5-2017

Bayesian Flexible Modeling of Interval-censored Failure Time Data

Sheng-Yang Wang
University of South Carolina

Bayesian Flexible Modeling of Interval-censored Failure Time Data

by

Sheng-Yang Wang

Bachelor of Science
TamKang University, 1995

Master of Science
The University of New Mexico, 2009

Master of Science
The University of New Mexico, 2011

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Statistics

College of Arts and Sciences

University of South Carolina

2017

Accepted by:

Lianming Wang, Major Professor

John Grego, Committee Member

David Hitchcock, Committee Member

Yi Sun, Committee Member

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

Acknowledgments

I would like to thank Dr. Lianming Wang sincerely. I appreciate for his encouragement, training, and advice for the necessary courses and knowledge required for this research.

I give special thanks to Dr. Grego for his incisive comments on my dissertation context. In addition, he helped me develop my statistical consulting experience in my career. I also thank to my committee members, Dr. Hitchcock and Dr. Sun for their insightful comments and suggestions in supporting my dissertation.

At last, not the least, I dedicate this work to my beloved wife, Esther, who has shown so much confidence in me and has given me great encouragement with her unconditional love and patient. Without her accompanying in the journey, I cannot move the important step to complete my Ph.D. In addition, I am proud of my lovely and cute children, Emunah and Emeth for their mature mindsets above their own age to fully support me over the past few years all the time.

As a believer in Jesus Christ, I thank my Lord and Savior, Jesus Christ, for His work on me through my graduate study and research. As it is said in the Bible, "Although the Lord gives you the bread of adversity and the water of affliction, your teacher will be hidden no more; with your own eyes you will see them. Whether you turn to the right or to the left, your eyes will hear a voice behind you, saying, 'This is the way; walk in it.'"

Abstract

Interval-censored data are a special type of survival data, in which the survival time is not accurately observed but known to fall within a specific time interval. Interval-censored data commonly arise in real-life epidemiological and medical studies that involve periodic examinations. In this dissertation, several semiparametric regression models are investigated to provide flexible modeling and robust inference for interval-censored data from Bayesian perspectives.

Chapter 1 provides a detailed description about interval-censored data and gives several examples. Existing models and methods for analyzing such interval-censored data are reviewed as well. Chapter 2 develops a unified Bayesian estimation approach under the framework of semiparametric linear transformation models for regression analysis of current status data, which is a special type of interval-censored data. This work provides an alternative estimation approach to the existing methods for the proportional hazards, proportional odds, and probit models. As a unified Bayesian estimation approach, the proposed method allows direct comparison of three different semiparametric regression models in the same framework of the Gibbs Sampler. Chapter 3 proposes a Bayesian estimation approach for analyzing general interval-censored data under the generalized odds-rate hazards (GORH) models. The GORH models are a general class of semiparametric regression models including the proportional hazards and proportional odds models as special cases. Submodels of GORH models can be specified by indexing a non-negative value ν, where the "sub" prefix refers to the fact that for each ν, a semiparametric regression model is well-specified for regression analysis of general interval-censored data. It is found that treating ν as

an unknown parameter leads to biased estimation, which in this case is a consistent research result for right-censored data in the literature. To solve this issue, a Bayesian approach with a known ν is proposed and has shown excellent performance in the simulation study. Chapter 4 extends the semiparametric probit model for regression analysis of arbitrarily censored data. The proposed method has been implemented using two sets of latent variables for posterior computation. The proposed method can be easy to implement in the estimation of regression parameters for two special types of arbitrarily censored data: right-censored data and general interval-censored data.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Interval-Censored Data

Observations in time-to-event data are subject to censoring when practitioners cannot know or exactly observe the occurrence of an event. The failure time is usually defined as the length of time until the occurrence of an event. Many clinicians and epidemiologists have designed and conducted their experiments in prospective cohort studies or longitudinal studies. In such studies, participants are often seen at pre-scheduled visits or regular check-ups or random examinations but the event of interest (e.g. failure) may occur in between visits. Interval-censored data naturally arise when each subject is observed at only one time or inspected at a one-time sacrifice. For example, in animal carcinogenicity studies, the onset of tumor cannot be known, but the presence of tumors can be diagnosed through a biopsy or some laboratory test in regular check-ups. Then the onset of tumor can be known to lie in a specific time interval. In the literature, this type of censored data is so-called current status data (a.k.a. case 1 interval-censored data) in time-to-event history data (Sun, 2007). Interval-censored data naturally arise when each subject proceeds a periodic check-ups or random examinations, for example, animal carcinogenicity studies and longitudinal studies. In the literature, this type of censored data is so-called interval-censored data (a.k.a. case 2 interval-censored data) in time-to-event history data (Groeneboom and Wellner, 1992; Huang and Wellner, 1997; Kalbfleisch and Prentice, 2002; Sun, 2007). This time interval is formed from a sequence of periodic pre-scheduled observation

time points. The failure can be known to occur before the first observation time, between two adjacent observation times, or after the last observation time. For example, the onset time of HIV for a participant is usually interval censored and the observed interval is formed by the last examination time with negative status and the first examination time with positive status for that participant. The structure of interval-censored data accommodates the incidence of drop-outs. For example, drop-outs occur when participants drop out or die before the end of the study or miss several check-ups in the study, as in the breast cosmesis dataset (Finkelstein, 1986). Current status data are a special case of interval-censored data. Current status data are yielded when a periodic scientific investigation is limited to one random examination for a diagnosis of whether or not the failure has been revealed. If failure is revealed, the observation is left-censored; otherwise, right-censored. Current status data are relatively more cost-effective and less time-consuming. The costs of studies are often reduced by alleviating the frequency of observation times.

1.2 Motivating Examples

1.2.1 Uterine Leiomyomas Data

Right From the Start (RFTS) is an on-going, prospective cohort study of early-pregnancy health that was conducted in three states (NC/TX/TN). In this cohort study, the onset time of uterine fibroids was unknown. However, the onset time of uterine fibroids was known to occur either before or after a one-time ultrasound examination that was performed at the exact enrollment time of each participant, in early pregnancy (prior to the seventh gestational week). Participants were diagnosed as positive when the ultrasound examination revealed leiomyomata diameter of 0.5 cm or larger. A more detailed description of this study is available from (Laughlin et al., 2009; Wang and Dunson, 2011).

1.2.2 Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial Data

The Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial sponsored by the U.S. National Cancer Institute (NCI) is a multicenter and randomized two-arm trial designed to assess the effect of a regular screening strategy for cancer-related mortality. This program was initiated in November 1993 and ended in September 2001. Our analysis takes account of the prostate cancer screening data collected on male participants in the intervention arm. The response variable of this study is the time to onset of prostate cancer. The onset time of prostate cancer is not observed, but is known to lie in two adjacent screenings because of the design of the study and the diagnosis mechanism of prostate cancer. Two adjacent screenings form one screening interval, and practitioners identify whether the onset of a tumor exists or not through laboratory testing. Thus, the censoring information of each subject can be obtained from a state of having a negative result in the previous test and a state of having a positive result at the current one. Two other cases could be: Participants have been diagnosed as positive at the enrollment time, or they could have negative results in all screenings in the program. Thus, the PLCO data are the interval-censored data. The primary goal of the present study is to estimate the association of risk factors from the intervention arm with the onset of prostate cancer. For a more detailed description of this screening trial, please see Andriole et al. (2012).

1.2.3 Diabetic Nephropathy

Diabetes mellitus is a common chronic disease caused due to a disturbance of normal production of insulin. There are two types of diabetes: Type 1 diabetes (a.k.a. insulin-dependent or juvenile-onset) and Type 2 diabetes(a.k.a. non-insulin-dependent or adult-onset). Type 1 diabetes is the most severe type and occurs primarily in pa-

tients at a young age, and Type 2 diabetes is a mild type and develops in those patients later in life. Nowadays, there are many obese children in the adolescent population; therefore, there are many young people with insulin resistance and Type 2 diabetes. Moreover, the characteristics of both Type 1 and Type 2 diabetes may be present in the same patient. Type 1 and Type 2 diabetes are difficult to distinguish.

The Steno Memorial Hospital in Copenhagen, Denmark served as a diabetes research hospital beginning in 1933. A study was conducted there between 1933 and 1972 of patients who had been diagnosed before age 31 with insulin-dependent diabetes mellitus for Type 1 diabetes (Andersen et al., 1983). Borch-Johnsen et al. (1985)'s research showed that the development of diabetic nephropathy (DN) was regarded as a highly associated prognostic biomarker for a low survival rate in Type 1 diabetics. Insulin treatment was adopted as the primary method to treat diabetes disease in 1922, but it is still possible for a patient treated with insulin to develop DN for Type 1 diabetes. DN is defined as persistent proteinuria and not an irreversible complication. It is mainly used to assess kidney failure, which is indicated as positive whenever a subject has at least four urine samples within 24 hours, during a time interval of at least one month, that each contain more than 0.5-gram of protein. The survival time is used as the basic time scale from onset of a patient's diabetes to the time when they transition from having diabetes without DN to having diabetes with DN. Subjects are diabetic patients who either enter this study with DN or develop DN before the end of the study. The primary research interest is to estimate the association of risk factors (e.g. gender and age at the onset of diabetes) with the onset of the development of DN.

1.3 Literature Review

The primary research interests focus on the estimation of regression parameters and the survival curves.

1.3.1 Survival Curves

Without covariates, many existing approaches have been developed and applied to estimate the survival curve for interval-censored data. For example, Peto (1973) proposed the Newton-Raphson method to estimate the (experimental) survival curve. Turnbull (1976) presented the self-consistent estimation algorithm, and Groeneboom and Wellner (1992) introduced the iterative convex minorant (ICM) algorithm to compute the nonparametric maximum likelihood estimator (NPMLE) for the distribution of failure time. Wellner and Zhan (1997) developed a hybrid algorithm to facilitate the expectation-maximization (EM) algorithm and the ICM algorithm to attain global convergence. Groeneboom and Wellner (1992)'s empirical studies showed that using the ICM algorithm to estimate NPMLE converges faster than Turnbull's algorithm. Turnbull's self-consistency algorithm is simple to implement, though. Wellner and Zhan (1997)'s simulation studies reported that their algorithm converges to the NPMLE faster than both the EM and ICM algorithms, with fewer iterations and less computation time required for current status data. Moreover, Wellner and Zhan (1997) emphasized that self-consistency equations do not determine the uniqueness of the NPMLE for interval-censored data. Finally, Gentleman and Geyer (1994) proved that the Karush-Kuhn-Tucker conditions are necessary and sufficient for optimization for a self-consistency procedure to apply standard convex optimization techniques.

1.3.2 Regression Analysis

Almost at the same time as researchers were estimating the survival curve, many approaches were developed for regression analysis of interval-censored data under semiparametric survival models. For example, the proportional hazards (PH) model (Cox, 1972, 1975), the proportional odds (PO) model (Bennett, 1983), and the probit (PB) model (Lin and Wang, 2010) were developed.

In the PH model, Andersen and Gill (1982) exploited counting process and martingale theory to provide an elegant proof for asymptotic normality of a consistent maximum likelihood estimator for the right-censored data. Finkelstein (1986), however, commented that the martingale techniques cannot be adapted to current status data because of the difficulty in defining an appropriately increasing sequence of sigma-algebras. Many existing approaches have been developed to estimate the regression coefficients. Huang (1996) demonstrated that the MLE of the finite dimensional regression parameter in the PH model is asymptotically efficient but the infinite-dimensional parameter converges slower than $\sqrt{n}$ for current status data. Pan (1999) extended the iterative convex minorant algorithm, which was developed by Groeneboom and Wellner (1992), to consider covariate effects in the PH model, and this method is called the generalized gradient projection (GGP) method. Spline-based methods have prevailed since the early 1990s for current status data analysis. Kooperberg and Clarkson (1997) introduced hazard regression, and Kooperberg et al. (1995) used linear splines and their tensor products in the estimation of the conditional log-hazard function for interval-censored data. Shiboski (1998) fitted the generalized additive model (GAM) and isotonic regression, and provided simultaneous estimation of regression coefficients and the baseline event time distribution. Cai and Betensky (2003) proposed a flexible locally parametric procedure to model the baseline log-hazard function and to obtain MLE via Penalized Spline for interval-censored data. Wang et al. (2015) developed a novel EM algorithm for regression analysis of bivariate case 1 interval-censored data under the Gamma-frailty PH model. Wang et al. (2016) developed a novel EM algorithm under the PH model. Compared to frequentist approaches, Bayesian methods are few. For example, Cai et al. (2011) proposed a Bayesian approach by using monotone splines.

Frequentist approaches for the PO model have also been introduced. Rossini and Tsiatis (1996) proposed a uniformly spaced step function method for approximating

the baseline function and the number of jumps in their methods, which are predetermined by a Lipschitz-continuity assumption. Huang and Rossini (1997) and Shen (1998) developed a random sieve likelihood method on both the baseline function and the regression coefficients. Chen et al. (2007) developed a maximum likelihood approach to fit the marginal PO model for multivariate interval-censored data.

The accelerated failure time (AFT) model presumes that the logarithm of the failure time is linearly related to the covariates, but also needs to take account of an unspecified random error (Cox and Oakes, 1984; Kalbfleisch and Prentice, 1980). The AFT model has an explicit interpretation and would be a useful alternative to the PH model in survival analysis. Rabinowitz et al. (1995) proposed an adaptive procedure based on score statistics for estimating the regression coefficients. Betensky et al. (2001) suggested an estimating equation approach, but it does not involve the NPMLE of the distribution at the residuals. Compared to Rabinowitz et al. (1995)'s approach, Betensky et al. (2001)'s approach is the simple and practical alternative to the computationally demanding procedure.

Lin and Wang (2010) proposed a semiparametric probit model from a Bayesian perspective. Their method derived a data augmentation approach based on normal latent variables and resulted in a very nice conjugate normal prior for general interval-censored data. In my dissertation, I have reviewed some newly existing methods for regression analysis of (arbitrarily) interval-censored failure time data. The proposed methods have sound theoretical justification and can be implemented with an efficient Bayesian sampling-based approach.

1.4 Preliminaries

1.4.1 The Proportional Hazards Model

The proportioanl hazards model is also known as the Cox model (Cox, 1972, 1975). The failure time random variable is denoted as T, and its survival function is denoted

as $S(t)$ with density function $f(t)$. The hazard function of T is defined as

$$\begin{aligned}\lambda(t) &= \lim_{\triangle t \to 0} \frac{P(T \leq t + \triangle t \mid T > t)}{\triangle t} \\ &= \frac{f(t)}{S(t)}.\end{aligned}$$

Now consider a semiparametric regression model. Let $\boldsymbol{x}_i = (x_1, \cdots, x_p)^{'}$ be a p-dimensional covariate vector for the i-th subject, and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^{'}$ is the corresponding vector of regression parameter. Given the covariate $\boldsymbol{x}$, the hazard function of T is represented as a proportional hazards model,

$$\begin{aligned}\lambda(t \mid \boldsymbol{x}) &= \lambda_0(t) \exp\Big\{ \sum_{j=1}^{n} x_j \beta_j \Big\} \\ &= \lambda_0(t) \exp\{\boldsymbol{x}^{'} \boldsymbol{\beta}\},\end{aligned}$$

where $\lambda_0(t)$ is a so-called baseline hazard function, which is usually unknown. A baseline hazard function could be parametrically specified (e.g. the Weibull class of hazard functions). It is much common to use nonparametric form and thus leads to a semiparametric regression model. The baseline hazard function is a hazard function when covariates are all taken to be zero. Taking $x_j = 1$ for the treatment group with the other $(p-1)$'s covariates fixed, the hazard function of T is expressed in the form,

$$\lambda(t \mid x_j = 1, \boldsymbol{x}_{(-j)}) = \lambda_0(t) \exp\Big\{ \beta_j + \sum_{k \neq j} x_k \beta_k \Big\}.$$

Similarly, taking $x_j = 0$ for the placebo group with the other $(p-1)$'s covariates fixed, the baseline of the hazard function of T is expressed in the form,

$$\lambda(t \mid x_j = 0, \boldsymbol{x}_{(-j)}) = \lambda_0(t) \exp\Big\{ \sum_{k \neq j} x_k \beta_k \Big\}.$$

For every t, the regression coefficient β_j satisfies the identity:

$$\exp(\beta_j) = \frac{\lambda(t \mid x_j = 1, \boldsymbol{x}_{(-j)})}{\lambda(t \mid x_j = 0, \boldsymbol{x}_{(-j)})}.$$

The quantity $\exp(\beta_j)$ is called the relative risk of the treatment group to the placebo group, and it is constant over time. If $\beta_j = 0$, the treatment group and the placebo

group have the same effect on the failure time. The positive risk of failure indicates an increase in the treatment group. The negative risk of failure indicates a decrease in the treatment group. The inference for the PH model treats the baseline hazard as a nuisance parameter, and primarily focuses on the estimate of regression coefficients.

1.4.2 The Proportional Odds Model

Bennett (1983)'s seminal paper relates the odds ratio function to the covariates. Let $\boldsymbol{x}_i = (x_1, \cdots, x_p)^{'}$ be a p-dimensional covariate vector for the i-th subject, and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^{'}$ is the corresponding vector of regression parameter. Given the regressor vector $\boldsymbol{x}$, the odds ratio of T is represented as a proportional odds model,

$$\frac{1 - S(t; \boldsymbol{x})}{S(t; \boldsymbol{x})} \quad = \quad \frac{1 - S_0(t)}{S_0(t)} \times \exp(\boldsymbol{x}^{'}\boldsymbol{\beta}),$$

where $S_0(t)$ is the baseline distribution function controlling all covariates equal to 0, and $S(t; \boldsymbol{x})$ is the survival function with the covariates. Equivalently, one can rewrite the equation in this form:

$$\texttt{logit}\{1 - S(t; \boldsymbol{x})\} \quad = \quad \texttt{logit}\{1 - S_0(t)\} + \boldsymbol{x}^{'}\boldsymbol{\beta},$$

where logit function $\texttt{logit}(p)$ is defined as $\log\left(\frac{p}{1-p}\right)$, and $\texttt{logit}\{1 - S_0(t)\}$ is the baseline log odds function at time t. The baseline log odds function is the log odds function in when covariates are all taken to be zero. Taking $x_j = 1$ for the treatment group with the other $(p-1)$'s covariates fixed, the log odds function of T is expressed in the form,

$$\log\left\{\frac{1 - S(t; x_j = 1, \boldsymbol{x}_{(-j)})}{S(t; x_j = 1, \boldsymbol{x}_{(-j)})}\right\} \quad = \quad \log\left\{\frac{1 - S(t)}{S(t)}\right\} + \beta + \sum_{k \neq j} x_k \beta_k.$$

Taking $x_j = 0$ for the placebo group with the other $(p-1)$'s covariates fixed, the log odds function of T is expressed in the form,

$$\log\left\{\frac{1 - S(t; x_j = 0, \boldsymbol{x}_{(-j)})}{S(t; x_j = 0, \boldsymbol{x}_{(-j)})}\right\} \quad = \quad \log\left\{\frac{1 - S(t)}{S(t)}\right\} + \sum_{k \neq j} x_k \beta_k.$$

Then one can obtain the difference of the log odds of T in the form,

$$\beta_j \;=\; \log\left\{\frac{1-S(t;x_j=1,\boldsymbol{x}_{(-j)})}{S(t;x_j=1,\boldsymbol{x}_{(-j)})}\right\}-\log\left\{\frac{1-S(t;x_j=0,\boldsymbol{x}_{(-j)})}{S(t;x_j=0,\boldsymbol{x}_{(-j)})}\right\}.$$

The interpretation of the quantity β_j is the increase in the log odds of failure by time t from the treatment group to the placebo group. The odds of failure by time t of the treatment group to the placebo group is $\exp(\beta_j)$ with the other $(p-1)$'s covariates fixed. If $\beta_j = 0$, the treatment group and the placebo group have the same effect on the failure time. The positive log odds of failure indicates an increase in the treatment group. The negative log odds of failure indicates a decrease in the treatment group. The inference for the PO model treats the baseline log odds as a nuisance parameter, and primarily focuses on the estimation of regression coefficients.

1.4.3 Gibbs Sampler

By convention, the joint, conditional, and marginal forms of the densities for random variables X and Y in the Gibbs sampler are indicated by square brackets, represented as $[X,Y]$, $[X \mid Y]$, and $[Y]$, respectively. Specifically, the marginalization can be used in the form $[X] = \int [X \mid Y]\cdot[Y]\, d\, Y$ by integration. Suppose that for a collection of n univariate random variables $[X_i, X_2, \cdots, X_n]$, their full conditional densities are represented as $[X_i \mid X_j; i \neq j]$, $\forall i, j = 1, 2, \cdots, n$ and marginal densities are denoted as $[X_i]$, where $i = 1, 2, \cdots, n$. Performing random variate samples of X_i, from $[X_i \mid X_j; i \neq j]$ is an iterative procedure that produces sample-based estimates. The Gibbs sampling method is a Markov Chain Monte Carlo (MCMC) algorithm for propagating and updating schemes as follows. Given an arbitrary initial set of values $\left(X_1^{(0)}, X_2^{(0)}, \cdots, X_k^{(0)}\right)$, for iteration t from 1 to M, the random variate sample $[X_1^{(t)}]$ can be sequentially drawn from $[X_1 \mid X_2^{(t-1)}, X_3^{(t-1)}, \cdots, X_n^{(t-1)}]$. Similarly, the random variate sample $[X_2^{(t)}]$ can be sequentially drawn from $[X_2 \mid X_1^{(t)}, X_3^{(t-1)}, \cdots, X_n^{(t-1)}]$, etc. Up to the last iteration, the random variate sample $[X_n^{(t)}]$ is drawn from $[X_n \mid X_1^{(t)}, X_2^{(t)}, \cdots, X_{n-1}^{(t)}]$ in $[X_i, X_2, \cdots, X_n]$. Each random

variate sample is drawn from its corresponding full conditional density.

$$
\begin{aligned}
[X_1^{(t)}] &\propto [X_1 \mid X_2^{(t-1)}, X_3^{(t-1)}, \cdots, X_n^{(t-1)}] \\
[X_2^{(t)}] &\propto [X_2 \mid X_1^{(t)}, X_3^{(t-1)}, \cdots, X_n^{(t-1)}] \\
&\vdots \\
[X_n^{(t)}] &\propto [X_n \mid X_1^{(t)}, X_2^{(t)}, \cdots, X_{n-1}^{(t)}]
\end{aligned}
$$

in $[X_i, X_2, \cdots, X_n]$. Each random variate sample is drawn from its corresponding full conditional density.

1.4.4 Model Selection Criteria

In this dissertation, three Bayesian model assessment criteria are used for model comparison: Monte Carlo estimation of conditional predictive ordinates (CPO) (Geisser and Eddy, 1979; Gelfand and Dey, 1994; Gelfand et al., 1992; Hanson and Yang, 2007), Bayesian Information Criterion (BIC)(Schwarz and others, 1978), and Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002).

Conditional Predictive Ordinate

Geisser and Eddy (1979) firstly proposed Conditional Predictive Ordinates (CPO) which is a Bayesian cross-validation approach. The quantity of CPO for model j involves prediction of the i-th subject of observed data D_i given that the i-th subject is removed from data $D_{(-i)}$ is $CPO_i^{(j)} = P(D_i \mid D_{(-i)}, M_j)$, where $D_{(-i)} = \{(C_j, \delta_j, \boldsymbol{x}_j) : \delta_j = I(T_i \in (0, C_i]), \text{for all } j \neq i, \boldsymbol{x}_j \in \mathbb{R}^p\}$. The CPO statistic is the posterior predictive probability of failure time T_i falling in the observed interval $(0, C_i]$ or (C_i, ∞) given that i-th subject is removed from the observed data under Model j. The evaluation of the conditional predictive density of $D_i \mid D_{(-i)}$ can be expressed in

the form,

$$\begin{aligned} CPO_i^{(j)} &= \int_\Theta f(D_i \mid \theta, D_{(-i)})\pi(\theta \mid D_{(-i)}, M_j)\ d\theta \\ &= \int_\Theta f(D_i \mid \theta)\pi(\theta \mid D_{(-i)}, M_j)\ d\theta, \end{aligned} \tag{1.1}$$

where $\pi(\boldsymbol{\theta} \mid D_{(-i)}, M_j)$ is posterior distribution of θ given $D_{(-i)}$ under Model j and parameter space $\boldsymbol{\theta} \in \Theta = \mathbb{R}^p$. It may be difficult to directly estimate (1.1) by the integration method. Rather, Gelfand et al. (1992), Dey et al. (1997), and Sinha et al. (1999) showed that using Monte Carlo integration estimates the CPO for interval-censored data in closed form,

$$\begin{aligned} CPO_i^{(j)} &= [E_{\boldsymbol{\theta}|\boldsymbol{D}_i}\{P(T_i \in D_i \mid \boldsymbol{\theta}, \boldsymbol{x}_i, M_j)\}^{-1}]^{-1} \\ &\approx \Big[\frac{1}{N_{\setminus B}} \sum_{l=1}^{N_{\setminus B}} \frac{1}{P(T_i \in D_i \mid \boldsymbol{\theta}^{(l)}, x_i, M_j)}\Big]^{-1}, \end{aligned}$$

where $N_{\setminus B}$ is the total number of iterations discarding the first B burn-ins, $P(T_i \in D_i \mid \boldsymbol{\theta}^{(l)}, x_i, M_j) = F(C_i \mid \boldsymbol{\theta}^{(l)}, x_i, M_j)$, and $\boldsymbol{\theta}^{(l)}$ is the sampled value of model parameters at the l-th iteration of MCMC algorithm. The characteristic of CPO is that the larger quantity of $CPO_i^{(j)}$ indicates how strong evidence to support the proposed model M_j when removed i-th subject from data.

Suppose that there are two models: M_1 and M_2 and the corresponding posterior predictive probabilities: $\pi_1(\cdot)$ and $\pi_2(\cdot)$, respectively. Given the remaining data $D_{(-i)}$, the ratio $CPO_i^{(2)}/CPO_i^{(1)}$ is used to measure how well or poorly the i-th observation supports M_2 relative to M_1. Furthermore, the product of the CPO ratios provides an overall aggregate summary of how well or poorly data D support M_2 relative to M_1. (1.2) is the so-called pseudo Bayes factor instead of the Bayes Factor (Kass and Raftery, 1995).

$$PBF^{(2):(1)} = \prod_{i=1}^{n} \frac{CPO_i^{(2)}}{CPO_i^{(1)}}\ . \tag{1.2}$$

For each model, the pseudo marginal likelihood (PML) is computed for model choice. PML is the product of the CPO statistics which provides an aggregate quantity of

all n observations under model j in the form,

$$PML^{(j)} = \prod_{i=1}^{n} CPO_i^{(j)}.$$

It is legitimate to take the logarithmic function on both sides, and thus, this so-called logarithm of the pseudo-marginal likelihood for model j is defined as

$$LPML^{(j)} = \sum_{i=1}^{n} \log(CPO_i^{(j)}).$$

Thus, the pseudo Bayes factor of model 2 to model 1, $PBF^{(2):(1)} = \exp(LPML^{(2)} - LPML^{(1)})$. LPML is used to model selection criteria: The model with larger LPML value is preferred to models with smaller values.

Bayesian Information Criteria

The BIC cannot directly evaluate the required maximized log likelihood from MCMC implementation. Instead, one can approximate the BIC quantity by averaging out the value of the log likelihood function at each MCMC evaluation. The BIC can be expressed in the form,

$$\widehat{BIC}^{(j)} = -\frac{2}{N_{\setminus B}} \sum_{l=1}^{N_{\setminus B}} \sum_{i=1}^{n} \log L_i(\boldsymbol{\theta}^{(l)} \mid \boldsymbol{x}_i, M_j) + p \log n,$$

where p is the dimension of the parameter space for model M_j, $N_{\setminus B}$ is the total number of iterations discarding the first B burn-ins, $\boldsymbol{\theta}^{(l)}$ is the sampled value of model parameters at the l-the iteration of MCMC algorithm, and $\boldsymbol{x}_i$ is the time-independent covariate vector.

Deviance Information Criteria

Spiegelhalter et al. (1998, 2002) proposed the DIC criteria for Bayesian model selection. The DIC quantity is measured on the posterior distribution of the deviance, $D(\boldsymbol{\theta})$ defined as

$$D(\boldsymbol{\theta}) = \bar{D}(\boldsymbol{\theta}) + P_D,$$

where $\bar{D}(\boldsymbol{\theta})$ is the posterior mean deviance for a measure of fit. P_D is a measure of model complexity, namely, a measure of the effective number of parameters in a model. Namely,

$$P_D \;=\; \bar{D}(\boldsymbol{\theta}) - D(\bar{\boldsymbol{\theta}}) \; .$$

Moreover, the DIC quantity can be evaluated from MCMC analysis, and thus, each iteration takes two times the value of the sample mean of the deviance $2\bar{D}$ minus the estimate of the deviance by using plug-in posterior mean of the parameters $\boldsymbol{\theta}$.

1.4.5 Monotone Splines

This subsection gives an overview of monotone splines from a computational perspective. The number of estimated parameters for the nonparametric function depends on the order of sample size. In practice, a large sample size often leads to estimation difficulties. A polynomial splines approach approximates nonparametric functions. A spline function is a linear combinations of piecewise polynomial basis functions that are convenient to manipulate; there are two commonly used spline functions: monotone splines developed by Ramsay (1988) and B-splines developed by Boor (1978). It is very common to apply a spline-based method to estimate nonparametric functions in the literature (Cai and Betensky, 2003; Grummer-Strawn, 1993; Lin and Wang, 2011; McMahan et al., 2013). This is because spline-based approaches lead to an estimation of a small to moderate number of parameters, while also maintaining adequate modeling flexibility without the assumption of a specific shape for the unknown nonparametric function.

Monotone splines can be used in the approximation of nonparametric functions. Monotone splines approximate an unknown function in an interval by utilizing a linear combination of basis functions with degree d, where specifying d to be 1, 2, 3 or a higher degree corresponds to the use of linear, quadratic, cubic or a higher order of

polynomial basis functions, respectively. An M-spline of degree d is defined as

$$M_j(x \mid d) = \frac{d[(x - t_j)M_j(x \mid d-1) + (t_{j+d} - x)M_{j+1}(x \mid d-1)]}{(d-1)(t_{j+d} - t_j)} \text{ if } t_j \leq x \leq t_{j+d},$$

with the boundary condition,

$$M_j(x \mid 1) = \frac{1}{(t_{j+1} - t_j)} \text{ if } t_j \leq x \leq t_{j+1},$$

where $t_1, t_2, \cdots, t_m$ is a sequence of increasing knots. Each $M_j(x \mid d)$ is zero outside of the interval $[t_j, t_{j+d}]$ and is nonzero over d intervals, and over each interval, there are d nonzero M-splines. Each M-spline is associated with an I-spline which is defined as

$$I_j(x \mid d) \quad = \quad \int_0^x M_j(y \mid d)\ dy.$$

M_j is a piecewise polynomial with degree $d-1$ and is associated with I_j, which is a piecewise polynomial of degree d defined as (if $t_j \leq x \leq t_{j+1}$)

$$I_h(x \mid d) = \begin{cases} 0 & \text{if } h > j \\ \sum_{l=h}^{j} (t_{l+d+1} - t_l) \frac{M_l(x \mid d+1)}{d+1} & \text{if } j - d + 1 \leq h \leq j \\ 1 & \text{if } h < j - d + 1. \end{cases}$$

M-splines are non-negative functions, and I-splines preserve monotonicity; consequently, the monotonicity constraint for a desired function represented on a basis of I-splines can be fulfilled by constraining the coefficients to be positive.

The placement of knots also contributes to model flexibility. The more knots that are allocated in a region of the observed data, the greater the model flexibility that can be attained in that region. One can allot as many knots as needed at each observed data interval; however, the computational price is often high when the number of observations grows exponentially. In general, the specification of the placement of knots and the degree of these basis functions affect the estimation of the spline coefficients. In his seminal paper, Ramsay (1988) noted that it is not necessary in a statistical

environment to use a large number of knots. Breiman (1988) also suggested that few knots suffice providing that they are in the right place after much experimentation. The final model can then be chosen according to a model selection criteria, e.g. log pseudo marginal likelihood (LPML). Similar strategies for determining knot placement are commonly used in the literature; e.g., see Sinha et al. (1999). In addition, several Bayesian approaches (Cai et al., 2011; Lin et al., 2014; Wang and Lin, 2011) adopted shrinkage priors for monotone spline coefficients and prevented over-fitting problems that may be potentially caused due to the use of excessively large number of knots.

1.5 Outline of this Dissertation

Chapter 1 has reviewed some semiparametric regression models for censored-type data from frequentist and Bayesian perspectives. The large class of transformation models is attractive and appealing from a model perspective, taking the PH, PO, AFT, and PB models as special cases. In the following chapters, two richly semiparametric transformation models are introduced: semiparametric linear transformation (LT) models (Cheng et al., 1995, 1997) and the generalized odds-rate hazards (GORH) models (Scharfstein et al., 1998). The standard features are taking semiparametric PH and PO models as special cases. Chapter 2 proposes a unified estimation approach for regression analysis of current status data under semiparametric linear transformation models. Chapter 3 proposes a Bayesian estimation approach for regression analysis of general interval-censored data under the GORH models with the fixed nuisance parameter. Chapter 4 extends Lin and Wang (2010)'s work to analyze arbitrarily censored data and applies the proposed method to the Steno Memorial Hospital diabetic nephropathy dataset.

Chapter 2

A Unified Bayesian Estimation Approach for Regression Analysis of Current Status Data Under Semiparametric Linear Transformation Models

Summary: Semiparametric linear transformation models are a broad class of semiparametric regression models taking the proportional hazards model, proportional odds model and probit model as special cases. Although semiparametric linear transformation models are widely used for analyzing right-censored survival data in the literature, their applications to current status data are limited. In this chapter, we propose a unified Bayesian estimation approach for regression analysis of current status data among three commonly used semiparametric regression models in the framework of semiparametric linear transformation models. The proposed method adopts monotone splines for modeling the unknown, increasing transformation functions in semiparametric linear transformation models. The proposed method facilitates shrinkage priors for monotone spline coefficients and prevents overfitting problems that potentially caused due to the use of excessively large number of knots. A novel unified estimation approach is proposed to facilitate posterior computation. The proposed method also allows for the estimation of regression coefficients and simultaneously, estimates the marginal survival function. The proposed method is generic for all semiparametric linear transformation models and allows model selec-

tion. We illustrate the process through an application to a uterine fibroid dataset from an epidemiological study.

Keywords: Current status data; monotone splines; semiparametric linear transformation models; semiparametric regression; uniform latent variable

2.1 Introduction

Current status data is also known as case 1 interval-censored data (Groeneboom and Wellner, 1992). Current status data are a special case of interval-censored data. Current status data naturally arise when the failure time is not observed (e.g. the onset time of a tumor), but the failure time occurs either before or after the one-time observation time point. The prominent feature of current status data is that each subject is observed only once in the study. The onset time of the tumor either precedes or follows a censoring time for each subject. Jewell and van der Laan (2002) gave a concrete example with carcinogenicity testing: practitioners conducted laboratory animal carcinogenicity experiments to investigate the concealed tumor onset time from exposure to a potential carcinogen until the first occurrence of the tumor. Practitioners can determine the presence or absence of the concealed tumor at the animal's time of death to provide the censoring information on the onset time of the tumor. As a result, all subjects are either left-censored when the onset of the tumor occurs before the observation time point or right-censored when the onset of the tumor occurs after the observation time point.

Many existing methods have been developed for regression analysis of current status data under the proportional hazards (PH) model. Huang (1996) demonstrated that the MLE of the finite dimensional regression parameter in the PH model is asymptotically efficient but the infinite-dimensional parameter converges slower than $\sqrt{n}$ for current status data. Pan (1999) extended the iterative convex minorant algo-

rithm, which was developed by Groeneboom and Wellner (1992), to consider covariate effects in the PH model, and this method is called the generalized gradient projection (GGP) method.

Spline-based methods have prevailed for regression analysis of current status data. Kooperberg and Clarkson (1997) introduced hazard regression, and Kooperberg et al. (1995) used linear splines and their tensor products in the estimation of the conditional log-hazard function for interval-censored data. Shiboski (1998) fitted the generalized additive model (GAM) and isotonic regression and provided simultaneous estimation of regression coefficients and the baseline survival function. Cai and Betensky (2003) proposed a flexible locally parametric procedure to model the baseline log-hazard function and to obtain MLE via Penalized Spline for interval-censored data. McMahan et al. (2013) exploited a novel data augmentation approach based on Poisson latent variables and utilized monotone splines. Compared to the flourishing research from frequentist perspectives, limited research from Bayesian perspectives is found in the literature for current status data or interval-censored data under the PH model. To name a few, Sinha et al. (1999) proposed a Bayesian approach for analyzing general interval-censored data. Cai et al. (2011) proposed a Bayesian approach by using monotone splines for analyzing current status data.

Under the PO model, many frequentist approaches have been proposed for regression analysis of current status data. Rossini and Tsiatis (1996) modified the standard maximum likelihood procedures and used the approximate likelihood to obtain consistent, asymptotically normal, and semiparametric efficient estimates. Rabinowitz et al. (2000) presented an approach which is efficiently implemented by using conditional logistic regression routines in standard software packages. Sun (2007), and Zhang and Sun (2009) reviewed existing approaches as a preliminary study of current status data. A few Bayesian approaches have been proposed. For example, Wang and Lin (2011) and Lin and Wang (2011) proposed adaptive monotone splines to estimate

the regression parameters for current status data.

Semiparametric linear transformation (LT) models are a broad class of regression models which take the PH, PO, and PB models as special cases. Semiparametric LT models presume that an unknown non-decreasing transformation function of failure time is linearly correlated to covariates plus the random error term (Cheng et al., 1995, 1997). The Box-Cox Transformation model can be regarded as a parametric version of semiparametric LT models which indexed by a finite-dimensional unknown parameter vector (Box and Cox, 1964). Fine et al. (1998) pointed out that the Box-Cox parametric transformation model may not give a straightforward interpretation of the estimation of regression coefficient in semiparametric LT models. Semiparametric LT models are specified with a general link function; more details are given in section 2.2.2. The Box-Cox parametric transformation model cannot generate the PH or PO model by specifying the general link function in semiparametric LT models. In addition, for the sake of possible misspecification of parametric models, we use semiparametric LT models with unknown and smooth transformation functions. Semiparametric LT models relax the parametric assumption in order to obtain flexibility and robustness against misspecified parametric LT models. A few approaches fit semiparametric LT models for regression analysis of current status data. Ma and Kosorok (2005) applied penalized log-likelihood estimation for partial LT models with current status data. Sun and Sun (2005) proposed a general inference procedure based on estimating functions for regression analysis of current status data under semiparametric LT models. Compared to the frequentist approaches, existing Bayesian approach for regression analysis of current status data in a semiparametric LT model is very limited. Some existing Bayesian methods are used for regression analysis of right-censored data (Mallick and Walker, 2003).

The aim of this chapter is to propose a data augmentation approach based on uniformly-distributed latent variables among these three semiparametric (PH/PO/PB)

regression models. The proposed method can simultaneously estimate regression coefficients and the baseline CDF. This augmented likelihood function can be easily derived from the observed likelihood function of current status data. This chapter is organized in the following sections: Section 2 provides the details of the proposed Bayesian approach, involving the introduction to semiparametric LT models, monotone splines for modeling the cumulative baseline hazard function in the PH model, the baseline odds function in the PO model, the transformed baseline CDF with probit link in the PB model, and prior specification and posterior computation. Section 3 presents the simulation results of the proposed method and model selection under the unified framework of semiparametric LT models. Section 4 illustrates the proposed method with the application of uterine fibroid data from an epidemiological study. Section 5 provides some concluding remarks and discussions.

2.2 The Proposed Method

2.2.1 Notation

Assume that there are n independent subjects in the study. Let T_i be the failure time, $\boldsymbol{x}_i$ be a $p \times 1$ vector of time-independent covariates for the i-th subject, and $F(t \mid \boldsymbol{x}_i) = P(T_i \leq t \mid \boldsymbol{x}_i)$ is denoted as the CDF of T_i given $\boldsymbol{x}_i$, for all $0 < t < \infty$ and $i = 1, \cdots, n$. The observed data are represented as $\{C_i, \delta_i, \boldsymbol{x}_i\}_{i=1}^n$, where C_i is the random censoring time, and δ_i is the indicator variable defined as $\delta_i = I(T_i \leq C_i)$ for the i-th subject. Given covariates $\boldsymbol{x}_i$, random censoring time C_i's and the failure time T_i's are independent. This assumption is also called the non-informative censoring (Betensky, 2000; Sun, 2007; Turnbull, 1976; Williams and Lagakos, 1977). The observed likelihood function can be expressed in the form,

$$L_{obs} = \prod_{i=1}^{n} \{F(C_i|\boldsymbol{x}_i)\}^{\delta_i} \{1 - F(C_i|\boldsymbol{x}_i)\}^{1-\delta_i}. \tag{2.1}$$

2.2.2 Semiparametric Linear Transformation Models

Cheng et al. (1995, 1997) and Fine et al. (1998) proposed semiparametric LT models for which a completely unspecified monotone transformation of event time, T is presumed to be linearly related to observed covariate vector, $\boldsymbol{x}$, with a specified random error, ϵ. Semiparametric LT models can be written as $\alpha(T) = -\boldsymbol{x}'\boldsymbol{\beta} + \epsilon$, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)'$ is a $p \times 1$ vector of regression parameters, and $\alpha(\cdot)$ is an unspecified non-decreasing function with $\lim_{t\to 0} \alpha(t) = -\infty$ and $\lim_{t\to\infty} \alpha(t) = \infty$ for $t \in \mathbf{R}^+$. The distribution of the error term ϵ is well-defined with a completely known cumulative distribution $G(\cdot)$. The resulting CDF function of T is $F(t \mid \boldsymbol{x}) = G(\alpha(t) + \boldsymbol{x}'\boldsymbol{\beta})$. In specifying semiparametric LT models, $G(\cdot)$ can take one of but not limit to the following cumulative distribution functions.

- PH model: G is the CDF of the standard extreme value distribution with $G(s) = P(\epsilon \le s) = 1 - \exp\{-\exp(s)\}$. In this situation, $\alpha(t)$ is the logarithm of the cumulative baseline hazard function, namely, $\alpha(t) = \log[-\log\{1-F_0(t)\}]$.

- PO model: G is the CDF of the standard logistic distribution with $G(s) = P(\epsilon \le s) = \{\exp(s)\}\{1 + \exp(s)\}^{-1}$. In this situation, $\alpha(t)$ is the logarithm of the baseline odds function, namely, $\alpha(t) = \log[F_0(t)/\{1 - F_0(t)\}]$.

- PB model: G is the CDF of the standard normal distribution function, $G(s) = P(\epsilon \le s) = \Phi(s)$. In this situation, $\alpha(t)$ is the transformed baseline CDF with probit link, namely, $\alpha(t) = \Phi^{-1}\{F_0(t)\}$.

Observe that $G(\cdot)$ is a strictly increasing function, and its inverse function $G^{-1}(\cdot)$ can regard as a link function. For example, one can derive the complementary log-log transformation $G^{-1}(s) = \log\{-\log(1 - s)\}$ (McCullagh, 1980) in the PH model, the logit transformation function $G^{-1}(s) = \log\{s/(1 - s)\}$ (Bennett, 1983; Pettitt, 1984) in the PO model, and the probit transformation $G^{-1}(s) = \Phi^{-1}(s)$ in the PB model.

2.2.3 Modeling $\alpha(\cdot)$ for Monotone Splines

The estimation of model parameters under a semiparametric regression model is difficult because of the existence of the infinite-dimensional nonparametric transformation function. The unspecified nonparametric transformation function $\alpha(\cdot)$ can be modeled by a linear combination of integrated spline (I-spline) basis functions Ramsay (1988). Following the work of Lin and Wang (2010), Cai et al. (2011), Wang and Dunson (2011), Wang and Lin (2011), Lin and Wang (2011), Wang et al. (2012), and Lin et al. (2014), the proposed approach leads to the following representation,

$$\alpha(t) = \gamma_0 + \sum_{l=1}^{k} \gamma_l b_l(t), \quad t \in \mathbf{R}^+, \tag{2.2}$$

where γ_0 is an unconstrained intercept of a monotone spline, $\boldsymbol{\gamma}_l$'s are a set of nonnegative spline coefficients, and $b_l(t)$'s are integrated spline (I-spline) basis functions with degree d, each of which is a non-decreasing function from 0 to 1. Nuisance parameters are γ_0 and γ_l's are used to specify the unknown non-decreasing transformation function. The shapes of the basis functions are predominantly determined by the placement of knots and the degree d of the basis function which controls the overall smoothness of the basis functions (e.g., specifying degree to be 1, 2, or 3 corresponds to the use of piecewise linear, quadratic, or cubic basis functions, respectively) (Ramsay, 1988). Thus, spline basis functions are piecewise polynomial functions. The construction of I-spline basis functions is determined by the degree d of the basis functions and m interior knots which are chosen in an increasing sequence of knots within a time range (Ramsay, 1988). Once the placement of knots and the degree of the basis functions are specified, the k spline basis functions are fully determined, where the total number of basis functions is $k = m + d$.

The placement of knots determines the overall modeling flexibility; therefore, the more knots that are allocated in a region of the observed data, the greater model flexibility that can be attained in that region. Lin and Wang (2010), Wang and Lin

(2011), and Wang and Lin (2011) recommended using approximately 10-30 equispaced knots in the application of monotone splines for analyzing interval-censored data. Our prior specification (see Section 2.2.4) showed that Bayesian regularization can penalize excessively large knot sets by shrinking spline coefficients of those unnecessary basis functions toward zero in the use of a shrinkage prior. Therefore, the proposed method utilizes the allocation of equispaced knots with a moderate number of knots to capture curvature information of the unspecified nonparametric transformation function.

2.2.4 Prior Specification & Posterior Computation

From a Bayesian perspective, one may directly apply the sampling method on the observed likelihood (2.1) after incorporating their prior distributions. However, this approach that is based on the complicated observed data likelihood (2.1) often leads to extremely difficult computations because none of the parameters have a standard full conditional distribution. To overcome this difficulty, the proposed approach augments the observed data with the introduction of uniformly distributed latent variables. Based on the augmented likelihood, we develop a unified Bayesian estimation approach, which can estimate model parameters for all semiparametric linear transformation models in the same framework of the Gibbs Sampler. The augmented likelihood function can be expressed in the form,

$$L_{aug}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \{I_{[0,G\{\alpha(c_i)+x_i'\beta\}]}(u_i)\}^{\delta_i} \{I_{[G\{\alpha(c_i)+x_i'\beta\},1]}(u_i)\}^{1-\delta_i}, \tag{2.3}$$

where each u_i is a uniformly distributed latent variable. Priors of unknown parameters $\theta = (\boldsymbol{\beta}', \boldsymbol{\gamma}_0, \boldsymbol{\gamma}_l')'$ in the augmented likelihood function (2.3) are specified as follows. Regression coefficient β_j is assigned independent vague normal priors $\pi(\beta_j) = \mathcal{N}(\beta_{j0}, \sigma_{j0}^2)$ with a large σ_{j0}^2, for $j = 1, 2, \cdots, p$. The intercept of the monotone spline γ_0 is unconstrained and is assigned a conventional vague normal prior for $\pi(\boldsymbol{\gamma}_0) = \mathcal{N}(\mu_0, \nu_0)$ with a large ν_0. Motivated by the widely used double ex-

ponential prior in Bayesian LASSO regression (Park and Casella, 2008), one can assign independent exponential priors $\mathcal{E}(\lambda)$ for the non-negative spline coefficients (the $\boldsymbol{\gamma}_l$'s). A gamma prior $\mathcal{G}(a_\lambda, b_\lambda)$ is assigned for λ with mean a_λ/b_λ and variance a_λ/b_λ^2. Such a prior specification is appealing because this prior specification allows to borrow of information among γ_l's and to shrink the spline coefficients of those unnecessary basis functions toward zero. This property allows us to use many knots to provide adequate modeling flexibility without additional computational costs and avoids over-fitting issues. Such prior specifications are equivalent to a penalized likelihood approach with a penalty on the sum of those nonnegative spline coefficients from a frequentist perspective. However, such penalized likelihood approach needs to select a proper tuning parameter with more computational costs by using generalized cross-validation (GCV). In contrast, our approach treats the tuning parameter λ as a parameter and allows to update it within the Bayesian posterior computation. A Markov Chain Monte Carlo (MCMC) algorithm iterates through the following steps.

1. Sample u_i from

$$\begin{aligned} U_{[0, G\{\alpha(c_i)+x_i'\beta\}]} \quad &\text{when} \quad \delta_i = 1, \\ U_{[G\{\alpha(c_i)+x_i'\beta\}, 1]} \quad &\text{when} \quad \delta_i = 0. \end{aligned}$$

2. Sample β_j from $\mathcal{N}(\beta_0, \sigma_0^2)1_{(a_j, b_j)}$ for $j = 1, 2, \cdots, p$, where $a_j = \max(a_{j,1}, a_{j,2})$ and $b_j = \min(b_{j,1}, b_{j,2})$. The truncation endpoints are as listed below.

$$\begin{aligned} a_{j,1} &= \max_{i \in A_1} x_{ij}^{-1} \Big\{ G^{-1}(u_i) - \alpha(c_i) - \sum_{k \neq j} \beta_k x_{ik} \Big\}, \\ a_{j,2} &= \max_{i \in A_2} x_{ij}^{-1} \Big\{ G^{-1}(1-u_i) - \alpha(c_i) - \sum_{k \neq j} \beta_k x_{ik} \Big\}, \\ b_{j,1} &= \min_{i \in B_1} x_{ij}^{-1} \Big\{ G^{-1}(u_i) - \alpha(c_i) - \sum_{k \neq j} \beta_k x_{ik} \Big\}, \\ b_{j,2} &= \min_{i \in B_2} x_{ij}^{-1} \Big\{ G^{-1}(1-u_i) - \alpha(c_i) - \sum_{k \neq j} \beta_k x_{ik} \Big\}, \end{aligned}$$

where

$$\begin{aligned} A_1 &= \{i : \delta_i = 1, x_{ij} > 0, i = 1, 2, \cdots, n\}, \\ A_2 &= \{i : \delta_i = 0, x_{ij} < 0, i = 1, 2, \cdots, n\}, \\ B_1 &= \{i : \delta_i = 1, x_{ij} < 0, i = 1, 2, \cdots, n\}, \\ B_2 &= \{i : \delta_i = 0, x_{ij} > 0, i = 1, 2, \cdots, n\}. \end{aligned}$$

3. Sample γ_0 from $\mathcal{N}(\mu_0, \nu_0)1_{(c,d)}$, where

$$\begin{aligned} c &= \max_{\delta_i=1} \left\{ G^{-1}(u_i) - \sum_{l=1}^{k} \gamma_l b_l(c_i) - x_i'\beta \right\}, \\ d &= \min_{\delta_i=0} \left\{ G^{-1}(1-u_i) - \sum_{l=1}^{k} \gamma_l b_l(c_i) - x_i'\beta \right\}. \end{aligned}$$

4. Sample γ_l from $\mathcal{E}(\lambda)1_{(e_l,f_l)}$, for $l = 1, 2, \cdots, k$, where

$$\begin{aligned} e_l &= \max_{\delta_i=1, b_l(c_i)>0} \{b_l(c_i)\}^{-1} \left\{ G^{-1}(u_i) - \gamma_o - \sum_{q \neq l} \gamma_q b_q(c_i) - x_i'\beta \right\}, \\ f_l &= \min_{\delta_i=0, b_l(c_i)>0} \{b_l(c_i)\}^{-1} \left\{ G^{-1}(1-u_i) - \gamma_o - \sum_{q \neq l} \gamma_q b_q(c_i) - x_i'\beta \right\}. \end{aligned}$$

5. Sample λ from $\mathcal{G}(a_\lambda + k, b_\lambda + \sum_{l=1}^{k} \gamma_l)$.

2.3 Simulation Evidence

An intensive simulation study was conducted to estimate the regression coefficients and assess the performance of the proposed approach across several settings: 200 datasets with 200 observations per dataset. The true cumulative distribution function of the failure time T_i is written as,

$$F(t \mid x_{i1}, x_{i2}) = G\{\alpha(t) + \beta_1 x_{i1} + \beta_2 x_{i2}\},$$

where $G(\cdot)$ takes the CDF for the considered models: PH, PO, and PB, and $\alpha(t) = 1 + t^{\frac{3}{2}} + \log(t)$. The covariates are $x_1 \sim \texttt{Bernoulli}(0.5)$ and $x_2 \sim \mathcal{N}(0, 1)$, for

$i = 1, \cdots, 200$. True β_1 takes on the values $\{1, 0\}$ and β_2 takes on the values {1,-1}, resulting in four parameter configurations under each model, respectively. The censoring time C_i was generated from a truncated exponential distribution $\mathcal{E}(1)I_{(0,10)}$, and then we generated censoring indicator y_i from a Bernoulli distribution with probability $G\{\alpha(C_i) + \beta_1 x_{i1} + \beta_2 x_{i2}\}$. To implement the posterior computation, we use independent normal priors with $\beta_{j0} = 0$ and $\sigma_{j0}^2 = 10^2$ for $j = 1$ and 2, vague univariate normal distribution with mean μ_0 = -4 and precision $\nu_0 = 0.1$ for γ_0, independent exponential distribution with rate 1 for γ_l, and a hyper gamma prior $\mathcal{G}(1, 1)$ for λ.

As seen in Table 2.1, right censoring rate (CR) refers to the average of the right-censoring rates across 200 datasets. `BIAS` is the average of the 200 posterior means minus the true value; `ESD` is the mean of the estimated standard deviation from their posterior distributions across 200 datasets; `SSD` is the sample standard deviation of the 200 point estimates; and `CP95` is the 95% coverage probability (i.e., the proportion of the 95% credible intervals which cover the true value). For all three regression models, the bias is small if any for all regression parameters in all the configurations. It is observed that the sample standard deviation `SSD` and the estimated standard error `ESE` are quite close, and the 95% coverage probabilities for β_1 and β_2 are close to the nominal level 0.95 in all parameter configurations. In a sensitivity analysis, we ran additional simulation studies to investigate the effect of the hyperparameters by using more vague priors, $(a_\eta, b_\eta) = (0.1, 0.1)$ and $(0.01, 0.01)$, respectively. The results of this sensitivity analysis demonstrate that the proposed method is robust and suggest that taking $a_\eta = b_\eta = 1$ is not overly informative. Thus, the proposed method is promising in the estimation of the regression coefficients under each semiparametric linear transformation model.

For the purpose of model selection, each true model is used to generate 200 datasets with 200 observations per dataset. Two model comparison criteria, MSE and LPML, are applied to compare with the other two candidate models. Define

a local mean squared error (MSE) of the baseline cumulative distribution function $\hat{F}_0(t)$ at time t as

$$MSE\{\hat{F}_0(t)\} = \frac{1}{100}\sum_{j=1}^{100}\{F_0(t) - \hat{F}_0^{(j)}(t)\}^2,$$

where $\hat{F}_0^{(j)}(t)$ is the estimate of $F_0(t)$ in our approach for the j-th dataset. Denote $\overline{\texttt{MSE}}$ (`maxMSE`) as the global mean (maximum) squared error of $\hat{F}_0$. The global mean (maximum) squared error of $\hat{F}_0$ is defined as the mean (maximum) of the local MSEs of $\hat{F}_0(t)$ and evaluated on a set of pre-specified equispaced grid points. The smaller values of global MSEs are, the more accurate estimates of the cumulative distribution function are. Table 2.2 demonstrates that the proposed approach can estimate the baseline CDF ($\hat{F}_0$) accurately for each parameter configuration when the true models are used for our approach.

LPML is a commonly used model comparison criterion in Bayesian semiparametric regression models (Hanson and Yang, 2007; Sinha et al., 1999; Wang and Lin, 2011). Large LPML values indicate strong evidence to support the proposed model. For each generated dataset, our proposed approach fit three candidate models and calculated individual LPML values for each parameter configuration. Then we aggregated the number of largest LPML values for each model over 200 generated datasets for each parameter configuration. Table 2.3 shows that the percentage of the largest LPML values for each model over 200 datasets in each parameter configuration. The results of model comparison show that around 50% of the time our method can choose the true model rather than the other two candidate models in each model configuration.

The proposed Gibbs sampler has shown a slow mixing in the Markov chains. Thus, we ran the MCMC with a total of 50,000 iterations and took the first 10,000 as a burn-in. The assessment was performed by using various convergence criteria, such as `gelman.diag` and `geweke.diag` in the R package coda. The summary results were obtained based on a sample which was taken every 40th sampled value of the MCMC sample from the latter 40,000 iterations. For the purpose of attaining efficiency, we

implemented seamless R/C++ integration via the Rcpp package. The running time of MCMC sampling can significantly decrease from 8 hours (implemented in MATLAB) to 30 minutes in the computer platform (Intel Core i7-3.4GHz 16GB DDR3 Memory).

2.4 Real Data Application

2.4.1 Uterine Leiomyomas Data

Right From the Start (RFTS) is an on-going, prospective cohort study of early-pregnancy health that was conducted in three states (NC/TX/TN). In this cohort study, the onset time of uterine fibroids was unknown. However, the onset time of uterine fibroids was known to occur either before or after a one-time ultrasound examination that was performed at the exact enrollment time of each participant, in early pregnancy (prior to the seventh gestational week). Participants were diagnosed as positive when the ultrasound examination revealed leiomyomata diameter of 0.5 cm or larger. The RFTS study was a prospective cohort study of early pregnancy. This study was composed of three individually funded sub-studies: RFTS 1, 2, and 3. RFTS 1 was found to have an under-reporting issue because of less intensive training of the ultrasonographers, and RFTS 3 is an on-going sub-study with a small dataset. The percentage of women who were diagnosed to have uterine fibroids is similar in RFTS 2 and 3 for both African American and white American women. This study is based on the RFTS 2 dataset only. Among 1604 participants, there are 1377 white American women and 227 African American women. The eligibility requirements were maternal age of 18 years or older, and enrollment in the program by 12 6/7 weeks of gestation based on last menstrual period.

An ultrasound examination was scheduled for each participant with the aim at the seventh week of gestation. The status of uterine fibroids was unknown before the ultrasound examination, and participants were not told to check their fibroid status in their ultrasound examinations; therefore, it is reasonable to presume that the onset

time of fibroids and examination time is independent. In this cohort study, practitioners diagnosed all participants' fibroid statuses through their one-time ultrasound examinations. Thus, the onset time of uterine fibroids for each participant was either left-censored or right-censored. Hence, the dataset forms typical current status data. The primary research interests are to estimate the cumulative incidence of uterine fibroids (a.k.a. uterine leiomyoma) and find out which risk factors significantly affect the incidence of fibroids between African-American and white American women. The potential risk factors of the investigator's interests are ethnicity (white versus African-American women), parity status (i.e. subject had given birth before), age of menarche (age when a participant had her first menstrual period), and BMI (body mass index) status. The details of the experimental design can be found in (Cai et al., 2011; Laughlin et al., 2009).

We vectorized $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \boldsymbol{x}_{i3}, \boldsymbol{x}_{i4})$ as covariates of subject i, with $\boldsymbol{x}_{i1} = 1$ indicating the i-th participant as an African-American Women, $\boldsymbol{x}_{i2} = 1$ indicating the i-th participant having pregnancy history, $\boldsymbol{x}_{i3}$ indicating the mean-centered age of menarche for the i-th participant, and $\boldsymbol{x}_{i4} = 1$ indicating a BMI exceeding 30, which is the cut point for obesity for the i-th participant. Under a Bayesian framework, we applied the proposed method to run 50,000 MCMC iterations with 10,000 as a burn-in for collecting one sample out of every 40 samples from the Markov chain. As a result, the resulting MCMC samples alleviated auto-correlation, and the Markov chains are stationary. The model selection criteria LPML for uterine fibroids data analysis is summarized in Table 2.4. The largest LPML for the PH model occurs when $m = 12$, the largest LPML for the PO model occurs when $m = 10$, and the largest LPML for the PH model occurs when $m = 16$, respectively. As seen in Table 2.5, the estimation results of risk factors appear that parity (subject had given birth before the enrollment of program), age of menarche, and race are significant risk factors in the development of uterine fibroids, but obesity status (BMI > 30) is not significant

at 5% level of significance. In particular, under the PH model, the hazard rate in the development of fibroids for African-American participants is $\exp(1.42) \approx 4.137$ times that for white participants, holding all other covariates at the same levels. Under the PO model, the odds of developing fibroids for African-American participants is about $\exp(1.58) \approx 4.855$ times that for white participants, holding the other risk factors at the same levels. Under the PB model, the transformed cumulative incidence of developing fibroids for African-American participants is about 0.91 less than that for white American participants under the probit link, holding the other risk factors at the same levels. Regarding the other two negative significant risk effects, participants who had given birth before their enrollment in the program and who had a late menarche age had reduced risk of developing uterine fibroids. Figure 2.1 presents the estimated cumulative incidences of fibroids under different semiparametric regression models for African women and white women participants, who were non obese, had the mean age of menarche, and had no pregnancy before the enrollment. As seen in Figure 2.1, the estimated cumulative incidences curves are close under three semiparametric regression models, and there is significantly different between African women and white women.

2.5 Discussion

In this chapter, we proposed a unified approach for current status data under semiparametric regression models. Monotone splines are used to approximate the unknown non-decreasing function for each model to reduce the number of parameters while maintaining adequate modeling flexibility. A quantile-based interior knot selection method does not achieve better performance than an equispaced knot selection method (Lin et al., 2014). We facilitated this unified approach by using data augmentation based on uniform latent variables in the Gibbs sampler. The proposed method does not involve any Metropolis steps in the MCMC algorithm. Each Gibbs sam-

pling step of the proposed method is a standard full conditional distribution. This work sheds light on how to apply a simple uniform latent variable for researchers to construct a unified sampling technique in semiparametric LT models from the Bayesian perspective. The simulation results are appealing in the estimation of regression parameters. Moreover, the proposed method can be extended to analyze general interval-censored data.

Table 2.1: Sensitivity Analysis: Estimation of the regression coefficients for different hyper-parameters, a_λ and b_λ based on 200 simulated datasets.

					Results on β_1				Results on β_2			
Model	CR	β_1	β_2	$a_\lambda = b_\lambda$	Bias	ESE	SSD	CP95	Bias	ESE	SSD	CP95
PH	18.29%	1	1	1	−0.0491	0.3752	0.3821	0.950	−0.0087	0.2297	0.2192	0.950
				0.1	−0.0562	0.3789	0.3741	0.950	−0.0135	0.2303	0.2156	0.960
				0.01	−0.0527	0.3816	0.3776	0.950	−0.0123	0.2311	0.2168	0.950
	18.79%	1	−1	1	−0.0101	0.3792	0.3336	0.980	0.0175	0.2287	0.2097	0.950
				0.1	−0.0142	0.3794	0.3276	0.980	0.0223	0.2287	0.2083	0.970
				0.01	−0.0196	0.3793	0.3275	0.980	0.0258	0.2283	0.2066	0.965
	23.78%	0	1	1	−0.0184	0.3260	0.3142	0.965	−0.0131	0.2135	0.1821	0.995
				0.1	−0.0148	0.3296	0.3130	0.960	−0.0171	0.2156	0.1812	0.990
				0.01	−0.0182	0.3283	0.3132	0.945	−0.0185	0.2150	0.1788	0.995
	24.15%	0	−1	1	0.0287	0.3272	0.2876	0.950	0.0205	0.2115	0.2131	0.970
				0.1	0.0264	0.3289	0.2849	0.965	0.0278	0.2121	0.2116	0.955
				0.01	0.0270	0.3300	0.2864	0.960	0.0240	0.2136	0.2125	0.955
PO	26.48%	1	1	1	−0.0010	0.4560	0.4400	0.950	−0.0096	0.2477	0.2254	0.960
				0.1	−0.0014	0.4639	0.4389	0.945	−0.0108	0.2482	0.2229	0.960
				0.01	−0.0004	0.4639	0.4404	0.950	−0.0113	0.2467	0.2219	0.965
	26.50%	1	−1	1	−0.0201	0.4607	0.4832	0.940	0.0279	0.2518	0.2563	0.960
				0.1	−0.0261	0.4640	0.4783	0.955	0.0256	0.2517	0.2522	0.950
				0.01	−0.0251	0.4638	0.4807	0.945	0.0255	0.2514	0.2517	0.955
	31.99%	0	1	1	0.0041	0.4258	0.3874	0.955	−0.0051	0.2356	0.2434	0.935
				0.1	0.0089	0.4330	0.3859	0.975	−0.0057	0.2380	0.2394	0.940
				0.01	0.0073	0.4313	0.3834	0.970	−0.0075	0.2382	0.2390	0.935
	32.29%	0	−1	1	0.0084	0.4241	0.4036	0.965	0.0325	0.2324	0.2032	0.965
				0.1	0.0111	0.4287	0.3951	0.970	0.0379	0.2335	0.2019	0.980
				0.01	0.0127	0.4296	0.3943	0.965	0.0379	0.2324	0.2003	0.970
PB	24.00%	1	1	1	−0.0345	0.3441	0.3320	0.950	−0.0345	0.2022	0.1836	0.970
				0.1	−0.0391	0.3457	0.3279	0.960	−0.0388	0.2031	0.1806	0.965
				0.01	−0.0388	0.3475	0.3275	0.950	−0.0394	0.2026	0.1811	0.960
	23.86%	1	−1	1	−0.0325	0.3407	0.3297	0.955	0.0481	0.2000	0.1972	0.920
				0.1	−0.0365	0.3446	0.3219	0.965	0.0511	0.2015	0.1952	0.925
				0.01	−0.0396	0.3447	0.3232	0.950	0.0531	0.2010	0.1971	0.920
	30.29%	0	1	1	−0.0074	0.3053	0.8756	0.975	−0.0407	0.1915	0.1826	0.965
				0.1	−0.0025	0.3060	0.2756	0.975	−0.0428	0.1915	0.1809	0.965
				0.01	−0.0050	0.3073	0.2746	0.970	−0.0449	0.1920	0.1806	0.965
	29.98%	0	−1	1	−0.0444	0.3062	0.2762	0.965	0.0379	0.1923	0.1892	0.925
				0.1	−0.0402	0.3130	0.2752	0.970	0.0418	0.1931	0.1842	0.920
				0.01	−0.0408	0.3120	0.2772	0.970	0.0423	0.1922	0.1845	0.920

Table 2.2: Mean square errors of the estimates of F_0 based on 200 datasets.

True Model	Fit Model	MSE	$(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$	(1,1)	(1,-1)	(0,1)	(0,-1)
PH	PH	$\overline{\texttt{MSE}}$		0.0011	0.0010	0.0012	0.0011
		maxMSE		0.0093	0.0076	0.0082	0.0082
	PO	$\overline{\texttt{MSE}}$		0.0101	0.0097	0.0095	0.0100
		maxMSE		0.0605	0.0587	0.0549	0.0584
	PB	$\overline{\texttt{MSE}}$		0.0050	0.0050	0.0050	0.0044
		maxMSE		0.0375	0.0372	0.0361	0.0314
PO	PH	$\overline{\texttt{MSE}}$		0.0051	0.0043	0.0045	0.0041
		maxMSE		0.0280	0.0240	0.0224	0.0202
	PO	$\overline{\texttt{MSE}}$		0.0016	0.0016	0.0017	0.0018
		maxMSE		0.0069	0.0068	0.0070	0.0066
	PB	$\overline{\texttt{MSE}}$		0.0018	0.0018	0.0017	0.0016
		maxMSE		0.0097	0.0095	0.0091	0.0072
PB	PH	$\overline{\texttt{MSE}}$		0.0042	0.0033	0.0039	0.0036
		maxMSE		0.0471	0.0370	0.0443	0.0405
	PO	$\overline{\texttt{MSE}}$		0.0050	0.0048	0.0048	0.0051
		maxMSE		0.0232	0.0220	0.0219	0.0240
	PB	$\overline{\texttt{MSE}}$		0.0013	0.0014	0.0015	0.0013
		maxMSE		0.0087	0.0090	0.0093	0.0078

Table 2.3: Model Comparison of LPML criteria based on 200 datasets.

True Model	Fit Model $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$	(1,1)	(1,-1)	(0,1)	(0,-1)
PH	PH	54.50%	55.50%	49.50%	51.00%
	PO	7.50%	10.00%	13.00%	10.50%
	PB	38.00%	34.50%	37.50%	38.50%
PO	PH	7.00%	9.50%	6.00%	7.00%
	PO	59.50%	61.00%	61.50%	58.50%
	PB	33.50%	29.50%	32.00%	34.00%
PB	PH	36.50%	29.00%	28.50%	23.50%
	PO	15.00%	10.00%	13.00%	14.50%
	PB	48.50 %	61.00%	58.50%	62.00%

Table 2.4: The optimal knots of three candidate models for estimating a non-parametric transformation function. The comparison of the Log Pseudo Marginal Likelihood (LPML) value in uterine fibroids data analysis.

Model	PH	PO	PB
Knots	LPML	LPML	LPML
3	-575.129	-575.329	-576.733
4	-574.875	-575.224	-578.029
5	-575.119	-575.400	-577.806
6	-575.561	-575.853	-578.437
7	-575.369	-575.567	-578.011
8	-574.589	-574.719	-576.503
9	-574.573	-574.229	-575.104
10	-574.473	-573.915	-574.885
11	-575.514	-575.186	-575.023
12	-573.621	-574.464	-575.122
13	-573.948	-574.063	-574.713
14	-574.669	-574.885	-576.303
15	-574.312	-574.926	-575.871
16	-574.078	-574.280	-574.301
17	-573.870	-575.189	-575.854
18	-574.471	-574.737	-574.739

Table 2.5: Results of fibroids data analysis: Posterior Mean (Mean) and 95% Credible Interval (CI) of fibroids when applying equispaced knots and degree 3 for monotone spline for three proposed models.

Model	PH		PO		PB	
Risk Factors	Mean	95%CI	Mean	95%CI	Mean	95%CI
Race	1.4233	(1.0972, 1.7220)	1.5851	(1.2520, 1.9404)	0.9167	(0.6916, 1.1244)
Parity	-0.2605	(−0.5567,−0.0168)	-0.3560	(−0.7745,−0.0855)	-0.1818	(−0.3176,−0.0447)
Menarche	-0.1464	(−0.2757,−0.0106)	-0.1810	(−0.3420,−0.0530)	-0.1071	(−0.2008,−0.0211)
BMI	0.0727	(−0.2775, 0.4525)	0.1197	(−0.2604, 0.4755)	0.0701	(−0.1329, 0.2690)
LPML		-573.621		-573.915		-574.301
Optimal Knots		12		10		16

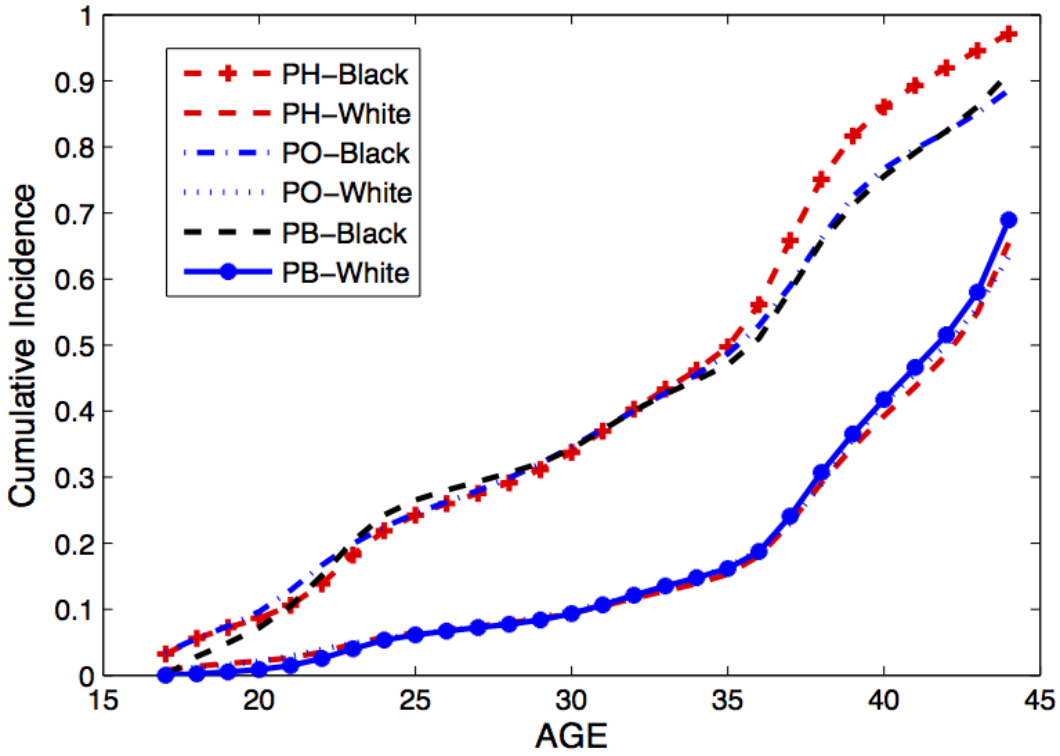

Figure 2.1: Estimated cumulative incidences functions for African and white American women under three difference semiparametric regression models.

Chapter 3

A Bayesian Approach for Regression Analysis of General Interval-censored Data under Generalized Odds-Rate Hazards Models

Summary: In this chapter, we investigate the generalized odds-rate hazards (GORH) models in the study of regression analysis of general interval-censored data. Valsecchi et al. (1996) showed that the proportional hazards assumption is often violated, particularly in the presence of long-term survival observations in clinical trials. The GORH models are a broad class of semiparametric regression models which relax the proportionality assumption for the hazard function and take the proportional hazards model and the proportional odds model as special cases. The GORH models have been widely applied in analyzing right-censored survival data, but not much in analyzing general interval-censored data. Compared to the frequentist approaches, the development of Bayesian approaches in the literature is sparse. Thus, we propos a Bayesian estimation approach for regression analysis of interval-censored data. Properties of the GORH models have been examined and a novel three-stage data augmentation has been developed for the deviation of our Gibbs sampler. The proposed Gibbs sampler is a computationally efficient because it does not involve imputations or complicated Metropolis-Hastings steps. The proposed method is illustrated by two applications to an HIV infection data and the prostate cancer screening data in the Prostate, Lung, Colorectal, and Ovarian (PLCO) study.

Keywords: Data augmentation; GORH models; interval-censored data; monotone splines; non-proportionality; semiparametric regression

3.1 Introduction

Right-censored data naturally arise when the failure time is either exactly observed or right-censored. Often in many epidemiological studies and clinical trials, participants are enrolled to the study randomly. Such a study often scheduled a periodic follow-up examinations. Each participant can be observed or examined multiple times in the study of the time to the first event. For their sake of convenience, participants may visit clinical centers on occasions rather than at pre-scheduled examination times. Participants may miss one or more pre-scheduled appointments and return with a changed of the event of interest (say some disease). In this scenario, interval-censored data arise since the failure time of interest is not exactly observed but is known to fall within some interval formed by two examination times. Interval-censored data arise in many real-life studies, such as infection and cancer studies. In this chapter, the proposed method is applied to the HIV infection data and the PLCO data which are a general interval-censored data. Such data are composed of a mixture of left-censored, right-censored, and interval-censored observations. A left-censored observation occurs when the failure time has already occurred before the first observation time. A right-censored observation occurs when the failure time is assumed to occur after the last observation time. Interval-censored observation occurs when the failure time may only be known to have taken place among two adjacent observation times.

The development of semiparametric proportional hazards (PH) regression models for regression analysis of interval-censored data remains a topic of active research interest. Many methods have been developed for regression analysis of interval-censored data. Goetghebeur and Ryan (2000) proposed an expectation-maximization (EM) al-

gorithm relying on a data augmentation of 0-1 counting processes. Finkelstein (1986) presented the maximum likelihood for the regression analysis and derived the log-rank test for the comparison of several survival curves. Satten (1996) developed a marginal likelihood approach by using a stochastic approximation scheme that was fulfilled by a Gibbs sampler. Satten et al. (1998) proposed the marginal approach by averaging overall rankings of imputed censored survival times.

Some spline-based methods are available for this topic. Cai and Betensky (2003) proposed a flexible locally parametric procedure to model the baseline log-hazard function with a piecewise-linear spline. Zhang et al. (2010) modeled the cumulative baseline hazard function with monotone B-splines and proposed a seive maximum likelihood method for the inference. Wang et al. (2016) proposed a flexible and computationally efficient EM algorithm based on the adoption of monotone I splines to approximate the unknown cumulative baseline hazard function.

The accelerated failure time (AFT) model is an alternative to the PH model. Odell et al. (1992) compared the application of an imputation technique in a Weibull-based AFT model with the use of midpoints estimates (MDEs). Rabinowitz et al. (1995) proposed an adaptive procedure based on estimating the optimal score from a class of score statistics which are used for estimating the regression coefficients. Hanson and Johnson (2004) developed a fully Bayesian nonparametric approach by utilizing a mixture of Dirichlet processes (MDP). There are also methods using the PO model for interval-censored data. Huang and Wellner (1997) presented theoretical conditions and finite-sample behavior for a sieve maximum likelihood estimator(MLE). From a Bayesian perspective, Wang and Lin (2011) proposed two Bayesian estimation methods based on two different data augmentations.

Overall, frequentist approaches for general interval-censored data are plentiful, but Bayesian methods are relatively limited. Such semiparametric regression models, however, often impose rather stringent assumptions, such as the proportionality of

hazard ratios in the PH model. Valsecchi et al. (1996) showed that the constant hazard ratio assumption is often violated, particularly in the presence of long-term survival observations in clinical trials. For example, the effect of treatment may decline or increase its effectiveness as time progresses. Many researchers turn to more flexible semiparametric regression models. One such semiparametric regression model is the generalized odds-rate hazards (GORH) models. The GORH models can be treated as the extension of the PH model to allow for non-proportional hazards (Royston and Parmar, 2002) and incorporate time varying regression coefficients (Hastie and Tibshirani, 1993; Hess, 1994). Submodels of GORH models can be specified by indexing a non-negative value ν, where the "sub" prefix refers to the fact that for each ν, a semiparametric regression model is well-specified for regression analysis of general interval-censored data. Many existing methods are applied in the estimation of regression parameter with a known ν for right-censored data under the GORH models (Scharfstein et al., 1998; Zeng et al., 2006b; Zucker and Yang, 2006). Compared to studying in right-censored data, the literature is relatively few in the GORH models for regression analysis of interval-censored data in the Bayesian framework. The goal of this chapter is to develop a Bayesian estimation approach that estimates the regression coefficients and the baseline survival function simultaneously. Section 2 provides the details of the proposed Bayesian approach, which involves a novel three-stage data augmentation, prior specification, and posterior computation. Section 3 evaluates the performance of the proposed approach through extensive simulation studies. Section 4 applies the proposed approach to two real life datasets. Section 5 provides concluding remarks and discussions.

3.2 The Proposed Method

3.2.1 Notation

It is assumed that all n subjects are independent in the study. Let T_i be the failure time for the i-th subject, and $\boldsymbol{x}_i$ be a $p \times 1$ vector of time-independent covariates for $i = 1, \cdots, n$. The failure time occurred within a certain observed interval. Such an interval is denoted as $(L_i, R_i]$ for the i-th subject, where L_i and R_i are the left- and right endpoints of the observed interval. Given the observed data $\{(L_i, R_i], \boldsymbol{x}_i, i = 1, \cdots, n\}$, the observed likelihood function can be expressed in the form,

$$L_{obs} = \prod_{i=1}^{n} \{S(L_i|\boldsymbol{x}_i) - S(R_i|\boldsymbol{x}_i)\},$$

where $S(t \mid \boldsymbol{x}_i) = P(T_i > t \mid \boldsymbol{x}_i)$ is the survival function of T_i given $\boldsymbol{x}_i$. To further distinguish different types of censoring, this observed likelihood can be rewritten as

$$L_{obs} = \prod_{i=1}^{n} \{1 - S(R_i|\boldsymbol{x}_i)\}^{\delta_{i1}} \{S(L_i|\boldsymbol{x}_i) - S(R_i|\boldsymbol{x}_i)\}^{\delta_{i2}} \{S(L_i|\boldsymbol{x}_i)\}^{\delta_{i3}}, \quad (3.1)$$

where censoring indicators $\delta_{i1}, \delta_{i2}, \delta_{i3}$ are denoted as left-, interval-, and right-censoring for the i-th subject, respectively, subject to the constraint $\delta_{i1} + \delta_{i2} + \delta_{i3} = 1$. The likelihood is constructed under an assumption that the failure time and the observational process are conditionally independent given covariates. This assumption is also called the non-informative censoring (Betensky, 2000; Sun, 2007; Turnbull, 1976; Williams and Lagakos, 1977).

3.2.2 GORH Models

The GORH models are a broad class of semiparametric regression models for analyzing time-to-event data. The survival function takes the form,

$$S(t \mid \boldsymbol{x}) = \{1 + \nu\alpha(t) \exp(\boldsymbol{x}'\boldsymbol{\beta})\}^{-\nu^{-1}}, t > 0, \nu > 0, \quad (3.2)$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)'$ is a $p \times 1$ vector of regression parameters, and $\alpha(t)$ is an unspecified nonparametric function, and ν is a non-negative constant. The charac-

teristic of nonparametric function is absolutely continuous and nondecreasing with $\lim_{t\to 0}\alpha(t) = 0$ and $\lim_{t\to\infty}\alpha(t) = \infty$, $t \in \mathbf{R}^+$. The GORH models encompass several well-known proportional survival models (Banerjee et al., 2007; Scharfstein et al., 1998). Note that the GORH model in (3.2) reduces to the PH model when $\nu \to 0$, and the GORH model in (3.2) becomes to the PO model when $\nu = 1$.

In many real-life applications, the constant hazard ratio assumption in the PH model is not realistic. The class of the GORH models is more desirable and flexible than well-known PH and PO models. The GORH models can be treated as an extension of the PH model to allow for non-proportional hazards (Royston and Parmar, 2002) and incorporate time varying regression coefficients (Hastie and Tibshirani, 1993; Hess, 1994). The GORH models can be regarded as a class of semiparametric linear transformation (LT) models (Cheng et al., 1995, 1997) in this form: $\alpha(T) = -\boldsymbol{x}'\boldsymbol{\beta} + \varepsilon_\nu$, where $\exp(\varepsilon_\nu)$ follows a generalized Pareto distribution with a known constant $\nu > 0$.

Most of the approaches in the literature for right-censored data have assumed a known ν as follows. Scharfstein et al. (1998) proposed a semiparametric efficient approach for regression analysis of right-censored data in the GORH models with a known ν. Zucker and Yang (2006) extended Yang and Prentice (1999)'s work to the generalized odds rate family and explored the inference of ν for treating as an unknown parameter and a known constant. Zucker and Yang (2006) pointed out the case of treating ν parameter as an unknown parameter, and it is necessary to assume that not all regression coefficients are identically zero. The unknown parameter ν is unidentifiable when all regression coefficients are zeros. Banerjee et al. (2007) proposed a Bayesian approach in the estimation of an unknown parameter ν in the GORH models. It is noteworthy that Banerjee et al. (2007)'s simulation results showed that a large bias in the estimation of regression parameters. Following the literature convention, we propose an estimation approach for regression analysis of

interval-censored data under the GORH models with a known ν. Submodels can be generated by taking many different ν's values in a broad class of the GORH models, and then adopted the logarithm of the pseudo-marginal likelihood (LPML) model selection criteria among the GORH models.

3.2.3 Modeling $\alpha(\cdot)$ with Monotone Splines

The estimation of model parameters under a semiparametric regression model is difficult because of the existence of the infinite-dimensional nonparametric transformation function. The unspecified nonparametric transformation function $\alpha(\cdot)$ can be modeled by a linear combination of integrated spline (I-spline) basis functions Ramsay (1988). Following the work of Lin and Wang (2010), Cai et al. (2011), Wang and Dunson (2011), Wang and Lin (2011), Lin and Wang (2011), Wang et al. (2012), and Lin et al. (2014), the proposed approach leads to the following representation,

$$\alpha(t) = \sum_{l=1}^{k} \gamma_l b_l(t), \ t \in \mathbf{R}^+, \tag{3.3}$$

where $\boldsymbol{\gamma}_l$'s are a set of non-negative spline coefficients, and $b_l(t)$'s are integrated spline (I-spline) basis functions with degree d, each of which is a non-decreasing function from 0 to 1. The shapes of the basis functions are predominantly determined by the placement of knots and the degree d of the basis function which controls the overall smoothness of the basis functions (e.g., specifying degree to be 1, 2, or 3 corresponds to the use of piecewise linear, quadratic, or cubic basis functions, respectively) (Ramsay, 1988). These spline basis functions are piecewise polynomial functions. The construction of I-spline basis functions is determined by the degree d of the basis functions and m interior knots which are chosen in an increasing sequence of knots within a time range (Ramsay, 1988). Once the placement of knots and the degree of the basis functions are specified, the k spline basis functions are fully determined, where the total number of basis functions is $k = m + d$.

The placement of knots determines the overall modeling flexibility; therefore, the

more knots that are allocated in a region of the observed data, the greater model flexibility that can be attained in that region. Lin and Wang (2010), Wang and Lin (2011), and Wang and Lin (2011) recommended using approximately 10-30 equispaced knots in the application of monotone splines for analyzing interval-censored data. Our prior specification (see Section 3.2.5) showed that Bayesian regularization can penalize excessively large knot sets by shrinking spline coefficients of those unnecessary basis functions toward zero in the use of a shrinkage prior. Therefore, the proposed method utilizes the allocation of equispaced knots with a moderate number of knots to capture curvature information of the unspecified nonparametric transformation function.

3.2.4 Data Augmentation

From Bayesian perspective, one may directly apply the sampling method on the observed likelihood (3.1) after incorporating their prior distributions. However, this approach based on the complicated observation likelihood (3.1) will result in extremely difficult computations because none of the parameters has a standard full conditional distribution. Here we propose a novel three-stage data augmentation to facilitate the posterior computation.

The first stage of the data augmentation exploits the relationship between the GORH models and the Gamma-frailty PH model. The survival function of T under the GORH models can be written as an integration of the conditional survival function of T under the Gamma-frailty PH model with respect to gamma frailty, i.e.,

$$S(t \mid \boldsymbol{x}) \;=\; \int \exp\{-\phi\alpha(t)\exp(\boldsymbol{x}'\boldsymbol{\beta})\}g(\phi)\; d\,\phi,$$

where $g(\phi)$ is the Gamma density function with both shape and rate parameter equal to ν^{-1}. Conditioning on the gamma frailty ϕ_i's, the augmented likelihood can be

expressed in the form,

$$L_{aug1}(\boldsymbol{\theta}) = \prod_{i=1}^{n} g(\phi_i)\{1 - S(t_{i1} \mid \boldsymbol{x}_i, \phi_i)\}^{\delta_{i1}}\{S(t_{i2} \mid \boldsymbol{x}_i, \phi_i)\}^{\delta_{i3}} \times [\{S(t_{i2} \mid \boldsymbol{x}_i, \phi_i)\} - \{S(t_{i1} \mid \boldsymbol{x}_i, \phi_i)\}]^{\delta_{i2}}, \tag{3.4}$$

where $\phi_i \sim g(\cdot)$, for $i = 1, \cdots, n$.

At the second stage of the data augmentation, an augmented data likelihood is established by using non-homogeneous Poisson latent variables as follows. The parameter vector is denoted as $\boldsymbol{\theta} = (\boldsymbol{\beta}', \alpha(\cdot))'$, where $\alpha(\cdot)$ are nuisance parameters for specifying the unknown non-decreasing transformation function. Let $N_i(t)$ be a non-homogeneous Poisson process with cumulative intensity function $\phi_i\alpha(t_i)\exp(\boldsymbol{x}_i'\boldsymbol{\beta})$, where ϕ_i is a Gamma frailty for the i-the subject which follows a Gamma probability distribution with the shape and rate parameters are equal to $\frac{1}{\nu}$. Let T_i be the first jump time of the counting process $N_i(t)$ for the i-th subject, for $i = 1, \cdots, n$, i.e., $T_i = \inf\{t_i : N_i(t) > 0\}$. It can be shown that T_i indeed follows the proportional hazards model with a conditional cumulative distribution function of T_i given the frailty ϕ_i, $F(t_i \mid \boldsymbol{x}_i, \phi_i) = 1 - \exp\{-\phi_i\alpha(t_i)\exp(\boldsymbol{x}_i'\boldsymbol{\beta})\} = 1 - S(t_i \mid \boldsymbol{x}_i, \phi_i)$, note for any $t_i \in (0, \infty)$ that $P(T_i > t_i \mid \boldsymbol{x}_i, \phi_i) = P(N_i(t) = 0 \mid \phi_i) = \exp\{-\phi_i\alpha(t_i)\exp(\boldsymbol{x}_i'\boldsymbol{\beta})\}$. Given that two observation time points t_{i1} and t_{i2}, $0 < t_{i1} < t_{i2} < \infty$, one can define $Z_i = N_i(t_{i1})$, where $t_{i1} = R_i 1_{(\delta_{i1}=1)} + L_i 1_{(\delta_{i1}=0)}$ and $W_i = N_i(t_{i2}) - N_i(t_{i1})$, where $t_{i2} = R_i 1_{(\delta_{i2}=1)} + L_i 1_{(\delta_{i3}=1)}$. Thus, Z_i and W_i are two Poisson random variables with mean parameter $\phi_i\alpha(t_{i1})\exp(\boldsymbol{x}'\boldsymbol{\beta})$ and $\phi_i\{\alpha(t_{i2})\exp(\boldsymbol{x}'\boldsymbol{\beta}) - \alpha(t_{i1})\exp(\boldsymbol{x}'\boldsymbol{\beta})\}$, respectively. The augmented likelihood can be expressed in the form by using the fact that two conditionally independent Poisson latent variables Z_i's and W_i's,

$$L_{aug2}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left(\mathcal{P}_{W_i}[w_i; \phi_i\{\alpha(t_{i2}) - \alpha(t_{i1})\}\exp(\boldsymbol{x}'\boldsymbol{\beta})]\right)^{\delta_{i2}+\delta_{i3}} [\mathcal{P}_{Z_i}\{z_i; \phi_i\alpha(t_{i1}) \exp(\boldsymbol{x}'\boldsymbol{\beta})\}] \times g(\phi_i) \times \{\delta_{i1}1_{(Z_i>0)} + \delta_{i2}1_{(Z_i=0,W_i>0)} + \delta_{i3}1_{(Z_i=0,W_i=0)}\}. \tag{3.5}$$

where $\mathcal{P}_A(\cdot)$ denotes the probability mass function associated with the Poisson random variable A. It is easy to see that one can obtain (3.4) by integrating out the Z_i's and W_i's of (3.5).

Lastly, at the last stage, one can exploit the monotone splines representation of $\alpha(\cdot)$ in section 3.2.3. Further, Z_i and W_i are independent given the frailty ϕ_i when $\delta_{i1} = 0$. Particularly, both Z_i's and W_i's are Poisson random variables, and then, one can decompose Z_i's and W_i's into the k's independent Poisson random variables, i.e., $Z_i = \sum_{i=1}^{n} Z_{il}$ and $W_i = \sum_{i=1}^{n} W_{il}$, for $l = 1, 2, \cdots, k$. Poisson latent variables Z_i's and W_i's have mean parameters $\phi_i\gamma_l b_l(t_{i1})\exp(\boldsymbol{x}'\boldsymbol{\beta})$ and $\phi_i\gamma_l\{b_l(t_{i2})\exp(\boldsymbol{x}'\boldsymbol{\beta}) - b_l(t_{i1})\exp(\boldsymbol{x}'\boldsymbol{\beta})\}$, respectively. The data augmented likelihood can be expressed in the form,

$$\begin{aligned} L_{aug3}(\boldsymbol{\theta}) \;=\; & \prod_{i=1}^{n} g(\phi_i) \prod_{l=1}^{k} \Big(\mathcal{P}_{W_{il}}[w_{il}; \phi_i\{b_l(t_{i2}) - b_l(t_{i1})\}\exp(\boldsymbol{x}'\boldsymbol{\beta})]\Big)^{\delta_{i2}+\delta_{i3}} [\mathcal{P}_{Z_{il}}\{z_{il}; \\ & \phi_i b_l(t_{i1})\exp(\boldsymbol{x}'\boldsymbol{\beta})\}]\{\delta_{i1}1_{(Z_{il}>0)} + \delta_{i2}1_{(Z_{il}=0,W_{il}>0)} + \delta_{i3}1_{(Z_{il}=0,W_{il}=0)}\}\,, \end{aligned} \tag{3.6}$$

where $Z_i > 0$ if $\delta_{i1} = 1$, $Z_i = 0, W_i > 0$ if $\delta_{i2} = 1$, and $Z_i = 0, W_i = 0$ if $\delta_{i3} = 1$, $Z_i = \sum_{l=1}^{k} Z_{il}$ and $W_i = \sum_{l=1}^{k} W_{il}$. Integrating out the latent variables Z_{il}'s, W_{il}'s, conditioning on frailty ϕ_i in (3.6) leads to the augmented likelihood (3.5) in the second stage. Consequently, the augmented data likelihood (3.6) can be viewed as the complete data likelihood with all the Z_i's, Z_{il}'s, W_i's, W_{il}'s, and ϕ_i's being regarded as missing data.

3.2.5 Prior Specification and Posterior Computation

Regression coefficient β_j is assigned a vague univariate normal prior $\pi(\beta_j) = \mathcal{N}(\beta_{j0}, \sigma_{j0}^2)$ by large σ_{j0}^2 , for $j = 1, 2, \cdots, p$. This leads to a log-concave conditional posterior distribution for each β_j , which can use automatic sampling method, such as the adaptive rejection sampling (ARS) (Gilks and Wild, 1992). Motivated by the widely

used double exponential prior in the Bayesian LASSO regression (Park and Casella, 2008), an exponential prior $\mathcal{E}(\eta)$ is assigned independently for all spline coefficients γ_l's. A Gamma prior $\mathcal{G}(a_\eta, b_\eta)$ is assigned for η with mean a_η/b_η and variance a_η/b_η^2. Such a prior specification is appealing because this prior specification allows to borrow of information among γ_l's and to shrink the spline coefficients of those unnecessary basis functions toward zero. This property allows us to use many knots to provide adequate modeling flexibility without additional computational costs and avoids over-fitting issues. Such prior specifications are equivalent to a penalized likelihood approach with a penalty on the sum of those nonnegative spline coefficients from a frequentist perspective. However, such penalized likelihood approach needs to select a proper tuning parameter with more computational costs by using generalized cross-validation (GCV). In contrast, our approach treats the tuning parameter λ as a parameter and allows to update it within the Bayesian posterior computation. We present details on the posterior full conditionals of the Gibbs sampler as below:

1. Sample Z_i's, Z_{il}'s, W_i's, and W_{il}'s for $i = 1, 2, \cdots, n$ and for $l = 1, 2, \cdots, k$. Initially, all set to zero. For each i,

   (1.1) if left-censored, i.e. $\delta_{i1} = 1$:

$$\begin{aligned} Z_i \mid \boldsymbol{\gamma}_l, \boldsymbol{\beta}_g, \boldsymbol{\phi}_i, \boldsymbol{x}_i &\sim \mathcal{P}(\phi_i \alpha(R_i) \exp(\boldsymbol{x}_i'\boldsymbol{\beta})) I_{(Z_i>0)}, \\ (Z_{i1}, \cdots, Z_{ik} \mid Z_i) &\sim \mathcal{M}(Z_i, \boldsymbol{p}_i), \text{ with } \boldsymbol{p}_i = (p_{i1}, p_{i2}, \cdots, p_{ik}), \\ p_{il} &= \frac{\gamma_l b_l(R_i)}{\sum_{s=1}^k \gamma_s b_s(R_i)} \text{ for } l = 1, 2, \cdots, k, \end{aligned}$$

   (1.2) if interval-censored, i.e. $\delta_{i2} = 1$:

$$\begin{aligned} W_i \mid \boldsymbol{\gamma}_l, \boldsymbol{\beta}_g, \boldsymbol{\phi}_i, \boldsymbol{x}_i &\sim \mathcal{P}(\phi_i \{\alpha(R_i) - \alpha(L_i)\} \exp(\boldsymbol{x}_i'\boldsymbol{\beta})) I_{(W_i>0)}, \\ (W_{i1}, \cdots, W_{ik} \mid W_i) &\sim \mathcal{M}(W_i, \boldsymbol{q}_i), \text{ with } \boldsymbol{q}_i = (q_{i1}, q_{i2}, \cdots, q_{ik}), \\ q_{il} &= \frac{\gamma_l b_l(R_i) - \gamma_l b_l(L_i)}{\sum_{s=1}^k \gamma_s b_s(R_i) - \gamma_s b_s(L_i)} \text{ for } l = 1, 2, \cdots, k. \end{aligned}$$

   where $\mathcal{M}(\cdot, \cdot)$ is multinomial distribution.

2. Sample β_j, for $j = 1, 2, \cdots, p$, from the following full conditional

$$\exp\Big[\sum_{i=1}^{n} \boldsymbol{x}_i'\boldsymbol{\beta}(z_i\delta_{i1} + w_i\delta_{i2}) - \sum_{i=1}^{n} \phi_i \exp\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)\{\alpha(R_i)(\delta_{i1} + \delta_{i2}) + \alpha(L_i)\delta_{i3}\}\Big]\pi(\beta_j),$$

using ARS or ARMS.

3. Sample γ_l from $\mathcal{G}(s_{\gamma_l}, r_{\gamma_l})$ for $l = 1, 2, \cdots, k$, where

$$\begin{aligned} s_{\gamma_l} &= 1 + \sum_{i=1}^{n}\{Z_{il}\delta_{i1} + W_{il}\delta_{i2}\}, \\ r_{\gamma_l} &= \eta + \sum_{i=1}^{n} \phi_i \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\{I_l(R_i)(\delta_{i1} + \delta_{i2}) + I_l(L_i)\delta_{i3}\}. \end{aligned}$$

4. Sample η from $\mathcal{G}(a_\eta + k,\ b_\eta + \sum_{l=1}^{k} \gamma_l)$.

5. Sample ϕ_i from $\mathcal{G}(a_{\phi_i}, b_{\phi_i})$ $i = 1, 2, \cdots, n$, where

$$\begin{aligned} a_{\phi_i} &= \nu + Z_i\delta_{i1} + W_i\delta_{i2}, \\ b_{\phi_i} &= \nu + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\{\alpha(R_i)(\delta_{i1} + \delta_{i2}) + \alpha(L_i)\delta_{i3}\}. \end{aligned}$$

The proposed Gibbs sampler is computationally efficient because all parameters and latent variables can be updated either from standard conjugate distribution or by using an automatic sampling method, such as ARS or ARMS.

3.3 Simulation Evidence

A simulation study was conducted to assess the performance of the proposed approach across several settings: 250 datasets with 100 observations per dataset. We generated T_i from the following GORH model and datasets are independent. The true survival function of the failure time T_i is

$$S(t \mid x_{i1}, x_{i2}, \nu) = \{1 + \nu\alpha(t)\exp(x_{i1}\beta_1 + x_{i2}\beta_2)\}^{-\nu^{-1}},$$

where $\alpha(t) = \log(1 + t) + \exp(t) - 1$, true ν takes 0.5, 1, 2, or 4, and the covariates $x_{i1} \sim \texttt{Bernoulli}(0.5)$ and $x_{i2} \sim \mathcal{N}(0, 1)$, for $i = 1, 2, \cdots, 250$. True β_1 and β_2 take

on the values $\{1, 0, -1\}$ and $\{0, 1\}$, respectively. Observation times were generated through an independent observational process for interval-censored failure time. The total number of observation times for each subject was generated according to 1 plus a Poisson random variable with mean 2. The gap times between two adjacent observations were sampled according to an exponential distribution with mean 1. This allowed at least one observation time for each subject and different subjects are allowed to have different numbers of observations. An increasing sequence of the observation times were generated by the cumulative sums of the gap times. Two endpoints for the i-th observed interval, L_i and R_i, were determined by examining which of the observation times bounded the failure time T_i with the convention that if T_i was smaller (larger) than the smallest (largest) observation time, then $L_i = 0 (R_i = \infty)$. The average right censoring rate (`CR`) of the 100 datasets varies from 8.2% to 43.75% across all settings.

For the monotone spline specifications, we took 20 equispaced knots within the minimum and maximum values of the observed interval excluding 0 and $+\infty$ and used the degree 3 of basis function for adequate smoothness. To implement the posterior computation, we use independent normal priors with $\beta_{j0} = 0$ and $\sigma_{j0}^2 = 10^2$ for $j = 1$ and 2, independent exponential distribution with rate 1 for γ_l, and a hyper gamma prior $\mathcal{G}(1, 1)$ for λ. The results were based on every 10th sample out of the total 12,000 iterations in Markov Chain Monte Carlo (MCMC) output after discarding the first 2000 iterations as a burn-in. The estimates of regression coefficients are shown in Table 3.1. `BIAS` is the average of the 100 posterior means minus the true value; `ESD` is the mean of the estimated standard deviation from their posterior distributions across 100 datasets; `SSD` is the sample standard deviation of the 100 point estimates; and `CP95` is the 95% coverage probability (i.e., the proportion of the 95% credible intervals which cover the true value of the parameter). As seen in Table 3.1, the bias is small if any for all regression parameters in all the configurations under the GORH models.

It is observed that the sample standard deviation `SSD` and the estimated standard error `ESE` are quite close. The 95% coverage probability for β_1 and β_2 are close to the nominal level 0.95 in all parameter configurations. In a sensitivity analysis, we ran additional simulation studies to investigate the effect of the hyperparameters by using more vague priors, $(a_\eta, b_\eta) = (0.1, 0.1)$ and $(0.01, 0.01)$, respectively. The results of this sensitivity analysis as seen in Table 3.2 and Table 3.3 demonstrate that the proposed method is robust, and suggest that taking $a_\eta = b_\eta = 1$ is not overly informative for the proposed approach. Thus, the proposed method is promising in the estimation of the regression coefficients under the GORH models.

3.4 Real Data Application

3.4.1 Human Immunodeficiency Virus (HIV) Infection Data

The study of HIV infection incidence was conducted in a 16 multicenter hemophilia in the United States and Europe from 1978 to 1990 (Goedert et al., 1989; Kroner et al., 1994; Kulkarni et al., 2003; Mccarthy et al., 2014). In 11 centers participants were seropositve, and in the remaining five centers, participants were recruited independent of diagnosis of HIV status. Each hemophilia patient who enrolled in the multicenter hemophilia cohort study was at-risk of HIV infection by being transfused in annual doses from plasma donors.

The primary research goal is to compare the HIV infection rates between these dose groups and to quantify the dose effect. Risk factor concentrates for recipients were categorized as high (> 50,000 U), moderate (20,001-50,000 U), low (1-20,000 U), and none (baseline) annual dose levels. The screening assays included licensed, commercially available whole-virus enzyme-linked immuno-assays (ELISAs). Most blood samples representing the last negative and first positive test were confirmed by Western blot or two radio-immuno-precipitation assays (Kroner et al., 1994; Robey et al., 1985). The onset time of infectivity was unknown for the development of de-

tectable HIV antibody, but was known to lie within a time window(Cai and Betensky, 2003; Goggins et al., 1999; Sun, 1995). Of 544 participants, the dataset contained 74 high-dose, 102 moderate-dose, 132 low-dose and 236 none-dose recipients. The summary of the number of participants in the study and censoring rates among 544 participants is: 63 (11.58%) were left-censored, 204 (37.50%) were interval-censored, and 277 (50.92%) were right-censored.

We applied the proposed method with different ν values, the numbers of interior knots, and the degree of basis functions. The optimal model is when $\nu = 1$ with the degree 3 of basis function, that is, the PO model with cubic basis function. As seen in Table 3.4, all the differences among these dose groups are significant because their 95% credible intervals do not cover zero. To provide the interpretation, the estimated odds of infection for the low dose group is $\exp(1.99) \approx 7.31$ times for that of baseline group.

3.4.2 PLCO dataset

The Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial sponsored by the U.S. National Cancer Institute (NCI) is a randomized two-arm trial designed to estimate the risk factors of cancer-related mortality. The participants were aged 55 through 74 years without cancer history of any PLCO cancer prior to enrollment in the screening program. Also, they were not participating in any other cancer screening and/or primary prevention trials. They were randomly assigned to either the intervention or the control arm at the time of enrollment. Male participants in the intervention arm received annual serum prostate-specific antigen (PSA) tests of screenings for prostate cancers during the first 6 years and were followed for an additional 7 years. In contrast, participants randomized to the control arm were offered no interventions and received their normal medical care. All their diagnoses of cancer, deaths, and causes of death were ascertained by an annual follow-up questionnaire

and periodic linkage to the National Death Index. Participants in the control arm followed up 13 years after enrollment or until December 31, 2009, whichever came earlier. A positive test was defined as a PSA level greater than four ng/mL. Then a prostate biopsy was applied to these patients to determined whether or not they had developed prostate cancer. For more details about the PLCO Cancer Screening Trial see Andriole et al. (2012).

The primary research interest focuses on the prostate cancer screening data collected in male participants in the intervention arm. The response variable of this study is the time to onset of prostate cancer. The onset time of prostate cancer was not observed, but was known to lie in two adjacent screenings because of the design of the study and the diagnosis mechanism of prostate cancer. Of 32720 observations having complete covariate information, 7 (0.02%) were left-censored, 2853 (8.7%) were interval-censored, and 29860 (91.3%) were right-censored. The covariates of interests: ethnic groups (`Caucasian` as baseline, `African American`, and other), education level (`1: college` or `above; 0: high school` or below), obesity (`BMI` > 30), heart attack, stroke, diabetes, ulcerative colitis, hepatitis, use Aspirin or Ibuprofen regularly (`1: Yes,0: No`), and standardized age at randomization (mean age: 62.6 years and standard deviation age: 5.30). A statistical summary of risk factors is listed in Table 3.5.

Our approach applied to PLCO dataset with the known values of ν in a set $\{0.5, 1, 2, 4\}$. The nonparametric transformation function was modeled using different values of the degree d of I-spline basis functions. A candidate knot set is in the list: 3, 5, 10, 15, 20, 25, 30, and 35. We adopted the LPML model selection criteria among the GORH models, and the optimal model was found when using 30 equal spaced interior knots in the time range (0,9.99). The summary of the estimated regression coefficients was shown in Table 3.5. The significant risk factors were ethnic group, family history, diabetes, and age at randomization, all of which are associated with

the development of prostate cancer; the other factors are not significant. Particularly, African American ethnicity, family history, and age at randomization were found to be positively associated with the odds of developing prostate cancer; on the other hand, diabetes and other American were negatively associated with the odds of developing prostate cancer. Our results showed that the estimated odds of developing prostate cancer for African American participant is $\exp(0.554) = 1.74$ times that of white American, holding all other risk factors at the same level.

3.5 Discussion

In this chapter, we propose a Bayesian estimation approach for analyze general interval-censored data under flexible semiparametric GORH models. Our approach is based on a novel three-stage data augmentation. The first stage of the data augmentation exploits the relationship between the GORH models and the Gamma-frailty PH model. At the second stage of the data augmentation, we expanded the likelihood using Poisson latent variables from a latent non-homogeneous Poisson process, which takes failure time of interest as the first jump time. At the last stage, we refined the augmented likelihood as a product of Poisson mass functions through introducing more conditional independent Poisson latent variables. The proposed method is a fully Bayesian approach and facilitates monotone splines to develop a flexible parametric formulation of semiparametric GORH models. The use of monotone splines reduces the computational costs even if too many knots are specified in the model in addition to providing a smooth estimate of the baseline survival function. Also, our method adapts a shrinkage prior to estimate of spline coefficients and simultaneously, prevents over-fitting issues that may cause excessively large knots.

A model selection procedure based on the log pseudo-marginal likelihood is proposed to handle the case in which ν is unknown. Simulation and real-life applications show that the proposed approach is robust to the choice of the number of knots and

the degree of I-spline basis functions. The proposed approach has enormous computational advantages for general interval-censored data over the existing Bayesian methods in the literature.

Table 3.1: Estimation of the regression parameters (β_1, β_2) based on 100 simulated datasets.

				Results on β_1				Results on β_2			
ν	CR†	β_1	β_2	Bias	ESE	SSD	CP95	Bias	ESE	SSD	CP95
0.5	09.95%	1	1	0.011	0.270	0.254	0.95	−0.024	0.152	0.157	0.93
	08.20%	1	0	0.005	0.272	0.286	0.91	−0.014	0.129	0.113	0.97
	12.94%	0	1	0.003	0.242	0.222	0.95	−0.029	0.143	0.144	0.99
	10.91%	0	0	0.013	0.236	0.241	0.95	0.027	0.119	0.130	0.93
	18.18%	−1	1	0.072	0.242	0.253	0.94	−0.045	0.137	0.124	0.95
	16.66%	−1	0	0.028	0.235	0.234	0.95	−0.001	0.116	0.125	0.91
1	13.55%	1	1	0.032	0.307	0.353	0.94	−0.023	0.169	0.177	0.93
	11.90%	1	0	0.036	0.308	0.343	0.90	−0.015	0.152	0.156	0.95
	17.81%	0	1	0.042	0.285	0.292	0.92	−0.002	0.159	0.151	0.96
	15.59%	0	0	−0.003	0.278	0.288	0.94	0.003	0.141	0.143	0.96
	22.61%	−1	1	0.060	0.279	0.287	0.94	−0.043	0.156	0.148	0.94
	20.77%	−1	0	0.014	0.273	0.236	0.99	0.027	0.136	0.123	0.98
2	21.72%	1	1	0.032	0.370	0.424	0.90	−0.013	0.203	0.200	0.93
	20.14%	1	0	0.051	0.380	0.388	0.95	0.027	0.189	0.197	0.96
	25.06%	0	1	0.043	0.355	0.373	0.95	−0.017	0.191	0.198	0.96
	24.16%	0	0	0.054	0.356	0.374	0.94	0.003	0.179	0.167	0.97
	30.97%	−1	1	0.074	0.391	0.342	0.96	−0.020	0.209	0.229	0.93
	29.74%	−1	0	−0.012	0.381	0.361	0.96	−0.007	0.194	0.198	0.95
4	35.24%	1	1	0.093	0.478	0.493	0.95	−0.067	0.240	0.259	0.90
	34.79%	1	0	0.024	0.486	0.457	0.96	0.003	0.249	0.267	0.91
	39.20%	0	1	0.044	0.466	0.470	0.95	−0.023	0.241	0.250	0.91
	39.02%	0	0	0.070	0.474	0.499	0.94	0.035	0.240	0.242	0.94
	43.75%	−1	1	0.068	0.460	0.437	0.99	−0.029	0.240	0.219	0.98
	43.37%	−1	0	0.075	0.460	0.446	0.95	−0.036	0.232	0.228	0.93

† right censoring rate

Table 3.2: Sensitivity Analysis: the estimated regression coefficients (β_1, β_2) for a Gamma hyper prior with parameters $(a_\eta, b_\eta) = (0.1, 0.1)$.

				Results on β_1				Results on β_2			
ν	CR	β_1	β_2	Bias	ESE	SSD	CP95	Bias	ESE	SSD	CP95
0.5	09.95%	1	1	0.016	0.274	0.263	0.96	−0.013	0.154	0.168	0.94
	08.20%	1	0	0.000	0.267	0.291	0.91	−0.011	0.128	0.112	0.98
	12.94%	0	1	−0.008	0.242	0.231	0.93	−0.027	0.143	0.151	0.95
	10.91%	0	0	−0.018	0.280	0.293	0.94	0.005	0.142	0.145	0.96
	18.18%	−1	1	0.055	0.241	0.247	0.95	−0.040	0.138	0.128	0.97
	16.66%	−1	0	0.014	0.237	0.236	0.94	−0.000	0.116	0.125	0.91
1	13.55%	1	1	−0.004	0.300	0.344	0.93	−0.021	0.166	0.176	0.90
	11.90%	1	0	0.028	0.310	0.346	0.92	−0.020	0.154	0.157	0.96
	17.81%	0	1	0.021	0.289	0.303	0.93	0.013	0.163	0.148	0.96
	15.59%	0	0	−0.014	0.283	0.292	0.95	0.005	0.142	0.145	0.96
	22.61%	−1	1	0.037	0.282	0.308	0.92	−0.036	0.155	0.146	0.96
	20.77%	−1	0	−0.016	0.275	0.242	0.99	0.028	0.135	0.125	0.96
2	21.72%	1	1	0.014	0.382	0.423	0.92	0.009	0.201	0.210	0.94
	20.14%	1	0	0.048	0.379	0.398	0.91	0.024	0.189	0.197	0.93
	25.06%	0	1	0.007	0.358	0.379	0.91	−0.011	0.189	0.197	0.92
	24.16%	0	0	0.022	0.358	0.379	0.95	0.002	0.179	0.166	0.97
	30.97%	−1	1	0.009	0.354	0.371	0.95	0.009	0.190	0.185	0.96
	29.74%	−1	0	0.008	0.350	0.383	0.92	0.020	0.175	0.189	0.95
4	35.24%	1	1	0.074	0.491	0.569	0.93	−0.043	0.238	0.262	0.90
	34.79%	1	0	−0.058	0.479	0.298	0.94	−0.007	0.241	0.258	0.91
	39.20%	0	1	0.001	0.473	0.469	0.98	−0.009	0.246	0.252	0.93
	39.02%	0	0	−0.016	0.476	0.494	0.96	0.034	0.240	0.246	0.94
	43.75%	−1	1	0.041	0.469	0.411	0.97	−0.016	0.242	0.271	0.94
	43.37%	−1	0	0.051	0.461	0.488	0.93	−0.044	0.234	0.238	0.94

Table 3.3: Sensitivity Analysis: the estimated regression coefficients (β_1, β_2) for a Gamma hyper prior with parameters $(a_\eta, b_\eta) = (0.01, 0.01)$.

				Results on β_1				Results on β_2			
ν	CR	β_1	β_2	Bias	ESE	SSD	CP95	Bias	ESE	SSD	CP95
0.5	09.95%	1	1	0.003	0.269	0.270	0.96	−0.018	0.150	0.162	0.92
	08.20%	1	0	0.001	0.271	0.293	0.93	−0.013	0.129	0.114	0.97
	12.94%	0	1	−0.010	0.245	0.232	0.95	−0.018	0.144	0.162	0.93
	10.91%	0	0	0.008	0.237	0.240	0.97	0.026	0.119	0.131	0.93
	18.18%	−1	1	0.045	0.245	0.256	0.96	−0.035	0.140	0.128	0.95
	16.66%	−1	0	0.012	0.238	0.238	0.95	0.001	0.116	0.126	0.91
1	13.55%	1	1	0.014	0.305	0.344	0.92	−0.023	0.167	0.170	0.90
	11.90%	1	0	0.029	0.312	0.351	0.92	−0.016	0.155	0.156	0.96
	17.81%	0	1	0.025	0.286	0.301	0.92	0.008	0.163	0.159	0.96
	15.59%	0	0	−0.020	0.278	0.297	0.93	0.004	0.142	0.144	0.97
	22.61%	−1	1	0.034	0.284	0.301	0.94	−0.035	0.157	0.145	0.96
	20.77%	−1	0	−0.018	0.277	0.235	0.99	0.028	0.135	0.124	0.96
2	21.72%	1	1	0.005	0.369	0.437	0.92	−0.001	0.191	0.198	0.93
	20.14%	1	0	0.036	0.377	0.386	0.95	0.026	0.191	0.207	0.91
	25.06%	0	1	0.004	0.363	0.395	0.93	−0.005	0.188	0.189	0.93
	24.16%	0	0	0.011	0.355	0.386	0.94	0.004	0.178	0.167	0.98
	30.97%	−1	1	−0.017	0.358	0.390	0.93	0.021	0.192	0.182	0.94
	29.74%	−1	0	0.003	0.351	0.391	0.94	−0.019	0.174	0.188	0.95
4	35.24%	1	1	−0.002	0.481	0.484	0.93	−0.042	0.244	0.267	0.93
	34.79%	1	0	−0.010	0.499	0.548	0.91	−0.000	0.254	0.256	0.95
	39.20%	0	1	−0.027	0.469	0.469	0.94	0.025	0.247	0.266	0.93
	39.02%	0	0	−0.020	0.488	0.498	0.96	0.034	0.244	0.248	0.93
	43.75%	−1	1	0.037	0.461	0.410	0.95	−0.008	0.247	0.284	0.92
	43.37%	−1	0	0.043	0.464	0.500	0.92	−0.037	0.233	0.228	0.96

Table 3.4: The estimated covariate effects and their corresponding 95% Credible Intervals from the proposed approach using quadratic and cubic splines and the number of knots 10 in the analysis of HIV data.

		d=2		d=3	
ν	Covariate	$\overline{\text{Mean}}$	95% CI	$\overline{\text{Mean}}$	95% CI
0.5	High	3.99	(3.43,4.55)	4.01	(3.49,4.55)
	Medium	3.31	(2.82,3.77)	3.35	(2.90,3.83)
	Low	1.78	(1.35,2.23)	1.78	(1.38,2.22)
	High-Medium	0.68	(0.18,1.14)	0.66	(0.21,1,14)
	Medium-Low	1.52	(1.10,1.95)	1.56	(1.14,1.98)
	LPML	-526.97		-525.43	
1	High	4.78	(4.13,5.42)	4.74	(4.12,5.38)
	Medium	3.88	(3.30,4.46)	3.83	(3.30,4.40)
	Low	2.04	(1.55,2.50)	1.99	(1.54,2.46)
	High-Medium	0.89	(0.33,1.51)	0.90	(0.37,1.48)
	Medium-Low	1.84	(1.31,2.36)	1.84	(1.26,2.39)
	LPML	-526.15		-524.30	
2	High	6.26	(5.19,7.18)	6.06	(5.26,7.02)
	Medium	4.91	(4.13,5.67)	4.71	(4.02,5.58)
	Low	2.59	(1.98,3.16)	2.46	(1.82,3.08)
	High-Medium	1.35	(0.44,2.16)	1.34	(0.54,2.06)
	Medium-Low	2.31	(1.60,3.04)	2.24	(1.62,2.96)
	LPML	-528.54		-527.28	
4	High	7.56	(6.80,8.39)	7.91	(7.01,9.31)
	Medium	6.06	(5.18,6.99)	5.97	(5.36,6.65)
	Low	3.33	(2.51,4.14)	3.22	(2.46,4.00)
	High-Medium	1.49	(0.57,2.49)	1.93	(1.14,2.98)
	Medium-Low	2.73	(1.83,3.58)	2.75	(1.94,3.56)
	LPML	-538.62		-539.17	

Table 3.5: Regression parameter estimates and their associated estimated standard error and 95% credible interval under GORH models by using quadratic basis function with 30 equally spaced knots in the analysis of PLCO data.

	$d = 2$											
Covariates	$\nu = 0.5$			$\nu = 1$			$\nu = 2$			$\nu = 4$		
Race (Afr.)*	0.536	(0.342,	0.720)	0.565	(0.379,	0.752)	0.580	(0.376,	0.811)	0.644	(0.399,	0.869)
Race (Oth.)*	−0.370	(−0.603,	−0.156)	−0.378	(−0.607,	−0.146)	−0.408	(−0.676,	−0.160)	−0.471	(−0.707,	−0.204)
Education	0.020	(−0.072,	0.110)	0.029	(−0.075,	0.128)	0.052	(−0.068,	0.162)	0.055	(−0.096,	0.189)
Obesity	−0.095	(−0.199,	0.020)	−0.101	(−0.226,	0.019)	−0.103	(−0.218,	0.012)	−0.086	(−0.241,	0.044)
Heart	−0.065	(−0.201,	0.066)	−0.056	(−0.208,	0.088)	−0.073	(−0.240,	0.059)	−0.038	(−0.198,	0.157)
Stroke	−0.176	(−0.487,	0.134)	−0.171	(−0.493,	0.131)	−0.186	(−0.570,	0.195)	−0.048	(−0.391,	0.333)
Diabetes*	−0.457	(−0.644,	−0.267)	−0.485	(−0.690,	−0.295)	−0.480	(−0.680,	−0.262)	−0.523	(−0.738,	−0.303)
Colitis	−0.104	(−0.612,	0.363)	−0.164	(−0.708,	0.291)	−0.140	(−0.641,	0.296)	−0.103	(−0.761,	0.391)
Hepatitis	−0.134	(−0.371,	0.128)	−0.140	(−0.425,	0.114)	−0.156	(−0.442,	0.101)	−0.184	(−0.491,	0.108)
Aspirin	−0.020	(−0.116,	0.076)	−0.024	(−0.115,	0.071)	0.001	(−0.108,	0.104)	−0.014	(−0.135,	0.115)
Ibuprofen	0.037	(−0.072,	0.145)	0.038	(−0.079,	0.143)	0.044	(−0.072,	0.158)	0.033	(−0.117,	0.161)
Family Hist.*	0.456	(0.313,	0.590)	0.485	(0.340,	0.615)	0.490	(0.330,	0.653)	0.616	(0.433,	0.788)
Age*	0.314	(0.265,	0.363)	0.328	(0.280,	0.378)	0.342	(0.288,	0.394)	0.376	(0.313,	0.444)
LPML	-9514			-9530			-9579			-9705		
	$d = 3$											
Race (Afr.)*	0.526	(0.331,	0.719)	0.554	(0.364,	0.767)	0.577	(0.348,	0.835)	0.648	(0.489,	0.854)
Race (Oth.)*	−0.365	(−0.573,	−0.134)	−0.384	(−0.623,	−0.156)	−0.420	(−0.638,	−0.199)	−0.424	(−0.666,	−0.180)
Education	0.023	(−0.069,	0.117)	0.027	(−0.061,	0.107)	0.048	(−0.045,	0.142)	0.110	(0.028,	0.241)
Obesity	−0.094	(−0.201,	0.023)	−0.100	(−0.230,	0.014)	−0.097	(−0.210,	0.031)	−0.060	(−0.188,	0.038)
Heart	−0.061	(−0.200,	0.070)	−0.062	(−0.215,	0.088)	−0.071	(−0.224,	0.095)	−0.054	(−0.205,	0.098)
Stroke	−0.153	(−0.463,	0.147)	−0.169	(−0.496,	0.145)	−0.138	(−0.522,	0.169)	−0.105	(−0.413,	0.190)
Diabetes*	−0.461	(−0.650,	−0.286)	−0.471	(−0.646,	−0.285)	−0.502	(−0.729,	−0.296)	−0.524	(−0.736,	−0.320)
Colitis	−0.119	(−0.632,	0.343)	−0.140	(−0.669,	0.335)	−0.133	(−0.742,	0.421)	−0.083	(−0.630,	0.389)
Hepatitis	−0.116	(−0.361,	0.128)	−0.146	(−0.429,	0.114)	−0.123	(−0.389,	0.131)	−0.199	(−0.499,	0.060)
Aspirin	−0.019	(−0.102,	0.075)	−0.018	(−0.123,	0.081)	−0.015	(−0.118,	0.091)	0.033	(−0.070,	0.121)
Ibuprofen	0.027	(−0.079,	0.127)	0.030	(−0.072,	0.136)	0.034	(−0.093,	0.147)	0.049	(−0.040,	0.149)
Family Hist.*	0.465	(0.316,	0.617)	0.479	(0.336,	0.624)	0.475	(0.328,	0.617)	0.595	(0.427,	0.748)
Age*	0.315	(0.270,	0.359)	0.326	(0.278,	0.375)	0.345	(0.300,	0.397)	0.384	(0.334,	0.444)
LPML	-9430			-9421			-9504			-9633		

* is denoted as the significant factor.

Chapter 4

Regression Analysis of Arbitrarily Censored Survival Data Under the Semiparametric Probit Model

Summary: Arbitrarily censored observations naturally appear when participants are under continuous monitoring at multiple specific time windows. If the survival event occurs within one such window, the failure time is exactly observed; otherwise, the incidence of such a survival event cannot be known; such an observation attributes to interval-censored data (strictly speaking, the observation could be left-censored, interval-censored, or right-censored). Such types of survival data are very generic and take both so-called right-censored data and interval-censored data as special cases. A Bayesian estimation approach is proposed to analyze such data under the semiparametric probit model. Specifically, monotone splines are used to approximate the unknown non-decreasing function in the model to reduce the number of parameters while maintaining adequate modeling flexibility. A novel two-stage data augmentation is developed with two sets of latent variables in order to facilitate posterior computation. The proposed Gibbs sampler is easy to implement because all the latent variables and parameters can be updated either from standard distributions or from automatic adaptive-rejection sampling. Simulation results suggest that the proposed method has good performance for estimating both the regression parameters and the baseline cumulative distribution function. Our method is illustrated with a real-life application to a dataset about diabetic nephropathy (DN).

Keywords: Arbitrary censored data; monotone splines; probit; right-censored

4.1 Introduction

Survival data, also known as time-to-event data, commonly arise in numerous real-life studies in various fields. The survival time of interest can be exactly observed if a participant is under continuous monitoring and meanwhile the survival event is symptomatic (such as death, power loss, or bankruptcy). Survival data usually contain incomplete observations due to the limitation of the study design. For example, right-censoring occurs if the survival event for a subject has not occurred at the end of a study or at the last examination time. Interval-censored data occur in studies where the survival event is asymptotic, and each participant is examined at multiple discrete times. In this chapter, we will study survival data with arbitrarily censored data. Such data naturally appear when participants are under continuous monitoring only at multiple specific time windows. If the survival event occurs within one such window, the survival time is exactly observed; otherwise, the incidence of such a survival event cannot be known; such an observation attributes to interval-censored data (more strictly speaking, the observation could be left-censored, interval-censored, or right-censored). Such types of survival data are very generic and take both so-called right-censored data and interval-censored data as special cases. We proposed a Bayesian estimation method for regression analysis of arbitrarily censored survival data in the semiparametric probit model. In many epidemiological studies and clinical trials, there is a need for regression models that can accommodate complex survival data, for example, arbitrarily censored data.

Take this hypothetical study as one case. In an animal experiment, a lab technician exposes experimental animals to toxic substances for the purpose of tolerance to assess how long animals can live after being exposed to toxic substances. The lab technician monitors the animal's time of death only within some specific time

windows. If the animal dies within such a time window, the animal's death time is exactly observed. On the other hand, if the animal dies beyond such a time window, the animal's death time is interval-censored. Such data are a mixture of exactly observed and interval-censored observations.

In the early study of arbitrarily censored data, researchers focused on the estimation of survival curves. For example, Peto (1973) proposed the Newton-Raphson method to estimate the (experimental) survival curve. Turnbull (1976) proposed a self-consistency algorithm (a.k.a. EM algorithm) for a nonparametric estimation of a survival curve with arbitrarily grouped, censored, and truncated data. For arbitrarily censored data, regression analysis has also been conducted under various survival models, for example, the PH and PO models. Joly et al. (1998) presented a penalized likelihood method for analyzing interval-censored data and left truncated data under the PH model. Kim (2003) used a maximum likelihood approach for analyzing arbitrarily censored data under the PH model. Zhang and Davidian (2008) developed a flexible quasi-parametric approach for arbitrarily censored data and introduced a family of distributions based on a class of polynomials to fit AFT, PH, and PO models. Zhou et al. (2015) developed a generalized AFT spatial frailty model for arbitrarily censored data.

The aim of this chapter is to develop a Bayesian estimation approach for arbitrarily censored data under the semiparametric probit model. The remainder of the chapter is organized as follows. Section 2 presents the notations in the proposed model, the property of monotone splines, the proposed Bayesian approach, and a likelihood-based method. Section 3 shows the intensive simulation results. Section 4 applies the proposed method to the Steno Memorial Hospital Diabetic Nephropathy dataset.

4.2 The Proposed Method

4.2.1 Notation

The survival time (or failure time) random variable is denoted as T. The cumulative distribution function of the survival time given the covariate vector $\boldsymbol{x}$ is denoted as $F(\cdot \mid \boldsymbol{x})$, and the corresponding probability density function of the survival time given the covariate vector is denoted as $f(\cdot \mid \boldsymbol{x})$. Survival time (or failure time) data are composed of either exact or interval-censored observations. The observed interval in time-to-event data is denoted as $[L, R]$, where L and R are denoted as the left and right bounds of the observed interval, respectively, with the constraint $L \leq R$. The observed interval $[R, R]$ is represented as the exact observation when L and R coincide. The survival time is left-censored with the observed interval $(0, R]$ if $L = 0$. The survival time is right-censored with the observed interval $[L, +\infty)$ if $R = +\infty$. Otherwise, the survival time is interval-censored with the observed interval $[L, R]$. Suppose that n independent subjects are observed on their survival times. Let $\boldsymbol{x}_i$ be a $p \times 1$ vector of time-independent covariates. It is assumed in this chapter that conditional on the covariates, the survival time and the observational process are independent (Liu and Shen, 2009; Sun, 2007). The structure of such data presents in the form $\{[L_i, R_i], \boldsymbol{x}_i\}_{i=1}^n$. The observed likelihood function can be expressed in the form,

$$L_{obs} \;=\; \prod_{i=1}^{n} \{f(R_i \mid \boldsymbol{x}_i)\}^{I(L_i=R_i)} [\{F(R_i|\boldsymbol{x}_i) - F(L_i|\boldsymbol{x}_i)\}]^{I(L_i<R_i)}.$$

To fully express the completely observed likelihood function, one can rewrite the observed likelihood in the following form by further distinguishing the censoring types,

$$L_{obs} \;=\; \prod_{i=1}^{n} \{f(R_i \mid \boldsymbol{x}_i)\}^{\delta_{i0}} \{F(R_i|\boldsymbol{x}_i)\}^{\delta_{i1}} \{F(R_i|\boldsymbol{x}_i) - F(L_i|\boldsymbol{x}_i)\}^{\delta_{i2}} \{1 - F(L_i|\boldsymbol{x}_i)\}^{\delta_{i3}},$$

where indicator variables $\delta_{i0}, \delta_{i1}, \delta_{i2}$, and δ_{i3} are used for the i-th subject denoting exactly observed, left-, interval-, and right-censored observation, respectively, subject

to the constraint $\delta_{i0} + \delta_{i1} + \delta_{i2} + \delta_{i3} = 1$.

4.2.2 Probit Model

The proposed semiparametric probit model specifies the cumulative distribution function (CDF) of survival time T in the form,

$$F(t \mid \boldsymbol{x}) = \Phi\{\alpha(t) + \boldsymbol{x}'\boldsymbol{\beta}\},\ t \in \mathbf{R}^+ ,$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)'$ is a $p \times 1$ vector of regression parameters. The corresponding probability density function (pdf) is in the form,

$$f(t; \boldsymbol{x}) = \phi\{\alpha(t) + \boldsymbol{x}'\boldsymbol{\beta}\}\alpha'(t),\ t \in \mathbf{R}^+ .$$

Here Φ and ϕ are the CDF and pdf of the standard normal distribution. A Probit model is a special case of semiparametric Linear Transformation (LT) models, $\alpha(T) = -\boldsymbol{x}'\boldsymbol{\beta} + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 1)$. The commonly used models in a class of LT models are the PH and PO models where random error ε_i are assumed to independently follow an extreme value distribution and a standard logistic distribution, respectively (Chen et al., 2002; Ma and Kosorok, 2005; Younes and Lachin, 1997). Many research papers have been published in the PH model and PO model for regression analysis of arbitrarily censored data (Joly et al., 1998; Kim, 2003; Zhang and Davidian, 2008; Zhou et al., 2015). There are no papers to investigate a semiparametric probit model for arbitrarily censored data to the best of our knowledge.

4.2.3 Modeling $\alpha(\cdot)$ with Monotone Splines

The estimation of model parameters under a semiparametric regression model is difficult because of the existence of the infinite-dimensional nonparametric transformation function. The unspecified nonparametric transformation function $\alpha(\cdot)$ can be modeled by a linear combination of integrated spline (I-spline) basis functions (Ramsay, 1988). Following the work of Lin and Wang (2010), Cai et al. (2011), Wang and

Dunson (2011), Wang and Lin (2011), Lin and Wang (2011), Wang et al. (2012), and Lin et al. (2014), the proposed approach leads to the following representation,

$$\alpha(t) = \gamma_0 + \sum_{l=1}^{k} \gamma_l b_l(t) \;\; t \in \mathbf{R}^+, \tag{4.1}$$

where γ_0 is an unconstrained intercept of a monotone spline, $\boldsymbol{\gamma}_l$ are a set of non-negative spline coefficients, and $b_l(t)$'s are integrated spline (I-spline) basis functions with degree d, each of which is a non-decreasing function from 0 to 1. The shapes of the basis functions are predominantly determined by the placement of knots and the degree d of the basis function which controls the overall smoothness of the basis functions (e.g., specifying degree to be 1, 2, 3 or a higher degree corresponds to the use of piecewise linear, quadratic, cubic or a higher order of polynomial basis functions, respectively) (Ramsay, 1988). These spline basis functions are piecewise polynomial functions. The construction of I-spline basis functions is determined by the degree d of the basis functions and m interior knots which are chosen in an increasing sequence of knots within a time range (Ramsay, 1988). Once the placement of knots and the degree of basis functions are specified, the k spline basis functions are fully determined, where the total number of basis functions is $k = m + d$.

The placement of knots determines the overall modeling flexibility; therefore, the more knots that are allocated in a region of the observed data, the greater model flexibility that can be attained in that region. Lin and Wang (2010),Wang and Lin (2011), and Wang and Lin (2011) recommended using approximately 10-30 equispaced knots in the application of monotone splines for analyzing interval-censored data. Our prior specification (see Section 4.2.4) showed that Bayesian regularization can penalize excessively large knot sets by shrinking spline coefficients of those unnecessary basis functions toward zero in the use of a shrinkage prior. Therefore, the proposed method utilizes the allocation of the equispaced knots with a moderate number of knots to capture curvature information of the unspecified nonparametric transformation function. For the purpose of estimating the probability density function of survival

time t, the first derivative of nonparametric transformation function $\alpha'(t)$ can be expressed in the form:

$$\alpha'(t) \;=\; \sum_{l=1}^{k} \gamma_l b_l'(t \mid d) \;\; t \in \mathbf{R}^+, \tag{4.2}$$

where the derivatives of $b_l'(t)$'s are also referred to as M-spline basis functions in the literature.

4.2.4 Posterior Computation

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \alpha(\cdot))'$, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)'$ denotes the corresponding vector of regression parameters, and $\alpha(\cdot)$ are nuisance parameters for specifying the unknown non-decreasing transformation function. Lin and Wang (2010) developed a Bayesian estimation method for analyzing general interval-censored data under the semiparametric probit model, and they introduced the following normal latent variables in their data augmentation,

$$Z_i \;\sim\; N(\alpha(t_i) + \boldsymbol{x}_i'\boldsymbol{\beta}, 1)\Big\{1_{(Z_i>0)}\Big\}^{\delta_{i1}}\Big\{1_{(\alpha(L_i)-\alpha(R_i)<Z_i<0)}\Big\}^{\delta_{i2}}\Big\{1_{(Z_i<0)}\Big\}^{\delta_{i3}},$$

where $t_i = R_i I(\delta_{i1} = 1) + L_i I(\delta_{i1} = 0)$ for $i = 1, \cdots, n$. Following their idea, we introduced a normal latent variable Z_i for each i in the same manner but incorporate the case of exactly observed failure time by restricting $z_i = 0$ when $\delta_{i0} = 1$. With the new data augmentation, the augmented likelihood function can be expressed as follows,

$$\begin{aligned} L_{aug1} \;&=\; \prod_{i=1}^{n}\{\phi(Z_i - \alpha(t_i) - \boldsymbol{x}_i'\boldsymbol{\beta})\} \times \Big\{\alpha'(t_i)\Big\}^{\delta_{i0}} \times \Big\{1_{(Z_i=0)}\Big\}^{\delta_{i0}} \\ &\times\; \Big\{1_{(Z_i>0)}\Big\}^{\delta_{i1}}\Big\{1_{(\alpha(L_i)-\alpha(R_i)<Z_i<0)}\Big\}^{\delta_{i2}}\Big\{1_{(Z_i<0)}\Big\}^{\delta_{i3}} \qquad (4.3) \\ &=\; \prod_{i=1}^{n}\{\phi(Z_i - \alpha(t_i) - \boldsymbol{x}_i'\boldsymbol{\beta})\} \times \Big\{\sum_{l=1}^{k}\gamma_l b_l'(t)\Big\}^{\delta_{i0}} \\ &\times\; \Big\{1_{(Z_i=0)}\Big\}^{\delta_{i0}} \times \Big\{1_{(Z_i>0)}\Big\}^{\delta_{i1}} \times \Big\{1_{(\alpha(L_i)-\alpha(R_i)<Z_i<0)}\Big\}^{\delta_{i2}} \\ &\quad \Big\{1_{(Z_i<0)}\Big\}^{\delta_{i3}}. \qquad (4.4) \end{aligned}$$

It is difficult to handle the summation of a linear combination of M-spline basis functions when $\delta_{i0} = 1$. Thus, one can introduce a set of multinomial latent variables, $\boldsymbol{v}_i = (v_{i1}, v_{i2}, \cdots, v_{ik})$ on the augmented likelihood function with

$$\boldsymbol{v}_i \sim \mathcal{M}(1, \mathbf{p}_i) ,$$

$\mathcal{M}(\cdot, \cdot)$ is multinomial distribution, and

$$\mathbf{p}_i = \left(\frac{1}{k}, \frac{1}{k}, \cdots, \frac{1}{k}\right).$$

Thus, the augmented likelihood function can be extended in this form,

$$\begin{aligned} L_{aug2} &= \prod_{i=1}^{n} \{\phi(Z_i - \alpha(t_i) - \boldsymbol{x}_i'\boldsymbol{\beta})\} \times k \left\{ \prod_{l=1}^{k} \gamma_l b_l'(t_i) \right\}^{\delta_{i0} v_{il}} \\ &\times \left\{1_{(Z_i=0)}\right\}^{\delta_{i0}} \times \left\{1_{(Z_i>0)}\right\}^{\delta_{i1}} \times \left\{1_{(\alpha(L_i)-\alpha(R_i)<Z_i<0)}\right\}^{\delta_{i2}} \\ &\quad \left\{1_{(Z_i<0)}\right\}^{\delta_{i3}} . \end{aligned} \tag{4.5}$$

Regression coefficient $\boldsymbol{\beta}$ is assigned a multivariate normal prior $\pi(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0^{-1})$. The intercept of the monotone spline γ_0 is unconstrained and is assigned a conventional vague normal prior for $\pi(\boldsymbol{\gamma}_0) = \mathcal{N}(\mu_0, \nu_0)$ with a large ν_0. Motivated by the widely used double exponential prior in Bayesian LASSO regression (Park and Casella, 2008), one can assign independent exponential priors $\mathcal{E}(\lambda)$ for the nonnegative spline coefficients (the $\boldsymbol{\gamma}_l$'s). A gamma prior $\mathcal{G}(a_\lambda, b_\lambda)$ is assigned for λ with mean a_λ/b_λ and variance a_λ/b_λ^2. Such a prior specification is appealing because this prior specification allows us to borrow of information among γ_l's and to shrink the spline coefficients of those unnecessary basis functions toward zero. This property allows us to use many knots to provide adequate modeling flexibility without additional computational costs and avoids over-fitting issues that potentially caused due to the use of excessively large number of knots. Such prior specifications are equivalent to a penalized likelihood approach with a penalty on the sum of those nonnegative spline coefficients from a frequentist perspective. However, such a penalized likelihood approach needs to select a proper tuning parameter with more computational costs by

using generalized cross-validation (GCV). In contrast, our approach treats the tuning parameter λ as a parameter and allows to update it within the Bayesian posterior computation. A Markov Chain Monte Carlo (MCMC) algorithm iterates through the following steps.

1. For each i, sample latent variable.

   - Exactly-observed observation, i.e. $\delta_{i0} = 1$, let $Z_i = 0$.
   - Left-censored observation, i.e. $\delta_{i1} = 1$:

$$Z_i \sim \mathcal{N}(Z_i; \alpha(t_i) + \boldsymbol{x}_i'\boldsymbol{\beta}, 1)1_{(Z_i>0)}.$$

   - Interval-censored observation, i.e. $\delta_{i2} = 1$:

$$Z_i \sim \mathcal{N}(Z_i; \alpha(t_i) + \boldsymbol{x}_i'\boldsymbol{\beta}, 1)1_{(\alpha(L_i)-\alpha(R_i)<Z_i<0)}.$$

   - Right-censored observation, i.e. $\delta_{i3} = 1$:

$$Z_i \sim \mathcal{N}(Z_i; \alpha(t_i) + \boldsymbol{x}_i'\boldsymbol{\beta}, 1)1_{(Z_i<0)}.$$

2. For $\delta_{i0} = 1$, sample latent variables $\boldsymbol{v}_i = (v_{i1}, v_{i2}, \cdots, v_{ik})$ from a multinomial distribution.

$$(v_{i1}, v_{i2}, \cdots, v_{ik}) \sim \mathcal{M}(1, \tilde{p}_i),$$

   where

$$\tilde{p}_i = \left(\frac{\gamma_1 b_l'(t_i)}{\sum_{l=1}^k \gamma_l b_l'(t_i)}, \frac{\gamma_2 b_2'(t_i)}{\sum_{l=1}^k \gamma_l b_l'(t_i)}, \cdots, \frac{\gamma_k b_k'(t_i)}{\sum_{l=1}^k \gamma_l b_l'(t_i)}\right),$$

   where $\mathcal{M}(\cdot, \cdot)$ is multinomial distribution.

3. Sample γ_0 from $\mathcal{N}(E_0, W_0^{-1})$, where

$$\begin{aligned} E_0 &= W_0^{-1}\Big[\nu_0 m_0 + \sum_{i=1}^n \Big\{\boldsymbol{Z}_i - \sum_{l=1}^k \gamma_l b_l(t_i) - \boldsymbol{x}_i'\boldsymbol{\beta}\Big\}\Big]. \\ W_0 &= n + \nu_0. \end{aligned}$$

4. Sample γ_l's using ARS or ARMS for each $l = 1, \cdots, k$, from

   - The full conditional distribution of γ_l is

$$\pi(\boldsymbol{\gamma}_l \mid \cdot) \propto \exp\left(-\frac{1}{2}\Big[\gamma_l^2 W_l - 2\gamma_l \sum_{i=1}^{n} b_l(t_i)\Big\{\boldsymbol{Z}_i - \gamma_0 - \sum_{j\neq l}\gamma_j b_j(t_i) - \boldsymbol{x}_i'\boldsymbol{\beta}\Big\}\Big] - \lambda\gamma_l\right)\gamma_l^{\sum_{i=1}^{n} v_{il}\delta_{i0}} 1(\gamma_l > d_l^*),$$

   where

   $W_l = \sum_{i=1}^{n} b_l^2(t_i)$, $d_l^* = \max(c_l^*, 0)$,

$$c_l^* = \max_{i:\delta_{i2}=1}\Big[\frac{-Z_i - \sum_{j\neq l}\gamma_j\{b_j(R_i) - b_j(L_i)\}}{b_l(R_i) - b_l(L_i)}\Big].$$

   Note that if $W_l = 0$, sample γ_l from $\mathcal{G}(\sum_{i=1}^{n} v_{il}\delta_{i0} + 1, \lambda)1(\gamma_l > d_l^*)$.

5. Sample $\boldsymbol{\beta}$ from $\mathcal{N}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\Sigma}}_\beta)$, where

$$\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\Sigma}}_\beta\Big[\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0 + \sum_{i=1}^{n}\boldsymbol{x}_i\Big\{\boldsymbol{Z}_i - \alpha(t_i)\Big\}\Big].$$
$$\tilde{\boldsymbol{\Sigma}}_\beta = \Big\{\boldsymbol{\Sigma}_0^{-1} + \sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\Big\}^{-1}.$$

6. Sample λ from $\mathcal{G}(a_\lambda + k, b_\lambda + \sum_{l=1}^{k}\gamma_l)$.

This method is a Bayesian Gibbs sampler and applies the adaptive rejection Metropolis sampling (ARMS) method to sample $\boldsymbol{\gamma}_l$'s because the full conditional distribution is log-concave at the fourth step.

4.2.5 Likelihood-Based Method

An existing method, such as a likelihood-based approach is applied to the same dataset for the purpose of comparison with the proposed Bayesian approach. Under the same specification of monotone spline (4.1)-(4.2), one can apply the maximum likelihood method to estimate a finite number of $(1 + k + p)$ parameters. Denote

$\boldsymbol{\theta} = (\boldsymbol{\beta}', \gamma_0, \boldsymbol{\gamma}')'$ as the unknown model parameter vector, where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_k)'$. Finding the maximum likelihood estimates $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ subject to the constrained parameter can be done by using a constrained optimization package, such as `fmincon` in `Matlab` and `nloptr` in `R`. Maximum likelihood method maximizes the observed likelihood function for estimating regression coefficients and then the variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ can be computed by the inverse of the observed information matrix, denoted as $\boldsymbol{I}^{-1}(\hat{\boldsymbol{\theta}})$. The observed information matrix can be approximated by numerical differentiation in the following form (Lin and Wang, 2010; Zeng et al., 2006a),

$$\boldsymbol{I}(i,j) \approx -\frac{\{\log l(\widehat{\boldsymbol{\theta}} + h_n\mathbf{1}_i + h_n\mathbf{1}_j) - \log l(\widehat{\boldsymbol{\theta}} + h_n\mathbf{1}_i) - \log l(\widehat{\boldsymbol{\theta}} + h_n\mathbf{1}_j) + \log l(\boldsymbol{\theta})\}}{h_n^2},$$

where $\mathbf{1}_i = \begin{bmatrix} 0 & 0 & \cdots & 1^{(i)} & 0 & \cdots \end{bmatrix}'$ and tuning constant $h_n = o(n^{-0.5})$, for $i = 1, \cdots, k+p+1$. In the simulation study, not all datasets attain the convergent numerical roots; rather, the optimization likelihood estimates sometimes fail to converge. As a result, the maximum likelihood method encounters the over-fitting issues for such datasets when specifying a large number of spline coefficients of monotone splines in the semiparametric regression model.

4.3 Simulation Evidences

An intensive simulation study was conducted to assess the performance of the proposed approach across several settings. In the first simulation scenario, the true cumulative distribution function of the survival time T_i was taken to be,

$$F(t \mid x_{i1}, x_{i2}) = \Phi\{\alpha(t) + x_{i1}\beta_1 + x_{i2}\beta_2\},$$

where $\alpha(t) = \log(t) + t^3 - 1$, and the covariates are $x_{i1} \sim \mathcal{N}(0,1)$, and $x_{i2} \sim$ `Bernoulli`(0.5), for $i = 1, \cdots, 200$. True β_1 takes on the values $\{1, 0, -1\}$, and true β_2 takes on the values $\{1, 0\}$. The sample size was 200, and the total number of 100 datasets were independently generated. For the purpose of simulating the observed

data, the true survival time T_i was generated by solving $F(t \mid \boldsymbol{x}_i) = u_i$ numerically, where $u_i \sim \mathcal{U}(0,1)$. We considered three scenarios according to the percentage of exactly observed observations, $p = 60\%$, 20%, and 5%. This is done by generating a Bernoulli random variable ω_i with success probability p. The failure time T_i will be exactly observed if $\omega_i = 1$; otherwise, T_i is used for an interval-censored observation. For non exact observations, we generated interval-censored observations in the following manner. First, for each subject, the number of observation times was generated in accordance with 1 plus a Poisson random variable having mean value as 2 and 3. Second, the gap times between adjacent observation times were sampled according to an exponential distribution with mean value as 1, $\frac{1}{2}$, and $\frac{1}{3}$. Simultaneously, this approach allowed the number of observations varies subject to subject. An increasing sequence of the observation times were obtained by taking the cumulative sums of the gap times. The two endpoints for the i-th observed interval, L_i and R_i, were determined by examining which of the observation times bounded the survival time T_i with the shortest length, with the convention that if T_i was smaller (larger) than the smallest (largest) observation time, then $L_i = 0$ ($R_i = \infty$). The initial values of priors in the Gibbs sampler were specified: $\boldsymbol{\beta}_0 = \mathbf{0}$ and $\boldsymbol{\Sigma}_0 = n(\boldsymbol{X}'\boldsymbol{X})^{-1}$, where $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2)'$ is the covariate matrix in which $\boldsymbol{x}_1$ were the 200 observations from $\mathcal{N}(0,1)$, and $\boldsymbol{x}_2$ were the 200 observations from $\texttt{Bernoulli}(0.5)$, $m_0 = 0$ and $\nu_0 = 0.1$, which allowed high precision for γ_0, and $a_\lambda = b_\lambda = 1$, which allowed $\lambda \sim \mathcal{G}(1,1)$.

For our proposed method, the nonparametric transformation function was modeled using two settings of basis splines: degree 3 and an interior knot set having cardinality 5, and degree 4 and an interior knot set having cardinality 10 on the interval for which minimum and maximum values of the observed interval generated from 100 dataset per parameter configuration. For each dataset, a MCMC algorithm ran 5500 iterations after discarding the first 500 iterations as a burn-in and then, proceeded with a systematic sampling by taking one sample out of every 5 samples

from the Markov chain. This was judged to be sufficient, given the excellent and fast convergence rate and mixing observed in the standard normal distribution for major model parameters. Table 4.1 - 4.4 presented the frequentist operating characteristics of the estimates of the regression parameters from the proposed Bayesian approach and maximum likelihood estimates. `BIAS` is the average of the 100 point estimates (posterior means in the Bayesian approach and MLEs in the likelihood approach) minus the true value; `ESD` is the mean of the estimated standard deviation from their posterior distributions across 100 datasets; `SSD` is the sample standard deviation of the 100 point estimates; and `CP95` is the 95% coverage probability (i.e., the proportion of the 95% credible or confidence intervals which cover the true value of the parameter). In general, both methods work very well; both methods have small bias (around 0.05%), the sample standard deviation `SSD` and the estimated standard error `ESE` are quite close with a smaller variance estimates, and the coverage probability is close to the nominal value 0.95 in all parameter configurations. As the percentage of exact observations decreases in the simulated dataset, the sampled standard deviation increases, i.e., the efficiency gets worse for both methods. Further, we computed the mean square errors of the estimates of $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{F}_0$. Define a mean square error (MSE) of the estimated parameter β_i, for $i = 1, 2$ as

$$MSE\{\hat{\boldsymbol{\beta}}_j\} \;=\; \frac{1}{100}\sum_{j=1}^{100}\{\boldsymbol{\beta}_i - \hat{\boldsymbol{\beta}}_i^{(j)}\}^2 \;,$$

where $\beta_i^{(j)}$ is the posterior mean of β_i for the j-th dataset. Define a local mean squared error (MSE) of the baseline cumulative distribution function $\hat{F}_0(t)$ at time t as

$$MSE\{\hat{F}_0(t)\} \;=\; \frac{1}{100}\sum_{j=1}^{100}\{F_0(t) - \hat{F}_0^{(j)}(t)\}^2 \;,$$

where $\hat{F}_0^{(j)}(t)$ is the estimate of $F_0(t)$ in our approach for the j-th dataset. The global mean (maximum) squared error of $\hat{F}_0(t)$ denotes as $\overline{\texttt{MSE}}(F_0)$ ($\texttt{maxMSE}(F_0)$), which is

computed as the mean (maximum) of the local MSEs of $F(t)$ and is evaluated on a set of pre-specified equispaced grid points. The smaller values of global MSEs are, the more accurate estimates of the cumulative distribution function are. The results are shown in Table 4.5 - 4.6. Both methods give good estimates of the regression coefficients and the baseline survival function in all settings, which is suggested by the MSEs of the baseline CDF estimates. It is noteworthy that the maximum likelihood method cannot always produce converged results. As the more basis functions are specified (e.g. from 8 basis functions to 14 basis functions), the percentage of non-convergence increases for estimating of regression coefficients by using the maximum likelihood method.

Right-censored data is a special case of arbitrarily censored data. In the second simulation scenario, 100 right-censored datasets were generated for each parameter configuration. True β_1 takes on the values {1,0,-1} and β_2 takes on the values {1,0}, resulting in six parameter configurations. The right-censoring rate varies from 73% to 83%. The failure times were generated from the same model as in scenario 1.

$$F(t \mid x_{i1}, x_{i2}) \;=\; \Phi\{\alpha(t) + x_{i1}\beta_1 + x_{i2}\beta_2\} \;,$$

where $\alpha(t) = \log(t) + t^3 - 1$, and the covariate effects are $x_{i1} \sim \mathcal{N}(0,1)$, and $x_{i2} \sim \texttt{Bernoulli}(0.5)$, for $i = 1, \cdots, 200$. For the purpose of simulating the observed data, the true survival time T_i can be computed by solving $F(t \mid \boldsymbol{x}_i) = u_i$ numerically, where $u_i \sim \mathcal{U}(0,1)$. The observation time C was generated from a truncated exponential distribution $\mathcal{E}(2)I_{(0,10)}$. The true failure time is $T^* = \min(T, C)$, and the right-censoring indicator is equal to one when the failure time T is less than or equal to the observation time C.

The initial values of priors in the Gibbs sampler were specified as in scenario 1. For our proposed method, the nonparametric transformation function was modeled using two settings of basis splines: degree 3 and an interior knot set having cardinality 5, and degree 4 and an interior knot set having cardinality 10 on the interval

for which minimum and maximum values of the observed interval generated from 100 dataset per parameter configuration. For each dataset, a MCMC algorithm ran 11000 iterations after discarding the first 1000 iterations as a burn-in and then, proceeded with a systematic sampling by taking one sample out of every 5 samples from the Markov chain. This was judged to be sufficient, given the excellent and fast convergence rate and mixing observed in the standard normal distribution for major model parameters. Table 4.7 presented the frequentist operating characteristics of the estimates of the regression parameters from the proposed Bayesian approach and using the spline model of Royston and Parmar (2002). We apply a parametric model in the `flexsurv` R package (e.g. log normal) to compare our approach shown in Table 4.7. In general, both methods work very well; both methods have small bias (around 0.05%), `ESD`s and `SSD`s are close with a smaller variance estimates. Our method is more efficient than flexible survival regression with log normal model because of the relatively smaller sample standard deviation `SSD` and the estimated standard error `ESE`.

4.4 Real Data Application

4.4.1 Steno Memorial Hospital dataset

The Steno Memorial Hospital in Copenhagen, Denmark, served as a diabetes research hospital beginning in 1933. A study was conducted between 1933 and 1972 for patients who had been diagnosed before age 31 with insulin-dependent diabetes mellitus for Type 1 diabetes (Andersen et al., 1983). Borch-Johnsen et al. (1985)'s research showed that the development of diabetic nephropathy (DN) was regarded as a highly associated prognostic biomarker for a low survival rate in Type 1 diabetics. Insulin treatment was adopted as the primary method to treat diabetes disease in 1922, but it is still possible for a patient treated with insulin to develop DN for Type 1 diabetes.

Diabetes mellitus is a common chronic disease caused due to a disturbance of

normal production to insulin. There are two types of diabetes: Type 1 diabetes (a.k.a. insulin-dependent or juvenile-onset) and Type 2 diabetes(a.k.a. non-insulin-dependent or adult-onset). Type 1 diabetes is the most severe type and occurs primarily in patients at a young age, and Type 2 diabetes is a mild type and develops in those patients later in life. Nowadays, there are many obese children in the adolescent population; therefore, there are many young people with insulin resistance and Type 2 diabetes. Moreover, the characteristics of both Type 1 and Type 2 diabetes may be present in the same patient. Type 1 and Type 2 diabetes are difficult to distinguish.

DN is defined as persistent proteinuria and not an irreversible complication. It is mainly used to assess kidney failure, which is indicated as positive whenever a subject has at least four urine samples within 24 hours, at the time interval of at least one month, that each contain more than 0.5-gram of protein in urine. Among the 732 patients in the study, 454 were males and 278 were females, and 596 of them had exactly observed DN onset time and 136 of them had interval-censored DN onset time. Subjects are diabetic patients who either enter this study with DN or develop DN before the end of the study. The survival time was used as the basic time scale from onset of a patient's diabetes to DN onset time when they transition from having diabetes without DN to having diabetes with DN. The medical records included the following information: gender of patient, age at diabetes onset, age at the first contact with the hospital, and age at the last seen was available. The primary research interest is to assess the association of risk factors (e.g. gender and age at the onset of diabetes) with the onset of the development of DN.

Age effect can be insightful to take account of the comparison of hazards between those who are under age 10 years and those who are relatively elder in age over 10 years. We dichotomize age group using 10 years as a cutoff; i.e. define a dummy variable $x_1 = 1$ for age < 10; and $x_1 = 0$ for age ≥ 10. In addition, male and female participants are expected to have different moralities of developing DN (Andersen

et al., 2012). Thus, four subgroups are formed to assess the mortalities show in Table 4.8. The number of burn-in steps in the MCMC procedure is 5000, and the total number of iterations is 25000. We proceeded with system sampling by taking one sample out of every 10 samples from the Markov chain. As a result, the resulting MCMC samples alleviated auto-correlation, and the Markov chains were stationary. Relative to four distinct initial values at the beginning of the MCMC algorithm in the long run, the estimated posterior means for both parameters were stationary. We presented a couple plots that are used in MCMC diagnostics and for graphical summary of posterior distribution β_1, β_2 in Figures 4.1.

The estimated regression coefficients for gender and age are shown in Table 4.9. For our proposed methodology under the Probit model, both risk factors are significant at 5% significant level. The interpretation of β_1, i.e. Gender (1:male) is that given patients in the same age group above 10 years, the transform CDF of DN incidence drops by 0.244 from female to male under Probit link[†]. Interestingly, a 95% confidence interval of gender effect does not cover zero by using the maximum likelihood method under the PH model (Kim, 2003).

The relative mortality was higher in women than men at all ages for those patients who developed persistent proteinuria. This result is consistent in this literature (Andersen et al., 1983; Borch-Johnsen et al., 1985). Male participants have a relatively high survival probability (low cumulative incidence) than female participants above 10 years old. Two groups of participants were created based on their ages being under 10 or above 10. Figure 4.2 provides four combinations of gender and age group in the estimated survival curves. It is noteworthy that Figure 4.3 provides a plot of the estimated survival functions from the proposed MCMC algorithm, when m = 17, at the different levels of gender and age group, superimposed with a model free estimator,

[†]the inverse of the CDF of the standard normal distribution to transform probabilities to the standard normal variable.

Turnbull estimator (Turnbull, 1976). As seen in Figure 4.3, the estimated survival curves are very close to the Turnbull estimates for all subgroups. This suggests that the probit model provides a good fit to DN data.

4.5 Discussion

We presented a Bayesian approach for regression analysis of arbitrarily censored data under the semiparametric probit model. Our approach adopts monotone splines for the unspecified nonparametric transformation functions and allows for estimating the regression coefficients and survival curves jointly. The use of monotone splines estimates the unspecified nonparametric transformation function and provides computational efficiency while maintaining adequate modeling flexibility. In the Bayesian framework, Bayesian regularization allows us to use many knots, and this action plays a role in penalizing excessively large knot sets while shrinking spline coefficients of those unnecessary basis functions toward zero in the use of a shrinkage prior. Frequentist approaches penalize the large values of the spline coefficients by enforcing a penalty term for the spline coefficients and require use of generalized cross validation (GCV) to select a proper tuning parameter. In contrast, Bayesian regularization treats the tuning parameter as a random variable and update it within a Bayesian MCMC algorithm, in which the data will provide information for the right value of this tuning parameter. This allows for automatic tuning with much less computational effort. The proposed approach can be extended to the PO model based on the relationship between normal and logistic distribution.

Table 4.1: Estimation of regression coefficients (β_1, β_2) based on 100 simulated datasets, sample size 200 per se, basis spline function degree 3 and interior knots 5.

						Results on β_1				Results on β_2			
PEO♠	LR♣	IR♡	CR♦	β_1	β_2	Bias	ESE	SSD	CP95	Bias	ESE	SSD	CP95
60%	6.5%	23.5%	10.0%	1	1	−0.015	0.098	0.094	0.98	−0.016	0.166	0.164	0.95
	6.0%	13.5%	20.5%	1	0	−0.002	0.102	0.107	0.95	0.002	0.162	0.155	0.97
	6.5%	22.0%	11.5%	0	1	−0.005	0.079	0.084	0.92	−0.017	0.164	0.168	0.95
	2.5%	17.0%	20.5%	0	0	−0.003	0.082	0.089	0.96	−0.013	0.158	0.153	0.98
	6.0%	14.0%	20.0%	−1	1	0.022	0.099	0.103	0.94	−0.024	0.168	0.150	0.96
	7.0%	19.0%	14.0%	−1	0	0.012	0.098	0.096	0.96	−0.001	0.158	0.142	0.97
20%	12.5%	42.5%	25.0%	1	1	0.020	0.191	0.180	0.99	−0.065	0.185	0.183	0.95
	6.5%	33.5%	40.0%	1	0	−0.021	0.124	0.117	0.96	0.003	0.194	0.189	0.95
	10.5%	19.0%	50.5%	0	1	−0.012	0.102	0.096	0.97	−0.017	0.207	0.240	0.94
	6.5%	34.5%	39.0%	0	0	−0.013	0.092	0.090	0.94	−0.034	0.175	0.183	0.92
	10.5%	31.0%	38.5%	−1	1	0.012	0.120	0.131	0.92	−0.039	0.198	0.204	0.93
	9.5%	38.5%	32.0%	−1	0	−0.037	0.118	0.110	0.97	−0.000	0.177	0.198	0.93
5%	11.5%	51.5%	32.0%	1	1	−0.004	0.122	0.107	0.97	−0.017	0.200	0.209	0.93
	9.0%	32.0%	54.0%	1	0	0.021	0.141	0.152	0.94	−0.024	0.208	0.205	0.94
	16.0%	50.0%	29.0%	0	1	0.002	0.094	0.185	0.96	−0.015	0.185	0.194	0.92
	6.5%	33.0%	55.5%	0	0	0.000	0.108	0.110	0.96	−0.038	0.205	0.232	0.90
	16.0%	28.0%	51.0%	−1	1	0.040	0.133	0.137	0.93	0.026	0.205	0.230	0.94
	16.0%	18.5%	60.5%	−1	0	0.003	0.159	0.137	0.98	0.002	0.241	0.220	0.92

♠ the percentage of exact observations
♣ left-censoring rate
♡ interval-censoring rate
♦ right-censoring rate

Table 4.2: Maximumlikelihood method for the estimation of regression coefficients (β_1, β_2) based on 100 simulated datasets, sample size 200 per se, basis spline function degree 3 and interior knots 5.

						Results on β_1				Results on β_2			
COR	LR	IR	CR	β_1	β_2	Bias	ESE	SSD	CP95	Bias	ESE	SSD	CP95
60%	6.5%	23.5%	10.0%	1	1	0.010	0.099	0.093	0.96	−0.008	0.168	0.191	0.94
	6.0%	13.5%	20.5%	1	0	0.031	0.102	0.106	0.95	−0.020	0.163	0.167	0.94
	6.5%	22.0%	11.5%	0	1	−0.009	0.079	0.086	0.91	−0.003	0.166	0.193	0.92
	3.0%	6.5%	30.5%	0	0	−0.000	0.084	0.088	0.93	−0.001	0.169	0.173	0.93
	6.0%	14.0%	20.0%	−1	1	−0.002	0.100	0.091	0.98	0.014	0.172	0.194	0.94
	7.0%	19.0%	14.0%	−1	0	−0.028	0.099	0.094	0.92	−0.018	0.160	0.169	0.93
20%	12.5%	42.5%	25.0%	1	1	0.020	0.192	0.219	0.90	−0.016	0.189	0.203	0.93
	6.5%	33.5%	40.0%	1	0	0.021	0.127	0.116	0.96	0.014	0.197	0.194	0.98
	10.5%	19.0%	50.5%	0	1	0.005	0.102	0.093	0.98	0.019	0.219	0.234	0.90
	6.5%	34.5%	39.0%	0	0	−0.008	0.093	0.092	0.92	−0.008	0.184	0.191	0.94
	10.5%	31.0%	38.5%	−1	1	−0.037	0.123	0.204	0.91	0.041	0.204	0.199	0.94
	9.5%	38.5%	32.0%	−1	0	−0.040	0.119	0.116	0.92	−0.011	0.186	0.185	0.93
5%	11.5%	51.5%	32.0%	1	1	0.033	0.126	0.127	0.93	0.039	0.207	0.227	0.93
	9.0%	32.0%	54.0%	1	0	0.047	0.145	0.150	0.94	−0.006	0.220	0.202	0.95
	16.0%	50.0%	29.0%	0	1	−0.001	0.094	0.086	0.98	0.006	0.199	0.200	0.92
	6.5%	33.0%	55.5%	0	0	−0.001	0.130	0.135	0.94	0.006	0.259	0.262	0.96
	16.0%	28.0%	51.0%	−1	1	−0.058	0.137	0.156	0.89	0.060	0.224	0.243	0.94
	16.0%	18.5%	60.5%	−1	0	−0.048	0.167	0.157	0.94	0.009	0.253	0.247	0.97

Table 4.3: Bayesian method for the estimation of the regression parameters (β_1, β_2) based on 100 simulated datasets, sample size 200 per se, basis spline function degree 4 and interior knots 10.

						Results on β_1				Results on β_2			
COR	LR	IR	CR	β_1	β_2	Bias	ESE	SSD	CP95	Bias	ESE	SSD	CP95
60%	6.5%	23.5%	10.0%	1	1	−0.011	0.098	0.095	0.97	−0.015	0.164	0.165	0.95
	6.0%	13.5%	20.5%	1	0	0.001	0.101	0.107	0.96	0.001	0.160	0.156	0.97
	6.5%	22.0%	11.5%	0	1	−0.005	0.079	0.083	0.93	−0.018	0.163	0.168	0.93
	2.5%	17.0%	20.5%	0	0	−0.004	0.082	0.089	0.95	0.011	0.160	0.152	0.97
	6.0%	14.0%	20.0%	−1	1	0.022	0.099	0.103	0.94	−0.024	0.168	0.150	0.96
	7.0%	19.0%	14.0%	−1	0	0.012	0.098	0.096	0.96	−0.001	0.158	0.142	0.97
20%	12.5%	42.5%	25.0%	1	1	0.027	0.190	0.183	0.98	−0.062	0.181	0.181	0.97
	6.5%	33.5%	40.0%	1	0	−0.012	0.125	0.118	0.97	0.003	0.190	0.192	0.94
	10.5%	19.0%	50.5%	0	1	−0.011	0.102	0.095	0.96	−0.021	0.211	0.237	0.95
	6.5%	34.5%	39.0%	0	0	−0.013	0.093	0.088	0.96	−0.032	0.179	0.179	0.94
	10.5%	31.0%	38.5%	−1	1	0.003	0.119	0.133	0.92	−0.035	0.193	0.204	0.93
	9.5%	38.5%	32.0%	−1	0	0.024	0.118	0.108	0.98	−0.002	0.180	0.196	0.93
5%	11.5%	51.5%	32.0%	1	1	0.005	0.123	0.105	0.98	−0.014	0.197	0.214	0.93
	9.0%	32.0%	54.0%	1	0	0.012	0.141	0.152	0.95	−0.020	0.211	0.205	0.94
	16.0%	50.0%	29.0%	0	1	0.002	0.094	0.083	0.98	−0.027	0.188	0.189	0.94
	6.5%	33.0%	55.5%	0	0	−0.001	0.108	0.108	0.96	−0.028	0.209	0.230	0.92
	16.0%	28.0%	51.0%	−1	1	0.020	0.132	0.132	0.92	0.010	0.212	0.226	0.93
	16.0%	18.5%	60.5%	−1	0	0.005	0.160	0.134	0.99	0.002	0.240	0.217	0.97

Table 4.4: Maximumlikelihood method for the estimation of the regression coefficients (β_1, β_2) based on 100 simulated datasets, sample size 200 per se, basis spline function degree 4 and interior knots 10.

						Results on β_1				Results on β_2			
COR	LR	IR	CR	β_1	β_2	Bias	ESE	SSD	CP95	Bias	ESE	SSD	CP95
60%	6.5%	23.5%	10.0%	1	1	0.011	0.099	0.095	0.94	−0.003	0.168	0.191	0.92
	6.0%	13.5%	20.5%	1	0	0.030	0.103	0.103	0.94	−0.018	0.163	0.164	0.92
	6.5%	22.0%	11.5%	0	1	−0.009	0.079	0.087	0.91	−0.011	0.167	0.189	0.92
	3.0%	6.5%	30.5%	0	0	−0.006	0.084	0.085	0.94	0.002	0.169	0.173	0.94
	6.0%	14.0%	20.0%	−1	1	−0.007	0.100	0.091	0.97	0.020	0.172	0.193	0.93
	7.0%	19.0%	14.0%	−1	0	−0.029	0.099	0.096	0.93	−0.018	0.160	0.175	0.93
20%	12.5%	42.5%	25.0%	1	1	0.018	0.192	0.224	0.92	−0.007	0.189	0.191	0.97
	6.5%	33.5%	40.0%	1	0	0.025	0.128	0.121	0.96	0.019	0.199	0.196	0.97
	10.5%	19.0%	50.5%	0	1	0.003	0.103	0.096	0.98	0.019	0.220	0.231	0.92
	6.5%	34.5%	39.0%	0	0	−0.010	0.093	0.097	0.91	−0.023	0.185	0.197	0.94
	6.5%	33.0%	55.5%	0	0	−0.048	0.125	0.142	0.90	0.056	0.206	0.211	0.92
	9.5%	38.5%	32.0%	−1	0	−0.032	0.119	0.106	0.97	−0.023	0.187	0.189	0.94
5%	11.5%	51.5%	32.0%	1	1	0.037	0.127	0.123	0.95	0.055	0.209	0.231	0.93
	9.0%	32.0%	54.0%	1	0	0.055	0.147	0.160	0.93	−0.015	0.223	0.207	0.95
	16.0%	50.0%	29.0%	0	1	−0.007	0.095	0.090	0.97	0.009	0.200	0.189	0.95
	6.5%	33.0%	55.5%	0	0	−0.001	0.132	0.135	0.93	0.006	0.261	0.266	0.96
	16.0%	28.0%	51.0%	−1	1	−0.075	0.140	0.160	0.86	0.079	0.227	0.259	0.91
	16.0%	18.5%	60.5%	−1	0	−0.062	0.171	0.165	0.93	0.004	0.256	0.251	0.95

Table 4.5: Simulation results of three different levels of the completely observed rate dataset for concerning the estimation on the baseline cumulative distribution function F_0. Provided results include the average ($\overline{\texttt{MSE}}$) and maximum ($\texttt{maxMSE}$) mean squared errors ($\times 10^{-3}$) of the estimates of the baseline cumulative distribution function $F_0(t)$ calculated over a set of pre-specified time points. Modeling the nonparametric transformation function is based on basis spline function degree 3 and interior knots 5.

		Proposed Bayesian Method				Maximum Likelihood Method				
COR	(β_1, β_2)	$\texttt{MSE}(\beta_1)$	$\texttt{MSE}(\beta_2)$	$\overline{\texttt{MSE}}(F_0)$	$\texttt{maxMSE}(F_0)$	$\texttt{MSE}(\beta_1)$	$\texttt{MSE}(\beta_2)$	$\overline{\texttt{MSE}}(F_0)$	$\texttt{maxMSE}(F_0)$	$\texttt{NCP}^\flat\%$
60%	(1, 1)	9.08	26.89	0.87	2.56	8.74	36.47	0.24	2.34	0
	(1, 0)	11.35	23.80	0.60	1.80	12.22	28.27	0.17	1.27	0
	(0, 1)	7.16	28.49	0.76	2.14	7.54	36.99	0.21	2.04	0
	(0, 0)	7.99	23.35	0.63	1.43	7.70	29.83	0.19	1.74	2
	(-1, 1)	10.99	23.12	0.76	2.43	8.21	37.65	0.22	1.97	0
	(-1, 0)	9.33	20.00	0.68	2.05	9.67	28.84	0.20	1.61	2
20%	(1, 1)	32.53	37.53	0.88	2.81	48.00	41.20	0.25	2.59	1
	(1, 0)	14.01	35.70	1.15	3.08	13.97	37.54	0.27	1.74	0
	(0, 1)	9.30	57.42	1.35	7.45	8.77	54.75	0.37	3.42	1
	(0, 0)	8.28	34.34	0.85	4.33	8.50	36.46	0.26	2.41	1
	(-1, 1)	17.29	42.89	1.12	5.87	20.50	41.15	0.29	2.12	0
	(-1, 0)	13.52	38.83	0.94	2.55	15.11	34.03	0.25	1.66	1
5%	(1, 1)	11.47	43.69	1.20	6.66	17.26	52.91	0.38	4.09	4
	(1, 0)	23.40	42.24	1.57	4.64	24.75	40.65	0.40	2.83	1
	(0, 1)	7.22	37.64	1.05	5.46	7.36	39.62	0.30	3.12	5
	(0, 0)	11.98	55.10	1.14	3.10	18.24	68.16	0.51	3.77	2
	(-1, 1)	20.49	53.49	1.63	5.26	27.61	62.15	0.49	3.95	3
	(-1, 0)	18.65	48.26	1.82	9.96	26.99	50.77	0.57	4.44	0

$^\flat$ NCP indicates the percentage of non-convergence; it is the percentage of not applicable results among 100 datasets. The summary of the estimation of regression coefficients is based on those convergent results only; therefore, the maximum likelihood estimates of the regression parameters failed to converge in some dataset.

Table 4.6: Simulation results of three different levels of the completely observed rate dataset for concerning the estimation on the baseline cumulative distribution function F_0. Provided results include the average ($\overline{\texttt{MSE}}$) and maximum ($\texttt{maxMSE}$) mean squared errors ($\times 10^{-3}$) of the estimates of the baseline cumulative distribution function $F_0(t)$ calculated over a set of pre-specified time points. Modeling the nonparametric transformation function is based on basis spline function degree 4 and interior knots 10.

		Proposed Bayesian Method				Maximum Likelihood Method				
COR	(β_1, β_2)	$\texttt{MSE}(\beta_1)$	$\texttt{MSE}(\beta_2)$	$\overline{\texttt{MSE}}(F_0)$	$\texttt{maxMSE}(F_0)$	$\texttt{MSE}(\beta_1)$	$\texttt{MSE}(\beta_2)$	$\overline{\texttt{MSE}}(F_0)$	$\texttt{maxMSE}(F_0)$	NCP %
60%	(1, 1)	9.13	27.24	0.93	2.99	9.20	36.40	0.24	2.30	2
	(1, 0)	11.44	24.14	0.66	1.95	11.50	27.00	0.18	1.40	5
	(0, 1)	6.96	28.24	0.83	2.36	7.60	35.60	0.22	2.00	7
	(0, 0)	8.01	23.04	0.70	1.89	7.30	29.80	0.20	1.80	6
	(-1, 1)	10.97	23.78	0.83	2.79	8.30	37.30	0.23	2.00	1
	(-1, 0)	9.47	20.35	0.76	2.52	10.00	31.00	0.22	1.80	8
20%	(1, 1)	33.99	36.52	0.95	2.89	50.10	36.10	0.26	2.20	24
	(1, 0)	14.10	36.87	1.21	3.45	15.30	38.70	0.27	1.80	17
	(0, 1)	9.18	56.25	1.41	4.22	9.20	53.50	0.40	3.30	7
	(0, 0)	8.01	33.06	0.89	2.87	9.50	38.70	0.30	2.50	15
	(-1, 1)	17.54	42.75	1.17	3.09	22.40	47.50	0.35	2.30	18
	(-1, 0)	12.15	38.12	1.00	2.87	12.30	35.80	0.28	2.00	28
5%	(1, 1)	11.07	45.60	1.31	4.51	16.50	55.90	0.43	4.60	35
	(1, 0)	23.24	42.06	1.61	4.95	28.60	42.60	0.45	3.10	27
	(0, 1)	6.91	36.41	1.07	2.48	8.00	35.70	0.32	2.50	29
	(0, 0)	11.65	53.24	1.22	3.41	18.20	70.10	0.55	4.10	5
	(-1, 1)	17.82	50.83	1.72	5.52	31.00	73.00	0.57	5.30	19
	(-1, 0)	18.04	46.74	1.81	5.31	31.00	62.40	0.70	5.60	8

Table 4.7: Estimation of the regression parameters (β_1, β_2) based on 100 simulated datasets, sample size 200 per se for right-censored data.

		Proposed Method				flexsurv R			
CR	β_1 β_2	Bias	ESE	SSD	CP95	Bias	ESE	SSD	CP95
73.75%	1	0.047	0.146	0.163	0.91	0.048	0.146	0.174	0.92
	1	0.044	0.246	0.240	0.98	0.061	0.252	0.273	0.94
80.09%	1	−0.016	0.162	0.149	0.97	0.019	0.165	0.162	0.97
	0	0.016	0.255	0.249	0.96	−0.024	0.260	0.267	0.96
76.22%	0	0.013	0.117	0.115	0.95	−0.000	0.118	0.121	0.93
	1	0.028	0.251	0.228	0.97	0.054	0.258	0.238	0.96
83.21%	0	0.036	0.138	0.141	0.93	0.018	0.141	0.144	0.95
	0	0.046	0.269	0.270	0.93	−0.007	0.278	0.285	0.94
73.89%	−1	0.038	0.142	0.132	0.94	−0.013	0.145	0.149	0.94
	1	0.004	0.243	0.240	0.96	0.037	0.250	0.273	0.92
80.25%	−1	0.064	0.160	0.140	0.96	−0.020	0.166	0.161	0.97
	0	0.053	0.254	0.239	0.96	0.022	0.260	0.259	0.96

Table 4.8: Summary characteristics for Steno Memorial Hospital Diabetic Nephropathy data

	age < 10	age ≥ 10	Total
male	129	325	454
female	93	185	278
Total	222	510	732

Table 4.9: Estimated covariate effects on the DN incidence from the proposed method under the PB model and from the likelihood approach under the PH model.

Model	PB Model		PH Model	
Risk Factors	Posterior Mean	95%CI	MLE	95% CI
Gender (1:Male)	-0.244	$(-0.404,-0.070)$	-0.145	$(-0.283,-0.006)$
Age < 10	-0.259	$(-0.418,-0.105)$	-0.096	$(-0.247,\ 0.055)$

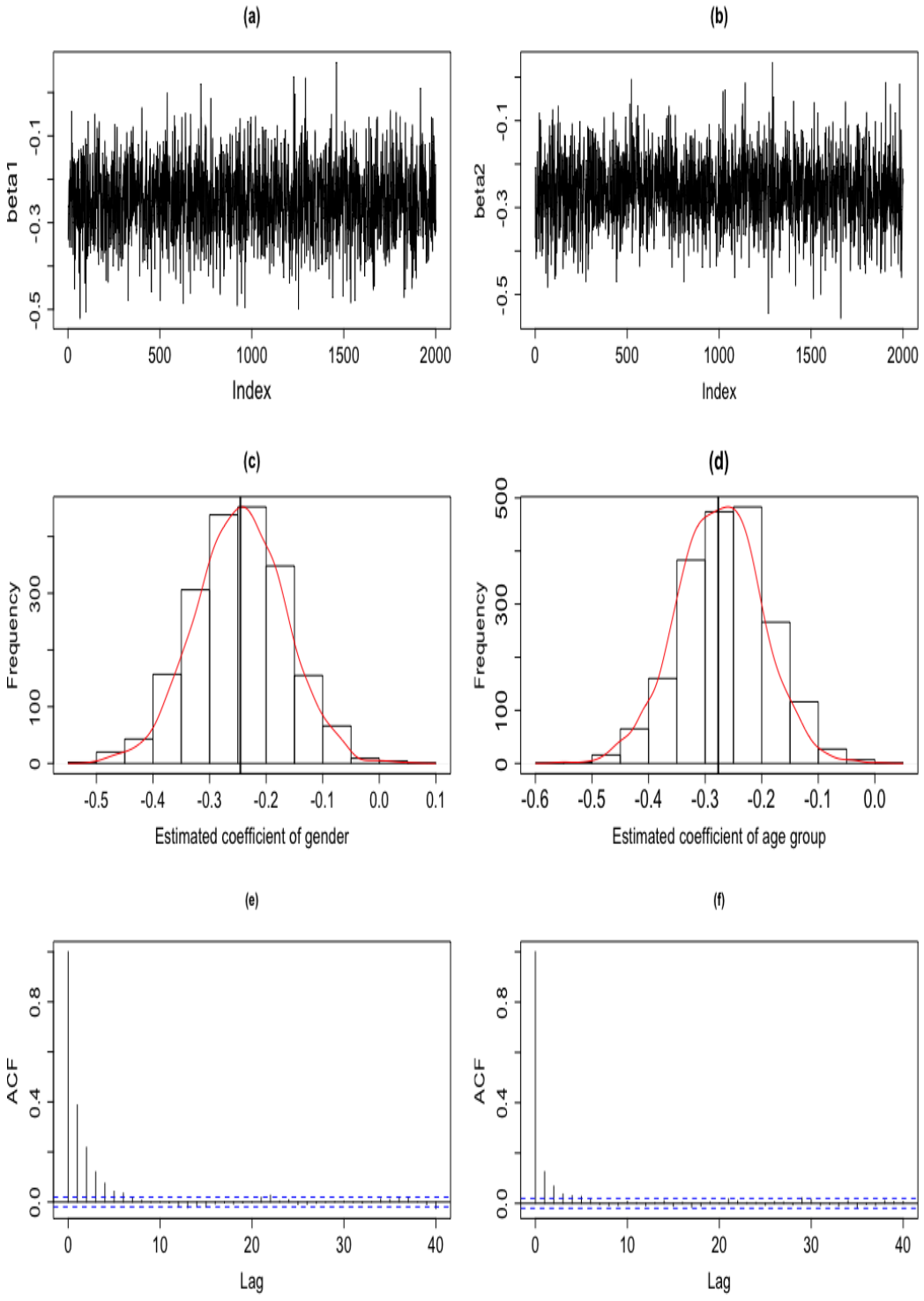

Figure 4.1: SMH diagnosis plots: Trace plot of $\hat{\beta}_1$ (a), Trace of $\hat{\beta}_2$ (b), Histogram of $\hat{\beta}_1$ (c), Histogram of $\hat{\beta}_2$ (d), Autocorrelation of $\hat{\beta}_1$ (e), and Autocorrelation of $\hat{\beta}_2$ (f).

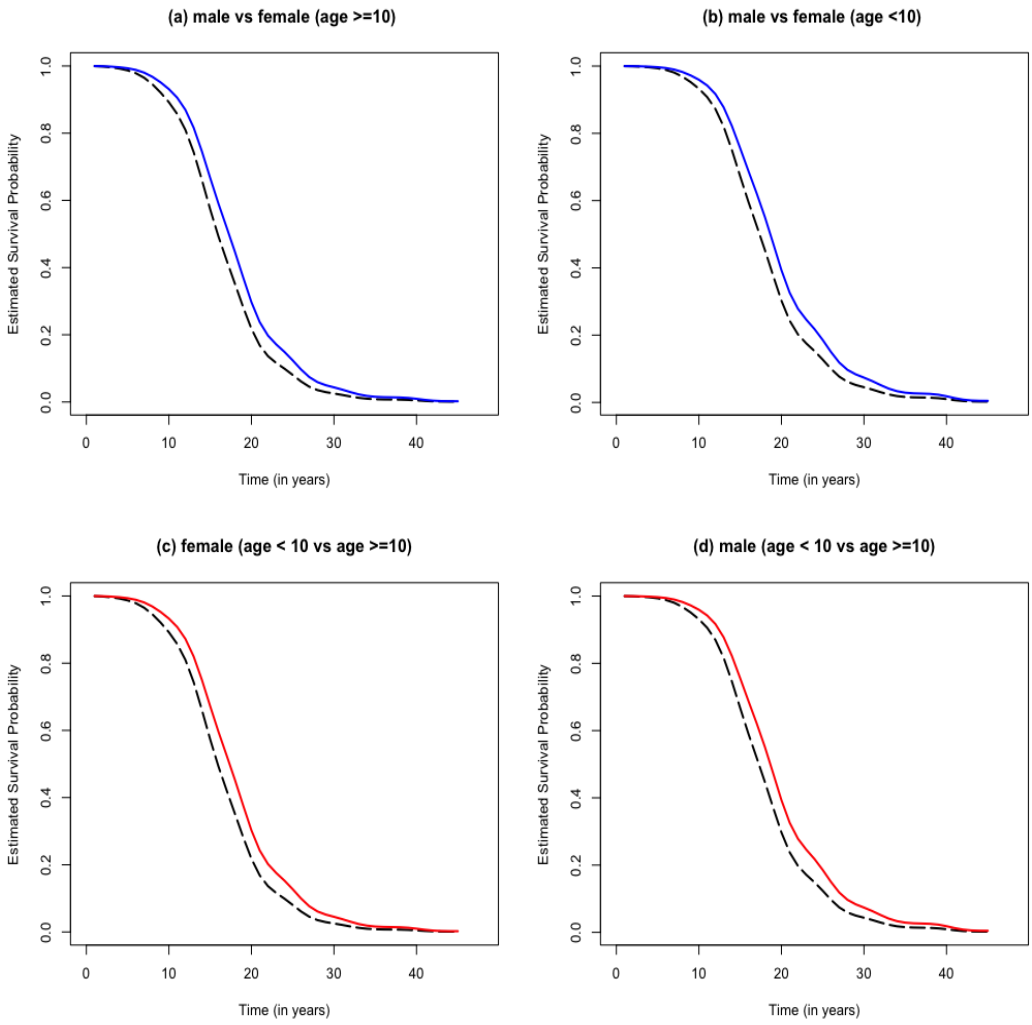

Figure 4.2: SMH data analysis: Estimates of the survival functions obtained by the proposed method (smooth red curves) at the different levels of gender and age group: Male vs. Female participants between ages 10 and 30 (a), Male vs Female participants under age 10 (b), Female participants under age 10 versus between ages 10 and 30 (c), and Male participants under age 10 versus between ages 10 and 30 (d). Smooth blue curves are the indicators for Male (a) and (b). Smooth red curves are the indicators for participants age under 10 (c) and (d).

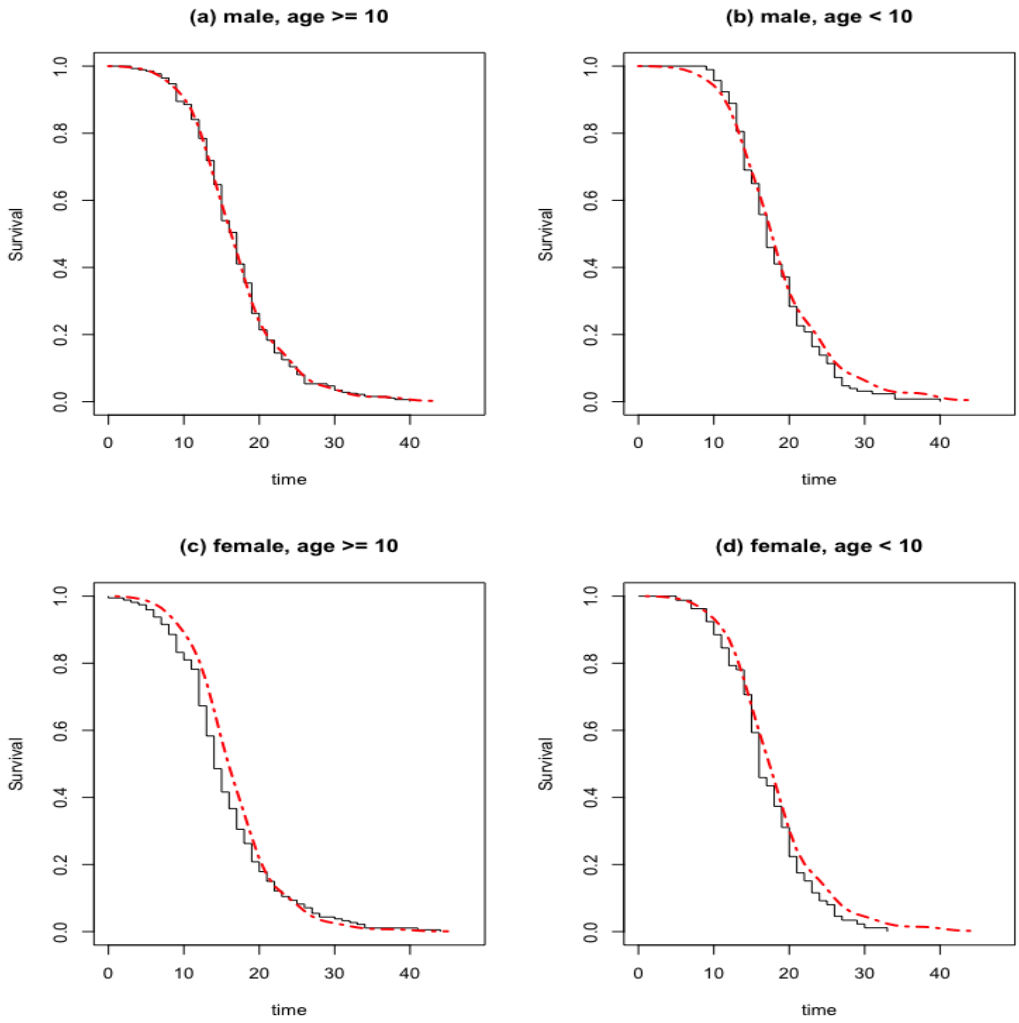

Figure 4.3: SMH data analysis: Estimates of the survival functions obtained by the proposed method (smooth red curves) and the Turnbull estimates (black step functions) at the different levels of gender and age group: Male participants above age 10 (a), Male participants under ages 10 (b), Female participants above ages 10 (c), and Female participants under age 10 (d).

Bibliography

Andersen, A. R., Christiansen, J. S., Andersen, J. K., Kreiner, S., and Deckert, T. (1983). Diabetic nephropathy in type 1 (insulin-dependent) diabetes: an epidemiological study. Diabetologia, 25(6):496–501.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (2012). Statistical models based on counting processes. Springer Science & Business Media.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. The annals of statistics, pages 1100–1120.

Andriole, G. L., Crawford, E. D., Grubb, R. L., Buys, S. S., Chia, D., Church, T. R., Fouad, M. N., Isaacs, C., Kvale, P. A., Reding, D. J., and others (2012). Prostate cancer screening in the randomized Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial: mortality results after 13 years of follow-up. Journal of the National Cancer Institute.

Banerjee, T., Chen, M.-H., Dey, D. K., and Kim, S. (2007). Bayesian analysis of generalized odds-rate hazards models for survival data. Lifetime data analysis, 13(2):241–260.

Bennett, S. (1983). Analysis of survival data by the proportional odds model. Statistics in Medicine, 2(2):273–277.

Betensky, R. A. (2000). Miscellanea. On nonidentifiability and noninformative censoring for current status data. Biometrika, 87(1):218–221.

Betensky, R. A., Rabinowitz, D., and Tsiatis, A. A. (2001). Computationally simple accelerated failure time regression for interval censored data. Biometrika, 88(3):703–711.

Boor, C. d. (1978). A Practical Guide to Splines. Applied Mathematical Sciences, New York: Springer, 1978.

Borch-Johnsen, K., Andersen, P. K., and Deckert, T. (1985). The effect of proteinuria on relative mortality in type 1 (insulin-dependent) diabetes mellitus. 28(8):590–596.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. Journal of the Royal Statistical Society. Series B (Methodological), pages 211–252.

Breiman, L. (1988). [Monotone Regression Splines in Action]: Comment. Statistical Science, 3(4):442–445.

Cai, B., Lin, X., and Wang, L. (2011). Bayesian proportional hazards model for current status data with monotone splines. Computational Statistics & Data Analysis, 55(9):2644–2651.

Cai, T. and Betensky, R. A. (2003). Hazard Regression for Interval-Censored Data with Penalized Spline. Biometrics, 59(3):570–579.

Chen, K., Jin, Z., and Ying, Z. (2002). Semiparametric analysis of transformation models with censored data. Biometrika, 89(3):659–668.

Chen, M.-H., Tong, X., and Sun, J. (2007). The proportional odds model for multivariate interval-censored failure time data. Statistics in medicine, 26(28):5147–5161.

Cheng, S. C., Wei, L. J., and Ying, Z. (1995). Analysis of transformation models with censored data. Biometrika, 82(4):835–845.

Cheng, S. C., Wei, L. J., and Ying, Z. (1997). Predicting Survival Probabilities with Semiparametric Transformation Models. Journal of the American Statistical Association, 92(437):227–235.

Cox, D. R. (1972). Regression Models and Life-Tables. Journal of the Royal Statistical Society. Series B (Methodological), 34(2):187–220.

Cox, D. R. (1975). Partial likelihood. Biometrika, 62(2):269–276.

Cox, D. R. and Oakes, D. (1984). Analysis of survival data, volume 21. CRC Press.

Dey, D. K., Chen, M.-H., and Chang, H. (1997). Bayesian approach for nonlinear random effects models. Biometrics, pages 1239–1252.

Fine, J. P., Ying, Z., and Wei, L. G. (1998). On the linear transformation model for censored data. Biometrika, 85(4):980–986.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. Biometrics, pages 845–854.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. Journal of the American Statistical Association, 74(365):153–160.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. Journal of the Royal Statistical Society. Series B (Methodological), pages 501–514.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. Technical report, DTIC Document.

Gentleman, R. and Geyer, C. J. (1994). Maximum likelihood for interval censored data: Consistency and computation. Biometrika, 81(3):618–623.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. Applied Statistics, pages 337–348.

Goedert, J. J., Kessler, C. M., Aledort, L. M., Biggar, R. J., Andes, W. A., White, G. C., Drummond, J. E., Vaidya, K., Mann, D. L., Eyster, M. E., and others (1989). A prospective study of human immunodeficiency virus type 1 infection and the development of AIDS in subjects with hemophilia. New England Journal of Medicine, 321(17):1141–1148.

Goetghebeur, E. and Ryan, L. (2000). Semiparametric Regression Analysis of Interval-Censored Data. Biometrics, 56(4):1139–1144.

Goggins, W. B., Finkelstein, D. M., and Zaslavsky, A. M. (1999). Applying the Cox Proportional Hazards Model when the Change Time of a Binary Time-Varying Covariate is Interval Censored. Biometrics, 55(2):445–451.

Groeneboom, P. and Wellner, J. A. (1992). Information bounds and nonparametric maximum likelihood estimation, volume 19. Springer Science & Business Media.

Grummer-Strawn, L. M. (1993). Regression Analysis of Current-Status Data: An Application to Breast-Feeding. Journal of the American Statistical Association, 88(423):758–765.

Hanson, T. and Johnson, W. O. (2004). A Bayesian Semiparametric AFT Model for Interval-Censored Data. Journal of Computational and Graphical Statistics, 13(2):341–361.

Hanson, T. and Yang, M. (2007). Bayesian Semiparametric Proportional Odds Models. Biometrics, 63(1):88–95.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. Journal of the Royal Statistical Society. Series B (Methodological), pages 757–796.

Hess, K. R. (1994). Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. Statistics in medicine, 13(10):1045–1062.

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. The Annals of Statistics, 24(2):540–568.

Huang, J. and Rossini, A. J. (1997). Sieve Estimation for the Proportional-Odds Failure-Time Regression Model with Interval Censoring. Journal of the American Statistical Association, 92(439):960–967.

Huang, J. and Wellner, J. A. (1997). Interval censored survival data: a review of recent progress. In Proceedings of the First Seattle Symposium in Biostatistics, pages 123–169. Springer.

Jewell, N. P. and van der Laan, M. J. (2002). Current status data: review, recent developments and open problems.

Joly, P., Commenges, D., and Letenneur, L. (1998). A Penalized Likelihood Approach for Arbitrarily Censored and Truncated Data: Application to Age-Specific Incidence of Dementia. Biometrics, 54(1):185–194.

Kalbfleisch, J. and Prentice, R. (1980). The Statistical Analysis of Failure Time Data. New York: John Wiley & Sons. KalbfleischThe Statistical Analysis of Failure Time Data1980.

Kalbfleisch, J. D. and Prentice, R. L. (2002). The statistical analysis of failure time data, volume 360. John Wiley & Sons.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. Journal of the american statistical association, 90(430):773–795.

Kim, J. S. (2003). Maximum likelihood estimation for the proportional hazards model with partly interval-censored data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65(2):489–502.

Kooperberg, C. and Clarkson, D. B. (1997). Hazard Regression with Interval-Censored Data. Biometrics, 53(4):1485–1494.

Kooperberg, C., Stone, C. J., and Truong, Y. K. (1995). Hazard Regression. Journal of the American Statistical Association, 90(429):78–94.

Kroner, B. L., Rosenberg, P. S., Aledort, L. M., Alvord, W. G., and Goedert, J. J. (1994). HIV-1 infection incidence among persons with hemophilia in the United States and Western Europe, 1978-1990. JAIDS Journal of Acquired Immune Deficiency Syndromes, 7(3):279–286.

Kulkarni, P. S., Butera, S. T., and Duerr, A. C. (2003). Resistance to HIV-1 infection: lessons learned from studies of highly exposed persistently seronegative (HEPS) individuals. AIDs Rev, 5(2):87–103.

Laughlin, S. K., Baird, D. D., Savitz, D. A., Herring, A. H., and Hartmann, K. E. (2009). Prevalence of Uterine Leiomyomas in the First Trimester of Pregnancy: An Ultrasound-Screening Study. Obstetrics & Gynecology, 113(3):630–635.

Lin, X., Cai, B., Wang, L., and Zhang, Z. (2014). A Bayesian proportional hazards model for general interval-censored data. Lifetime Data Analysis, pages 1–21.

Lin, X. and Wang, L. (2010). A semiparametric probit model for case 2 interval-censored failure time data. Statistics in Medicine, 29(9):972–981.

Lin, X. and Wang, L. (2011). Bayesian proportional odds models for analyzing current status data: univariate, clustered, and multivariate. Communications in Statistics-Simulation and Computation, 40(8):1171–1181.

Liu, H. and Shen, Y. (2009). A semiparametric regression cure model for interval-censored data. Journal of the American Statistical Association, 104(487):1168–1178.

Ma, S. and Kosorok, M. R. (2005). Penalized log-likelihood estimation for partly linear transformation models with current status data. Annals of Statistics, pages 2256–2290.

Mallick, B. K. and Walker, S. (2003). A Bayesian semiparametric transformation model incorporating frailties. Journal of Statistical Planning and Inference, 112(1ÂŰ2):159–174.

Mccarthy, J. M., Shea, P. R., Goldstein, D. B., and Allen, A. S. (2014). Testing for risk and protective trends in genetic analyses of HIV acquisition. Biostatistics, page kxu044.

McCullagh, P. (1980). Regression models for ordinal data. Journal of the Royal Statistical Society, Series B, 42(2):109–142.

McMahan, C. S., Wang, L., and Tebbs, J. M. (2013). Regression analysis for current status data using the EM algorithm. Statistics in Medicine, 32(25):4452–4466.

Odell, P. M., Anderson, K. M., and D'Agostino, R. B. (1992). Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. Biometrics, pages 951–959.

Pan, W. (1999). Extending the iterative convex minorant algorithm to the Cox model for interval-censored data. Journal of Computational and Graphical Statistics, 8(1):109–120.

Park, T. and Casella, G. (2008). The Bayesian Lasso. Journal of the American Statistical Association, 103(482):681–686.

Peto, R. (1973). Experimental survival curves for interval-censored data. Applied Statistics, pages 86–91.

Pettitt, A. N. (1984). Proportional Odds Models for Survival Data and Estimates Using Ranks. Journal of the Royal Statistical Society. Series C (Applied Statistics), 33(2):169–175.

Rabinowitz, D., Betensky, R. A., and Tsiatis, A. A. (2000). Using conditional logistic regression to fit proportional odds models to interval censored data. Biometrics, 56(2):511–518.

Rabinowitz, D., Tsiatis, A., and Aragon, J. (1995). Regression with interval-censored data. Biometrika, 82(3):501–513.

Ramsay, J. O. (1988). Monotone Regression Splines in Action. Statistical Science, 3(4):425–441.

Robey, W. G., Safai, B., Oroszlan, S., Arthur, L. O., Gonda, M. A., Gallo, R. C., and Fischinger, P. J. (1985). Characterization of envelope and core structural gene products of HTLV-III with sera from AIDS patients. Science, 228(4699):593–595.

Rossini, A. J. and Tsiatis, A. A. (1996). A Semiparametric Proportional Odds Regression Model for the Analysis of Current Status Data. Journal of the American Statistical Association, 91(434):713–721.

Royston, P. and Parmar, M. K. B. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Statistics in Medicine, 21(15):2175–2197.

Satten, G. A. (1996). Rank-based inference in the proportional hazards model for interval censored data. Biometrika, 83(2):355–370.

Satten, G. A., Datta, S., and Williamson, J. M. (1998). Inference based on imputed failure times for the proportional hazards model with interval-censored data. Journal of the American Statistical Association, 93(441):318–327.

Scharfstein, D. O., Tsiatis, A. A., and Gilbert, P. B. (1998). Semiparametric efficient estimation in the generalized odds-rate class of regression models for right-censored time-to-event data. Lifetime data analysis, 4(4):355–391.

Schwarz, G. and others (1978). Estimating the dimension of a model. The annals of statistics, 6(2):461–464.

Shen, X. (1998). Propotional odds regression and sieve maximum likelihood estimation. Biometrika, 85(1):165–177.

Shiboski, S. C. (1998). Generalized Additive Models for Current Status Data. Lifetime Data Analysis, 4(1):29–50.

Sinha, D., Chen, M.-H., and Ghosh, S. K. (1999). Bayesian Analysis and Model Selection for Interval-Censored Survival Data. Biometrics, 55(2):585–590.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van der Linde, A. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Technical report, Research Report, 98-009.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(4):583–639.

Sun, J. (1995). Empirical estimation of a distribution function with truncated and doubly interval-censored data and its application to AIDS studies. Biometrics, pages 1096–1104.

Sun, J. (2007). The statistical analysis of interval-censored failure time data. Springer Science & Business Media.

Sun, J. and Sun, L. (2005). Semiparametric linear transformation models for current status data. Canadian Journal of Statistics, 33(1):85–96.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. Journal of the Royal Statistical Society. Series B (Methodological), pages 290–295.

Valsecchi, M. G., Silvestri, D., and Sasieni, P. (1996). Evaluation of Long-Term Survival: Use of Diagnostics and Robust Estimators with Cox's Proportional Hazards Model. Statistics in Medicine, 15(24):2763–2780.

Wang, L. and Dunson, D. B. (2011). Semiparametric Bayes' Proportional Odds Models for Current Status Data with Underreporting. Biometrics, 67(3):1111–1118.

Wang, L. and Lin, X. (2011). A Bayesian approach for analyzing case 2 interval-censored data under the semiparametric proportional odds model. Statistics & Probability Letters, 81(7):876–883.

Wang, L., McMahan, C. S., Hudgens, M. G., and Qureshi, Z. P. (2016). A flexible, computationally efficient method for fitting the proportional hazards model to interval-censored data. Biometrics.

Wang, L., Xiaoyan, L., and Bo, C. (2012). Bayesian semiparametric regression analysis of interval-censored data with monotone splines. Interval-Censored Time-to-Event Data: Methods and Applications, page 149.

Wang, N., Wang, L., and McMahan, C. S. (2015). Regression analysis of bivariate current status data under the Gamma-frailty proportional hazards model using the EM algorithm. Computational Statistics & Data Analysis, 83:140–150.

Wellner, J. A. and Zhan, Y. (1997). A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data. Journal of the American Statistical Association, 92(439):945–959.

Williams, J. S. and Lagakos, S. W. (1977). Models for censored survival analysis: Constant-sum and variable-sum models. Biometrika, 64(2):215–224.

Yang, S. and Prentice, R. L. (1999). Semiparametric Inference in the Proportional Odds Regression Model. Journal of the American Statistical Association, 94(445):125–136.

Younes, N. and Lachin, J. (1997). Link-based models for survival data with interval and continuous time censoring. Biometrics, pages 1199–1211.

Zeng, D., Cai, J., and Shen, Y. (2006a). Semiparametric additive risks model for interval-censored data. Statistica Sinica, pages 287–302.

Zeng, D., Yin, G., and Ibrahim, J. G. (2006b). Semiparametric transformation models for survival data with a cure fraction. Journal of the American Statistical Association, 101(474):670–684.

Zhang, M. and Davidian, M. (2008). Smooth semiparametric regression analysis for arbitrarily censored time-to-event data. Biometrics, 64(2):567–576.

Zhang, Y., Hua, L., and Huang, J. (2010). A Spline-Based Semiparametric Maximum Likelihood Estimation Method for the Cox Model with Interval-Censored Data. Scandinavian Journal of Statistics, 37(2):338–354.

Zhang, Z. and Sun, J. (2009). Interval censoring. Statistical Methods in Medical Research.

Zhou, H., Hanson, T., and Zhang, J. (2015). Generalized accelerated failure time spatial frailty model for arbitrarily censored data. Lifetime data analysis, pages 1–21.

Zucker, D. M. and Yang, S. (2006). Inference for a family of survival models encompassing the proportional hazards and proportional odds models. Statistics in medicine, 25(6):995.