda^* - H^*(Y_i)}. \end{aligned} \tag{A.43}$$

We simply denote (A.41) by $o_n(1)\text{II}_1 + o_n(1)\text{II}_2$ and then calculate II_1, II_2 separately. In fact, we have

$$\begin{aligned} \text{II}_1 &= \sum_{i=1}^{n} \hat{F}_n\{Y_i\} = 1 \\ \text{II}_2 &= \frac{1}{n} \sum_{i=1}^{n} \frac{I(Y_i < \infty) \Delta_i}{\lambda^* - H^*(Y_i)} \to E(\frac{\Delta I(Y < \infty)}{\lambda^* - H^*(Y)}) \leq 1. \end{aligned} \tag{A.44}$$

Therefore, $\text{II} \to 0$ by (A.42) and (A.43). So far, we have proved in (A.40)

$$\begin{aligned} \text{I} &\to 0, \\ \text{II} &\to 0, \\ \text{III} &\to E\left(\Delta I(Y < \infty) \log \frac{f^*(Y)}{f_0(Y)} \right). \end{aligned}$$

Therefore,

$$\frac{1}{n} \sum_{i=1}^{n} \Delta_i I(Y_i < \infty) \log \frac{\hat{F}_n\{Y_i\}}{\tilde{F}_n\{Y_i\}} \to E\left(\Delta I(Y < \infty) \log \frac{f^*(Y)}{f_0(Y)} \right). \blacksquare$$

Now, let $n \to \infty$ in (A.36), we have $E\left(\log \frac{L(\beta^*, F^*)}{L(\beta_0, F_0)} \right) \geq 0$, where $L(.)$ is the likelihood function,

$$\begin{aligned} L(\beta, F) &= [(-G'(\eta(\beta, X)F(Y))\eta(\beta, F)f(Y))^{\Delta} \\ &\quad (G(\eta(\beta, X)F(Y)))^{1-\Delta}]^{I(Y<\infty)} [G(\eta(\beta, X))]^{I(Y=\infty)} \end{aligned} \tag{A.45}$$

By Jenson inequality, we have

$$\log E\left(\frac{L(\beta^*, F^*)}{L(\beta_0, F_0)} \right) \geq E\left(\log \frac{L(\beta^*, F^*)}{L(\beta_0, F_0)} \right) \geq 0, \tag{A.46}$$

where "=" holds if and only if $L(\beta^*, F^*) \equiv L(\beta_0, F_0)$, which concludes $\beta^* = \beta_0,\ F^* = F_0$ since the model is identifiable. Therefore, We only need to show $E\left(\frac{L(\beta^*,F^*)}{L(\beta_0,F_0)}\right) \leq 1$.

Theorem 1. Under conditions (C1)-(C6),

$$E\left(\frac{L(\beta^*, F^*)}{L(\beta_0, F_0)}\right) = 1.$$

The maximum likelihood estimates $(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)$ based on the modified likelihood function are strongly consistent, that is

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

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 $E\left(\frac{L(\beta^*,F^*)}{L(\beta_0,F_0)}\right) \leq 1$. In fact, we have

$$\begin{aligned} E\left(\frac{L(\beta^*, F^*)}{L(\beta_0, F_0)}\right) &= E\left(\Delta I(Y<\infty)\frac{G'(\eta(\beta^*,X)F^*(Y))\eta(\beta^*,X)f^*(Y)}{G'(\eta(\beta_0,X)F_0(Y))\eta(\beta_0,X)f_0(Y)}\right) \\ &\quad +E\left((1-\Delta)I(Y<\infty)\frac{G(\eta(\beta^*,X)F^*(Y))}{G(\eta(\beta_0,X)F_0(Y))}\right) \\ &\quad +E\left(I(Y=\infty)\frac{G(\eta(\beta^*,X))}{G(\eta(\beta_0,X))}\right) \\ &\quad = \text{I+II+III.} \end{aligned} \tag{A.47}$$

The three terms in (A.47) are denoted by I, II and III, respectively. We will then calculate each term separately. We have

$$\begin{aligned} \text{I} &= E\left(I(T\leq C)I(Y<\infty)\frac{G'(\eta(\beta^*,X)F^*(Y))\eta(\beta^*,X)f^*(Y)}{G'(\eta(\beta_0,X)F_0(Y))\eta(\beta_0,X)f_0(Y)}\right) \\ &= E_X\left(\int_0^\infty S_C(T|X)(-G'(\eta(\beta^*,X)F^*(T))\eta(\beta,X)f^*(T)dT)\right) \\ &= E_X\left(-\int_0^\infty S_C(T|X)dG(\eta(\beta^*,X)F^*(T))\right) \\ &= 1-E_X\left(S_C(\infty|X)G(\eta(\beta^*,X)F^*(\infty))\right) \\ &\quad +E_X\left(\int_0^\infty G(\eta(\beta^*,X)F^*(T))dS_C(T|X)\right), \end{aligned} \tag{A.48}$$

$$\begin{aligned} \text{II} &= E\left(I(T>C)I(C<\infty)\frac{G(\eta(\beta^*,X)F^*(C))}{G(\eta(\beta_0,X)F_0(C))}\right) \\ &= -E_X\left(\int_0^\infty G(\eta(\beta^*,X)F^*(C))dS_C(C|X)\right), \end{aligned} \tag{A.49}$$

$$\begin{aligned} \text{III} &= E\left(I(Y=\infty)\frac{G(\eta(\beta^*,X))}{G(\eta(\beta_0,X))}\right) \\ &= -E_X\left(\int_0^\infty G(\eta(\beta^*,X)F^*(C))dS_C(C|X)\right). \end{aligned} \tag{A.50}$$

Therefore,

$$\begin{aligned} E\left(\frac{L(\beta^*,F^*)}{L(\beta_0,F_0)}\right) &= 1 - E_X(S_C(\infty|X)G(\eta(\beta^*,X)F^*(\infty))) \\ &\quad +E_X(S_C(\infty|X)G(\eta(\beta^*,X))). \end{aligned} \tag{A.51}$$

Because $F^*(\infty) \leq 1$ and G is a monotone decreasing function, therefore in (A.51) we have

$$E\left(\frac{L(\beta^*,F^*)}{L(\beta_0,F_0)}\right) \leq 1. \tag{A.52}$$

On the other hand, we know that $E\left(\frac{L(\beta^*,F^*)}{L(\beta_0,F_0)}\right) \geq 1$ by the Jenson inequality in (A.44). So $E\left(\frac{L(\beta^*,F^*)}{L(\beta_0,F_0)}\right) = 1$, which concludes $\beta^* = \beta_0$,$F^* = F_0$ and also concludes $F^*(\infty) = 1$ because of (A.51).

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

For any $\epsilon > 0$, there exist a y_1, such that for any $y > y_1$ we have $F_0(y) \geq F_0(y_1) \geq 1 - \epsilon$ because $F_0(y)$ is continuous and $\lim_{y\to\infty} F_0(y) = 1$. Therefore, for any $y \geq y_1$, we have

$$\begin{aligned} & \hat{F}_n(y_1) - F_0(y_1) - \epsilon \\ \leq\ & \hat{F}_n(y_1) - 1 \\ \leq\ & \hat{F}_n(y) - F_0(y) \\ \leq\ & 1 - F_0(y_1) \\ \leq\ & \epsilon. \end{aligned} \tag{A.53}$$

On the other hand, there also exist a y_2, such that for any $y > y_2$ we have $|\hat{F}_n(y) - F_0(y)| < \epsilon$ because $\hat{F}_n(y) \to F_0(y)$ pointwise. Take $y_3 = \max(y_1, y_2)$, we then have $\hat{F}_n(y)$ converges to $F_0(y)$ uniformly in y for any $y > y_3$. Therefore, the uniform convergence holds for any $y \in [0, \infty)$. ■

A.2 Asymptotic Normality

In this section, we will prove the asymptotic normality of $(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{F}}_n)$ based on our proposed model using the Theorem 3.3.1 in van der Vaart and Wellner [23]. To study the joint distribution of $(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{F}}_n)$, we consider $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0, \hat{F}_n - F_0)$ as a linear operator in $l^\infty(\mathbb{R}^d \times V_0)$, where d is the dimension of $\boldsymbol{\beta}$ and V_0 is the set of functions with bounded total variation on $[0, \infty)$. Under some regular conditions, $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0, \hat{F}_n - F_0)$ will converge weakly to a mean zero Gaussian process when $n \to \infty$.

We prove Theorem 2 under the same conditions used in the proof of Theorem 1 except condition (C5). The condition (C5) is modified as the following,

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

We consider the likelihood function

$$\begin{aligned} L(\boldsymbol{\beta}, F) &= [\{-G'(\eta(\boldsymbol{\beta}, \boldsymbol{X})F(Y))\eta(\boldsymbol{\beta}, \boldsymbol{X})f(Y)\}^{\Delta} \\ &\quad \{G(\eta(\boldsymbol{\beta}, \boldsymbol{X})F(Y))\}^{1-\Delta}]^{I(Y<\infty)}[G(\eta(\boldsymbol{\beta}, \boldsymbol{X}))]^{I(Y=\infty)}, \end{aligned} \tag{A.54}$$

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

$$\begin{aligned} K(\boldsymbol{\beta}, F) &= [\{-G'(\eta(\boldsymbol{\beta}, \boldsymbol{X})F(Y))\eta(\boldsymbol{\beta}, \boldsymbol{X})\}^{\Delta} \\ &\quad \{G(\eta(\boldsymbol{\beta}, \boldsymbol{X})F(Y))\}^{1-\Delta}]^{I(Y<\infty)}[G(\eta(\boldsymbol{\beta}, \boldsymbol{X}))]^{I(Y=\infty)}. \end{aligned}$$

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

$$l(\boldsymbol{\beta}, F) = \Delta I(Y < \infty) \log f(Y) + \log K(\boldsymbol{\beta}, F). \tag{A.55}$$

In this section, we will still use the notations introduced in section A.1. Let $\mathbf{E}_n$ denote the empirical measure of n independent observations and $\mathbf{E}$ denote the expectation.

First, we prove that the likelihood function reaches its maximum when $\boldsymbol{\beta}$ and F take their true values. The result is stated in the following lemma.

Lemma 4. For any $\boldsymbol{\beta}$ and any distribution function F with a density, we have

$$\mathbf{E}(\log L(\boldsymbol{\beta}_0, F_0)) \geq \mathbf{E}(\log L(\boldsymbol{\beta}, F)), \tag{A.56}$$

where $L(\boldsymbol{\beta}, F)$ is the likelihood function given in (A.54), $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$ and F_0 is the true promotion time cumulative distribution function.

Proof: The proof of (A.56) is quite straight forward. By the Jensen inequality we have

$$\mathbf{E}\left(\log\frac{L(\boldsymbol{\beta},F)}{L(\boldsymbol{\beta}_0,F_0)}\right)\leq\log\mathbf{E}\left(\frac{L(\boldsymbol{\beta},F)}{L(\boldsymbol{\beta}_0,F_0)}\right).$$

Thus, it suffices to show that $\mathbf{E}\left(\frac{L(\beta,F)}{L(\beta_0,F_0)}\right)=1$. The following calculation is similar to part of the proof in Theorem 1, but slightly different. In fact, we have,

$$\begin{aligned}\mathbf{E}\left(\frac{L(\boldsymbol{\beta},F)}{L(\boldsymbol{\beta}_0,F_0)}\right) &= \mathbf{E}\left(\Delta I(Y<\infty)\frac{-G'(\eta(\boldsymbol{\beta},\boldsymbol{X})F(Y))\eta(\boldsymbol{\beta},\boldsymbol{X})f(Y)}{-G'(\eta(\boldsymbol{\beta}_0,\boldsymbol{X})F_0(Y))\eta(\boldsymbol{\beta}_0,\boldsymbol{X})f_0(Y)}\right) \\ &\quad +\mathbf{E}\left((1-\Delta)I(Y<\infty)\frac{G(\eta(\boldsymbol{\beta},\boldsymbol{X})F(Y))}{G(\eta(\boldsymbol{\beta}_0,\boldsymbol{X})F_0(Y))}\right) \\ &\quad +\mathbf{E}\left(I(Y=\infty)\frac{G(\eta(\boldsymbol{\beta},\boldsymbol{X}))}{G(\eta(\boldsymbol{\beta}_0,\boldsymbol{X}))}\right) \\ &= \text{I} +\text{II} +\text{III}.\end{aligned} \tag{A.57}$$

We denote the three terms in (A.57) by I, II, and III, respectively, and do the calculation separately.

From (3.14) in section 3.4, the survival function in our proposed model is given by

$$S(t)=G(\eta(\boldsymbol{\beta}_0,\boldsymbol{X})F_0(t)). \tag{A.58}$$

The corresponding density function, denoted by $f_T(t)$, is

$$f_T(t)=-G'(\eta(\boldsymbol{\beta}_0,\boldsymbol{X})F_0(T))\eta(\boldsymbol{\beta}_0,\boldsymbol{X})f_0(T). \tag{A.59}$$

From (A.58), we have

$$\begin{aligned}\mathbf{E}_T(I(T>C)|C) &= S(C)=G(\eta(\boldsymbol{\beta}_0,\boldsymbol{X})F_0(C)), \\ \mathbf{E}_T(I(T=\infty)|C) &= G(\eta(\boldsymbol{\beta}_0,\boldsymbol{X})).\end{aligned}$$

Therefore, the second and third terms in (A.57) can be simplified as the following,

$$\begin{aligned}\text{II} &= \mathbf{E}\left(I(T>C)I(C<\infty)\frac{G(\eta(\boldsymbol{\beta},\boldsymbol{X})F(C))}{G(\eta(\boldsymbol{\beta}_0,\boldsymbol{X})F_0(C))}\right) \\ &= \mathbf{E}\left(I(C<\infty)\frac{G(\eta(\boldsymbol{\beta},\boldsymbol{X})F(C))}{G(\eta(\boldsymbol{\beta}_0,\boldsymbol{X})F_0(C))}\mathbf{E}_T(\,I(T>C)|\,C)\right) \\ &= \mathbf{E}(I(C<\infty)G(\eta(\boldsymbol{\beta},\boldsymbol{X})F(C))), \\ \text{III} &= \mathbf{E}\left(I(C=\infty)I(T=\infty)\frac{G(\eta(\boldsymbol{\beta},\boldsymbol{X}))}{G(\eta(\boldsymbol{\beta}_0,\boldsymbol{X}))}\right) \\ &= \mathbf{E}\left(I(C=\infty)\frac{G(\eta(\boldsymbol{\beta},\boldsymbol{X}))}{G(\eta(\boldsymbol{\beta}_0,\boldsymbol{X}))}\mathbf{E}_T(\,I(T=\infty)|\,C)\right) \\ &= \mathbf{E}(I(C=\infty)G(\eta(\boldsymbol{\beta},\boldsymbol{X}))).\end{aligned} \tag{A.60}$$

Now, because of (A.59) we have

$$\begin{aligned}
&\mathbf{E}_T\left(I(T\le C)\frac{-G'(\eta(\boldsymbol{\beta},\boldsymbol{X})F(T))\eta(\boldsymbol{\beta},\boldsymbol{X})f(T)}{-G'(\eta(\boldsymbol{\beta}_0,\boldsymbol{X})F_0(T))\eta(\boldsymbol{\beta}_0,\boldsymbol{X})f_0(T)}\middle| C\right)\\
=&\int_0^C -G'(\eta(\boldsymbol{\beta},\boldsymbol{X})F(T))\eta(\boldsymbol{\beta},\boldsymbol{X})f(T)dT.
\end{aligned}$$

Noticing that

$$\begin{aligned}
\Delta I(Y<\infty) &= I(T\le C)I(T<\infty) \qquad (A.61)\\
&= I(T\le C)I(C<\infty)+I(T<\infty)I(C=\infty),
\end{aligned}$$

we obtain

$$\begin{aligned}
\mathrm{I} &= \mathbf{E}\left(I(T\le C)I(C<\infty)\frac{-G'(\eta(\boldsymbol{\beta},\boldsymbol{X})F(T))\eta(\boldsymbol{\beta},\boldsymbol{X})f(T)}{-G'(\eta(\boldsymbol{\beta}_0,\boldsymbol{X})F_0(T))\eta(\boldsymbol{\beta}_0,\boldsymbol{X})f_0(T)}\right) \qquad (A.62)\\
&\quad +\mathbf{E}\left(I(T<\infty)I(C=\infty)\frac{-G'(\eta(\boldsymbol{\beta},\boldsymbol{X})F(T))\eta(\boldsymbol{\beta},\boldsymbol{X})f(T)}{-G'(\eta(\boldsymbol{\beta}_0,\boldsymbol{X})F_0(T))\eta(\boldsymbol{\beta}_0,\boldsymbol{X})f_0(T)}\right)\\
&= \mathbf{E}\left(I(C<\infty)\mathbf{E}_T\left(I(T\le C)\frac{-G'(\eta(\boldsymbol{\beta},\boldsymbol{X})F(T))\eta(\boldsymbol{\beta},\boldsymbol{X})f(T)}{-G'(\eta(\boldsymbol{\beta}_0,\boldsymbol{X})F_0(T))\eta(\boldsymbol{\beta}_0,\boldsymbol{X})f_0(T)}\middle| C\right)\right)\\
&\quad +\mathbf{E}\left(I(C=\infty)\mathbf{E}_T\left(I(T<\infty)\frac{-G'(\eta(\boldsymbol{\beta},\boldsymbol{X})F(T))\eta(\boldsymbol{\beta},\boldsymbol{X})f(T)}{-G'(\eta(\boldsymbol{\beta}_0,\boldsymbol{X})F_0(T))\eta(\boldsymbol{\beta}_0,\boldsymbol{X})f_0(T)}\middle| C\right)\right)\\
&= \mathbf{E}\left(I(C<\infty)\int_0^C -G'(\eta(\boldsymbol{\beta},\boldsymbol{X})F(T))\eta(\boldsymbol{\beta},\boldsymbol{X})f(T)dT\right)\\
&\quad +\mathbf{E}\left(I(C=\infty)\int_0^\infty -G'(\eta(\boldsymbol{\beta},\boldsymbol{X})F(T))\eta(\boldsymbol{\beta},\boldsymbol{X})f(T)dT\right)\\
&= \mathbf{E}(I(C<\infty)(1-G(\eta(\boldsymbol{\beta},\boldsymbol{X})F(C))))+\mathbf{E}(I(C=\infty)(1-G(\eta(\boldsymbol{\beta},\boldsymbol{X}))))\\
&= \mathbf{E}(I(C<\infty)+I(C=\infty))-\mathbf{E}(I(C<\infty)G(\eta(\boldsymbol{\beta},\boldsymbol{X})F(C)))\\
&\quad -\mathbf{E}(I(C=\infty)G(\eta(\boldsymbol{\beta},\boldsymbol{X})))\\
&= 1-\mathbf{E}(I(C<\infty)G(\eta(\boldsymbol{\beta},\boldsymbol{X})F(C)))-\mathbf{P}(I(C=\infty)G(\eta(\boldsymbol{\beta},\boldsymbol{X}))).
\end{aligned}$$

Combining (A.57), (A.60), and (A.62), we have $\mathbf{E}\left(\frac{L(\beta,F)}{L(\beta_0,F_0)}\right)=\mathrm{I}+\mathrm{II}+\mathrm{III}=1$. Therefore, by Jensen inequality $\mathbf{E}\left(\log\frac{L(\beta,F)}{L(\beta_0,F_0)}\right)\le\log\mathbf{E}\left(\frac{L(\beta,F)}{L(\beta_0,F_0)}\right)$, (A.56) holds .■

Remark: In section 2.4, we have proved the proposed transformation model is identifiable. Therefore, the two sides of (A.56) is equal if and only if $\boldsymbol{\beta}=\boldsymbol{\beta}_0$, $F=F_0$. ■

Now, from (A.56) 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 $\nu \in \mathbb{R}$ 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 $F_0(t) + \nu H(t)$. By Lemma 4, for any $\boldsymbol{\beta} \in \mathbb{R}^d$ and any distribution function F with a density, we have

$$\mathbf{E}(\log L(\boldsymbol{\beta}_0, F_0)) \geq \mathbf{E}(\log L(\boldsymbol{\beta}, F)).$$

Also, for any $\boldsymbol{\alpha} \in \mathbb{R}^d$, where d is the dimension of $\boldsymbol{\beta}$, we have $\boldsymbol{\beta}_0 + \nu\boldsymbol{\alpha} \in \mathbb{R}^d$. Therefore,

$$\mathbf{E}(\log L(\boldsymbol{\beta}_0, F_0)) \geq \mathbf{E}(\log L(\boldsymbol{\beta}_0 + \nu\boldsymbol{\alpha}, F_0 + \nu H)) \tag{A.63}$$

when $|\nu|$ is small. Now, for any $\boldsymbol{\alpha} \in \mathbb{R}^d$ and any function $H(t)$ such that conditions (1)-(3) hold, we have

$$\frac{d}{d\nu}\left(\mathbf{E}(\log L(\boldsymbol{\beta}_0 + \nu\boldsymbol{\alpha}, F_0 + \nu H))\right)|_{\nu=0} = 0. \tag{A.64}$$

By (A.55), (A.64) becomes

$$\mathbf{E}\left(\Delta I(Y < \infty)\frac{h}{f_0} + \frac{\partial}{\partial F}\log K(\boldsymbol{\beta}_0, F_0)H + \frac{\partial}{\partial \boldsymbol{\beta}}\log K(\boldsymbol{\beta}_0, F_0)\boldsymbol{\alpha}\right) = 0. \tag{A.65}$$

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

$$\|h(t)\|_V = \sup \sum_{i=1}^{m} |h(t_{i+1} - h(t_i))|, \tag{A.66}$$

where the supreme is taken over all finite partitions $0 = t_1 < t_2 < ... < t_{m+1} = \infty$.

For the proposed model, we define

$$\begin{aligned} Q_{F_0}h(y) &= h(y) - \int_0^\infty h(y)dF_0(y), \\ H(y) &= \int_{[0,y]} Q_{F_0}h(s)dF_0(s). \end{aligned} \tag{A.67}$$

We will show $H(y)$ defined in (A.67) satisfies conditions (1)-(3). First, it is obvious that $H(0) = 0$ and $H(\infty) = 0$. In fact,

$$\begin{aligned} H(\infty) &= \int_0^\infty Q_{F_0} h(s) dF_0(s) \\ &= \int_0^\infty \left(h(s) - \int_0^\infty h(s) dF_0(s) \right) dF_0(s) \\ &= \int_0^\infty h(s) dF_0(s) - \int_0^\infty h(s) dF_0(s) \int_0^\infty dF_0(s) \\ &= 0. \end{aligned}$$

Secondly, from the definition of $H(y)$ in (A.67) we can see that $H(y)$ is continuously differentiable with $H'(y) = Q_{F_0} h(y) f_0(y)$. Also, because $h(y)$ has bounded total variation, function $Q_{F_0} h(y)$ is bounded. Therefore, when $|\nu|$ is small enough we have $1 + \nu Q_{F_0} h(y) > 0$ and hence, $(1 + \nu Q_{F_0} h(y)) f_0(y) > 0$ is a density function with corresponding distribution $F_0(y) + \nu H(y)$.

With the specially constructed function $H(.)$ in (A.67), equation (A.65) can be written as

$$\mathbf{E} \left(\Delta I(Y < \infty) Q_{F_0} h + \frac{\partial}{\partial F} \log K(\boldsymbol{\beta}_0, F_0) \int_{[0,Y]} Q_{F_0} h dF_0 + \frac{\partial}{\partial \boldsymbol{\beta}} \log K(\boldsymbol{\beta}_0, F_0) \boldsymbol{\alpha} \right) = 0. \quad \text{(A.68)}$$

To study the asymptotic behavior of $(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0, \hat{F}_n - F_0)$, we also need to derive a differential equation with $(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)$ in the similar format of (A.68). The derivation is summarized in the following Lemma and some discussions after it. Before we state and prove the Lemma, we need to point out that here we will consider a modified semiparametric version likelihood function,

$$\begin{aligned} L_1(\boldsymbol{\beta}, F_n) &= [\{-G'(\eta(\boldsymbol{\beta}, \boldsymbol{X}) F_n(Y)) \eta(\boldsymbol{\beta}, \boldsymbol{X}) F_n\{Y\}\}^\Delta \quad \text{(A.69)} \\ &\quad \{G(\eta(\boldsymbol{\beta}, \boldsymbol{X}) F_n(Y))\}^{1-\Delta}]^{I(Y<\infty)} [G(\eta(\boldsymbol{\beta}, \boldsymbol{X}))]^{I(Y=\infty)} \\ &= F_n\{Y\}^{\Delta I(Y<\infty)} K(\boldsymbol{\beta}, F_n), \end{aligned}$$

where

$$\begin{aligned} K(\boldsymbol{\beta}, F_n) &= [\{-G'(\eta(\boldsymbol{\beta}, \boldsymbol{X}) F_n(Y)) \eta(\boldsymbol{\beta}, \boldsymbol{X})\}^\Delta \quad \text{(A.70)} \\ &\quad \{G(\eta(\boldsymbol{\beta}, \boldsymbol{X}) F_n(Y))\}^{1-\Delta}]^{I(Y<\infty)} [G(\eta(\boldsymbol{\beta}, \boldsymbol{X}))]^{I(Y=\infty)}. \end{aligned}$$

In the likelihood function (A.69), function $F_n(.)$ is a monotonic step function and belongs to a class of functions $\mathcal{F}_n$ based on the observations $\{(\Delta_i, Y_i) : i = 1, 2, ..., n\}$,

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

For any y, we define $F_n(y) = \sum_{Y_i \le y} F_n\{Y_i\}$.

For any $\boldsymbol{\beta}$ and any step function $F_n(.) \in \mathcal{F}_n$, letting $(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)$ be the maximum likelihood estimate of $(\boldsymbol{\beta}, F_n)$ based on (A.69) and $\mathbf{E}_n$ be the empirical measure of a random sample $\{(\Delta_i, Y_i) : i = 1, 2, ..., n\}$, we have

$$\mathbf{E}_n(\log L_1(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)) \ge \mathbf{E}_n(\log L_1(\boldsymbol{\beta}, F_n)), \tag{A.71}$$

where $L_1(\boldsymbol{\beta}, F_n)$ is the modified semiparametric version likelihood function given in (A.69).

Similar to the continuous case, now we can derive a differential equation with $(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)$. 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, $\hat{F}_n\{Y_i\} + \nu H_n\{Y_i\} \ge 0$ for any Y_i.

Under condition (1)', function $H_n(.)$ is a step function but it does not have to be monotonic. For any y, we define $H_n(y) = \sum_{Y_i \le y} H_n\{Y_i\}$. If a function $H_n(.)$ satisfies conditions (1)'-(3)', then $\hat{F}_n + \nu H_n$ is a qualified distribution function for likelihood (A.69). Take $\boldsymbol{\alpha} \in \mathbb{R}^d$ such that $\hat{\boldsymbol{\beta}}_n + \nu\boldsymbol{\alpha} \in \mathbb{R}^d$. Therefore, because of (A.71) we have

$$\mathbf{E}_n(\log L_1(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)) \ge \mathbf{E}_n(\log L_1(\hat{\boldsymbol{\beta}}_n + \nu\boldsymbol{\alpha}, \hat{F}_n + \nu H_n)), \tag{A.72}$$

when $|\nu|$ is small enough.

Immediately from (A.72), we get a differential equation with $(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)$,

$$\frac{d}{d\nu}\mathbf{E}_n(\log L_1(\hat{\boldsymbol{\beta}}_n + \nu\boldsymbol{\alpha}, \hat{F}_n + \nu H_n))|_{\nu=0} = 0. \tag{A.73}$$

After some algebra, we obtain

$$\frac{1}{n}\sum_{i=1}^{n} \Delta_i I(Y_i < \infty)\frac{H_n\{Y_i\}}{\hat{F}_n\{Y_i\}} + \mathbf{E}_n\left(\frac{\partial}{\partial F}\log K(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)H_n + \frac{\partial}{\partial \boldsymbol{\beta}}\log K(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)\boldsymbol{\alpha}\right) = 0. \tag{A.74}$$

Equation (A.74) is comparable with equation (A.65). In fact, what we derived here is an empirical version of (A.65) with the maximum likelihood estimates $(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)$.

Again, we are particularly interested in step function $H_n(.)$ constructed through a function $h(.)$ with bounded total variation. Using the notations of Stieltjes integral, we define

$$\begin{aligned} Q_{\hat{F}_n} h(y) &= h(y) - \int_0^\infty h(s) d\hat{F}_n(s) \\ &= h(y) - \sum_{i=1}^{\infty} h(Y_i) \hat{F}_n\{Y_i\}, \\ H_n(y) &= \int_{[0,y]} Q_{\hat{F}_n} h(s) d\hat{F}_n(s) \\ &= \sum_{Y_i \le y} Q_{\hat{F}_n} h(Y_i) \hat{F}_n\{Y_i\}. \end{aligned} \tag{A.75}$$

It is easy to check function $H_n(.)$ in (A.75) is a step function with a jump of size

$$H_n\{Y_i\} = Q_{\hat{F}_n} h(Y_i) \hat{F}_n\{Y_i\} \tag{A.76}$$

at Y_i when $\Delta_i = 1$. Also, the summation of $H_n(.)$ over all Y_i's is zero, which is

$$\sum_{i=1}^{n} H_n\{Y_i\} = \sum_{i=1}^{n} h(Y_i) \hat{F}_n\{Y_i\} - \sum_{i=1}^{n} h(Y_i) \hat{F}_n\{Y_i\} \sum_{i=1}^{n} \hat{F}_n\{Y_i\} = 0. \tag{A.77}$$

Plugging (A.76) and (A.77) in (A.74), and noticing that

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \Delta_i I(Y_i < \infty) \frac{H_n\{Y_i\}}{\hat{F}_n\{Y_i\}} &= \frac{1}{n} \sum_{i=1}^{n} \Delta_i I(Y_i < \infty) Q_{\hat{F}_n} h\{Y_i\} \\ &= \mathbf{E}_n(\Delta I(Y < \infty) Q_{\hat{F}_n} h), \end{aligned} \tag{A.78}$$

we have

$$\mathbf{E}_n \left(\Delta I(Y < \infty) Q_{\hat{F}_n} h + \frac{\partial}{\partial F} \log K(\hat{\boldsymbol{\beta}}_n, \hat{F}_n) \int_{[0,Y]} Q_{\hat{F}_n} d\hat{F}_n + \frac{\partial}{\partial \boldsymbol{\beta}} \log K(\hat{\boldsymbol{\beta}}_n, \hat{F}_n) \boldsymbol{\alpha} \right) = 0. \tag{A.79}$$

Now, let us consider functions h with bounded total variation such that

$$\int_0^\infty h(y) dF_0(y) = 0$$

and define a set of such functions as V_0,

$$V_0 = \left\{ h \in V \,\middle|\, \int_0^\infty h dF_0 = 0 \right\}.$$

We introduce two notations, Ω_β and Ω_F, which can be view as the second order derivatives with respect to $\boldsymbol{\beta}$ and F, respectively. For any $(\boldsymbol{\alpha}, h) \in \mathbb{R}^d \times V_0$, we define

$$\begin{aligned} \Omega_\beta(\boldsymbol{\alpha}, h) &= \mathbf{E}\left(\frac{\partial^2}{\partial \boldsymbol{\beta}^2} \log K(\boldsymbol{\beta}_0, F_0)\right)\boldsymbol{\alpha} + \mathbf{E}\left(\frac{\partial^2}{\partial F \partial \boldsymbol{\beta}} \log K(\boldsymbol{\beta}_0, F_0) \int_{[0,Y]} h dF_0\right), \quad \text{(A.80)} \\ \Omega_F(\boldsymbol{\alpha}, h) &= \omega - \int_0^\infty \omega dF_0, \end{aligned}$$

where

$$\begin{aligned} \omega &= -\mathbf{E}\left(\Delta I(Y < \infty) + \frac{\partial}{\partial F} \log K(\boldsymbol{\beta}_0, F_0)(F_0(Y) - I(Y \geq s))\right) h \\ &\quad + \mathbf{E}\left(\frac{\partial^2}{\partial F^2} \log K(\boldsymbol{\beta}_0, F_0) I(Y \geq s) \int_{[0,Y]} h dF_0\right) \\ &\quad + \mathbf{E}\left(\frac{\partial^2}{\partial F \partial \boldsymbol{\beta}} \log K(\boldsymbol{\beta}_0, F_0) I(Y \geq s)\right)\boldsymbol{\alpha}. \end{aligned}$$

To study the asymptotic behavior of $(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)$, we need to measure the size of $\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0$ and $\hat{F}_n - F_0$. Some of the results are summarized in Lemma 5.

Lemma 5. With the notations defined in (A.80), we have

$$\begin{aligned} &\sqrt{n}\left(\Omega_\beta(\boldsymbol{\alpha}, h)(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + \int_0^\infty \Omega_F(\boldsymbol{\alpha}, h) d(\hat{F}_n - F_0)\right) \quad \text{(A.81)} \\ = \; &-\sqrt{n}(\mathbf{E}_n - \mathbf{E})\left(\frac{\partial}{\partial \boldsymbol{\beta}} \log L(\boldsymbol{\beta}_0, F_0)\boldsymbol{\alpha} + \frac{\partial}{\partial F} \log L(\boldsymbol{\beta}_0, F_0) \int_{[0,Y]} h dF_0\right) \\ &+ o_p(\sqrt{n}\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| + \sqrt{n}\|\hat{F}_n - F_0\|_{L^\infty}) + o_p(1), \end{aligned}$$

where $h \in V_0$ and $L(.,.)$ is the likelihood function given in (A.54).

Proof: In (A.68) and (A.79), we have derived the differential equation with $(\boldsymbol{\beta}_0, F_0)$

$$\mathbf{E}\left(\Delta I(Y < \infty) Q_{F_0} h + \frac{\partial}{\partial F} \log K(\boldsymbol{\beta}_0, F_0) \int_{[0,Y]} Q_{F_0} h dF_0 + \frac{\partial}{\partial \boldsymbol{\beta}} \log K(\boldsymbol{\beta}_0, F_0)\boldsymbol{\alpha}\right) = 0, \quad \text{(A.82)}$$

and the differential equation with $(\hat{\boldsymbol{\beta}}, \hat{F}_n)$

$$\mathbf{E}_n\left(\Delta I(Y < \infty) Q_{\hat{F}_n} h + \frac{\partial}{\partial F} \log K(\hat{\boldsymbol{\beta}}_n, \hat{F}_n) \int_{[0,Y]} Q_{\hat{F}_n} d\hat{F}_n + \frac{\partial}{\partial \boldsymbol{\beta}} \log K(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)\boldsymbol{\alpha}\right) = 0. \quad \text{(A.83)}$$

From (A.82) and (A.83), we can obtain the following equation after some simple algebra,

$$
\begin{aligned}
& -\sqrt{n}(\mathbf{E}_n - \mathbf{E}) \qquad\qquad (A.84)\\
& \left(\Delta I(Y<\infty) Q_{\hat{F}_n} h + \frac{\partial}{\partial F}\log K(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)\int_{[0,Y]} Q_{\hat{F}_n} h d\hat{F}_n + \frac{\partial}{\partial \boldsymbol{\beta}}\log K(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)\boldsymbol{\alpha}\right)\\
= \;& \sqrt{n}\mathbf{E}\left(\Delta I(Y<\infty) Q_{\hat{F}_n} h + \frac{\partial}{\partial F}\log K(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)\int_{[0,Y]} Q_{\hat{F}_n} h d\hat{F}_n + \frac{\partial}{\partial \boldsymbol{\beta}}\log K(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)\boldsymbol{\alpha}\right)\\
& -\sqrt{n}\mathbf{E}\left(\Delta I(Y<\infty) Q_{F_0} h + \frac{\partial}{\partial F}\log K(\boldsymbol{\beta}_0, F_0)\int_{[0,Y]} Q_{F_0} h dF_0 + \frac{\partial}{\partial \boldsymbol{\beta}}\log K(\boldsymbol{\beta}_0, F_0)\boldsymbol{\alpha}\right).
\end{aligned}
$$

Equation (A.84) is an important step in the proof of asymptotic normality. We will calculate the left hand side of (A.84) by using Donsker class and Theorem 3.3.1 in van der Vaart and Wellner [23] and simplify the right hand side of (A.84) by using Taylor expansion.

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

$$
\begin{aligned}
\mathcal{A} \;=\; & \left\{\Delta I(Y<\infty) Q_F h + \frac{\partial}{\partial F}\log K(\boldsymbol{\beta}, F)\int_{[0,Y]} Q_F h dF \qquad\qquad (A.85)\right.\\
& \left. + \frac{\partial}{\partial \boldsymbol{\beta}}\log K(\boldsymbol{\beta}, F)\boldsymbol{\alpha} : \|\boldsymbol{\alpha}\| \le 1, \|h\|_V \le 1, \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| + \|F - F_0\|_{L^\infty} \le 1\right\}.
\end{aligned}
$$

By the same arguments as in the consistency proof, we can show that the class $\mathcal{A}$ is a Donsker class. Therefore, by Theorem 3.3.1 in van der Vaart and Wellner [23] the left hand side of (A.84) equals

$$
\begin{aligned}
& -\sqrt{n}(\mathbf{E}_n - \mathbf{E}) \qquad\qquad (A.86)\\
& \left(\Delta I(Y<\infty) Q_{F_0} h + \frac{\partial}{\partial F}\log K(\boldsymbol{\beta}_0, F_0)\int_{[0,Y]} Q_{F_0} h dF_0 + \frac{\partial}{\partial \boldsymbol{\beta}}\log K(\boldsymbol{\beta}_0, F_0)\boldsymbol{\alpha}\right) + o_p(1)\\
= \;& -\sqrt{n}(\mathbf{E}_n - \mathbf{E})\left(\frac{\partial}{\partial F}\log K(\boldsymbol{\beta}_0, F_0)\int_{[0,Y]} Q_{F_0} h dF_0 + \frac{\partial}{\partial \boldsymbol{\beta}}\log K(\boldsymbol{\beta}_0, F_0)\boldsymbol{\alpha}\right) + o_p(1).
\end{aligned}
$$

The calculation of the right hand side of (A.84) is long, but the idea is very simple. Basically, we use Taylor expansion to estimate the size of $\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0$ and the size of $\hat{F}_n - F_0$. For simplicity, let us ignore the factor $\sqrt{n}$ for now, group each two similar terms together, and write the expectation as a summation of three differences. Therefore, the expectation

in the right hand side of (A.84) equals

$$\begin{aligned} & \mathbf{E}\left(\Delta I(Y<\infty)Q_{\hat{F}_n}h - \Delta I(Y<\infty)Q_{F_0}h\right) \qquad (A.87)\\ & +\mathbf{E}\left(\frac{\partial}{\partial F}\log K(\hat{\boldsymbol{\beta}}_n,\hat{F}_n)\int_{[0,Y]}Q_{\hat{F}_n}hd\hat{F}_n - \frac{\partial}{\partial F}\log K(\boldsymbol{\beta}_0,F_0)\int_{[0,Y]}Q_{F_0}hdF_0\right)\\ & +\mathbf{E}\left(\frac{\partial}{\partial \boldsymbol{\beta}}\log K(\hat{\boldsymbol{\beta}}_n,\hat{F}_n)\boldsymbol{\alpha} - \frac{\partial}{\partial \boldsymbol{\beta}}\log K(\boldsymbol{\beta}_0,F_0)\boldsymbol{\alpha}\right)\\ = \;& \text{I+II+III.} \end{aligned}$$

Definitions of $Q_{F_0}h$ and $Q_{\hat{F}_n}h$ in (A.67) and (A.75) will be used when we simplify the first term, I, in equation (A.87). Expand the third term, III, at $(\boldsymbol{\beta}_0, F_0)$ by using Taylor expansion.

$$\begin{aligned} \text{I} &= \mathbf{E}\left(\Delta I(Y<\infty)Q_{\hat{F}_n}h - \Delta I(Y<\infty)Q_{F_0}h\right) \qquad (A.88)\\ &= -\mathbf{E}\left(\Delta I(Y<\infty)\int_0^\infty hd(\hat{F}_n - F_0)\right)\\ &= -\mathbf{E}\left(\Delta I(Y<\infty)\int_0^\infty Q_{F_0}hd(\hat{F}_n - F_0)\right)\\ &= -\int_0^\infty \mathbf{E}\left(\Delta I(Y<\infty)\right)Q_{F_0}h(s)d\left(\hat{F}_n(s) - F_0(s)\right), \end{aligned}$$

(A.89)

$$\begin{aligned} \text{III} &= \mathbf{E}\left(\frac{\partial}{\partial \boldsymbol{\beta}}\log K(\hat{\boldsymbol{\beta}}_n,\hat{F}_n)\boldsymbol{\alpha} - \frac{\partial}{\partial \boldsymbol{\beta}}\log K(\boldsymbol{\beta}_0,F_0)\boldsymbol{\alpha}\right)\\ &= \mathbf{E}\left(\frac{\partial^2}{\partial \boldsymbol{\beta}^2}\log K(\boldsymbol{\beta}_0,F_0)(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0,\boldsymbol{\alpha}) + \frac{\partial^2}{\partial F\partial \boldsymbol{\beta}}\log K(\boldsymbol{\beta}_0,F_0)(\hat{F}_n(Y) - F_0(Y),\boldsymbol{\alpha})\right)\\ &\quad +o_p(\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| + \|\hat{F}_n - F_0\|_{L^\infty}))\\ &= \mathbf{E}\left(\frac{\partial^2}{\partial \boldsymbol{\beta}^2}\log K(\boldsymbol{\beta}_0,F_0)(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0,\boldsymbol{\alpha})\right)\\ &\quad +\mathbf{E}\left(\frac{\partial^2}{\partial F\partial \boldsymbol{\beta}}\log K(\boldsymbol{\beta}_0,F_0)\left(\int_0^\infty I(Y\geq s)d\left(\hat{F}_n(s) - F_0(s)\right),\boldsymbol{\alpha}\right)\right)\\ &\quad +o_p(\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| + \|\hat{F}_n - F_0\|_{L^\infty}))\\ &= \mathbf{E}\left(\frac{\partial^2}{\partial \boldsymbol{\beta}^2}\log K(\boldsymbol{\beta}_0,F_0)\right)(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0,\boldsymbol{\alpha})\\ &\quad +\int_0^\infty \mathbf{E}\left(\frac{\partial^2}{\partial F\partial \boldsymbol{\beta}}\log K(\boldsymbol{\beta}_0,F_0)I(Y\geq s)\right)\boldsymbol{\alpha}d\left(\hat{F}_n(s) - F_0(s)\right)\\ &\quad +o_p(\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| + \|\hat{F}_n - F_0\|_{L^\infty})), \end{aligned}$$

In the calculation of II of (A.87), we insert a term $\mathbf{E}\left(\frac{\partial}{\partial F}\log K(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)\int_{[0,Y]} Q_{F_0} h dF_0\right)$, group the four terms as II_1 and II_2, and then use Taylor expansion again.

$$\begin{aligned}
\mathrm{II} &= \mathbf{E}\left(\frac{\partial}{\partial F}\log K(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)\int_{[0,Y]} Q_{\hat{F}_n} h d\hat{F}_n - \frac{\partial}{\partial F}\log K(\boldsymbol{\beta}_0, F_0)\int_{[0,Y]} Q_{F_0} h dF_0\right) \quad \text{(A.90)}\\
&= \mathbf{E}\left(\left(\frac{\partial}{\partial F}\log K(\hat{\boldsymbol{\beta}}_n, \hat{F}_n) - \frac{\partial}{\partial F}\log K(\boldsymbol{\beta}_0, F_0)\right)\int_{[0,Y]} Q_{F_0} h dF_0\right)\\
&\quad +\mathbf{E}\left(\frac{\partial}{\partial F}\log K(\hat{\boldsymbol{\beta}}_n, \hat{F}_n)\left(\int_{[0,Y]} Q_{\hat{F}_n} h d\hat{F}_n - \int_{[0,Y]} Q_{F_0} h dF_0\right)\right)\\
&= \mathrm{II}_1 + \mathrm{II}_2.
\end{aligned}$$

The first term equals

$$\begin{aligned}
\mathrm{II}_1 &= \mathbf{E}\left(\frac{\partial^2}{\partial F \partial \boldsymbol{\beta}}\log K(\boldsymbol{\beta}_0, F_0)(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\int_{[0,Y]} Q_{F_0} h dF_0\right) \quad \text{(A.91)}\\
&\quad +\mathbf{E}\left(\frac{\partial^2}{\partial F^2}\log K(\boldsymbol{\beta}_0, F_0)(\hat{F}_n - F_0)\int_{[0,Y]} Q_{F_0} h dF_0\right)\\
&\quad +o_p(\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| + \|\hat{F}_n - F_0\|_{L^\infty})\\
&= \mathbf{E}\left(\frac{\partial^2}{\partial F \partial \boldsymbol{\beta}}\log K(\boldsymbol{\beta}_0, F_0)\int_{[0,Y]} Q_{F_0} h dF_0\right)(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\\
&\quad +\int_0^\infty \mathbf{E}\left(\frac{\partial^2}{\partial F^2}\log K(\boldsymbol{\beta}_0, F_0) I(Y \geq s)\int_{[0,Y]} Q_{F_0} h dF_0\right) d\left(\hat{F}_n(s) - F_0(s)\right)\\
&\quad +o_p(\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| + \|\hat{F}_n - F_0\|_{L^\infty}).
\end{aligned}$$

Before we calculate II_2, let us set up an equation that will be used in the next step. By the definition of $Q_{F_0}h$ and $Q_{\hat{F}_n}h$ in (A.67) and (A.75), we have

$$Q_{\hat{F}_n}h - Q_{F_0}h = \int_0^\infty h d\hat{F}_n - \int_0^\infty h dF_0 = \int_0^\infty Q_{F_0} h d(\hat{F}_n - F_0). \quad \text{(A.92)}$$

In fact, we have seen (A.92) is the calculation of (A.88). Using (A.92) and also noticing that

$$\left|\int_0^\infty Q_{F_0} h d(\hat{F}_n - F_0)\right| = \left|\int_0^\infty (\hat{F}_n - F_0) dQ_{F_0} h\right| \leq \|\hat{F}_n - F_0\|_{L^\infty}\|h\|_V, \quad \text{(A.93)}$$

we obtain

$$
\begin{aligned}
& \int_{[0,Y]} Q_{\hat{F}_n} h d\hat{F}_n - \int_{[0,Y]} Q_{F_0} h dF_0 \qquad (A.94)\\
= & \int_{[0,Y]} Q_{F_0} h d(\hat{F}_n - F_0) + \int_{[0,Y]} (Q_{\hat{F}_n} h - Q_{F_0} h) d\hat{F}_n \\
= & \int_0^\infty I(Y \geq s) Q_{F_0} h(s) d(\hat{F}_n(s) - F_0(s)) - \hat{F}_n(Y) \int_0^\infty Q_{F_0} h d(\hat{F}_n - F_0) \\
= & \int_0^\infty I(Y \geq s) Q_{F_0} h(s) d(\hat{F}_n(s) - F_0(s)) - F_0(Y) \int_0^\infty Q_{F_0} h d(\hat{F}_n - F_0) \\
& + (F_0(Y) - \hat{F}_n(Y)) \int_0^\infty Q_{F_0} h d(\hat{F}_n - F_0) \\
= & \int_0^\infty (I(Y \geq s) - F_0(Y)) Q_{F_0} h(s) d(\hat{F}_n(s) - F_0(s)) + o_p(\|\hat{F}_n - F_0\|_{L^\infty}).
\end{aligned}
$$

From (A.94), we can prove the second term II_2 equals

$$
\begin{aligned}
\mathrm{II}_2 = & \mathbf{E}\left(\frac{\partial}{\partial F} \log K(\boldsymbol{\beta}_0, F_0) \int_0^\infty (I(Y \geq s) - F_0(Y)) Q_{F_0} h(s) d(\hat{F}_n(s) - F_0(s)) \right) \qquad (A.95)\\
& + o_p(\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| + \|\hat{F}_n - F_0\|_{L^\infty}) \\
= & \int_0^\infty \mathbf{E}\left(\frac{\partial}{\partial F} \log K(\boldsymbol{\beta}_0, F_0) (I(Y \geq s) - F_0(Y)) Q_{F_0} h(s) d(\hat{F}_n(s) - F_0(s)) \right) \\
& + o_p(\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| + \|\hat{F}_n - F_0\|_{L^\infty}).
\end{aligned}
$$

Recall the notations, Ω_β and Ω_F, we introduced in (A.80). For any $(\boldsymbol{\alpha}, h) \in \mathbb{R}^d \times V_0$, we define

$$
\begin{aligned}
\Omega_\beta(\boldsymbol{\alpha}, h) &= \mathbf{E}\left(\frac{\partial^2}{\partial \boldsymbol{\beta}^2} \log K(\boldsymbol{\beta}_0, F_0) \right) \boldsymbol{\alpha} + \mathbf{E}\left(\frac{\partial^2}{\partial F \partial \boldsymbol{\beta}} \log K(\boldsymbol{\beta}_0, F_0) \int_{[0,Y]} h dF_0 \right), \\
\Omega_F(\boldsymbol{\alpha}, h) &= \omega - \int_0^\infty \omega dF_0,
\end{aligned}
$$

where

$$
\begin{aligned}
\omega = & -\mathbf{E}\left(\Delta I(Y < \infty) + \frac{\partial}{\partial F} \log K(\boldsymbol{\beta}_0, F_0)(F_0(Y) - I(Y \geq s)) \right) h \\
& + \mathbf{E}\left(\frac{\partial^2}{\partial F^2} \log K(\boldsymbol{\beta}_0, F_0) I(Y \geq s) \int_{[0,Y]} h dF_0 \right) \\
& + \mathbf{E}\left(\frac{\partial^2}{\partial F \partial \boldsymbol{\beta}} \log K(\boldsymbol{\beta}_0, F_0) I(Y \geq s) \right) \boldsymbol{\alpha}.
\end{aligned}
$$

Combining equation (A.86)-(A.91) and (A.95) and using notations (A.80), equation (A.84) can finally be written as

$$\begin{aligned} & \sqrt{n}\left(\Omega_\beta(\boldsymbol{\alpha},h)(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0)+\int_0^\infty \Omega_F(\boldsymbol{\alpha},h)d(\hat{F}_n-F_0)\right) \qquad (A.96)\\ = & -\sqrt{n}(\mathbf{E}_n-\mathbf{E})\left(\frac{\partial}{\partial\boldsymbol{\beta}}\log L(\boldsymbol{\beta}_0,F_0)\boldsymbol{\alpha}+\frac{\partial}{\partial F}\log L(\boldsymbol{\beta}_0,F_0)\int_{[0,Y]}hdF_0\right)\\ & +o_p(\sqrt{n}\|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\|+\sqrt{n}\|\hat{F}_n-F_0\|_{L^\infty})+o_p(1). \end{aligned}$$

■

In the next step, we want to prove the linear operator $(\boldsymbol{\alpha},h)\to(\Omega_\beta(\boldsymbol{\alpha},h),\Omega_F(\boldsymbol{\alpha},h))$ is invertible from $\mathbb{R}^d\times V_0$ to itself. If this is true, we can show the asymptotic normality of $(\hat{\boldsymbol{\beta}}_n,\hat{F}_n)$ by Theorem 3.3.1 in van der Vaart and Wellner [23]. To do so, we need to prove

(1). Ω can be decomposed as $\Omega=A+K$, where A is an invertible operator and K is a compact operator.

(2). The kernel set of Ω only contains zero, that is $Ker(\Omega)=\{0\}$.

Lemma 6. The linear operator $(\boldsymbol{\alpha},h)\to(\Omega_\beta(\boldsymbol{\alpha},h),\Omega_F(\boldsymbol{\alpha},h))$ is invertible from $\mathbb{R}^d\times V_0$ to itself.

Proof: Since the first component of Ω, $(\boldsymbol{\alpha},h)\to\Omega_\beta(\boldsymbol{\alpha},h)$, is a map to a finite-dimensional space $\mathbb{R}^d$, let us only focus on the second component for now, that is $(\boldsymbol{\alpha},h)\to\Omega_F(\boldsymbol{\alpha},h)$.

Recall the definition of $\Omega_F(\boldsymbol{\alpha},h)$ in (A.80)

$$\Omega_F(\boldsymbol{\alpha},h) \;=\; \omega-\int_0^\infty \omega dF_0, \qquad (A.97)$$

where

$$\begin{aligned} \omega \;=\; & -\mathbf{E}\left(\Delta I(Y<\infty)+\frac{\partial}{\partial F}\log K(\boldsymbol{\beta}_0,F_0)(F_0(Y)-I(Y\geq s))\right)h\\ & +\mathbf{E}\left(\frac{\partial^2}{\partial F^2}\log K(\boldsymbol{\beta}_0,F_0)I(Y\geq s)\int_{[0,Y]}hdF_0\right)\\ & +\mathbf{E}\left(\frac{\partial^2}{\partial F\partial\boldsymbol{\beta}}\log K(\boldsymbol{\beta}_0,F_0)I(Y\geq s)\right)\boldsymbol{\alpha}, \end{aligned}$$

and denote

$$g(s)=\mathbf{E}\left(\Delta I(Y<\infty)+\frac{\partial}{\partial F}\log K(\boldsymbol{\beta}_0,F_0)(F_0(Y)-I(Y\geq s))\right). \qquad (A.98)$$

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

$$\Omega_F(\boldsymbol{\alpha}, h) = \Omega_1(h) + \Omega_2(\boldsymbol{\alpha}, h), \tag{A.99}$$

where

$$\begin{aligned} \omega_1 &= gh, \\ \Omega_1(h) &= \omega_1 - \int_0^\infty \omega_1 dF_0 = -gh + \int_0^\infty ghdF_0, \\ \omega &= \omega_1 + \omega_2, \\ \Omega_2(\boldsymbol{\alpha}, h) &= \omega_2 - \int_0^\infty \omega_2 dF_0. \end{aligned} \tag{A.100}$$

We can show $\Omega_1(h)$ is an invertible operator from V_0 to V_0. The proof is done in three steps.

Firstly, the function $g(s)$ defined in (A.98) has bounded total variation. In fact, the total variation of $g(s)$, denoted by $\mathrm{Var}(g)$, is less then $\mathbf{E}\left(\Delta I(Y < \infty) \left| \frac{\partial}{\partial F} \log K(\boldsymbol{\beta}_0, F_0) \right| \right)$, because for any $0 \leq s_1 < s_2 < ... < s_n < \infty$ we have

$$\begin{aligned} \sum_{i=1}^n |g(s_i) - g(s_{i-1})| &\leq \sum_{i=1}^n \mathbf{E}\left(\Delta I(Y < \infty) \left| \frac{\partial}{\partial F} \log K(\boldsymbol{\beta}_0, F_0) \right| I(s_i \geq Y \geq s_{i-1}) \right) \\ &= \mathbf{E}\left(\Delta I(Y < \infty) \left| \frac{\partial}{\partial F} \log K(\boldsymbol{\beta}_0, F_0) \right| \right). \end{aligned} \tag{A.101}$$

Secondly, we will prove there exist a constant $a > 0$, such that $g(s) \geq a$ for any s. If this is true, then the function $\frac{1}{g(s)}$ also has bounded total variation. We consider an indicator function $I_\delta(Y)$, such that $I_\delta(Y) = 1$ when $Y \geq s$, $I_\delta(Y) = 0$ when $Y \leq s - \delta$, and $I_\delta(Y)$ is linear when $Y \in [s - \delta, s]$. Using the differential equation (A.65) and taking $H = F_0 - I_\delta$ and $\boldsymbol{\alpha} = 0$, we obtain

$$\mathbf{E}\left(\Delta I(Y < \infty) \frac{f_0 - I_\delta'}{f_0} + \frac{\partial}{\partial F} \log K(\boldsymbol{\beta}_0, F_0)(F_0(Y) - I_\delta(Y)) \right) = 0, \tag{A.102}$$

which is equivalent to

$$\begin{aligned} &\mathbf{E}\left(\Delta I(Y < \infty) + \frac{\partial}{\partial F} \log K(\boldsymbol{\beta}_0, F_0)(F_0(Y) - I_\delta(Y)) \right) \\ = \ &\frac{1}{\delta} \mathbf{E}\left(\Delta I(s - \delta \leq Y \leq s) \frac{1}{f_0} \right). \end{aligned} \tag{A.103}$$

When $\delta \to 0$, the left hand side of (A.103) will go to

$$\mathbf{E}\left(\Delta I(Y<\infty)+\frac{\partial}{\partial F}\log K(\boldsymbol{\beta}_0,F_0)(F_0(Y)-I(Y\geq s))\right), \tag{A.104}$$

which is the function $g(s)$ defined in (A.97). The right hand side of (A.103) equals

$$\begin{aligned}
& \frac{1}{\delta}\mathbf{E}\left(\Delta I(s-\delta\leq Y\leq s)\frac{1}{f_0}\right) \qquad (A.105)\\
= \;& \frac{1}{\delta}\mathbf{E}\left(I(T\leq C)I(s-\delta\leq T\leq s)\frac{1}{f_0(T)}\right)\\
= \;& \frac{1}{\delta}\int_{s-\delta}^{s}E_X\left(S_C(T|X)\frac{1}{f_0(T)}(-G'(\eta(\boldsymbol{\beta}_0,\boldsymbol{X})F_0(T))\eta(\boldsymbol{\beta}_0,\boldsymbol{X})f_0(T))\right)dT\\
\to \;& -E_X\left(S_C(s|X)G'(\eta(\boldsymbol{\beta}_0,\boldsymbol{X})F_0(T))\eta(\boldsymbol{\beta}_0,\boldsymbol{X})\right).
\end{aligned}$$

Noticing that $G'(\eta(\boldsymbol{\beta}_0,\boldsymbol{X})F_0(T))<0$ and $S_C(s|X)\geq S_C(\infty|X)>0$, therefore there exist a constant $a>0$ such that

$$-E_X\left(S_C(s|X)G'(\eta(\boldsymbol{\beta}_0,\boldsymbol{X})F_0(T))\eta(\boldsymbol{\beta}_0,\boldsymbol{X})\right)\geq a>0 \tag{A.106}$$

because of conditions (C1) and (C2). Combining equations (A.103)-(A.106), we proved there exist a constant $a>0$ such that $g(s)\geq a$ for any s.

Finally, let us calculate the inverse of $\Omega_1(h)$ in V_0. By the definition of $\Omega_1(h)$, we have

$$k=\Omega_1(h)=-gh+\int_0^\infty ghdF_0. \tag{A.107}$$

We want to solve h from (A.107) in terms of k. Dividing g and integrating on both sides of (A.107), we obtain

$$\begin{aligned}
\int_0^\infty \frac{1}{g}kdF_0 &= -\int_0^\infty hdF_0+\int_0^\infty \frac{1}{g}dF_0\int_0^\infty ghdF_0 \qquad (A.108)\\
&= \int_0^\infty \frac{1}{g}dF_0\int_0^\infty ghdF_0.
\end{aligned}$$

Combining (A.107) and (A.108), we have

$$\begin{aligned}
h &= -\frac{1}{g}k+\frac{1}{g}\int_0^\infty ghdF_0 \qquad (A.109)\\
&= -\frac{1}{g}k+\frac{\frac{1}{g}\int_0^\infty \frac{1}{g}kdF_0}{\int_0^\infty \frac{1}{g}dF_0},
\end{aligned}$$

which is a bounded linear operator from V_0 to V_0. Therefore, we proved $\Omega_1(h)$ has bounded inverse and $\Omega_1(h)$ is invertible from V_0 to itself.

Now, let us rewrite Ω as the following

$$\Omega = (\Omega_\beta, \Omega_F) = (I_d, \Omega_1) + (\Omega_\beta - I_d, \Omega_2), \tag{A.110}$$

where I_d is the identity map from $\mathbb{R}^d$ to $\mathbb{R}^d$. Recall that we want to prove Ω can be decomposed as $\Omega = A + K$, where A is an invertible operator and K is a compact operator. Let $A = (I_d, \Omega_1)$. Because I_d is invertible in $\mathbb{R}^d$ and Ω_1 is invertible in V_0, A is an invertible operator from $(\mathbb{R}^d, V_0)$ to itself. Now, let $K = (\Omega_\beta - I_d, \Omega_2)$ and we want to show K is a compact operator. To do so, we only need to prove Ω_2 is a compact operator.

Recall the definition of $\Omega_2(\boldsymbol{\alpha}, h)$ in (A.99),

$$\begin{aligned} \Omega_2(\boldsymbol{\alpha}, h) &= \omega_2 - \int_0^\infty \omega_2 dF_0, \\ \omega_2(\boldsymbol{\alpha}, h) &= \mathbf{E}\left(\frac{\partial^2}{\partial F^2}\log K(\boldsymbol{\beta}_0, F_0) I(Y \geq s) \int_{[0,Y]} h dF_0\right) \\ &\quad + \mathbf{E}\left(\frac{\partial^2}{\partial F \partial \boldsymbol{\beta}}\log K(\boldsymbol{\beta}_0, F_0) I(Y \geq s)\right)\boldsymbol{\alpha}. \end{aligned} \tag{A.111}$$

It can be shown that the map $h \mapsto p(s) = \int_0^s h dF_0$, where $s \in [0, \infty)$, is compact from V_0 to $C[0, \infty)$, the set of continuous functions on $[0, \infty)$. Therefore, for $h_n \in V_0, \|h_n\|_V \leq 1$ by choosing a subsequence still indexed by $\{n\}$, we have

$$p_n(s) = \int_0^s h_n dF_0 \to H_0(s), \tag{A.112}$$

where $H_0(s) \in C[0, \infty)$. By choosing a further subsequence, for $\boldsymbol{\alpha}_n \in \mathbb{R}^d, \|\boldsymbol{\alpha}_n\| \leq 1$, we assume $\boldsymbol{\alpha}_n \to \boldsymbol{\alpha}_0$, where $\boldsymbol{\alpha}_0 \in \mathbb{R}^d$. For the chosen subsequence, we can prove in V_0

$$\begin{aligned} \omega_2(\boldsymbol{\alpha}_n, h_n) \to \quad & \mathbf{E}\left(\frac{\partial^2}{\partial F^2}\log K(\boldsymbol{\beta}_0, F_0) I(Y \geq s) H_0(Y)\right) \\ & + \mathbf{E}\left(\frac{\partial^2}{\partial F \partial \boldsymbol{\beta}}\log K(\boldsymbol{\beta}_0, F_0) I(Y \geq s)\right)\boldsymbol{\alpha}_0. \end{aligned} \tag{A.113}$$

Denote the right hand side of (A.113) by ω_0. The total variation of $\omega_2(\boldsymbol{\alpha}_n, h_n) - \omega_0$ equals

$$
\begin{aligned}
& \operatorname{Var}\left(\omega_2(\boldsymbol{\alpha}_n, h_n) - \omega_0\right) \qquad (A.114)\\
\leq\ & \operatorname{Var}\left(\mathbf{E}\left(\frac{\partial^2}{\partial F^2}\log K(\boldsymbol{\beta}_0, F_0) I(Y \geq s)\left(\int_{[0,Y]} h_n dF_0 - H_0(Y)\right)\right)\right)\\
& + \|\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_0\| \operatorname{Var}\left(\mathbf{E}\left(\frac{\partial^2}{\partial F \partial \boldsymbol{\beta}}\log K(\boldsymbol{\beta}_0, F_0) I(Y \geq s)\right)\right)\\
\leq\ & \sup_{Y \in [0,\infty)} \left|\int_{[0,Y]} h_n dF_0 - H_0(Y)\right| \mathbf{E}\left(\left|\frac{\partial^2}{\partial F^2}\log K(\boldsymbol{\beta}_0, F_0)\right|\right)\\
& + \|\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_0\| \mathbf{E}\left(\frac{\partial^2}{\partial F \partial \boldsymbol{\beta}}\log K(\boldsymbol{\beta}_0, F_0)\right).
\end{aligned}
$$

When when $n \to \infty$, the right hand side of inequality (A.114) will go to zero. Therefore, we proved $\Omega_2(\boldsymbol{\alpha}, h)$ is a compact operator because of (A.113). Then immediately, $K = (\Omega_\beta - I_d, \Omega_2)$ is also an compact operator.

So far we have shown that Ω can be decomposed as $\Omega = A + K$, where A is an invertible operator and K is a compact operator. To prove $\Omega = (\Omega_\beta, \Omega_F)$ is invertible, we still need to show the kernel set of Ω only contains zero, that is $Ker(\Omega) = \{0\}$. Suppose we have

$$
\Omega_\beta(\boldsymbol{\alpha}, h)\boldsymbol{\alpha} + \int_0^\infty \Omega_F(\boldsymbol{\alpha}, h) h dF_0 = 0, \qquad (A.115)
$$

we want to prove that $\boldsymbol{\alpha} = 0$ and $h = 0$. The left hand side of equation (A.115) equals

$$
\begin{aligned}
& \Omega_\beta(\boldsymbol{\alpha}, h)\boldsymbol{\alpha} + \int_0^\infty \Omega_F(\boldsymbol{\alpha}, h) h dF_0 \qquad (A.116)\\
= \; & \mathbf{E}\left(\frac{\partial^2}{\partial F \partial \boldsymbol{\beta}} \log K(\boldsymbol{\beta}_0, F_0) \int_{[0,Y]} h dF_0\right)\boldsymbol{\alpha} + \mathbf{E}\left(\frac{\partial^2}{\partial \boldsymbol{\beta}^2} \log K(\boldsymbol{\beta}_0, F_0)\right)(\boldsymbol{\alpha}, \boldsymbol{\alpha})\\
& - \int_0^\infty \mathbf{E}\left(\Delta I(Y < \infty) + \frac{\partial}{\partial F} \log K(\boldsymbol{\beta}_0, F_0)(F_0(Y) - I(Y \geq s))h^2(s)\right) dF_0\\
& + \int_0^\infty \mathbf{E}\left(\frac{\partial^2}{\partial F^2} \log K(\boldsymbol{\beta}_0, F_0) I(Y \geq s) \int_{[0,Y]} h dF_0\right) h(s) dF_0\\
& + \int_0^\infty \mathbf{E}\left(\frac{\partial^2}{\partial F \partial \boldsymbol{\beta}} \log K(\boldsymbol{\beta}_0, F_0) I(Y \geq s)\right)\boldsymbol{\alpha} h(s) dF_0\\
= \; & 2\mathbf{E}\left(\frac{\partial^2}{\partial F \partial \boldsymbol{\beta}} \log K(\boldsymbol{\beta}_0, F_0) \int_{[0,Y]} h dF_0\right)\boldsymbol{\alpha} + \mathbf{E}\left(\frac{\partial^2}{\partial \boldsymbol{\beta}^2} \log K(\boldsymbol{\beta}_0, F_0)\right)(\boldsymbol{\alpha}, \boldsymbol{\alpha})\\
& - \mathbf{E}\left(\Delta I(Y < \infty) h^2(Y)\right) + \mathbf{E}\left(\frac{\partial^2}{\partial F^2} \log K(\boldsymbol{\beta}_0, F_0)(\int_{[0,Y]} h dF_0)^2\right)\\
= \; & \frac{d^2}{dt^2}\mathbf{E}\left(\log L(\boldsymbol{\beta}_0 + t\boldsymbol{\alpha}, F_0 + t\int_{[0,Y]} h dF_0)\right)|_{t=0}\\
= \; & -\mathbf{E}\left(\frac{d}{dt} \log L(\boldsymbol{\beta}_0 + t\boldsymbol{\alpha}, F_0 + t\int_{[0,Y]} h dF_0)|_{t=0}\right)^2\\
= \; & -\mathbf{E}\left(\frac{\partial}{\partial \boldsymbol{\beta}} \log(\boldsymbol{\beta}_0, F_0)\boldsymbol{\alpha} + \frac{\partial}{\partial F} \log(\boldsymbol{\beta}_0, F_0) \int_{[0,Y]} h dF_0\right)^2.
\end{aligned}
$$

Combining (A.115) and (A.116), we obtain

$$
\mathbf{E}\left(\frac{\partial}{\partial \boldsymbol{\beta}} \log(\boldsymbol{\beta}_0, F_0)\boldsymbol{\alpha} + \frac{\partial}{\partial F} \log(\boldsymbol{\beta}_0, F_0) \int_{[0,Y]} h dF_0\right)^2 = 0. \qquad (A.117)
$$

Therefore, with probability 1 we have

$$
\frac{\partial}{\partial \boldsymbol{\beta}} \log(\boldsymbol{\beta}_0, F_0)\boldsymbol{\alpha} + \frac{\partial}{\partial F} \log(\boldsymbol{\beta}_0, F_0) \int_{[0,Y]} h dF_0 = 0. \qquad (A.118)
$$

Taking $Y = \infty$ in (A.118), because $h \in V_0$ we obtain

$$
\begin{aligned}
& \frac{\partial}{\partial \boldsymbol{\beta}} \log(\boldsymbol{\beta}_0, F_0)\boldsymbol{\alpha} + \frac{\partial}{\partial F} \log(\boldsymbol{\beta}_0, F_0) \int_{[0,Y]} h dF_0 \qquad (A.119)\\
= \; & \frac{\partial}{\partial \boldsymbol{\beta}} \log(\boldsymbol{\beta}_0, F_0)\boldsymbol{\alpha}\\
= \; & \frac{G'(\eta(\boldsymbol{\beta}_0, \boldsymbol{X}))}{G(\eta(\boldsymbol{\beta}_0, \boldsymbol{X}))}\eta'(\boldsymbol{\beta}_0, \boldsymbol{X}) \cdot \boldsymbol{X} \cdot \boldsymbol{\alpha}\\
= \; & 0.
\end{aligned}
$$

Equation (A.119) holds for any $\boldsymbol{X}$ almost surely, hence $\boldsymbol{\alpha} = 0$. Then taking $Y < \infty$ and $\Delta = 0$ in (A.118), we have

$$\begin{aligned} & \frac{\partial}{\partial \boldsymbol{\beta}} \log(\boldsymbol{\beta}_0, F_0)\boldsymbol{\alpha} + \frac{\partial}{\partial F} \log(\boldsymbol{\beta}_0, F_0) \int_{[0,Y]} h dF_0 \qquad (A.120) \\ = & \frac{\partial}{\partial F} \log(\boldsymbol{\beta}_0, F_0) \int_{[0,Y]} h dF_0 \\ = & \frac{G'(\eta(\boldsymbol{\beta}_0, \boldsymbol{X})F_0(Y))}{G(\eta(\boldsymbol{\beta}_0, \boldsymbol{X})F_0(Y))} \eta(\boldsymbol{\beta}_0, \boldsymbol{X}) \int_{[0,Y]} h dF_0 \\ = & 0, \end{aligned}$$

Equation (A.120) implies $\int_{[0,Y]} h(s) dF_s = 0$ for any $Y < \infty$ and $\Delta = 0$. Therefore, $h = 0$ almost surely. Back to the definition of $\Omega_F(\boldsymbol{\alpha}, h)$ in (A.80), we have $\Omega_F(\boldsymbol{\alpha}, h) = -gh$, and therefore $h \equiv 0$ in V_0. ■

Now we have proved $(\boldsymbol{\alpha}, h) \to (\Omega_\beta(\boldsymbol{\alpha}, h), \Omega_F(\boldsymbol{\alpha}, h))$ is invertible from $\mathbb{R}^d \times V_0$ to itself. Then by using Theorem 3.3.1 in van der Vaart and Wellner [23], we will obtain the following Theorem 2.

Theorem 2. Under condition (C1)-(C6), $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0, \hat{F}_n - F_0)$ converges weakly to a Gaussian process in $l^\infty(\mathbb{R}^d \times V_0)$.

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.96) can be written as

$$\begin{aligned} & \sqrt{n}\left(\boldsymbol{\alpha}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + \int_0^\infty h d(\hat{F}_n - F_0)\right) \qquad (A.121) \\ = & -\sqrt{n}(\mathbf{P}_n - \mathbf{P})\left(\frac{\partial}{\partial \boldsymbol{\beta}} \log L(\boldsymbol{\beta}_0, F_0)\tilde{\Omega}_\beta(\boldsymbol{\alpha}, h) + \frac{\partial}{\partial F} \log L(\boldsymbol{\beta}_0, F_0) \int_{[0,Y]} \tilde{\Omega}_F(\boldsymbol{\alpha}, h) dF_0\right) \\ & + o_p(\sqrt{n}\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| + \sqrt{n}\|\hat{F}_n - F_0\|_{L^\infty}) + o_p(1). \end{aligned}$$

Immediately from (A.96) and (A.121), we have

$$\sqrt{n}(\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| + \|\hat{F}_n - F_0\|) = o_p(1). \qquad (A.122)$$

Back to (A.121), we obtain

$$\begin{aligned} & \sqrt{n}\left(\boldsymbol{\alpha}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + \int_0^\infty h d(\hat{F}_n - F_0)\right) \qquad (A.123) \\ = & -\sqrt{n}(\mathbf{P}_n - \mathbf{P})\left(\frac{\partial}{\partial \boldsymbol{\beta}} \log L(\boldsymbol{\beta}_0, F_0)\tilde{\Omega}_\beta(\boldsymbol{\alpha}, h) + \frac{\partial}{\partial F} \log L(\boldsymbol{\beta}_0, F_0) \int_{[0,Y]} \tilde{\Omega}_F(\boldsymbol{\alpha}, h) dF_0\right) \\ & + o_p(1). \end{aligned}$$

Equation (A.123) holds uniformly for any $\|\boldsymbol{\alpha}\| \leq 1$ and $\|h\|_V \leq 1$. By using Theorem 3.3.1 in van der Vaart and Wellner [23], $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0, \hat{F}_n - F_0)$ converges weakly to a Gaussian process in $l^\infty(\mathbb{R}^d \times V_0)$. ■

REFERENCES

[1] J. Berkson and R. P. Cage. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47(259):501–515, 1952. 1, 2.1.1

[2] A. Y. Yakovlev and A. D. Tsodikov. *Stochastic Models of Tumor Latency and Their Biostatistical Application*. World Scientific, New Jersey, 1996. 1, 2.1.2

[3] M.-H. Chen, J. G. Ibrahim, and D. Sinha. A new bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, 94(447):909–919, 1999. 1, 2.1.2, 2.1.2, 2.1.2

[4] J. G. Ibrahim, M.-H. Chen, and D. Sinha. *Bayesian Survival Analysis*. Springer, New York, 2001. 1, 2.1.1

[5] D. Zeng, G. Yin, and J. G. Ibrahim. Semiparametric transformation models for survival data with a cure fraction. *Journal of the American Statistical Association*, 101(474):670–684, 2006. 1, 2.1.2, 2.1.2, 2.1.2, 2.1.2, 2.1.2, 2.2, 2.3, 2.4, 2.5, 4.1, 4.1.2, 5, A, A.1, A.1

[6] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972. 1, 3.3

[7] V. T. Farewell. The use of mixtures models for the analysis of survival data with long-term survivors. *Biometrics*, 38(4):1041–1046, 1982. 2.1.1

[8] R. J. Gray and A. A. Tsiatis. A linear rank test for use when the main interest is in differences in cure rates. *Biometrics*, 45(3):899–904, 1989. 2.1.1

[9] R. Sposto, H. N. Sather, and S. A. Baker. A comparison of tests of the difference in the proportion of patients who are cured. *Biometrics*, 48(1):87–99, 1992. 2.1.1

[10] E. M. Laska and M. J. Meisner. Nonparametric estimation and testing in a cure model. *Biometrics*, 48(4):1223–1234, 1992. 2.1.1

[11] J. P. Sy and J. M. G. Taylor. Estimation in a cox proportional hazards cure model. *Biometrics*, 56(1):227–236, 2000. 2.1.1

[12] W. Lu and Z. Ying. On semiparametric transformation cure models. *Biometrika*, 91(2):331–343, 2004. 2.1.1

[13] N. Mantel and W. Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cance Institute*, 22:719–748, 1959. 2.1.1

[14] A. Tsodikov. A proportional hazards model taking account of long-term survivors. *Biometrics*, 54(4):1508–1516, 1998. 2.1.2

[15] C. J. Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, 13(2):689–705, 1985. 2.2

[16] P. Royston and D. G. Altman. Regression using fractional polynomials of continuous covariates: Parsimonious parametric modeling. *Applied Statistics*, 43(3):429–467, 1994. 2.2, 2.2, 2.2, 3.1, 4.1.3, 5

[17] J. M. Kirkwood, J. G. Ibrahim, V. K. Sondak, J. Richards, L. E. Flaherty, M. S. Ernstoff, T. J. Smith, U. Rao, M. Steele, and R. H. Blum. High- and low-dose interferon alfa-2b in high-risk melanoma: First analysis of intergroup trial e1690/s9111/c9190. *Journal of Clinical Oncology*, 18:2444–2458, 2000. 4.1

[18] M. A. Cleves, W. W. Gould, and R.G. Gutierrez. *In Introduction to Survival Analysis*. Stata Corportation, 2004. 4.1.1

[19] G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3, 1950. 4.1.4

[20] M. May, P. Roystion, M. Egger, A. C. Justice, and J. A. Sterne. Development and validation of a prognostic model for survival time data: Application to prognosis of hiv positve patients treated with antiretroviral therapy. *Statistics in Medicine*, 23:2375–2398, 2004. 4.1.4, 4.1.4, 5

[21] The diverse population collaboration. *http://www.biostat.stat.fsu.edu/Diverse.htm*. 4.2

[22] Plan and operation of the nhanes i epidemiologic follow-up study, 1987. *Vital and Health Statistics. Series 1*, 27, 1992. 4.2

[23] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, 1996. A, A.1, A.1, A.1, A.2, A.2, A.2, A.2, A.2, A.2

[24] G. R. Shorack. *Probability for Statisticians*. Springer-Verlag, 2000. A.1

BIOGRAPHICAL SKETCH

Yang Liu

Yang Liu was born in Heilongjiang Province in 1979 and grew up in Beijing, China. In the summer of 2002 she completed her Bachelor of Science degree in Mathematics at University of Science and Technology of China. In the fall of 2002 she was admitted to the University of Tennessee, Knoxville, and obtained her Master of Science degree in Mathematics in 2004. She enrolled in the Florida State University in the fall of 2004 and completed her Master of Science degree in Statistics in 2006. Her doctoral program started at the Florida State University in 2006, and she defended her dissertation in the spring of 2009.

She is married to Xiaohu Tang in January 2009 in Sichuan Province, China. She will move to New Jersey and start a new job there.