

ABSTRACT

DICASOLI, CARL MATTHEW. Bayesian Regression Methods for Crossing Survival Curves. (Under the direction of Dr. Sujit Ghosh and Dr. Subhashis Ghosal.)

In survival data analysis, the proportional hazards (PH), accelerated failure time (AFT), and proportional odds (PO) models are commonly used semiparametric models for the comparison of survivability in subjects. These models assume that the survival curves do not cross. However, in some clinical applications, the survival curves pertaining to the two groups of subjects under the study may cross each other, especially for long-duration studies. Hence, these three models stated above may no longer be suitable for making inference. Yang and Prentice (2005) proposed a model which separately models the short-term and long-term hazard ratios nesting both PH and PO. This feature allows for the survival functions to cross. First, we study the estimation procedure in the Yang-Prentice model with regards to the two-sample case. We propose two different approaches: (1) Bayesian bootstrap and (2) smoothing methods. The first approach involves Bayesian bootstrap with likelihoods corresponding to binomial and Poisson forms while the second approach involves kernel smoothing methods as well as smoothing spline methods. A simulation is conducted to compare various methods under the two-sample case. Next, we extend the Yang-Prentice model to a regression version involving predictors and examine three likelihood approaches including Poisson form, pseudo-likelihood, and Bayesian smoothing. The effects of model misspecification on asymptotic relative efficiency are also studied empirically. The results from simulation studies indicate that the PH, AFT, and PO models are not robust to model misspecifications when the survival functions are allowed to cross.

Finally, we calculate the marginal density via variational methods to determine the Bayes factor. Either a full Bayesian or Bayesian approach is implemented to perform model selection. Both approaches accurately identify the correct model, even under slight misspecification, and are computationally more efficient than MCMC techniques.

Bayesian Regression Methods for Crossing Survival Curves

by
Carl Matthew DiCasoli

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2009

APPROVED BY:

Dr. Wenbin Lu

Dr. Brian Reich

Dr. Sujit Ghosh
Chair of Advisory Committee

Dr. Subhashis Ghosal
Co-Chair of Advisory Committee

DEDICATION

To my parents, friends, and extended family

BIOGRAPHY

Carl DiCasoli was born in Palm Beach Gardens, Florida, USA. He graduated as valedictorian from the Alexander Dreyfoos School of the Arts at West Palm Beach, Florida, USA in 1999. He started postsecondary studies at the New England Conservatory of Music for piano performance and then finished his Bachelor of Science degree in Statistics from the University of Central Florida at Orlando, Florida, USA in 2004. He obtained a Master of Statistics degree from North Carolina State University in 2006 and continued graduate work thereafter.

ACKNOWLEDGMENTS

First, I will express my gratitude and greatest respect for the terrific guidance and training given by my advisors Dr. Sujit Ghosh and Dr. Subhashis Ghosal. I am greatly honoured and privileged to work with them throughout my studies. Furthermore, I would also like to thank the other committee members, Dr. Wenbin Lu, Dr. Brian Reich, and Dr. Charles Apperson, for their invaluable advice and suggestions. They have truly helped point me toward the correct direction in my research. I would also like to thank Dr. Song Yang from the National Institutes of Health / National Heart, Lung, and Blood Institute for providing me with both the motivation to pursue further graduate studies following my masters degree as well as the inspiration for my dissertation. Furthermore, I would like to thank my parents, Susan and Sebastian DiCasoli, grandparents, aunts, and uncles, as well as Sharon Dorias, Chunfu Chen, and Hoang Vo for their continuous and endearing support.

TABLE OF CONTENTS

LIST OF TABLES	vii
LIST OF FIGURES	x
1 Regression Models for Censored Data	1
1.1 Some Popular Models	2
1.1.1 The Proportional Hazards (PH) Model	2
1.1.2 The Accelerated Failure Time (AFT) Model	3
1.1.3 The Proportional Odds (PO) Model	4
1.2 Limitations of Popular Models	5
1.3 The Yang-Prentice (YP) Model	6
1.4 The Profile Likelihood Approach	7
1.4.1 The Proportional Hazards (PH) Model	8
1.4.2 The Accelerated Failure Time (AFT) Model	9
1.4.3 The Proportional Odds (PO) Model	10
1.4.4 The Yang-Prentice (YP) Model	11
1.5 Likelihood for Censored Data	12
1.6 Bayesian Bootstrap	14
1.7 The Adaptive Rejection Metropolis Sampling (ARMS) Algorithm	15
1.8 Outline of the Thesis	18
2 The Two-Sample Case	19
2.1 Likelihood-Based Inference	19
2.1.1 The Binomial form	20
2.1.2 The Poisson form	21
2.2 Smoothing Methods	22
2.2.1 Kernel-Smoothing Method	22
2.2.2 Smoothing-Spline Method	23
2.3 A Simulation Studies	25
2.3.1 Comparing Various Methods	25
2.4 An Application: The Gastrointestinal Tumor Study Group	32
3 A New Class of Regression Models	34
3.1 Likelihood approaches for the regression case	34
3.1.1 Poisson form	34
3.1.2 A Pseudo-likelihood approach	37
3.1.3 The Bayesian smoothing approach	38
3.2 Simulation studies	39
3.2.1 Comparing Various Methods	40
3.2.2 Investigating Model Misspecification	41

4 Model Selection	47
4.1 Approximating the Bayes factor corresponding to marginal densities	47
4.2 Variational Methods	53
4.2.1 Variational methods for Dirichlet process mixture models	54
4.2.2 A Variational Method Based on Bayesian Booststrap	58
4.3 Simulation Studies	61
Bibliography	67
Appendices	71
Appendix A	72
Appendix B	73

LIST OF TABLES

Table 1.1 The four main models used to analyze two-sample survival data where $r(t)$ denotes the hazard function ratio while $R(t)$ denotes the survival function ratio. . .	7
Table 2.1 A simulation study for each method regarding the non-crossing survival curves case ($\beta = (0, 0)$) based on 1000 repetitions. Bias of $\hat{\beta}$ as well as the estimated 95% coverage probability of the confidence intervals, YP = Yang-Prentice, BB = Bayesian Bootstrap	30
Table 2.2 A simulation study for each method regarding the crossing survival curves case ($\beta = (\frac{1}{2}, -\frac{1}{2})$) based on 1000 repetitions. Bias of $\hat{\beta}$ as well as the estimated 95% coverage probability of the confidence intervals, YP = Yang-Prentice, BB = Bayesian Bootstrap	31
Table 2.3 Results corresponding to fitting the Gastrointestinal Tumor Study Group Data under the Bayesian Bootstrap - Binomial, Bayesian Bootstrap - Poisson, Kernel Smoothing, and Smoothing-Spline Methods. Estimates of $\hat{\theta}_1$ and $\hat{\theta}_2$, corresponding posterior standard deviations, and 95% credible intervals. CrI = Credible Interval, s.d. = standard deviation.	32
Table 3.1 Simulation results for the likelihood given by the empirical likelihood method (EL) (with Poisson form (eq. (3.8)) / Pseudo-likelihood (eq. (3.14))) and Bayesian smoothing (eq. (3.16)), under 30% and 50% censoring at $n = 200$ when both simulating and fitting data from the YP model.	41
Table 3.2 Simulation results for the likelihood given by the empirical likelihood method (EL) (with Poisson form (eq. (3.8)) / Pseudo-likelihood (eq. (3.14))) and Bayesian smoothing (eq. (3.16)), when simulating data from the YP model but fitting the PH model under 30% and 50% censoring at $n = 200$	42
Table 3.3 Simulation results for the likelihood given by the empirical likelihood method (EL) (with Poisson form (eq. (3.8)) / Pseudo-likelihood (eq. (3.14))) and Bayesian	

smoothing (eq. (3.16)), when simulating data from the PH model but fitting the YP model under 30% and 50% censoring at $n = 200$	43
Table 3.4 Simulation results for the likelihood given by the empirical likelihood method (EL) (with Poisson form (eq. (3.8)) / Pseudo-likelihood (eq. (3.14))) and Bayesian smoothing, equation (eq. (3.16)), when both simulating and fitting data from the PH model under 30% and 50% censoring at $n = 200$	44
Table 3.5 Mean squared error (MSE) and asymptotic relative efficiency (ARE) defined as $\frac{\text{MSE}(\text{YP} \text{PH})}{\text{MSE}(\text{PH} \text{PH})}$, for the empirical likelihood (EL) method (with Poisson form (eq. (3.8)) / Pseudo-likelihood (eq. (3.14))) and Bayesian smoothing (eq. (3.16)), Par. = Parameter, PF = Poisson form, PL = Pseudo-likelihood, Bay. = Bayesian.	45
Table 3.6 Mean squared error (MSE) and asymptotic relative efficiency (ARE) defined as $\frac{\text{MSE}(\text{YP} \text{YP})}{\text{MSE}(\text{PH} \text{YP})}$, for the empirical likelihood (EL) method (with Poisson form (eq. (3.8)) / Pseudo-likelihood (eq. (3.14))) and Bayesian smoothing (eq. (3.16)), Par. = Parameter, PF = Poisson form, PL = Pseudo-likelihood, Bay. = Bayesian.	46
Table 4.1 Results corresponding to the probability of selecting a particular model when simulating data from the assumptions of four different cases under the full Bayesian variational approach: $\beta \neq 0, \gamma \neq 0$; $\gamma = 0, \beta \neq 0$; $\gamma \neq 0, \beta = 0$; $\beta = \gamma = 0$. Prob. = probability of selecting the specified model, Assump. = assumption, Sim. = simulated, Mar. = marginal, MC s.e. = Monte-Carlo standard error.	63
Table 4.2 Results corresponding to the probability of selecting a particular model conducting pairwise comparisons when simulating data from the assumptions of four different cases under the full Bayesian variational approach: $\beta \neq 0, \gamma \neq 0$; $\gamma = 0, \beta \neq 0$; $\gamma \neq 0, \beta = 0$; $\beta = \gamma = 0$. Assump. = assumption, med. = median, BF = Bayes Factor = Sim. Marginal Prob. / Fitted Marginal Prob., % correct = percentage that the correct model is identified ($\log \text{BF} \geq 0$). Range refers to distance between the 25th and 75th percentiles.	63
Table 4.3 Biases of $\beta = (\beta_0, \beta_1, \beta_2)$ under both the true and misspecified models using the full Bayesian variational approach, s.e. = standard error	65
Table 4.4 Results corresponding to the probability of selecting a particular model when simulating data from the assumptions of four different cases under the Bayesian	

bootstrap variational approach: $\beta \neq 0, \gamma \neq 0$; $\gamma = 0, \beta \neq 0$; $\gamma \neq 0, \beta = 0$; $\beta = \gamma = 0$. Prob. = probability of selecting the specified model, Assump. = assumption, Sim. = simulated, Mar. = marginal, MC s.e. = Monte-Carlo standard error.	65
Table 4.5 Results corresponding to the probability of selecting a particular model conducting pairwise comparisons when simulating data from the assumptions of four different cases under the Bayesian bootstrap variational approach: $\beta \neq 0, \gamma \neq 0$; $\gamma = 0, \beta \neq 0$; $\gamma \neq 0, \beta = 0$; $\beta = \gamma = 0$. Assump. = assumption, Med. = median, BF = Bayes Factor = Sim. Marginal Prob. / Fitted Marginal Prob., % correct = percentage that the correct model is identified ($\log \text{BF} \geq 0$). Range refers to distance between the 25th and 75th percentiles.....	66
Table 4.6 Biases of $\beta = (\beta_0, \beta_1, \beta_2)$ under both the true and misspecified models using the Bayesian bootstrap variational approach, s.e. = standard error	66

LIST OF FIGURES

Figure 1.1 The left panel is the estimated survival functions, the middle panel is the plot of the estimated survival function ratio, and the right panel is the estimated hazard function ratio for the tongue cancer study.....	4
Figure 1.2 Estimated survival functions for the Gastrointestinal Tumor Group Study data.	6
Figure 1.3 A plot of the pseudo log-likelihood corresponding to the PH model versus θ when fitting the proportional hazards model to the Gastrointestinal Tumor Study Group data.	9
Figure 1.4 A plot of the log-likelihood corresponding to the AFT model versus θ when fitting the AFT model empirically to the Gastrointestinal Tumor Study Group data.	10
Figure 1.5 A plot of the log-likelihood corresponding to the PO model versus θ when fitting the PO model empirically to the Gastrointestinal Tumor Study Group data.	11
Figure 1.6 A plot illustrating survival curves for the PH, AFT, PO, and YP cases for the Gastrointestinal Tumor Study Group data.....	13
Figure 2.1 A plot of the biases for β_1 under the case of no treatment effect, $\beta = (0,0)$	26
Figure 2.2 A plot of the biases for β_1 under the case where the treatment effect starts negative but eventually becomes positive, $\beta = (1/2, -1/2)$	26
Figure 2.3 A plot of the biases for β_2 under the case of no treatment effect, $\beta = (0,0)$	27
Figure 2.4 A plot of the biases for β_2 under the case where the treatment effect starts negative but eventually becomes positive, $\beta = (1/2, -1/2)$	27
Figure 2.5 A plot of the 95% credible probability of the confidence intervals for β_1 under the case of no treatment effect, $\beta = (0,0)$	28
Figure 2.6 A plot of the 95% credible probability of the confidence intervals for β_1 under the case where the treatment effect starts negative but eventually becomes positive, $\beta = (1/2, -1/2)$	28
Figure 2.7 A plot of the 95% credible probability of the confidence intervals for β_2 under the case of no treatment effect, $\beta = (0,0)$	29

Figure 2.8 A plot of the 95% credible probability of the confidence intervals for β_2 under the case where the treatment effect starts negative but eventually becomes positive, $\beta = (1/2, -1/2)$ 29

Chapter 1

Regression Models for Censored Data

In survival data analysis, researchers are often interested in describing the distribution of the time to a certain event for a given population of subjects. Additionally, we often try to find a relationship between “time to event” to a given set of predictors. A clinical trial is usually conducted over a given period of time and the “time to event” may or may not be completely observed for each individual enrolled in the clinical trial. If the complete “time to event” is unobserved, then we refer to that event as being censored. In our research, we will focus on Type I (right) censoring, where a particular event is observed only if it occurs before a prespecified time. To understand this concept more thoroughly, let us consider a study where, for both males and females, 300 cats are randomly partitioned into six dose-level groups. Here, we follow each cat until it dies or a finite, predetermined time (100 or 200 weeks) is attained. These predetermined times are chosen for the purpose of reducing the cost of raising the cats, but allow for some type of information regarding the survival experience of the cats that live longer. If a subject (cat) experiences death prior to the end of the study, then the event time is completely observed and hence the time to event remains uncensored. However, it is possible that the subjects have experienced death after the end of the study. But during the study, these subjects are only known to be alive and are hence labeled as “censored” observations. Survival analysis is a research area that specifically deals with this issue.

First, we will introduce some basic notations. Let T be a non-negative valued

random variable that represents survival (or failure) time to an event. Next, we define $S_T(t) = \Pr[T > t]$, $F_T(t) = 1 - S_T(t)$, $f_T(t) = \frac{dF_T(t)}{dt}$, and $h_T(t) = \frac{f_T(t)}{S_T(t)} = \lim_{\delta \rightarrow 0} \Pr[t < T \leq t + \delta | T > t] / \delta$ as the survival function, cumulative distribution function (cdf), probability density function (pdf), and the hazard function of T , respectively. It is well-known from standard survival analysis that $S_T(t) = \exp\left\{-\int_0^t h_T(u) du\right\} = \exp\{-H_T(t)\}$, where $H_T(t) = \int_0^t h_T(u) du$ is the cumulative hazard for the random variable T . As an example, let us consider the pdf of a Weibull distribution; that is, $f_T(t) = \alpha \lambda t^{\alpha-1} \exp(-\lambda t^\alpha)$. After using the above definitions, we can easily obtain that $S_T(t) = \exp(-\lambda t^\alpha)$, $h_T(t) = \alpha \lambda t^{\alpha-1}$, and $H_T(t) = \lambda t^\alpha$, where $\alpha, \lambda > 0$ are the parameters.

Next, we will consider the simple case of comparing survival rates based on two, independent samples. Let $S_i(t) = \Pr(T_i > t)$, $F_i(t) = 1 - S_i(t)$, $f_i(t) = \frac{dF_i(t)}{dt}$, $h_i(t) = \frac{f_i(t)}{S_i(t)}$ denote the survival function, cumulative distribution function (cdf), probability density function (pdf), and hazard function of the i th group for $i=1,2$. Assume that $T_{ij} \stackrel{iid}{\sim} S_i(\cdot)$ for $j = 1, \dots, n_i$ and $i = 1,2$. Due to censorship, it is not possible to observe the complete data; instead, we may only observe $X_{ij} = \min(T_{ij}, C_{ij})$ for $i = 1,2$, and $j=1, \dots, n_i$, where C_{ij} are the censoring times, independent of T_{ij} . Let $\Delta_{ij} = I(T_{ij} \leq C_{ij})$ denote the censoring indicator. This type of censoring is usually known as random right censoring. Our goal is to estimate the ratio of the hazards, that is, $r(t) = \frac{h_1(t)}{h_2(t)}$, and the ratio of the survival functions, $R(t) = \frac{S_1(t)}{S_2(t)}$. Finally, note that $r(t)R(t) = \frac{f_1(t)}{f_2(t)}$; thus, if the ratio of densities is known, then $r(t)$ is determined by $R(t)$ and vice versa.

1.1 Some Popular Models

Several models have been developed to analyze two-sample survival data. Section 1.1 discusses the proportional hazards model for the two-sample case. Sections 1.1.2 and 1.1.3 describe the accelerated failure time (AFT) and proportional odds models, respectively. Finally, Section 1.3 provides discussion regarding the newest model for fitting two-sample survival data, the Yang-Prentice (2005) model.

1.1.1 The Proportional Hazards (PH) Model

The proportional hazards (PH) model was introduced by Cox (1972) in a more general setting, involving a vector of predictors. In the two-sample case, we can express the

hazard function of the first group in terms of the hazard function of the second group as

$$h_1(t) = \theta h_2(t), \quad (1.1)$$

where θ is the ratio of hazards. That is, $r(t) = \theta > 0$ for all t . Recall that $S(t) = \exp\left\{-\int_0^t h(u)du\right\}$. Under (1.1), we have $S_1(t) = \exp\left\{-\int_0^t h_1(u)du\right\} = \exp\left\{-\int_0^t \theta h_2(u)du\right\} = S_2^\theta(t)$. Thus $R(t) = S_2^{\theta-1}(t)$. This representation suggests that $R(t) \leq 1$ for $\theta \geq 1$ and $R(t) > 1$ for $\theta < 1$, which implies that $S_1(t)$ and $S_2(t)$ do not cross for any $t \in (0, \infty)$ for a given value of θ . Also, note that $h_1(t)$ and $h_2(t)$ do not cross for any $t \in (0, \infty)$ for a given value of θ .

As an illustration, we consider a clinical trial that studied the effects of ploidy on the diagnosis of patients with cancer of the tongue. After survival data was obtained on each patient, tissue samples were taken via a flow cytometer that determined whether the tumor had an aneuploid (abnormal) or diploid (normal) DNA profile. Figure 1 shows that the estimated Kaplan-Meier survival curves (Kaplan and Meier, 1952) do not cross. Here, θ , as in (1.1), was estimated to be 0.64, with standard error 0.28, after performing standard analysis using PROC PHREG in SAS. $\hat{R}(t)$ is computed as the ratio of two Kaplan-Meier Estimates. Also, we can notice that $\hat{R}(t) > 1$ for all values of t , and the ratio of estimated hazard functions, $\hat{r}(t)$ does not appear to be constant.

1.1.2 The Accelerated Failure Time (AFT) Model

Although the proportional hazards model is widely used in survival analysis, it may not be appropriate in certain circumstances. The hazard ratio may not appear to be constant, such as if the study involves either a clinical trial, an observational study, or even a cohort study. To address this problem, the accelerated failure time (AFT) model has been proposed by Kalbfleisch and Prentice (1980). In the two sample case, the AFT model states that there exists a constant $\theta > 0$ such that

$$S_1(t) = S_2(\theta t) \text{ for some } \theta > 0 \text{ and all } t \geq 0. \quad (1.2)$$

For instance, let $S_1(t)$ be the survival function for a population of cats and $S_2(t)$ represent the survival function for the human population. From general knowledge, we will assume that one year in a cat's life is equivalent to about nine years for a human. Hence, this representation would imply that $\theta = 9$ and $S_1(t) = S_2(9t)$.

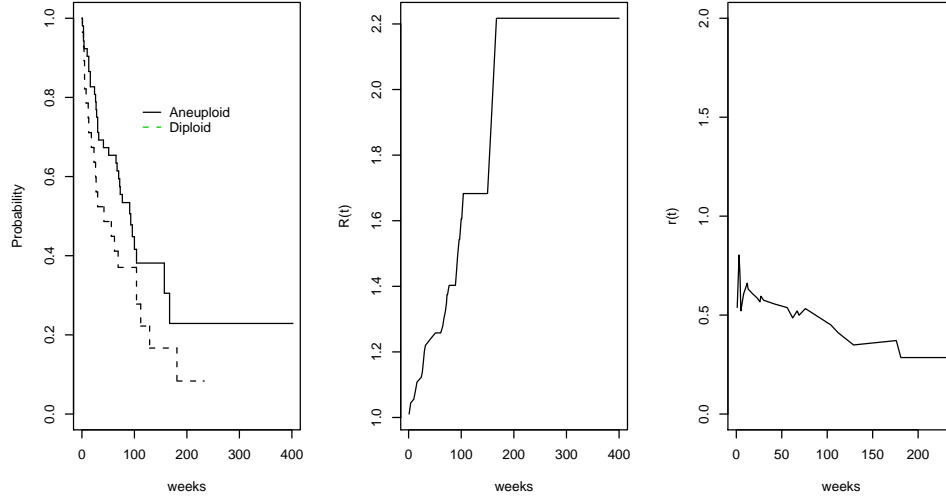


Figure 1.1: The left panel is the estimated survival functions, the middle panel is the plot of the estimated survival function ratio, and the right panel is the estimated hazard function ratio for the tongue cancer study.

Notice that under the above AFT model $R(t) = \frac{S_1(t)}{S_2(t)} = \frac{S_2(\theta t)}{S_2(t)}$ and hence $R(t) \leq 1$ for all t if and only if $\theta > 1$. Hence again the survival curves do not cross for any value of θ . Also, notice that $r(t) = \theta \frac{h_2(\theta t)}{h_2(t)}$ is a non-constant function of t unless T_2 has a Weibull distribution. The next model that we shall study (the proportional odds model) can also account for a non-constant hazard ratio.

1.1.3 The Proportional Odds (PO) Model

Another alternative method is to implement a proportional odds model which was first developed by Bennett (1983). The main idea behind the proportional odds model is to illustrate how certain predictive factors affect the odds against survival multiplicatively for any given time. The ratio, $\frac{F_i(t)}{1-F_i(t)} = \frac{F_i(t)}{S_i(t)}$ represents the odds that the event occurs at time t for group i . The proportional odds model assumes that the ratio between the two groups is constant over time. In other words,

$$\frac{F_1(t)/S_1(t)}{F_2(t)/S_2(t)} = \theta \text{ for some } \theta > 0. \quad (1.3)$$

Rearranging the above equation gives us

$$\frac{F_1(t)}{S_1(t)} = \frac{F_2(t)}{S_2(t)}\theta, \quad (1.4)$$

or equivalently,

$$S_1(t) = \frac{S_2(t)}{S_2(t) + \theta(1 - S_2(t))} = \left[1 + \theta \frac{1 - S_2(t)}{S_2(t)}\right]^{-1}. \quad (1.5)$$

This implies that

$$R(t) = \frac{S_1(t)}{S_2(t)} = \frac{1}{S_2(t) + \theta(1 - S_2(t))}, \quad (1.6)$$

which implies that $R(t) \geq 1$ iff $\theta \leq 1$. Therefore, $S_1(t)$ and $S_2(t)$ do not cross each other for any $t > 0$ for a given value of θ . To express the proportional odds model in terms of the ratio of hazard functions, we will first observe that $h_i(t) = -\frac{d\log[S_i(t)]}{dt}$ for $i = 1, 2$. Using this fact, along with equations (1.5) and (1.6), gives us

$$r(t) = \frac{h_1(t)}{h_2(t)} = \theta R(t). \quad (1.7)$$

Thus, $r(t) \geq \theta$ for all $t > 0$ iff $\theta \leq 1$.

In the next section, we will study the Yang-Prentice (2005) model, which is an extension of the PH and PO models and can account for both crossing survival curves as well as a non-constant hazard ratio.

1.2 Limitations of Popular Models

In many real applications, the survival curves may cross. In a clinical trial conducted by the Gastrointestinal Tumor Study Group (1982), we are interested in comparing chemotherapy alone versus chemotherapy combined with radiotherapy for treating gastric cancer that is locally unresectable. To conduct the clinical trial that took place over a time period of eight years, the researchers randomized forty-five patients to each of the two treatment arms. In Figure 1.2, we notice that the estimated Kaplan-Meier survival curves cross for the two groups at about 1000 days. This suggests that early differences favoring the chemotherapy only group are eventually negated by a later survival advantage with regards to the chemotherapy plus radiotherapy group (Klein and Moeschberger, 2003). It is clear that none of the previous three, well-known models (PH, AFT, and PO) can adequately capture this feature of crossing survival curves.

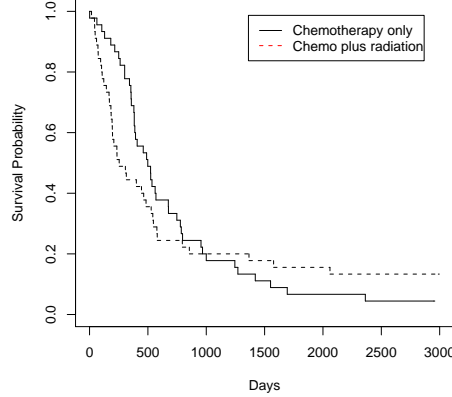


Figure 1.2: Estimated survival functions for the Gastrointestinal Tumor Group Study data.

1.3 The Yang-Prentice (YP) Model

To model the feature of crossing survival curves, Yang and Prentice (2005) developed a semiparametric model based on two samples that correspond to the short-term and long-term hazard ratios with the baseline distribution left completely unspecified. The Yang and Prentice (2005) model can be viewed as a generalization of both the Cox (1972) proportional hazards model and the Bennett (1983) proportional odds model; that is,

$$r(t) = \frac{\theta_1 \theta_2}{\theta_1 + (\theta_2 - \theta_1) S_2(t)} = \frac{\theta_1}{S_2(t) + \frac{\theta_1}{\theta_2} (1 - S_2(t))}, \quad (1.8)$$

for all t , where $\theta_1, \theta_2 > 0$. For $h_1(t)$ to be a proper hazard function, we need

$$h_1(t) \geq 0 \text{ and } \int_0^\infty h_1(t) dt = \infty. \quad (1.9)$$

If we assume that $h_2(t)$ is a proper hazard function, then the above two properties hold true for $h_1(t)$ as $r(t) \geq (\theta_1^{-1} + \theta_2^{-1})^{-1}$ for all t in the YP model.

Unlike the Cox (1972) model, we can notice that the hazard ratio between the treatment and placebo groups is not a constant and it depends on θ_1, θ_2 , and t . Furthermore, if $\theta_1 < \theta_2$, the hazard ratio $r(t)$ is monotonically increasing, whereas the case in which $\theta_1 > \theta_2$ implies a monotonically decreasing hazard ratio. In addition, note that

$$\theta_1 = \lim_{t \downarrow 0} \frac{h_1(t)}{h_2(t)}, \quad \theta_2 = \lim_{t \uparrow \infty} \frac{h_1(t)}{h_2(t)}. \quad (1.10)$$

Hence, we can think of θ_1 as the short-term hazard ratio and θ_2 as the long-term hazard ratio (Yang and Prentice, 2005). Using equations (1.8) and (1.10), we can also observe that the Yang-Prentice (2005) model reduces to the Cox (1972) model if $\theta_1 = \theta_2$ and the proportional odds model if $\theta_2 = 1$.

Additionally, from (1.8), after some algebra (see Appendix, Theorem 1), we can obtain the survival functions of both groups; that is,

$$S_1(t) = [1 + \frac{\theta_1}{\theta_2}K(t)]^{-\theta_2} \text{ and } S_2(t) = [1 + K(t)]^{-1}, \quad (1.11)$$

where $K(t) = \frac{F_2(t)}{S_2(t)}$ is the odds for the second group at t .

Using equation (1.11), we can obtain $R(t)$, the ratio of the two survival functions; that is,

$$R(t) = \frac{S_1(t)}{S_2(t)} = \frac{1 + K(t)}{[1 + \frac{\theta_1}{\theta_2}K(t)]^{\theta_2}}. \quad (1.12)$$

From equations (1.11) and (1.12), we can observe that the survival curves for the treatment and control groups cross in the cases where either $\theta_2 < 1 < \theta_1$ or $\theta_1 < 1 < \theta_2$ (Yang and Prentice, 2005); that is, $1 \in (\theta_1 \wedge \theta_2, \theta_1 \vee \theta_2)$ when $\theta_1 \neq \theta_2$. Notice that in each case of PH, AFT, and PO, $\theta = 1$ indicates that there is no difference between the groups. In summary, we can describe the models discussed so far in Table 1.1:

Table 1.1: The four main models used to analyze two-sample survival data where $r(t)$ denotes the hazard function ratio while $R(t)$ denotes the survival function ratio.

Model	$r(t)$	$R(t)$
PH	θ	$S_2(t)^{\theta-1}$
AFT	$\frac{\theta h_2(\theta t)}{h_2(t)}$	$\frac{S_2(\theta t)}{S_2(t)}$
PO	$\frac{\theta}{S_2(t) + \theta(1 - S_2(t))}$	$\frac{1}{S_2(t) + \theta(1 - S_2(t))}$
YP	$\frac{\theta_1 \theta_2}{\theta_1 + (\theta_2 - \theta_1)(S_2(t))}$	$\frac{S_2(t)^{\theta_2-1}}{[S_2(t) + \frac{\theta_1}{\theta_2}(1 - S_2(t))]^{\theta_2}}$

1.4 The Profile Likelihood Approach

In this section, we describe semiparametric inferential methods to obtain estimates of θ (or θ_1, θ_2), $r(t)$, and $R(t)$ based on the sample $\{(X_{ij}, \Delta_{ij}) : i = 1, 2; j = 1, \dots, n_i\}$. We

assume that $T_{2j} \stackrel{iid}{\sim} f_2(\cdot)$ and $T_{1j} \stackrel{iid}{\sim} f_1(\cdot)$ where $h_1(t) = g(h_2(t), \theta)$ for some known function $g(\cdot)$ and unknown parameter θ ; e.g., under the PH model $h_1(t) = \theta h_2(t)$ whereas under the AFT model $h_1(t) = \frac{\theta h_2(\theta t)}{h_2(t)}$. Let $\hat{h}_2(t)$ be a consistent estimate of $h_2(t)$ based on the data $\{(X_{2j}, \Delta_{2j}); j = 1, \dots, n_2\}$ and let $\hat{S}_2(t) = \exp\left\{-\int_0^t \hat{h}_2(u) du\right\}$. Then we estimate θ using the following pseudo-likelihood:

$$L_{n_1}(\theta) = \prod_{j=1}^{n_1} g(\hat{h}_2(x_{1j}), \theta)^{\Delta_{1j}} e^{-\int_0^{x_{1j}} g(\hat{h}_2(u), \theta) du}. \quad (1.13)$$

Notice that $h_1(t) = g(h_2(t), \theta)$ implies that $S_1(t) = \exp\left\{-\int_0^t g(h_2(u), \theta) du\right\}$.

1.4.1 The Proportional Hazards (PH) Model

Using the pseudo-likelihood approach described above, we can derive $\hat{\theta}$ by maximizing the log pseudo-likelihood. It follows that

$$\hat{\theta} = \frac{\sum_{j=1}^{n_1} \Delta_{1j}}{-\sum_{j=1}^{n_1} \log \hat{S}_2(x_{1j})}, \quad (1.14)$$

where $\hat{S}_2(t)$ is the Kaplan-Meier estimator of $S_2(t)$ based on the data $\{(X_{2j}, \Delta_{2j}), j = 1, \dots, n_2\}$. Notice that (in case of no ties),

$$\hat{S}_2(t) = \prod_{j: x_{2(j)} \leq t} \left(1 - \frac{1}{n_2 - j + 1}\right)^{\delta_{2(j)}}, \quad (1.15)$$

where $x_{2(1)} < x_{2(2)} < \dots < x_{2(n_2)}$ are ordered observations and $\delta_{2(j)}$ corresponds to $x_{2(j)}$. Using the Gastrointestinal Tumor Study Group data, we obtain $\hat{\theta} = 0.87$, with a bootstrap standard error of 0.17. This result is illustrated from the graph of the empirical log-likelihood function for the PH model versus θ given by Figure 1.3. After fitting the proportional hazards (PH) model directly using R, we obtain estimates close to those obtained via the pseudo-likelihood approach; $\hat{\theta} = 0.89$, with a standard error of 0.20. To determine whether there is a difference between the two groups, we will test $H_0: \theta = 1$ and calculate the test-statistic

$$Z = \frac{\hat{\theta} - 1}{s.e.(\hat{\theta})}. \quad (1.16)$$

Using the estimates and standard errors given above, we obtain that $|Z| = 0.76$ or 0.55 , which corresponds to the pseudo-likelihood and direct methods, respectively. In both cases, $|Z| < 1$ implies that the null hypothesis (no difference between the two groups) cannot be rejected.

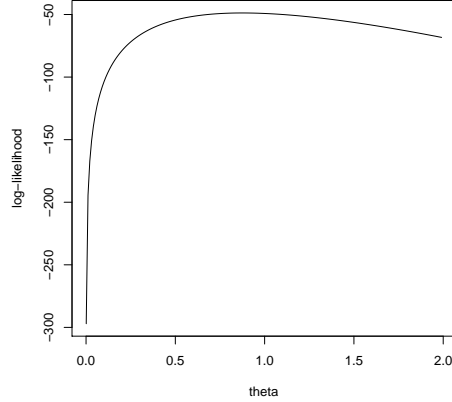


Figure 1.3: A plot of the pseudo log-likelihood corresponding to the PH model versus θ when fitting the proportional hazards model to the Gastrointestinal Tumor Study Group data.

1.4.2 The Accelerated Failure Time (AFT) Model

Next, we will use the pseudo-likelihood method to estimate θ for the AFT model. When fitting the Gastrointestinal Tumor Study Group data to the AFT model, there seems to be several local maxima due to the fact that we are using a non-smooth estimate of $h_2(\cdot)$. The derived log-likelihood is represented by:

$$\log L(\theta) = \sum_{j=1}^{n_1} \Delta_{1j} \log \theta + \sum_{j=1}^{n_1} \Delta_{1j} \log \hat{h}_2(\theta x_{1j}) + \sum_{j=1}^{n_1} \log \hat{S}_2(\theta x_{1j}), \quad (1.17)$$

where

$$\hat{h}_2(t) = \sum_{j=1}^{n_2} \left(1 - \frac{\delta_{(j)}}{n_2 - j + 1} I(x_{2(j-1)} \leq t \leq x_{2(j)}) \right) \quad (1.18)$$

Here, we notice that the log-likelihood is a function with several maxima, with the global maximum at $\hat{\theta} = 0.81$. Additionally, the corresponding bootstrap standard error is 0.36. This fact is shown by the graph of the log-likelihood for the AFT model versus θ given by Figure 1.4. After fitting the accelerated failure time (AFT) model directly using the “rankreg” procedure in R, we obtain the estimates $\hat{\theta} = 0.78$, with a standard error of 0.33. Using equation (1.16), we obtain that $|Z| = 0.53$ or 0.67 , which corresponds to the pseudo-likelihood and direct methods, respectively. These results ($|Z| < 1$) imply that the null

hypothesis ($H_0: \theta = 1$) cannot be rejected indicating no substantial difference between the groups.

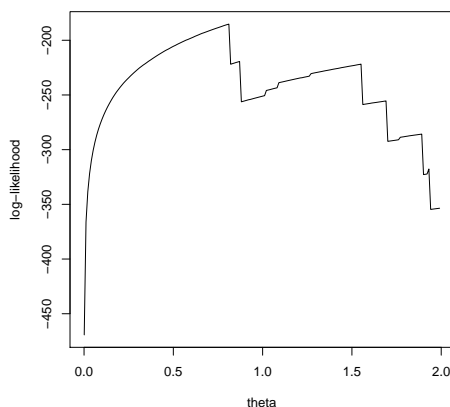


Figure 1.4: A plot of the log-likelihood corresponding to the AFT model versus θ when fitting the AFT model empirically to the Gastrointestinal Tumor Study Group data.

1.4.3 The Proportional Odds (PO) Model

When implementing the pseudo-likelihood method for the proportional odds model using the data from the Gastrointestinal Tumor Study Group, we first obtain the pseudo log-likelihood, which is given by the following equation:

$$\begin{aligned} \log L(\theta) = & \sum_{j=1}^{n_1} \Delta_{1j} \log \theta - \sum_{j=1}^{n_1} \Delta_{1j} \log(\hat{S}_2(x_{1j}) + \theta(1 - \hat{S}_2(x_{1j}))) \\ & - \sum_{j=1}^{n_1} \log(\hat{S}_2(x_{1j}) + \theta(1 - \hat{S}_2(x_{1j}))) \end{aligned} \quad (1.19)$$

Since this likelihood is not in closed form, we utilize the “optimize” function in R to obtain the pseudo maximum likelihood estimate to be $\hat{\theta} = 0.72$ with a bootstrap standard error of 0.31. In Figure 1.5, the graph of the log-likelihood corresponding to the PO model confirms a global maximum at the above-specified value of θ . After fitting the proportional odds (PO) model directly using R, we obtain estimates close to those obtained via the pseudo-likelihood approach; $\hat{\theta} = 0.73$, with a standard error of 0.29. Using equation (1.16), we obtain that $|Z| = 0.90$ or 0.93 , which corresponds to the pseudo-likelihood and direct methods, respectively.

Again, $|Z| < 1$ implies that the null hypothesis (no difference between the two groups) cannot be rejected. Thus, all three models (PH, AFT, and PO) would conclude that there is no substantial difference between the two treatment groups.

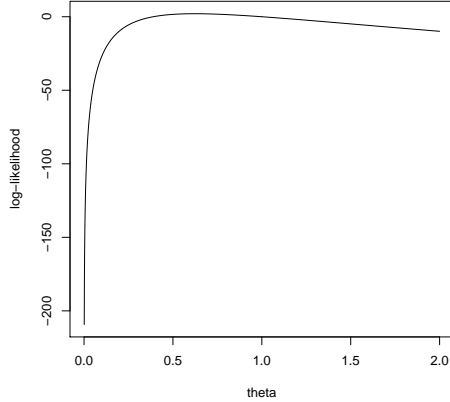


Figure 1.5: A plot of the log-likelihood corresponding to the PO model versus θ when fitting the PO model empirically to the Gastrointestinal Tumor Study Group data.

1.4.4 The Yang-Prentice (YP) Model

The final model that we will study requires maximization over two parameters corresponding to the short-term and long-term hazard ratios. Previously, Yang and Prentice (2005) had developed a likelihood method to estimate the two parameters based on martingale theory; they employ a pseudo maximum likelihood approach that can be written as simple estimating equations. Furthermore, Yang and Prentice (2005) also show that the pseudo maximum likelihood estimator is a weighted martingale residual estimator. Their estimates were $\hat{\theta}_1 = 4.97$ and $\hat{\theta}_2 = 0.39$ with 95% CI of (1.80, 13.70) and (0.24, 0.65), respectively. Alternatively, we implement a method based on the pseudo-likelihood method. Hence, we derive the log-likelihood:

$$\log L(\theta) = \sum_{j=1}^{n_1} \Delta_{1j} \log \left[\frac{\theta_1 \theta_2 \hat{h}_2(x_{1j})}{\theta_1 + (\theta_2 - \theta_1) \hat{S}_2(x_{1j})} \right] - \theta_2 \sum_{j=1}^{n_1} \log \left[1 + \frac{\theta_1 \hat{F}_2(x_{1j})}{\theta_2 \hat{S}_2(x_{1j})} \right] \quad (1.20)$$

Since the log-likelihood lacks a closed form, we implement the “mle” function in R, which can maximize a multidimensional vector of parameters, to obtain the values, $\hat{\theta}_1 = 5.00$

and $\hat{\theta}_2 = 0.48$ and corresponding 95% CI of (1.22, 8.71) and (0.21, 0.75) by bootstrapping. Because of its ability to find the global, optimal value of θ_1 and θ_2 even on a rough surface, the “SAAN” method (Belisle, 1992), a stochastic global optimization method that utilizes only function values, was implemented. A drawback of this method is that its convergence is relatively slow and can be quite dependent on the choice of initial parameters, although it will work well for non-differentiable functions. Notice that by either estimation approach, the null hypothesis $H_0: \theta_1 = \theta_2 = 1$ (no difference between the groups) is rejected. This is in sharp contrast to the previous three models, all of which concluded that there are no substantial differences between the two groups, which motivates us to further explore the Yang Prentice model beyond comparing two groups. Furthermore, it appears that the null hypothesis of $H_0: \theta_1 = \theta_2$ or $H_0: \theta_2 = 1$ cannot be rejected for this dataset, which indicates that both PH and PO models are not suitable for this dataset.

Next, we will graph the estimate of the first survival function (in this case, chemotherapy case only), using the second survival function (chemotherapy plus radiation) along with the estimate of $\hat{\theta}$ computed from the pseudo-likelihood given above, for each of the four cases (PH, AFT, PO, YP). For example, recall that in the PH case, $S_1(t|\theta) = S_2(t)^\theta$, and in the AFT case, $S_1(t|\theta) = S_2(\theta t)$. Likewise, we can use equation (1.5) for the PO case and (1.11) for the YP case. $\hat{S}_2(t)$ will be computed via the Kaplan-Meier estimate. We will graph both $S_1(t|\hat{\theta})$ and $\hat{S}_2(t)$ based on all four models, using a real data example. The graph is provided in Figure 1.6. From the graph, we can observe that the Yang-Prentice model is the only method that can account for the crossing of survival curves and identify substantial differences between two groups that would have been lost if the PH, AFT, or PO models were used incorrectly.

1.5 Likelihood for Censored Data

Finally, we will again consider a sample of iid right-censored observations $\{(X_i, \Delta_i); i = 1, \dots, n\}$. Here, Andersen *et al.* (1993) explain that if there is noninformative censoring, we can write a binomial form; that is,

$$\prod_{i=1}^n \prod_{x \in [0, \tau]} \left\{ (Y_i(x) dA(x))^{dN_i(x)} (1 - Y_i(x) dA(x))^{1 - dN_i(x)} \right\}, \quad (1.21)$$

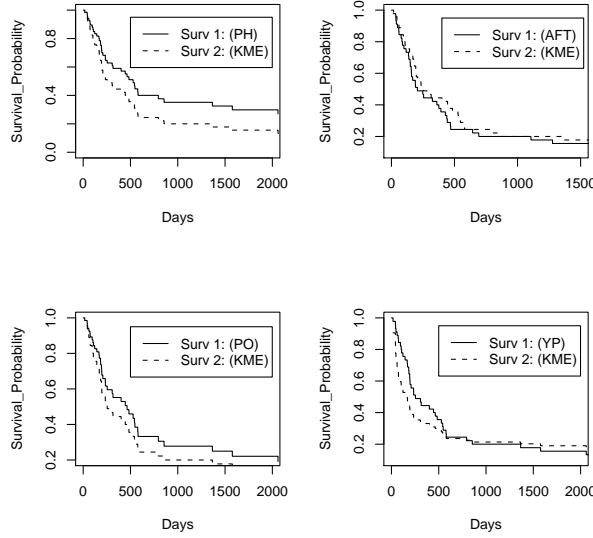


Figure 1.6: A plot illustrating survival curves for the PH, AFT, PO, and YP cases for the Gastrointestinal Tumor Study Group data.

where $Y_i = I(X_i \geq x)$, $N_i(x) = I(X_i \leq x, \Delta_i = 1)$, $A(x)$ represents a cumulative hazard function, d represents the derivative, and $\tau = \inf\{t : P[X > t] = 0\}$. Another possibility involves a Poisson form representation; that is,

$$\prod_{i=1}^n \prod_{x \in [0, \tau]} (Y_i(x) dA(x))^{dN_i(x)} \exp\left(-\int_0^\tau Y_i(x) dA(x)\right). \quad (1.22)$$

In the next chapter, these forms will be modified to be applicable for the two-sample case. Additionally, we will explore different possibilities besides the basic pseudo-likelihood method with frequentist bootstrap standard error estimation. In the standard bootstrap, sampling with replacement assigns to each sequence of censoring patterns certain weights, which are usually drawn from a multinomial distribution. If we employ Bayesian bootstrap instead, these sequences are assigned weights drawn from a Dirichlet distribution, which vary continuously, virtually eliminating the chance of a zero weight (Price *et al.*, 2005). The main advantage of using Bayesian bootstrap is that it is not affected by the replica bias which is inherent in the frequentist bootstrap.

1.6 Bayesian Bootstrap

In this section, we will be estimating the standard error by Bayesian bootstrap rather than the frequentist bootstrap. As stated earlier, the Bayesian bootstrap serves as an alternative to sampling from replacement using the frequentist bootstrap, which in effect assign each sequence of integer weights to be drawn from the multinomial distribution. Instead, in the Bayesian bootstrap approach, we sample from an undetermined distribution to which we associate a noninformative prior. We combine this prior with the sample likelihood using Bayes' theorem, to arrive at a posterior distribution (Dirichlet) based on the portion of the original population which each sampled sequence represents (Price *et al.*, 2005). Hence, we can think of these Bayesian bootstrap replicas as samples coming from a Bayesian posterior distribution (Durbin *et al.*, 1998). We will now explore three different methods which utilize the Bayesian bootstrap approach.

First, we will describe a general description of the Bayesian bootstrap method for a set of data with no censoring. Consider a sample of independent and identically distributed observations X_1, \dots, X_n from a cdf $F(\cdot)$. Let $k = k(n)$ be defined as the number of distinct, ordered observations, $t_{(1)} < \dots < t_{(k)}$. For these observations, we define the weights $w_j, j = 1, \dots, k$ to be distributed as Dirichlet($k; 1, \dots, 1$). More explicitly,

$$w_1 = \frac{v_1}{v_1 + \dots + v_k}, \dots, w_k = \frac{v_k}{v_1 + \dots + v_k}, \quad (1.23)$$

where v_1, \dots, v_k are k iid $\exp(1)$ variables. Then, the Bayesian bootstrap distribution of $F(\cdot)$ is given by $\hat{F}(x) = \sum_{j=1}^k w_j I(t_{(j)} \leq x)$. The Bayesian bootstrap is the noninformative limit of the Dirichlet process posterior (Rubin, 1981).

The Bayesian bootstrap method was extended by Lo (1993) to account for censored data. Susarla and van Ryzin (1976) had originally proposed a nonparametric Bayesian estimator of a survival curve based on incomplete right-censored data. Lo's (1993) Bayesian bootstrap is simply the noninformative limit of the posterior. The procedure can be described in terms of n i.i.d. standard exponential random variables v_1, \dots, v_n , and then implementing the mass-shifting algorithm used in defining the Kaplan-Meier estimator. Note that the original Kaplan-Meier estimator based on right-censored data $\{(X_i, \delta_i); i = 1, \dots, n\}$ can

be written as:

$$\hat{S}(t) = \prod_{j:t(j) \leq t} \left(1 - \frac{\#D(j)}{\#R(j)} \right), \quad (1.24)$$

where $t_{(1)} < t_{(2)} < \dots < t_{(k)}$ represent the distinct ordered values for times to death (uncensored data), and for $j = 1, \dots, k$, $D(j) = \{i : X_i = t_{(j)}, \Delta_i = 1\}$ and $R(j) = \{i : X_i \geq t_{(j)}\}$. In the Bayesian bootstrap method, the cardinalities $\#D(j)$ and $\#R(j)$ are replaced by certain sums of the standard exponential random variables “ v_q ’s” to obtain:

$$S^*(t) = \prod_{j:t(j) \leq t} \left(1 - \frac{\sum_{q \in D(j)} v_q}{\sum_{q \in R(j)} v_q} \right). \quad (1.25)$$

We will denote the k values that are obtained from computing $S^*(t)$ at the uncensored data points as weights w_1^*, \dots, w_k^* ; that is, $w_j^* = S^*(t_{(j)})$.

1.7 The Adaptive Rejection Metropolis Sampling (ARMS) Algorithm

Adaptive Rejection Metropolis Sampling (ARMS), developed by Gilks, Best and Tan (1995), is an extension of adaptive rejection sampling (ARS) (Gilks, 1992). To perform rejection sampling, ARS constructs a particular function, $h_n^*(x)$ of the log target density, $f(x)$. Under the assumption of log-concavity for $f(x)$, $h_n^*(x)$ is considered to be an envelope; that is, $h_n^*(x) \geq \log f(x)$ everywhere in D , where D represents the domain of f as an interval on the real line. When a particular point is rejected, ARS updates the envelope density such that it is nearer to the target log density, $\log f(x)$. This envelope is constructed where secants intersect at $\log f(x)$. The purpose of updating the envelope density is to make rejections less likely, thereby reducing computational time in Gibbs sampling applications. Further updates will also generate independent samples from f . The log-concavity assumption often holds in many cases, even under very complicated full conditional distributions. Unfortunately, the assumption of log-concavity does not hold under distributions that are not from an exponential family as well as in many non-linear models. In our case, the Yang and Prentice model is not from an exponential family. Hence, we will need to implement ARMS instead, which includes an additional Metropolis step to accommodate non-concave

log densities. Without log-concavity, $h_n^*(x)$, in general, will no longer be considered to be an envelope function for $\log f(x)$. Furthermore, ARMS will no longer produce independent samples from f . The general algorithm can be described as the following:

Define $S_n = \{x_i; i = 0, \dots, n+1\}$ as an ascending, current set of abscissae, with x_0 and x_{n+1} representing the possibly nonfinite lower and upper limits of the domain D . Next, we let $L_{ij}(x; S_n)$ represent a straight line through the set of points $[x_i, \log f(x_i)]$ and $[x_j, \log f(x_j)]$ for $1 \leq i < j \leq n$. For the other (i, j) , $L_{ij}(x; S_n)$ will be undefined. We denote a piecewise function of the following form:

$$h_n^*(x) = \max [L_{i,i+1}(x, S_n), \min \{L_{i-1,i}(x, S_n), L_{i+1,i+2}(x, S_n)\}], \quad x_i \leq x \leq x_{i+1}. \quad (1.26)$$

Furthermore, we define the log-concavity of the function f as

$$\log f(a) - 2 \log f(b) + \log f(c) < 0, \quad (1.27)$$

for all $a, b, c \in D$ such that $a < b < c$. If b is undefined, $\min(a, b) = \min(b, a) = \max(a, b) = \max(b, a) = a$. The sampling density $g_n^*(x)$ is denoted by

$$g_n^*(x) = \frac{1}{r_n^*} \exp h_n^*(x), \quad (1.28)$$

where

$$r_n^* = \int \exp h_n^*(x) dx. \quad (1.29)$$

We will denote X_{cur} as the current value of x , X_R as the new value from f that will either replace X_{cur} at the Hastings-Metropolis rejection step or X_A at the Hastings-Metropolis acceptance step, and X_A as the new value from f that will replace X at the ARS acceptance step. Finally, we execute the following steps as outlined by Gilks, Best, and Tan (1995) to implement ARMS:

step 0, initialize n and S_n independently of X_{cur} ;
step 1, sample X from $g_n^*(x)$;
step 2, sample U from $\text{uniform}(0, 1)$;
step 3, if $U > f(X)/\exp h_n^*(X)$ then {
 ARS rejection step:
 set $S_{n+1} = S_n \cup \{X\}$;
 relabel points in S_{n+1} in ascending order;
 increment n and go back to step 1;}
 else {
 ARS acceptance step:
 set $X_A = X$;}
step 4, sample U from $\text{uniform}(0, 1)$;
step 5, if $U > \min \left[1, \frac{f(X_A) \min \{f(X_{cur}), \exp h_n^*(X_{cur})\}}{f(X_{cur}) \min \{f(X_A), \exp h_n^*(X_A)\}} \right]$ then {
 Hastings – Metropolis rejection step:
 set $X_R = X_{cur}$;}
 else {
 Hastings – Metropolis acceptance step:
 set $X_R = X_A$;}
step 6, return X_R .

If f is log-concave, step 5 will always accept since h_n^* will be reduced to

$$\min [L_{i-1,i}(x, S_n), L_{i+1,i+2}(x, S_n)], \quad x_i \leq x \leq x_{i+1}, \quad (1.30)$$

forming an envelope for $\log f$. Hence, ARMS reduces to ARS under the condition that f is log-concave. In later chapters, the general ARMS algorithm stated above will be used to sample from the posterior distribution.

1.8 Outline of the Thesis

In the next chapter, we will explore smoothing the hazard function, which has a more convenient interpretation, rather than simply fitting the Nelson-Aalen estimate given by equation (1.18). We will study the two-sample problem and examine various methods such as Bayesian bootstrap involving likelihoods with Bayesian and Poisson forms. Additionally, we will explore the kernel-smoothing method which enables us to directly estimate the bandwidth, as well as another approach involving smoothing splines. A simulation study outlined in Yang and Prentice (2005) will be conducted to compare the former martingale approach with our proposed methods. In Chapter 3, we will extend the two-sample problem to a new class of regression models for censored data and examine issues of model misspecification and asymptotic relative efficiency (ARE). Finally, in Chapter 4, we will explore variational methods to calculate the marginal density of interest using either a full Bayesian approach or Bayesian bootstrap via the normal approximation to the likelihood. These variational methods may be more computationally feasible than MCMC techniques to obtain the marginal density and perform model selection via the Bayes factor.

Chapter 2

The Two-Sample Case

2.1 Likelihood-Based Inference

In this chapter, we will examine the two-sample problem, and explore several Bayesian approaches. To estimate the short-term and long-term hazard ratios, Yang and Prentice (2005) had only considered a frequentist approach using martingales. Although this is one conceivable possibility, there are perhaps other plausible estimation methods with regards to the two-sample problem. We will now consider additional possibilities and evaluate their effectiveness. These include both the Bayesian bootstrap approach as well as Bayesian smoothing methods.

First, we will consider the simple case of comparing survival rates based on two, independent samples. Let $S_i(t) = \Pr(T_i > t)$, $F_i(t) = 1 - S_i(t)$, $f_i(t) = \frac{dF_i(t)}{dt}$, $h_i(t) = \frac{f_i(t)}{S_i(t)}$ denote the survival function, cumulative distribution function (cdf), probability density function (pdf), and hazard function of the i th group for $i=1,2$. Assume that $T_{ij} \stackrel{iid}{\sim} S_i(\cdot)$ for $j = 1, \dots, n_i$ and $i = 1,2$. We observe $X_{ij} = \min(T_{ij}, C_{ij})$ for $i = 1,2$, and $j=1, \dots, n_i$, where C_{ij} are the censoring times, independent of T_{ij} . Let $\Delta_{ij} = I(T_{ij} \leq C_{ij})$ denote the censoring indicator. Now, we will consider ordered observations from the second group; that is, $X_{2(1)} \leq X_{2(2)} \leq \dots \leq X_{2(n_2)}$, with corresponding censoring indicators $\Delta_{2(1)}, \Delta_{2(2)}, \dots, \Delta_{2(n_2)}$. We will also denote the times $x_1^0 < \dots < x_{k_n}^0$ to represent the observed, distinct $X_{2(j)}$'s corresponding to $\Delta_{2(j)} = 1$, where $k_n \leq \sum_{j=1}^{n_2} \Delta_{2j}$.

Next, we will need to sample from the posterior distribution. Kim and Lee (2003)

develop a Bayesian bootstrap for proportional hazards models. In our case, we will extend that to also encompass the Yang-Prentice (2005) model. We will consider two forms of the likelihood which utilize product-integration: binomial and Poisson.

2.1.1 The Binomial form

The first possibility implements an extension to a binomial form. In the context of our problem, we can specify the binomial form given in (1.21) to the case of two samples; that is,

$$L^B(\theta, H_2) = \prod_{i=1}^2 \prod_{j=1}^{n_i} \prod_{x \in [0, \tau_i]} (Y_{ij}(x) dH_i(x))^{dN_{ij}(x)} (1 - Y_{ij}(x) dH_i(x))^{1 - dN_{ij}(x)}, \quad (2.1)$$

where $\tau_i = \inf\{t > 0 : P_n[X_i > t] = 0\}$ represents the time period of interest, $N_{ij}(x) = I(X_{ij} \leq x, \Delta_{ij} = 1)$, $Y_{ij}(x) = I(X_{ij} \geq x)$, and H_1, H_2 represent the cumulative hazard functions from the first and second group, respectively. First let us notice the following relationship:

$$\begin{aligned} 1 - dH_1(x) &= \left(1 - \frac{dH_2(x)}{1 + (\frac{\theta_2}{\theta_1} - 1)S_2(x)} \right)^{\theta_2} \\ &\approx 1 - \frac{dH_2(x) \theta_2}{1 + (\frac{\theta_2}{\theta_1} - 1)S_2(x)}, \end{aligned} \quad (2.2)$$

where the approximation holds as long as the point masses corresponding to $dH_2(x)$ are small. We substitute this result into the binomial form log-likelihood, which gives us:

$$\begin{aligned} &\left\{ \sum_{j=1}^{n_1} \sum_{x \in \mathfrak{S}_{n_1}} \left[\Delta N_{1j}(x) \log \left(\frac{\theta_2 Y_{1j}(x) dH_2(x)}{1 + (\frac{\theta_2}{\theta_1} - 1)S_2(x)} \right) \right] \right\} \\ &+ \left\{ \sum_{j=1}^{n_1} \sum_{x \in \mathfrak{S}_{n_1}} (1 - \Delta N_{1j}(x)) \log \left(1 - \frac{\theta_2 Y_{1j}(x) dH_2(x)}{1 + (\frac{\theta_2}{\theta_1} - 1)S_2(x)} \right) \right\} \\ &\quad + \left\{ \sum_{j=1}^{n_2} \sum_{x \in \mathfrak{S}_{n_2}} [\Delta N_{2j}(x) \log (Y_{2j}(x) dH_2(x))] \right\} \\ &+ \left\{ \sum_{j=1}^{n_2} \sum_{x \in \mathfrak{S}_{n_2}} [(1 - \Delta N_{2j}(x)) \log (1 - Y_{2j}(x) dH_2(x))] \right\} \end{aligned} \quad (2.3)$$

where \mathfrak{S}_{n_i} is defined as $\{x : \Delta N_i(x) \geq 1\}$, $\Delta N_i(x) = N_i(x) - N_i(x-)$, and $N_i(x) = \sum_{j=1}^{n_i} N_{ij}(x)$, for $i = 1, 2$. We will use the resulting w_l 's computed from Lo's (1993) method to write a closed form for each of the two expressions. Note that we can express $S_2(x)$ as:

$$S_2(x) = \begin{cases} 1 & x = 0 \\ w_1, \dots, w_l & x_{l-1}^0 \leq x \leq x_l^0 \end{cases} \quad (2.4)$$

where $l = 2, \dots, k_{n_2}$ and $w \in [0, 1]^{k_{n_2}}$. Next, we can notice that when T is discrete, $h_T(x_l) = \Pr(T = x_l | T \geq x_l) = \frac{f_T(x_l)}{S_T(x_{l-1})}$, for $l = 1, 2, \dots, k$, where $S_T(x_0) = 1$ and $f_T(x_l) = \Pr(T = x_l)$. Because $f_T(x_l) = S_T(x_{l-1}) - S_T(x_l)$, it follows that $h_T(x_l) = 1 - \frac{S_T(x_l)}{S_T(x_{l-1})}$, $l = 1, 2, \dots, k$. Hence, we can surmise that $\Delta H_2(x) = 1 - \frac{w_l}{w_{l-1}}$ if $x \in (x_l^0, x_{l+1}^0]$, and $dH_2(x_l) = \left(1 - \frac{w_l}{w_{l-1}}\right)$. The log-likelihood now becomes:

$$\begin{aligned} & \sum_{j=1}^{n_1} \sum_{l=1}^{k_{n_1}} \Delta N_{1j}(x_l^0) \log \left(\theta_2 \frac{Y_{1j}(x_l^0) \left(1 - \frac{w_l}{w_{l-1}}\right)}{1 + \left(\frac{\theta_2}{\theta_1} - 1\right)w_l} \right) \\ & + \sum_{j=1}^{n_1} \sum_{l=1}^{k_{n_1}} (1 - \Delta N_{1j}(x_l^0)) \log \left(1 - \theta_2 \frac{Y_{1j}(x_l^0) \left(1 - \frac{w_l}{w_{l-1}}\right)}{1 + \left(\frac{\theta_2}{\theta_1} - 1\right)w_l} \right) \\ & + \sum_{j=1}^{n_2} \sum_{l=1}^{k_{n_2}} \Delta N_{2j}(x_l^0) \log \left(Y_{2j}(x_l^0) \left(1 - \frac{w_l}{w_{l-1}}\right) \right) \\ & + \sum_{j=1}^{n_2} \sum_{l=1}^{k_{n_2}} (1 - \Delta N_{2j}(x_l^0)) \log \left(1 - Y_{2j}(x_l^0) \left(1 - \frac{w_l}{w_{l-1}}\right) \right) \end{aligned} \quad (2.5)$$

Finally, to obtain the posterior distribution, we can place $\text{Ga}(a, b)$ priors on θ_1 and θ_2 and implement adaptive rejection metropolis sampling (ARMS), which was developed by Gilks *et al.* (1995). Next, we will examine the Poisson form likelihood.

2.1.2 The Poisson form

There is another formulation besides the binomial form that can allow for any cumulative hazard function H , and not simply an absolutely continuous one. That next possibility is a Poisson form likelihood. Andersen *et al.* (1993) point out that for continuous H , the product-integrals of (2.1) and the binomial form both evaluate to the exponential in the Poisson form. In our case, the Poisson form likelihood, based on equation (1.22), can

be written as:

$$L^P(\theta, H_2) = \prod_{i=1}^2 \prod_{j=1}^{n_i} \prod_{x \in [0, \tau_i]} (Y_{ij}(x) dH_i(x))^{dN_{ij}(x)} \exp \left(- \int_0^{\tau_i} Y_{ij}(x) dH_i(x) \right), \quad (2.6)$$

After substituting the result given by equation (2.2), we obtain the following log-likelihood:

$$\begin{aligned} & \sum_{j=1}^{n_1} \sum_{x \in \mathfrak{S}_{n_1}} \left\{ \Delta N_{1j}(x) \log \left\{ \left[1 - \left(1 - \frac{dH_2(x)}{1 + \left(\frac{\theta_2}{\theta_1} - 1 \right) S_2(x)} \right)^{\theta_2} \right] \right\} \right\} \\ & \quad + \sum_{j=1}^{n_1} \sum_{x \in \mathfrak{S}_{n_1}} \left(-\theta_2 \int_0^{\tau_1} \frac{Y_{1j}(x) dH_2(x)}{1 + \left(\frac{\theta_2}{\theta_1} - 1 \right) S_2(x)} \right) \\ & + \sum_{j=1}^{n_2} \sum_{x \in \mathfrak{S}_{n_2}} \left\{ \Delta N_{2j}(x) \log (Y_{2j}(x) dH_2(x)) - \int_0^{\tau_2} Y_{2j}(x) dH_2(x) \right\} \end{aligned} \quad (2.7)$$

After using the properties above given by equation (2.4), the simplified log-likelihood can be represented as:

$$\begin{aligned} & \sum_{l=1}^{k_{n_1}} \left\{ \sum_{j \in D_1(x_l^0)} \log \left[1 - \left(1 - \frac{\left(1 - \frac{w_l}{w_{l-1}} \right)}{1 + \left(\frac{\theta_2}{\theta_1} - 1 \right) w_l} \right)^{\theta_2} \right] - \left(\sum_{j \in R_1(x_l^0)} \theta_2 \frac{\left(1 - \frac{w_l}{w_{l-1}} \right)}{1 + \left(\frac{\theta_2}{\theta_1} - 1 \right) w_l} \right) \right\} \\ & \quad + \sum_{l=1}^{k_{n_2}} \left\{ \sum_{j \in D_2(x_l^0)} \log \left[\left(1 - \frac{w_l}{w_{l-1}} \right) \right] - \sum_{j \in R_2(x_l^0)} \left(1 - \frac{w_l}{w_{l-1}} \right) \right\} \end{aligned} \quad (2.8)$$

where $R_i(x_l^0) = \{j : Y_{ij}(x_l^0) = 1\}$ and $D_i(x_l^0) = \{j : \Delta N_{ij}(x_l^0) = 1\}$ for $i = 1, 2$. Note that the set $D_i(x_l^0)$ represents the set of observations which fail at time x_l^0 and $R_i(x_l^0)$ represents the set of observations still at risk at time x_l^0 for each of the i groups.

2.2 Smoothing Methods

2.2.1 Kernel-Smoothing Method

We assume that the distributions of T_1 and T_2 are absolutely continuous with respect to the Lebesgue measure. This method allows us to directly estimate the bandwidth, with an alternative representation for the hazard functions based on a mixture of Weibull and beta distributions. Our bandwidth will be represented as an unknown parameter η

with an imposed prior. In this approach, we will find optimal estimates by implementing the following log-likelihood:

$$\begin{aligned} \sum_{j=1}^{n_1} \left[\Delta_{1j} \log \left(h_2^*(x_{1j}, \eta) \frac{\theta_1 \theta_2}{\theta_1 + (\theta_2 - \theta_1) e^{-H_2^*(x_{1j}, \eta)}} \right) - \theta_2 \log \left(1 - \frac{\theta_1}{\theta_2} + \frac{\theta_1}{\theta_2} e^{H_2^*(x_{1j}, \eta)} \right) \right] \\ + \sum_{j=1}^{n_2} [\Delta_{2j} \log h_2^*(x_{2j}, \eta) - H_2^*(x_{2j}, \eta)] \end{aligned} \quad (2.9)$$

where

$$h_2^*(t) = \sum_{i=1}^{k_n} \left[\eta \left(\frac{t}{x_i^0} \right)^{\eta-1} \prod_{j \leq i} (1 - \nu_j) \right] \quad (2.10)$$

and

$$H_2^*(t) = \sum_{i=1}^{k_n} \left[\left(\frac{t}{x_i^0} \right)^\eta x_i^0 \prod_{j \leq i} (1 - \nu_j) \right]. \quad (2.11)$$

Here, ν_j represents samples drawn from independent $\text{beta}(d_j, r_j - d_j)$ random variables, where d_j equals the number of deaths at time x_j^0 , and r_j equals the number of units at risk just before time x_j^0 . We will impose three independent gamma priors; that is, $\theta_1 \sim \text{Ga}(a, b)$, $\theta_2 \sim \text{Ga}(a, b)$, and $\eta \sim \text{Ga}(a, b)$. As before, we will implement the ARMS procedure to obtain posterior estimates of β_1 and β_2 with corresponding posterior standard deviations, where $\beta_1 = \log \theta_1$ and $\beta_2 = \log \theta_2$.

2.2.2 Smoothing-Spline Method

Next, we will further assume that:

$$h_2(t) = -\frac{d \log S_0(t)}{dt} = \sum_{j=0}^{k_n} \lambda_j I(x_j^0 \leq t \leq x_{j+1}^0), \quad (2.12)$$

where $x_1^0 < \dots < x_{k_n}^0$ are distinct ordered observed survival times from the second group and $x_0^0 = 0$ and $x_{k_n+1}^0 = \infty$. It also follows that

$$H_2(t) = \sum_{j=0}^{m(t)-1} \lambda_j (x_{j+1}^0 - x_j^0) + \lambda_{m(t)} (t - x_{m(t)}^0) \quad (2.13)$$

if $t \in [x_{(m(t))}, x_{(m(t)+1)})$, $m(t) = \max\{j : x_j^0 < t\}$. Our goal is to estimate β_1 , β_2 , and $\lambda = (\lambda_0, \lambda_1, \dots, \lambda_{k_n})^T \in [0, \infty)^{k_n+1}$ given the log-likelihood:

$$\log L(\beta_1, \beta_2, \lambda) = \sum_{i=1}^{n_1} [\Delta_{1i} \log h_1(x_i) + \log S_1(x_i)] + \sum_{i=1}^{n_2} [\Delta_{2i} \log h_2(x_i) + \log S_2(x_i)], \quad (2.14)$$

where $h_1(x_i) = \frac{\theta_1 \theta_2 h_2(x_i)}{\theta_2 + \theta_1 (1 - S_2(x_i))}$, $S_2(x_i) = e^{-H_2(x_i)}$, $H_2(\cdot)$ is defined in (2.13), $S_1(x_i) = \left[1 + \frac{\theta_1}{\theta_2} \frac{1 - S_2(x_i)}{S_2(x_i)}\right]^{-\theta_2}$, $\theta_1 = e^{\beta_1}$, and $\theta_2 = e^{\beta_2}$. Here, we will let $0 < \lambda_j = e^{\xi_j}$ and generate $\xi_0, \xi_1, \dots, \xi_{k_n}$ by an autoregressive first-order process. To accomplish this, let $0 \leq \alpha < 1$, and $\tau^2 > 0$. Define

$$\sigma_\xi^2 = \frac{\tau^2}{1 - \alpha^2} \quad (2.15)$$

To generate ξ , let the first value, ξ_0 be normal with expected value 0 and variance σ_ξ^2 . The next step is to simulate the other ξ_j 's for $j = 1, \dots, k_n$ according to the setup

$$\xi_j = \alpha \xi_{j-1} + \epsilon_j, \quad (2.16)$$

where the ϵ_j are independent and normal with mean 0 and variance τ^2 . Notice that the joint distribution of $(\xi_0, \xi_1, \dots, \xi_{k_n})$ is multivariate normal with mean vector m equal to $\mathbf{0}$ and autocovariance matrix V with diagonal elements $\frac{\tau^2}{1 - \alpha^2}$ and off-diagonal elements, $\frac{\tau^2}{1 - \alpha^2} \alpha^{|l-1|}$ for $l = 2, \dots, k_n + 1$. Hence, the joint density of $(\lambda_0, \lambda_1, \dots, \lambda_{k_n})$ is multivariate lognormal with the following representation:

$$f_\lambda(\lambda) = (2\pi)^{-(k_n+1)/2} |V|^{-1/2} [\lambda_0, \lambda_1, \dots, \lambda_{k_n}]^{-1} \exp \left[-(\ln \lambda - m)^T V^{-1} (\ln \lambda - m) / 2 \right], \quad (2.17)$$

where $\ln \lambda = (\ln \lambda_0, \ln \lambda_1, \dots, \ln \lambda_{k_n})$ is a $k_n + 1$ -component column vector and $\lambda = \exp(\xi)$. The mean of $(\lambda_0, \lambda_1, \dots, \lambda_{k_n})$ is $\exp(m_i + 0.5v_{ii})$ and each element of the corresponding autocovariance matrix equals $[\exp[(m_i + m_j) + (v_{ii} + v_{jj})/2]] [\exp(v_{ij}) - 1]$, where v_{ii} represents the i th diagonal element of the matrix V and v_{ij} represents the ij th element of V . We place $\text{Ga}(a, b)$ priors on τ^2 , θ_1 , and θ_2 as well as a uniform(0, 1) prior on α to obtain samples from the posterior distribution by implementing ARMS.

2.3 A Simulation Studies

2.3.1 Comparing Various Methods

Yang and Prentice (2005) conducted a simulation study that had examined two particular cases: the case where there is no treatment effect (corresponding to $\beta = (0, 0)^T$) and the case where the effect of treatment is negative but eventually becomes positive (corresponding to $\beta = (\frac{1}{2}, -\frac{1}{2})^T$); that is, the survival curves cross. Yang and Prentice (2005) estimated β_1 and β_2 and their corresponding standard errors using a martingale approach. Here, we essentially replicate their simulation study to compare the estimates based on the various Bayesian methods described in earlier sections. Following Yang and Prentice (2005), we set $S_2(t) = \frac{1}{1+t}$ and $S_1(t) = \frac{1+S_2(t)}{[1+\theta S_2(t)]^{\frac{\theta_1}{\theta_2}}}$. We generate censoring variables from a lognormal distribution with parameters $\mu = 0.85$ and $\sigma = 0.5$, which gives us mean $e^{\mu+0.5\sigma^2} = 2.65$ and variance $e^{2(\mu+\sigma^2)} - e^{2\mu+2\sigma^2} = 2.00$. The sample sizes for each of the two groups are set to be equal and tested at $n = 40$ and 80 for each of the two censoring rates under the cases $(\beta = (0, 0)^T$ and $\beta = (\frac{1}{2}, -\frac{1}{2})^T$) and 1000 replicates are generated. Our simulation will implement ARMS with 1000 samples to obtain the 95% confidence intervals. The results of the three methods (martingale, reproduced from Yang and Prentice (2005), Bayesian Bootstrap Method (Binomial and Poisson forms), and Bayesian Smoothing Methods (Kernel-Smoothing and Smoothing-Spline) are given in Tables 2.1, 2.2 and Figures 2.1-2.8:

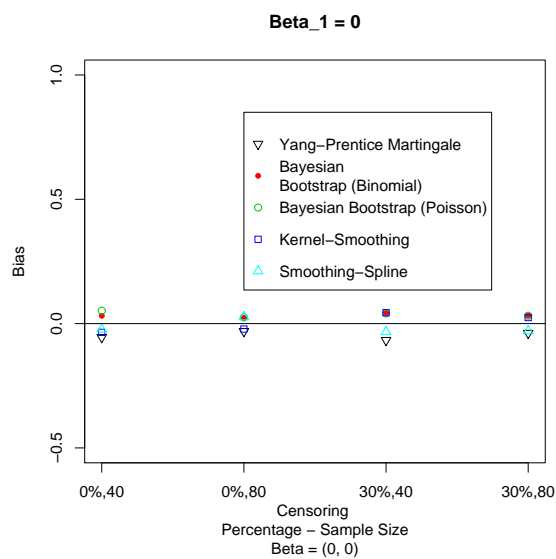


Figure 2.1: A plot of the biases for β_1 under the case of no treatment effect, $\beta = (0,0)$.

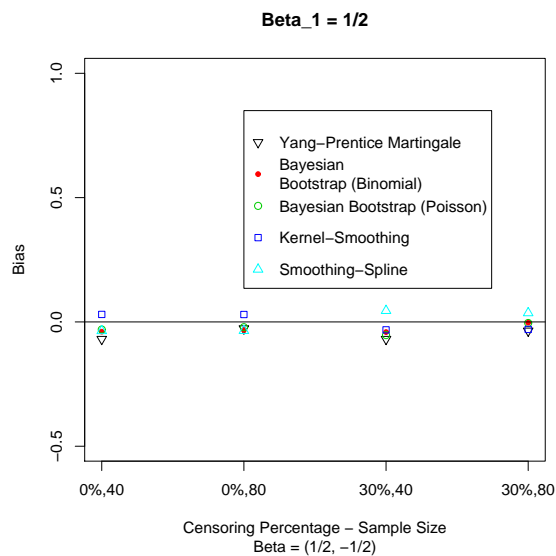


Figure 2.2: A plot of the biases for β_1 under the case where the treatment effect starts negative but eventually becomes positive, $\beta = (1/2, -1/2)$.

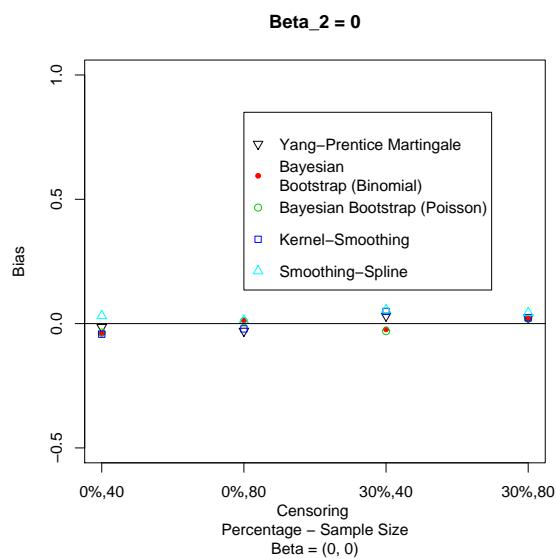


Figure 2.3: A plot of the biases for β_2 under the case of no treatment effect, $\beta = (0,0)$.

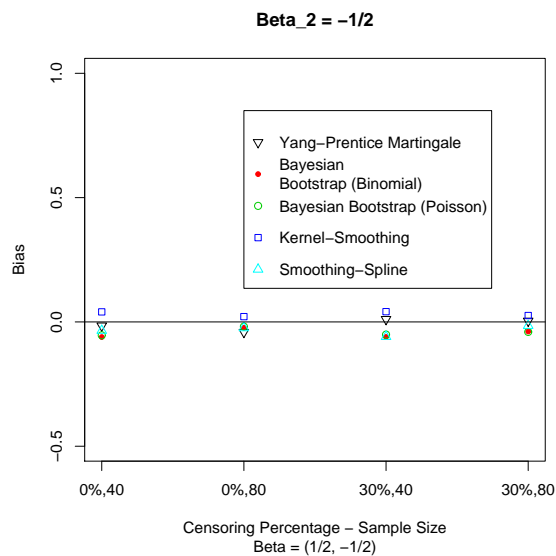


Figure 2.4: A plot of the biases for β_2 under the case where the treatment effect starts negative but eventually becomes positive, $\beta = (1/2, -1/2)$.

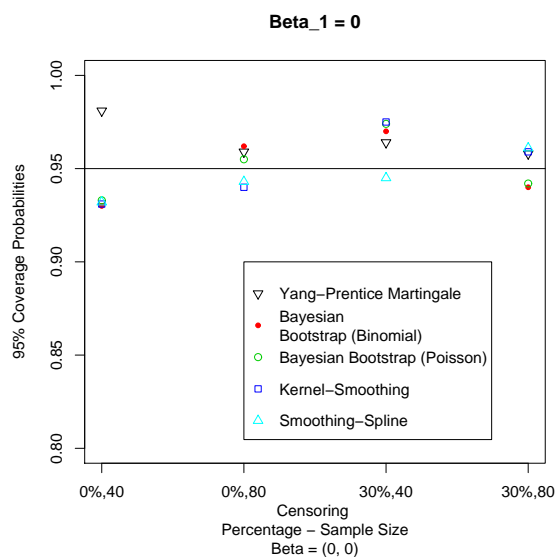


Figure 2.5: A plot of the 95% credible probability of the confidence intervals for β_1 under the case of no treatment effect, $\beta = (0,0)$.

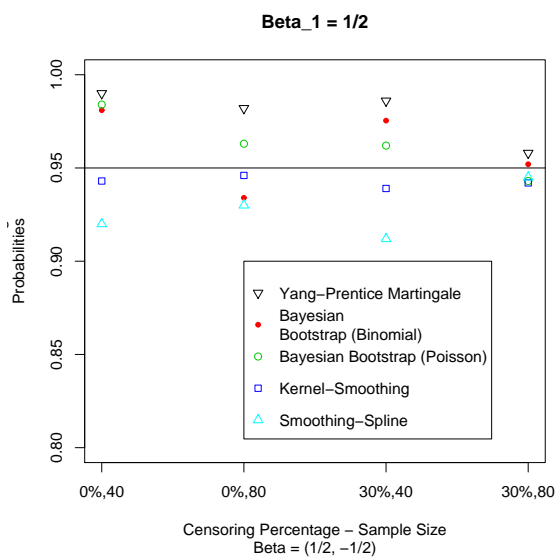


Figure 2.6: A plot of the 95% credible probability of the confidence intervals for β_1 under the case where the treatment effect starts negative but eventually becomes positive, $\beta = (1/2, -1/2)$.

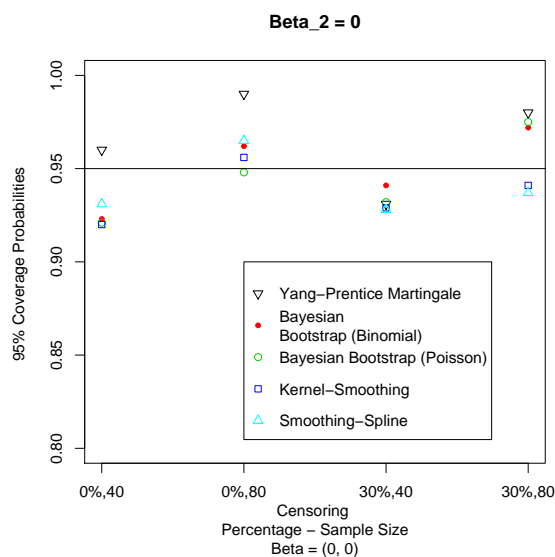


Figure 2.7: A plot of the 95% credible probability of the confidence intervals for β_2 under the case of no treatment effect, $\beta = (0,0)$.

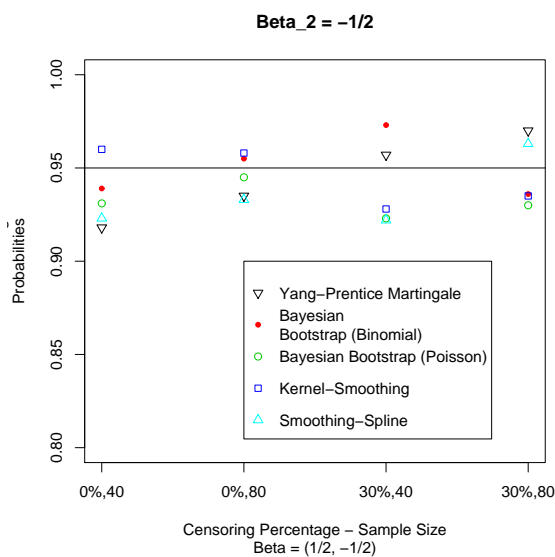


Figure 2.8: A plot of the 95% credible probability of the confidence intervals for β_2 under the case where the treatment effect starts negative but eventually becomes positive, $\beta = (1/2, -1/2)$.

Table 2.1: A simulation study for each method regarding the non-crossing survival curves case ($\beta = (0, 0)$) based on 1000 repetitions. Bias of $\hat{\beta}$ as well as the estimated 95% coverage probability of the confidence intervals, YP = Yang-Prentice, BB = Bayesian Bootstrap

Method	Censoring rate	n	Bias		Est. 95% CI cov. prob	
			β_1	β_2	β_1	β_2
YP Martingale	0%	40	-0.0551	-0.0145	0.9810	0.9600
	0%	80	-0.0309	-0.0297	0.9590	0.9900
	30%	40	-0.0668	0.0313	0.9640	0.9310
	30%	80	-0.0382	0.0241	0.9580	0.9800
BB - Binomial	0%	40	0.0313	-0.0402	0.9300	0.9230
	0%	80	0.0231	0.0129	0.9620	0.9430
	30%	40	0.0426	-0.0231	0.9700	0.9410
	30%	80	0.0321	0.0225	0.9400	0.9720
BB - Poisson	0%	40	0.0522	-0.0362	0.9330	0.9210
	0%	80	0.0251	0.0092	0.9550	0.9480
	30%	40	0.0412	-0.0304	0.9740	0.9320
	30%	80	0.0317	0.0199	0.9420	0.9570
Kernel-Smoothing	0%	40	-0.0352	-0.0427	0.9310	0.9200
	0%	80	-0.0219	-0.0214	0.9400	0.9560
	30%	40	0.0438	0.0491	0.9750	0.9290
	30%	80	0.0244	0.0230	0.9590	0.9410
Smoothing-Spline	0%	40	-0.0231	0.0310	0.9320	0.9310
	0%	80	0.0263	0.0125	0.9430	0.9650
	30%	40	-0.0332	0.0545	0.9450	0.9280
	30%	80	-0.0298	0.0433	0.9610	0.9370

From these results, we can notice that all three methods give accurate results with good coverage. Each of the methods are comparable in terms of bias, for both sample sizes ($n = 40, 80$) and censoring rates. Higher censoring rates usually tend to create more bias, while increasing the sample size usually helps to reduce bias as well as improve coverage. The Monte-Carlo standard errors range from 0.05 to 0.10 for the Bayesian bootstrap method / binomial form, 0.04 to 0.09 for the Bayesian bootstrap method / Poisson form, 0.02 to 0.09 for the kernel-smoothing method, and 0.02 to 0.08 for the smoothing-spline method. In the next chapter, we will examine the regression case, rather than assume two samples.

Table 2.2: A simulation study for each method regarding the crossing survival curves case ($\beta = (\frac{1}{2}, -\frac{1}{2})$) based on 1000 repetitions. Bias of $\hat{\beta}$ as well as the estimated 95% coverage probability of the confidence intervals, YP = Yang-Prentice, BB = Bayesian Bootstrap

Method	Censoring rate	n	Bias		Est. 95% CI cov. prob	
			β_1	β_2	β_1	β_2
YP Martingale	0%	40	-0.0693	-0.0155	0.9900	0.9180
	0%	80	-0.0267	-0.0413	0.9820	0.9350
	30%	40	-0.0693	0.0107	0.9860	0.9570
	30%	80	-0.0359	0.0040	0.9580	0.9700
BB - Binomial	0%	40	-0.0392	-0.0609	0.9810	0.9390
	0%	80	-0.0321	-0.0238	0.9340	0.9550
	30%	40	-0.0409	-0.0605	0.9754	0.9730
	30%	80	-0.0030	-0.0386	0.9520	0.9360
BB - Poisson	0%	40	-0.0305	-0.0538	0.9840	0.9310
	0%	80	-0.0212	-0.0186	0.9630	0.9450
	30%	40	-0.0531	-0.0508	0.9620	0.9230
	30%	80	-0.0043	-0.0402	0.9430	0.9300
Kernel-Smoothing	0%	40	0.0302	0.0404	0.9430	0.9600
	0%	80	0.0298	0.0213	0.9460	0.9580
	30%	40	-0.0319	0.0413	0.9390	0.9280
	30%	80	-0.0301	0.0257	0.9420	0.9350
Smoothing-Spline	0%	40	-0.0361	-0.0348	0.9200	0.9230
	0%	80	-0.0354	-0.0244	0.9300	0.9330
	30%	40	0.0451	-0.0596	0.9120	0.9220
	30%	80	0.0360	-0.0151	0.9450	0.9630

2.4 An Application: The Gastrointestinal Tumor Study Group

Now we will apply each of the techniques in the previous section (Bayesian Bootstrap and Smoothing Methods) to the Gastrointestinal Tumor Study Group dataset introduced in Chapter 1. We will estimate $\hat{\theta}_1$ and $\hat{\theta}_2$ along with their corresponding posterior standard deviations and 95% credible intervals by sampling from the posterior distribution via ARMS as described earlier. Recall that in Chapter 1, the estimates obtained by Yang and Prentice via martingales were $\hat{\theta}_1 = 4.97$ and $\hat{\theta}_2 = 0.39$ with a 95% CI of (1.80, 13.70) and (0.24, 0.65) while the pseudo-likelihood approach provided estimates of $\hat{\theta}_1 = 5.00$ and $\hat{\theta}_2 = 0.48$ with a 95% CI of (1.22, 8.71) and (0.21, 0.75). The results of our methods are as follows:

Table 2.3: Results corresponding to fitting the Gastrointestinal Tumor Study Group Data under the Bayesian Bootstrap - Binomial, Bayesian Bootstrap - Poisson, Kernel Smoothing, and Smoothing-Spline Methods. Estimates of $\hat{\theta}_1$ and $\hat{\theta}_2$, corresponding posterior standard deviations, and 95% credible intervals. CrI = Credible Interval, s.d. = standard deviation

Method	Parameter	Estimate	Posterior s.d.	95% CrI
Bayesian Bootstrap - Binomial	θ_1	4.743	1.783	(2.201, 8.975)
	θ_2	0.424	0.047	(0.326, 0.535)
Bayesian Bootstrap - Poisson	θ_1	4.861	1.821	(2.240, 9.308)
	θ_2	0.419	0.044	(0.335, 0.529)
Kernel-Smoothing	θ_1	5.063	2.053	(1.604, 9.102)
	θ_2	0.386	0.051	(0.251, 0.508)
Smoothing-Spline	θ_1	5.041	1.874	(2.248, 8.542)
	θ_2	0.404	0.050	(0.302, 0.511)

From these results, we notice that in each of the four methods, the null hypothesis of no difference between the groups ($H_0: \theta_1 = \theta_2 = 1$) is rejected. The estimates and confidence intervals for each of these estimation methods are quite close to each other. Overall, the new Bayesian methods produce estimates comparable to those of the frequentist methods produced in Chapter 1, but with narrower intervals which implies somewhat greater precision.

Chapter 3

A New Class of Regression Models

In this chapter, we will extend the Yang-Prentice model, which originally specified for only two samples, to the regression case. Let us set $\beta = \beta_2$ and $\gamma = \beta_2 - \beta_1$. The regression version of the Yang-Prentice model is now defined and reparameterized as:

$$h(x|z) = \frac{e^{\beta^T Z} h_0(x)}{1 + (e^{\gamma^T Z} - 1) S_0(x)}, \quad (3.1)$$

where $h_0(\cdot)$ represents the baseline hazard function. Furthermore, we can state the model in terms of the survival function with a baseline survival; that is,

$$S(x|z) = \left[1 + \frac{\theta_1(z)}{\theta_2(z)} \frac{1 - S_0(x)}{S_0(x)} \right]^{-\theta_2(z)}, \quad (3.2)$$

where we choose $\theta_1(z) = e^{\beta_1^T Z}$, $\theta_2(z) = e^{\beta_2^T Z}$, and $S_0(\cdot)$ represents the baseline survival function.

3.1 Likelihood approaches for the regression case

3.1.1 Poisson form

Here, we will develop an empirical likelihood approach for the regression case of the Yang-Prentice model. First, let us introduce some notation regarding the observed data. This data is represented as triplets; that is, $\{(X_i, \Delta_i, Z_i), i = 1, \dots, n\}$, $Z_i = (Z_{i1}, \dots, Z_{ip})^T$, $X_i = \min(T_i, C_i)$, T_i are the survival times, C_i are the censoring times, and $\Delta_i = I(T_i \leq C_i)$. We will assume that for each i , given Z_i , the survival time T_i and the censoring time C_i are

conditionally independent. To start, we will implement the Poisson form likelihood which is borrowed from Andersen, Borgan, Gill, and Keiding (1993); that is,

$$L^P(\beta, \gamma, H_0) = \prod_{i=1}^n \prod_{x \in [0, \tau]} (Y_i(x) dH_i(x))^{dN_i(x)} \exp \left(- \int_0^\tau Y_i(x) dH_i(x) \right), \quad (3.3)$$

where $Y_i = I(X_i \geq x)$, $N_i(x) = I(X_i \leq x, \delta_i = 1)$, $x \in [0, \tau]$ is the time period of interest, H_0 represents the cumulative baseline hazard function, d represents the derivative, and τ is determined as the largest observation time in the range where there is data. To construct the new likelihood, we will first notice the following relationship:

$$\begin{aligned} 1 - dH_i(x) &= \left(1 - \frac{dH_0(x)}{1 + (e^{\gamma^T Z_i} - 1) \prod_{s \leq x} (1 - dH_0(s))} \right)^{e^{\beta^T Z_i}} \\ &\approx 1 - \frac{dH_0(x) e^{\beta^T Z_i}}{1 + (e^{\gamma^T Z_i} - 1) \prod_{s \leq x} (1 - dH_0(s))}. \end{aligned} \quad (3.4)$$

Note that the approximation holds as long as the point masses corresponding to $dH_0(x)$ are small. Substituting this equality into the Poisson likelihood gives us

$$\begin{aligned} &\prod_{i=1}^n \prod_{x \in [0, \tau]} \left\{ Y_i(x) \left[1 - \left(1 - \frac{dH_0(x)}{1 + (e^{\gamma^T Z_i} - 1) \prod_{s \leq x} (1 - dH_0(s))} \right)^{e^{\beta^T Z_i}} \right] \right\}^{dN_i(x)} \\ &\times \exp \left\{ - \int_0^\tau Y_i(x) \frac{e^{\beta^T Z_i} dH_0(x)}{1 + (e^{\gamma^T Z_i} - 1) \prod_{s \leq x} (1 - dH_0(s))} \right\} \end{aligned} \quad (3.5)$$

Next, we will rewrite the product integral term $\prod_{x \in [0, \tau]}$ as $\prod_{x \in \mathfrak{S}_n}$, where \mathfrak{S}_n is defined as $\{x : \Delta N(x) \geq 1\}$, and $\Delta N(x) = N(x) - N(x-)$. It follows that

$$\begin{aligned} & \prod_{i=1}^n \prod_{x \in \mathfrak{S}_n} \left\{ 1 - \left(1 - \frac{dH_0(x)}{1 + (e^{\gamma^T Z_i} - 1) \prod_{s \leq x} (1 - dH_0(s))} \right)^{e^{\beta^T Z_i}} \right\}^{\Delta N_i(x)} \\ & \times \exp \left\{ -e^{\beta^T Z_i} \int_0^\tau \frac{Y_i(x) dH_0(x)}{1 + (e^{\gamma^T Z_i} - 1) \prod_{s \leq x} (1 - dH_0(s))} \right\} \end{aligned} \quad (3.6)$$

The final step is to reparameterize $H_0(x)$ in order to arrive at a suitable closed form that can be optimized by numerical procedures in software packages such as R. We will rewrite our baseline hazard function as:

$$H_0(x) = \begin{cases} 0, & x = 0, \\ w_1, & 0 < x \leq x_1^0, \\ w_1 + w_2, & x_1^0 < x \leq x_2^0, \\ w_1 + \dots + w_k, & x_{k-1}^0 < x \leq x_{k_n}^0, \end{cases} \quad (3.7)$$

where $0 \leq w_i \leq 1$, and $w \in [0, 1]^{k_n}$. From this representation, note that $\Delta H_0(x) = w_j$ if $x \in (x_j, x_{j+1}^0]$, and $H_0(x) = \sum_{j=1}^k w_j I(x \leq x_j^0)$. Finally, let us order all observations; that is, $X_{(1)} < X_{(2)} < \dots < X_{(n)}$, with corresponding censoring indicators $\delta_{(1)}, \delta_{(2)}, \dots, \delta_{(n)}$. We will define the times $t_{(1)} < \dots < t_{(k_n)}$ as the particular $X_{(j)}$'s where $\delta_{(j)} = 1$. After putting all of these facts together, we obtain our final likelihood:

$$\begin{aligned} & \prod_{j=1}^{k_n} \left[\prod_{i \in D(t_{(j)})} \left\{ 1 - \left(1 - \sum_{j=1}^k \frac{w_j}{1 + (e^{\gamma^T Z_i} - 1) \prod_{j' \leq j} (1 - w_{j'})} \right)^{e^{\beta^T Z_i}} \right\} \right] \\ & \times \exp \left(- \sum_{i \in R(t_{(j)})} e^{\beta^T Z_i} \sum_{j=1}^k \frac{w_j}{1 + (e^{\gamma^T Z_i} - 1) \prod_{j' \leq j} (1 - w_{j'})} \right), \end{aligned} \quad (3.8)$$

where $R(t_{(j)}) = \{i : Y_i(t_{(j)}) = 1\}$ and $D(t_{(j)}) = \{i : \Delta N_i(t_{(j)}) = 1\}$. Note that the set $D(t_{(j)})$ represents the set of observations which fail at time $t_{(j)}$ and $R(t_{(j)})$ represents the set of observations still at risk at time $t_{(j)}$.

3.1.2 A Pseudo-likelihood approach

The previous likelihood given by equation (3.8) relied on implementing the Poisson process. Now, we will derive a new likelihood that relies on the assumption that the hazard function exists and can be estimated. Notice that under the Yang-Prentice model,

$$S(t|\mathbf{z}) = \Pr[T > t|\mathbf{z}] = \left[1 + \frac{\theta_1(\mathbf{z})}{\theta_2(\mathbf{z})} K_0(t) \right]^{-\theta_2(\mathbf{z})}, \quad (3.9)$$

where $K_0(t) = \frac{1-S_0(t)}{S_0(t)}$ represents the baseline odds and $S_0(t) = \Pr[T > t|\mathbf{z} = 0]$ denotes the baseline survival function, $\theta_1(\mathbf{z})$ and $\theta_2(\mathbf{z})$ are non-negative functions of \mathbf{z} satisfying $\theta_1(\mathbf{0}) = \theta_2(\mathbf{0}) = 1$, and we assume that $\theta_j(\mathbf{z}) = \exp\{\beta_j^T \mathbf{z}\}$ for $j = 1, 2$. Let $h(t|\mathbf{z}) = -\frac{d \log S(t|\mathbf{z})}{dt}$ denotes the hazard function. Hence, the likelihood is given by

$$L(\beta_1, \beta_2, S_0(\cdot)|\mathbf{D}) = \prod_{i=1}^n h(x_i|\mathbf{z}_i)^{\Delta_i} S(x_i|\mathbf{z}_i). \quad (3.10)$$

Combining these facts, along with the representation of $dH_i(x)$ given by equation (3.4), we can now write

$$\prod_{i=1}^n \left[\frac{dH_0(x) e^{\beta^T Z_i}}{1 + (e^{\gamma^T Z_i} - 1) \prod_{s \leq x} (1 - dH_0(s))} \right]^{\Delta_i} \left[1 - e^{-\gamma^T Z_i} \left(1 + \frac{dH_0(x)}{dS_0(x)} \right) \right]^{-e^{\beta^T Z_i}}. \quad (3.11)$$

To reparameterize H_0 in the above equation, we can employ the baseline hazard function given by equation (3.7). For reparameterizing S_0 , we can notice that when $H_0(x)$ is discrete,

$$S_0(x) = \prod_{i:x_i \leq x} (1 - \Delta H_0(x_i)) \quad (3.12)$$

(Andersen *et al*, 1993), where $\Delta H_0(x) = w_j$ if $x \in (x_j, x_{j+1}^0]$. Using these facts, we can obtain that

$$S_0(x) = \begin{cases} 1, & x = 0, \\ (1 - w_1), & 0 < x \leq x_1^0, \\ (1 - w_1)(1 - w_2), & x_1^0 < x \leq x_2^0, \\ (1 - w_1)(1 - w_2) \dots (1 - w_k), & x_{k-1}^0 < x \leq x_{k_n}^0, \end{cases} \quad (3.13)$$

where $0 \leq w_i \leq 1$, and $w \in [0, 1]^{k_n}$. Hence, $\Delta S_0(x) = \left(\prod_{j' < j} (1 - w_{j'}) \right) w_j$ if $x \in (x_j, x_{j+1}^0]$,

and $S_0(x) = \sum_{j=1}^k \left(\prod_{j' < j} (1 - w_{j'}) \right) w_j I(x = x_j^0)$. Using these expressions, we can obtain:

$$\begin{aligned} & \prod_{i=1}^n \left(\sum_{j=1}^k \frac{e^{\beta^T Z_i} w_j I(x_i = x_j^0)}{1 + (e^{\gamma^T Z_i} - 1) \prod_{j' < j} (1 - w_{j'})} \right)^{\Delta_i} \\ & \times \left[1 - e^{-\gamma^T Z_i} \left(1 + \frac{\sum_{j=1}^k w_j I(x_i = x_j^0)}{\sum_{j=1}^k \left(\prod_{j' < j} (1 - w_{j'}) \right) w_j I(x_i = x_j^0)} \right) \right]^{-e^{\beta^T Z_i}}. \end{aligned} \quad (3.14)$$

Finally, we can write this likelihood into a more simplified form giving us:

$$\begin{aligned} & \left[\prod_{j=1}^{k_n} \prod_{i \in D(t_{(j)})} \left(\sum_{j=1}^k \frac{e^{\beta^T Z_i} w_j}{1 + (e^{\gamma^T Z_i} - 1) \prod_{j' < j} (1 - w_{j'})} \right) \right] \\ & \times \prod_{i=1}^n \left[1 - e^{-\gamma^T Z_i} \left(1 + \frac{\sum_{j=1}^k w_j I(x_i = x_j^0)}{\sum_{j=1}^k \left(\prod_{j' < j} (1 - w_{j'}) \right) w_j I(x_i = x_j^0)} \right) \right]^{-e^{\beta^T Z_i}}. \end{aligned} \quad (3.15)$$

For each of the two empirical likelihood methods, we will implement ARMS by obtaining samples from the posterior distribution by placing uniform(0, 1) priors on the w_j 's and normal(0, 1) priors on β, γ .

3.1.3 The Bayesian smoothing approach

To implement the Bayesian smoothing method, we will reparameterize equation (2.9) as:

$$\prod_{j=1}^n \left\{ h_0^*(x_j) \frac{e^{\beta^T Z_j}}{1 + (e^{\gamma^T Z_j} - 1) e^{-H_0^*(x_j)}} \right\}^{\Delta_j} \left\{ 1 - e^{-\gamma^T Z_j} + e^{-\gamma^T Z_j} e^{H_0^*(x_j)} \right\}^{-e^{\beta^T Z_j}}, \quad (3.16)$$

where

$$h_0^*(t) = \sum_{i=1}^{m'} \left[\eta \left(\frac{t}{u_i} \right)^{\eta-1} \prod_{j \leq i} (1 - \nu_j) \right] \quad (3.17)$$

and

$$H_0^*(t) = \sum_{i=1}^{m'} \left[\left(\frac{t}{u_i} \right)^{\eta} u_i \prod_{j \leq i} (1 - \nu_j) \right]. \quad (3.18)$$

Here, m' represents the number of uncensored observations, t refers to time, u_i refers to the uncensored observations, and ν_j represents samples drawn from independent $\text{beta}(d_j, r_j - d_j)$ random variables, where d_j equals the number of deaths at time t_j and r_j equals the number of units at risk just before time t_j . Next, we place $\text{uniform}(0, 1)$ priors on the w_j 's, $\text{normal}(0, 1)$ priors on β , γ , and let $\eta \sim \text{Ga}(.2, 5)$. To obtain samples from the posterior distribution, we will implement Adaptive Rejection Metropolis Sampling (ARMS).

3.2 Simulation studies

When conducting the simulation, we will need to generate T_i , C_i , and Z_i . First, we will set the baseline survival distribution, as described by equation (3.12), to a Weibull distribution; that is, $S_0(t) = e^{-t^\delta/\alpha}$ with unknown shape parameter, δ , and fixed scale parameter set at $\alpha = 1$. In this simulation, we will also set $\delta = 1$. We will then generate T_i by implementing the probability integral transform (inverse transform sampling). After implementing this procedure, it follows (after a little algebra) that we will simulate T_i using

$$T_i = \left\{ \log \left[(R_i^{-1/\theta_{2i}} - 1) \left(\frac{\theta_{2i}}{\theta_{1i}} \right) + 1 \right] \right\}^{1/\delta}, \quad (3.19)$$

where θ_{1i} and θ_{2i} are given by

$$\theta_{1i} = \exp \{ \gamma_0 + \gamma_1 z_{1i} + \dots + \gamma_p z_{pi} \} \quad (3.20)$$

and

$$\theta_{2i} = \exp \{ \beta_0 + \beta_1 z_{1i} + \dots + \beta_p z_{pi} \}, \quad (3.21)$$

respectively, and $R_i \sim \text{Unif}(0, 1)$. To start, we can take $p = 2$, and fix $\gamma_0 = 0.7$, $\gamma_1 = 0.4$, $\gamma_2 = 0.3$, $\beta_0 = 0$, $\beta_1 = 0$, and $\beta_2 = 0.1$. We will also simulate z_{1i} from a $\text{Bernoulli}(0.5)$ distribution and z_{2i} from a $N(0, 1)$ distribution.

The next step is to generate our censoring times C_i and achieve a reasonable amount of censoring (about 30%). To achieve 30% censoring, we generate $C_i \sim \text{Unif}(a, b)$, with $a = 0.3$ and $b = 1.8$. Finally, we generate $X_i = \min(T_i, C_i)$, with 100 Monte Carlo replicates.

After obtaining the necessary data, specifically (X_i, Δ_i, Z_i) , using the expressions given above in the previous section, we can proceed to obtain samples from the posterior for each of the methods described above (equations (3.8), (3.14), and (3.16)) by implementing ARMS. The simulation will be replicated 100 times for $n = 200$ under both 30% and 50% censoring, respectively. From this simulation, we will calculate coverage probabilities, as well as posterior estimates computed from the posterior mean of the β 's, Monte-Carlo standard error, and estimated standard error.

3.2.1 Comparing Various Methods

After running the simulation regarding the three methods shown above, we obtain the following results:

Table 3.1: Simulation results for the likelihood given by the empirical likelihood method (EL) (with Poisson form (eq. (3.8)) / Pseudo-likelihood (eq. (3.14))) and Bayesian smoothing (eq. (3.16)), under 30% and 50% censoring at $n = 200$ when both simulating and fitting data from the YP model

Method (Censoring %)	Parameter	Bias	Est. s.e.	MC s.e.	Est. 95% CrI	p-value
EL - Poisson form (30%)	β_0	-0.09	0.10	0.14	0.92	0.18
	β_1	0.09	0.10	0.13	0.97	0.18
	β_2	0.05	0.05	0.09	0.93	0.16
EL - Poisson form (50%)	β_0	-0.10	0.14	0.13	0.91	0.24
	β_1	0.07	0.10	0.16	0.93	0.24
	β_2	0.06	0.06	0.09	0.94	0.16
EL - Pseudo-likelihood (30%)	β_0	-0.09	0.08	0.11	0.93	0.13
	β_1	-0.01	0.12	0.18	0.93	0.47
	β_2	0.05	0.06	0.10	0.93	0.20
EL - Pseudo-likelihood (50%)	β_0	-0.10	0.12	0.13	0.92	0.20
	β_1	-0.02	0.13	0.18	0.96	0.44
	β_2	0.05	0.07	0.11	0.93	0.24
Bayesian smoothing (30%)	β_0	-0.01	0.03	0.01	0.95	0.37
	β_1	-0.03	0.04	0.04	0.93	0.23
	β_2	0.03	0.03	0.05	0.97	0.16
Bayesian smoothing (50%)	β_0	-0.03	0.03	0.03	0.91	0.16
	β_1	-0.03	0.05	0.04	0.97	0.27
	β_2	0.03	0.05	0.06	0.93	0.27

In this table, we have included are the bias of the Monte-Carlo average of the estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ along with the standard error, Monte-Carlo (MC) standard error (s.e. = standard error), the 95% coverage of credible interval probability, and p-values regarding the significance of the biases. The true values of $\beta_0, \beta_1, \beta_2$ are set at 0, 0, and 0.1, respectively and we have used the abbreviations: Est. = estimated, CrI = credible interval. From the results, we can notice that this method works very well. All of the biases are insignificant at $\alpha = 0.05$, and most of the estimated standard errors are quite close to the Monte-Carlo standard errors. The significance of the biases are comparable for both censoring cases. The estimated standard error is a little higher in the 50% censoring case. The Monte-Carlo standard error is a little higher for β in the 50% case as well.

3.2.2 Investigating Model Misspecification

Up to this point, we have simulated data from the Yang-Prentice model, and subsequently fit the same model (Table 3.1). Recall, that if $\theta_1 = \theta_2$ (which implies $\gamma = 0$), then our Yang-Prentice model reduces to the Cox Proportional Hazards (PH) Model. Next, we will investigate the following cases under each modeling approach (empirical likelihood

(with Poisson form / Pseudo-likelihood) and Bayesian smoothing):

1. We simulate data from the Yang-Prentice model, but fit the Cox PH model (setting $\gamma = 0$).
2. We simulate data from the Cox PH model (by setting $\theta_{1i} = \theta_{2i}$ in equation (3.19), i.e. ($\frac{\theta_{2i}}{\theta_{1i}} = 1$)), but fit the YP model.
3. We simulate data from the Cox PH model, and subsequently fit the same model.

Here, we will set the true values of $\beta_1 = \beta_2 = 0$, $\beta_3 = 0.1$, and follow the same general simulation procedure as before (ARMS sampling with 100 replications, $n = 200$). First, we will present the results for Case 1 (simulate YP, fit PH model):

Table 3.2: Simulation results for the likelihood given by the empirical likelihood method (EL) (with Poisson form (eq. (3.8)) / Pseudo-likelihood (eq. (3.14))) and Bayesian smoothing (eq. (3.16)), when simulating data from the YP model but fitting the PH model under 30% and 50% censoring at $n = 200$.

Method (Censoring %)	Parameter	Bias	Est. s.e.	MC s.e.	Est. 95% CrI	p-value
EL - Poisson form (30%)	β_0	0.12	0.06	0.09	0.36	0.02
	β_1	0.21	0.04	0.07	0.15	<0.01
	β_2	0.11	0.03	0.05	0.68	<0.01
EL - Poisson form (50%)	β_0	0.17	0.07	0.10	0.55	0.01
	β_1	0.20	0.05	0.08	0.09	<0.01
	β_2	0.10	0.06	0.08	0.45	0.04
EL - Pseudo-likelihood (30%)	β_0	0.10	0.07	0.09	0.74	0.07
	β_1	0.11	0.06	0.11	0.42	0.03
	β_2	0.09	0.05	0.09	0.65	0.03
EL - Pseudo-likelihood (50%)	β_0	0.15	0.10	0.12	0.62	0.06
	β_1	0.14	0.07	0.10	0.38	0.02
	β_2	0.07	0.05	0.08	0.58	0.08
Bayesian smoothing (30%)	β_0	-0.05	0.02	0.03	0.38	<0.01
	β_1	-0.05	0.02	0.01	0.42	<0.01
	β_2	0.04	0.02	0.03	0.70	0.02
Bayesian smoothing (50%)	β_0	-0.04	0.02	0.03	0.20	0.02
	β_1	-0.06	0.03	0.05	0.38	0.02
	β_2	-0.07	0.03	0.04	0.54	<0.01

From this table, we can notice that simulating data from the Yang-Prentice model, and then fitting the Cox model provides very biased results and poor coverage regardless of the method or censoring percentage. All of the given p-values are all less than 0.10 and many are < 0.01 . Next, we will explore Case 2, which is the opposite situation. Here,

we will simulate data from the Cox model, and then fit the Yang-Prentice model. The simulation results are given by Table 3.3.

Table 3.3: Simulation results for the likelihood given by the empirical likelihood method (EL) (with Poisson form (eq. (3.8)) / Pseudo-likelihood (eq. (3.14))) and Bayesian smoothing (eq. (3.16)), when simulating data from the PH model but fitting the YP model under 30% and 50% censoring at $n = 200$.

Method (Censoring %)	Parameter	Bias	Est. s.e.	MC s.e.	Est. 95% CrI	p-value
EL - Poisson form (30%)	β_0	-0.15	0.13	0.17	0.91	0.13
	β_1	-0.02	0.14	0.17	0.96	0.44
	β_2	0.01	0.07	0.09	0.94	0.44
EL - Poisson form (50%)	β_0	-0.15	0.14	0.19	0.92	0.14
	β_1	0.04	0.10	0.09	0.96	0.34
	β_2	0.04	0.06	0.09	0.95	0.25
EL - Pseudo-likelihood (30%)	β_0	-0.21	0.22	0.18	0.97	0.17
	β_1	-0.01	0.10	0.13	0.94	0.46
	β_2	0.03	0.07	0.10	0.93	0.34
EL - Pseudo-likelihood (50%)	β_0	-0.33	0.29	0.33	0.93	0.13
	β_1	-0.13	0.16	0.20	0.92	0.21
	β_2	0.01	0.08	0.11	0.94	0.44
Bayesian smoothing (30%)	β_0	-0.01	0.02	0.01	0.96	0.31
	β_1	0.01	0.02	0.01	0.97	0.31
	β_2	-0.01	0.02	0.02	0.93	0.31
Bayesian smoothing (50%)	β_0	0.01	0.01	0.01	0.91	0.16
	β_1	0.01	0.01	0.02	0.92	0.16
	β_2	0.02	0.02	0.04	0.97	0.16

From the results, this method appears to work very well. All of the biases in each of the three modeling techniques are insignificant at any reasonable level of α , and the estimated standard errors are quite close to the Monte-Carlo standard errors. In the higher censoring case, there is slightly more bias, and larger standard errors. The results when simulating data from the PH model, but fitting the YP model should be favorable since every PH model can be written as a YP model with $\theta_1 = \theta_2$. However, implementing the opposite situation (simulating data from the YP model, but fitting the PH model), should be biased since it is not true that every YP model is a PH model. Hence, these results seem plausible. Finally, we will repeat the above procedure by simulating data from a PH model, and subsequently fitting a PH model (Case 3). The results are shown below:

Table 3.4: Simulation results for the likelihood given by the empirical likelihood method (EL) (with Poisson form (eq. (3.8)) / Pseudo-likelihood (eq. (3.14))) and Bayesian smoothing, equation (eq. (3.16)), when both simulating and fitting data from the PH model under 30% and 50% censoring at $n = 200$.

Method (Censoring %)	Parameter	Bias	Est. s.e.	MC s.e.	Est. 95% CrI	p-value
EL - Poisson form (30%)	β_0	0.06	0.08	0.13	0.92	0.23
	β_1	0.05	0.07	0.11	0.91	0.24
	β_2	-0.02	0.05	0.09	0.97	0.34
EL - Poisson form (50%)	β_0	0.01	0.11	0.18	0.95	0.46
	β_1	0.13	0.16	0.17	0.91	0.21
	β_2	-0.05	0.06	0.10	0.91	0.20
EL - Pseudo-likelihood (30%)	β_0	-0.05	0.06	0.09	0.94	0.20
	β_1	0.02	0.12	0.13	0.94	0.43
	β_2	-0.05	0.06	0.10	0.92	0.20
EL - Pseudo-likelihood (50%)	β_0	-0.10	0.09	0.12	0.92	0.13
	β_1	-0.07	0.14	0.18	0.94	0.31
	β_2	-0.05	0.06	0.10	0.96	0.20
Bayesian smoothing (30%)	β_0	-0.01	0.02	0.02	0.92	0.31
	β_1	-0.01	0.02	0.01	0.94	0.31
	β_2	0.02	0.03	0.02	0.96	0.25
Bayesian smoothing (50%)	β_0	-0.01	0.01	0.01	0.94	0.16
	β_1	0.01	0.02	0.02	0.93	0.31
	β_2	0.02	0.02	0.02	0.93	0.16

From the results, this method works quite well. The biases are insignificant at any reasonable level of α , and the estimated standard errors are rather close to the Monte-Carlo standard errors. There is sometimes a little more bias and variability in the 50% censoring case as compared to the 30% censoring case; this is expected, although the bias is not significant. The coverage is rather good throughout. Overall, the simulation shows that when we generate data from the Cox model and fit the same model, we obtain favorable results.

The next issue that we will examine is the efficiency of the four methods. We will now determine whether there is a gain and/or loss of efficiency under two cases:

1. fitting the YP model / generating data from the PH model versus fitting the PH model / generating data from the PH model
2. fitting the YP model / generating data from the YP model versus fitting the PH model / generating data from the YP model,

by examining both the MSE (mean-squared error) and the asymptotic relative efficiency (ARE). In the following tables, we use the following notation: let $\text{MSE}(A|B)$ denote the MSE of $\hat{\beta}$ by fitting a model A to data generated from model B. The results of the first

case are shown below:

Table 3.5: Mean squared error (MSE) and asymptotic relative efficiency (ARE) defined as $\frac{\text{MSE}(\text{YP}|\text{PH})}{\text{MSE}(\text{PH}|\text{PH})}$, for the empirical likelihood (EL) method (with Poisson form (eq. (3.8)) / Pseudo-likelihood (eq. (3.14))) and Bayesian smoothing (eq. (3.16)), Par. = Parameter, PF = Poisson form, PL = Pseudo-likelihood, Bay. = Bayesian

Method	Par.	30% censoring			50% censoring		
		MSE(YP PH)	MSE(PH PH)	ARE	MSE(YP PH)	MSE(PH PH)	ARE
EL - PF	β_0	0.0394	0.0100	3.9400	0.0580	0.0112	5.1785
	β_1	0.0200	0.0074	2.7027	0.0116	0.0425	0.2729
	β_2	0.0050	0.0029	1.7241	0.0052	0.0061	0.8524
EL - PL	β_0	0.0970	0.0161	6.0248	0.0925	0.0181	5.1105
	β_1	0.0101	0.0148	0.6824	0.0173	0.0101	1.7128
	β_2	0.0050	0.0061	0.8197	0.0050	0.0061	0.8197
Bay. Smoothing	β_0	0.0005	0.0005	1.0000	0.0018	0.0002	9.0000
	β_1	0.0005	0.0005	1.0000	0.0018	0.0005	3.6000
	β_2	0.0029	0.0013	2.2307	0.0013	0.0008	1.6250

From the results, we can notice that in both cases for the empirical likelihood method, when data arises from the PH model, we lose efficiency regarding β_0 when we fit the YP model. This may have occurred since we generated our data from relatively small values of γ ; that is, $\gamma = (0.7, 0.4, 0.3)$. However, for β_1 , and β_2 , we can sometimes gain efficiency even when fitting the YP model after the data arises from the PH model. This is due to the fact that every PH model can be expressed as a YP model when $\theta_1 = \theta_2$. For the Bayesian smoothing method, we can notice that it is still more efficient to fit the PH model, rather than the YP model, when the data arises from the PH model. However, some of the larger asymptotic relative efficiency values (e.g., 9.0000) may have arisen due to very small MSE values (e.g., 0.0018 versus 0.0002). In summary, we can sometimes gain or lose efficiency in either of the two cases.

Next, we will examine the effect of the second case regarding the gain/loss of asymptotic relative efficiency; that is, we will simulate data from the YP model and then compare the asymptotic relative efficiency (ARE) when subsequently fitting the YP versus PH models. The results of this procedure are given below:

Table 3.6: Mean squared error (MSE) and asymptotic relative efficiency (ARE) defined as $\frac{\text{MSE}(\text{YP}|\text{YP})}{\text{MSE}(\text{PH}|\text{YP})}$, for the empirical likelihood (EL) method (with Poisson form (eq. (3.8)) / Pseudo-likelihood (eq. (3.14))) and Bayesian smoothing (eq. (3.16)), Par. = Parameter, PF = Poisson form, PL = Pseudo-likelihood, Bay. = Bayesian

Method	Par.	30% censoring			50% censoring		
		MSE(YP YP)	MSE(PH YP)	ARE	MSE(YP YP)	MSE(PH YP)	ARE
EL - PF	β_0	0.0164	0.0180	0.9111	0.0296	0.0338	0.8757
	β_1	0.0181	0.0457	0.3961	0.0149	0.0425	0.3506
	β_2	0.0050	0.0130	0.3846	0.0072	0.0136	0.5294
EL - PL	β_0	0.0145	0.0149	0.9732	0.0244	0.0325	0.7508
	β_1	0.0145	0.0157	0.9236	0.0173	0.0245	0.7061
	β_2	0.0061	0.0106	0.5562	0.0058	0.0074	0.7838
Bay. smoothing	β_0	0.0010	0.0029	0.3448	0.0018	0.0020	0.9000
	β_1	0.0025	0.0029	0.8621	0.0034	0.0045	0.7556
	β_2	0.0018	0.0020	0.9000	0.0034	0.0058	0.5862

From these results, we can observe that in most cases, there is a substantial loss in efficiency after generating data from the YP model and fitting the PH model instead. In fact, the asymptotic relative efficiency, defined as $\frac{\text{MSE}(\text{YP}|\text{YP})}{\text{MSE}(\text{PH}|\text{YP})}$, is ≤ 1 for each parameter in all cases. This is due to the fact that not every YP model can be expressed as a PH model. However, notice that some of the ARE's are still around 0.90 although others are close to 0.35. There probably would have been a greater loss of efficiency had we simulated the data using larger values of γ , such as $\gamma = (2, 1.5, 1.4)$.

Chapter 4

Model Selection

4.1 Approximating the Bayes factor corresponding to marginal densities

In this chapter, we will perform model selection by computing the Bayes factor corresponding to marginal densities. To approximate the Bayes factor, there are a few authors who have devised methods. The first method is that of Chib (1995), who simplified the computationally challenging problem of integrating the likelihood with respect to the parameters corresponding to the prior density. As an alternative, both the posterior, as well as the likelihood, is estimated at a particular point of high density. Next, let \mathbf{x} represent the observed data, ψ represent a vector of unknown parameters, $p(\mathbf{x}|\psi)$ denote the likelihood function, $\pi(\psi)$ denote the prior density, and notice that the marginal density $m(\mathbf{x})$ can be expressed as

$$m(\mathbf{x}) = \frac{p(\mathbf{x}|\psi)\pi(\psi)}{\pi(\psi|\mathbf{x})}, \quad (4.1)$$

It follows that the proposed estimate of the log of the marginal density is

$$\log \hat{m}(\mathbf{x}) = \log p(\mathbf{x}|\psi^*) + \log \pi(\psi^*) - \log \hat{\pi}(\psi^*|\mathbf{x}) \quad (4.2)$$

In this setup, ψ^* is chosen as a given high density point regarding the posterior distribution that serves to improve the accuracy of the approximation. Chib (1995) specifically focuses on a method of dividing the unknown parameter vector ψ into an arbitrary number of vector blocks. The purpose of this “blocking” is to facilitate computation by mandating only the

complete conditional densities. We will briefly study this method as applied to the simplest, base case of two vector blocks:

$$\pi(\psi|\mathbf{x}, \mathbf{u}); p^*(\mathbf{u}|\mathbf{x}, \psi), \quad (4.3)$$

where \mathbf{u} is an auxiliary variable. First, notice that the posterior density, $\pi(\psi|\mathbf{x})$ can be expressed as

$$\int \pi(\psi|\mathbf{x}, \mathbf{u})p^*(\mathbf{u}|\mathbf{x})d\mathbf{u}, \quad (4.4)$$

which means that a corresponding Monte Carlo estimate at ψ^* is

$$\hat{\pi}(\psi^*|\mathbf{x}) = \frac{1}{G} \sum_{g=1}^G \pi(\psi^*|\mathbf{x}, \mathbf{u}^{(g)}), \quad (4.5)$$

where $\mathbf{u}^{(g)}$ is drawn from the conditional distribution regarding $\mathbf{u}|\mathbf{x}$. After substituting the estimate from the above equation into equation (4.2), it follows that

$$\log \hat{m}(\mathbf{x}) = \log p(\mathbf{x}|\psi^*) + \log \pi(\psi^*) - \log \left\{ \frac{1}{G} \sum_{g=1}^G \pi(\psi^*|\mathbf{x}, \mathbf{u}^{(g)}) \right\}. \quad (4.6)$$

To obtain the necessary expressions for computing the Bayes factor, we can simply repeat the above calculation for all models and exponentiate. This approach is easily extended from two to an arbitrary number B of blocks. The posterior density at ψ^* can be written as:

$$\pi(\psi^*|\mathbf{x}) = \pi(\psi_1^*|\mathbf{x}) \times \pi(\psi_2^*|\mathbf{x}, \psi_1^*) \times \dots \times \pi(\psi_B^*|\mathbf{x}, \psi_1^*, \dots, \psi_{B-1}^*) \quad (4.7)$$

In the case of $B = 3$, equation (4.6) is now expressed as:

$$\log \hat{m}(\mathbf{x}) = \log p(\mathbf{x}|\psi^*) + \log \pi(\psi^*) - \log \hat{\pi}(\psi_1^*|\mathbf{x}) - \log \hat{\pi}(\psi_2^*|\mathbf{x}, \psi_1^*). \quad (4.8)$$

For the general case, this expression can be extended to

$$\log \hat{m}(\mathbf{x}) = \log p(\mathbf{x}|\psi^*) + \log \pi(\psi^*) - \sum_{r=1}^B \log \hat{\pi}(\psi_r^*|\mathbf{x}, \psi_s^* (s < r)). \quad (4.9)$$

To estimate the required quantities, we need to compute $(B - 1)$ iterations from G Gibbs samples. For example, when $B = 3$, $\hat{\pi}(\psi_2^*|\mathbf{x}, \psi_1^*)$ can be estimated with the same basic procedure as outlined in equation (4.5), where the draws are obtained from a Gibbs algorithm that

only implements the full conditional distributions $\pi(\theta_2|\mathbf{x}, \psi_1^*, \theta_3, \mathbf{u})$ and $\pi(\theta_3|\mathbf{x}, \psi_1^*, \theta_2, \mathbf{u})$. This efficient partitioning helps to reduce computing time while preserving accuracy.

Ritter and Tanner (1992) developed an alternative method for estimating the posterior distribution that also relies on Gibbs sampling. Let

$$K_G(\psi, \psi^*|\mathbf{x}) = \prod_{k=1}^B \pi(\psi_k^*|\mathbf{x}, \psi_1^*, \dots, \psi_{k-1}^*, \psi_{k+1}, \dots, \psi_B) \quad (4.10)$$

represent the Gibbs transition kernel. Using the fact that the invariance condition is satisfied; that is, $\pi(\psi^*|\mathbf{x}) = \int K_G(\psi, \psi^*|\mathbf{x})\pi(\psi|\mathbf{x})d\psi$, we can estimate the posterior ordinate. This is accomplished by averaging equation (4.10) over draws from the full Gibbs run, which gives us

$$\hat{\pi}(\psi^*|\mathbf{x}) = \frac{1}{M} \sum_{g=1}^M K_G(\psi^{(g)}, \psi^*|\mathbf{x}) \quad (4.11)$$

Unfortunately, this technique produces less accurate estimation than Chib's (1995) estimate when ψ is of high dimension. Furthermore, neither of the methods can be applied when sampling is from a single-block Metropolis sampler or at least one of the normalizing constants from the full conditional distributions is unknown. Chib and Jeliazkov (2001) extend the approach of Chib (1995) to remedy these problems. Suppose that we can update ψ , our parameter vector, in only one block, and let

$$\alpha(\psi, \psi'|\mathbf{x}) = \min \left\{ 1, \frac{p(\psi'|\mathbf{x})q(\psi', \psi|\mathbf{x})}{p(\psi|\mathbf{x})q(\psi', \psi|\mathbf{x})} \right\} \quad (4.12)$$

This quantity represents the chance of accepting the Metropolis-Hastings proposed value ψ' which is simulated from a candidate generating density $q(\psi, \psi'|\mathbf{x})$. Next, Chib and Jeliazkov (2001) set $p^*(\psi, \psi'|\mathbf{x}) = \alpha(\psi, \psi'|\mathbf{x})q(\psi, \psi'|\mathbf{x})$. It follows that for an arbitrary point ψ^*

$$p^*(\psi, \psi^*|\mathbf{x})\pi(\psi|\mathbf{x}) = \pi(\psi^*|\mathbf{x})p^*(\psi^*, \psi|\mathbf{x}) \quad (4.13)$$

After integrating each side of the expression, Chib and Jeliazkov (2001) arrive at the expression

$$\pi(\psi^*|\mathbf{x}) = \frac{E_1 \{\alpha(\psi, \psi^*|\mathbf{x})q(\psi, \psi^*|\mathbf{x})\}}{E_2 \{\alpha(\psi, \psi^*|\mathbf{x})\}}, \quad (4.14)$$

where E_1 and E_2 represent the expectations with respect to the posterior and candidate distributions, $\pi(\psi|\mathbf{x})$ and $q(\psi^*, \psi|\mathbf{x})$, respectively. The first expectation, E_1 , is calculated by sampling draws from the posterior distribution and averaging the numerator of equation

(4.14), while the second expectation, E_2 , is calculated by drawing from the candidate density $q(\psi^*, \psi|\mathbf{x})$. From this procedure, we can obtain the estimate of the log marginal likelihood; that is,

$$\log \hat{m}(\mathbf{x}) = \log p(\mathbf{x}|\psi^*) + \log \pi(\psi^*) - \log \hat{\pi}(\psi^*|\mathbf{x}) \quad (4.15)$$

This idea is easily applied to the case of more than one block by extending this procedure via multiple MCMC runs.

However, both of the methods of Chib (1995) and Chib and Jeliazkov (2001) have a main limitation. If we need to compare a Dirichlet Process Mixture (DPM) model with an embedded parametric alternative, the previous methods will suffice in calculating the marginal likelihood of the parametric model. But if the alternative DPM model is not entirely parametric, we need to examine the method of Basu and Chib (2003). Their method is the first that can be implemented when given a semiparametric model with covariates and hierarchical prior structures.

Next, we will provide the reader with some background on Dirichlet process mixture models. To start, let us define ζ as a continuous random variable, J_0 as a non-atomic probability distribution on ζ , and let φ be a scalar quantity. Also, let J be a random probability distribution on ζ , which is distributed by the Dirichlet process (DP). Hence, we can write the following for all k and k -partitions in the ζ -space:

$$(J(\zeta \in B_1), J(\zeta \in B_2), \dots, J(\zeta \in B_k)) \sim \text{Dir}(\varphi J_0(B_1), \varphi J_0(B_2), \dots, \varphi J_0(B_k)). \quad (4.16)$$

If we integrate out J , we will obtain a joint distribution based on a clustered set of variables $\zeta_{1:n}$; that is, ζ_1, \dots, ζ_n . If we condition on $n - 1$ draws, we will be able to determine that the distribution of the n th value will be equivalent to the following:

$$p(\zeta|\zeta_{1:n-1}) \propto \varphi p(\zeta|J_0) + \sum_{i=1}^{n-1} \delta(\zeta, \zeta_i). \quad (4.17)$$

Suppose that there are k unique values of $\zeta_{1:n-1}$ represented by $\zeta_{1:k}^*$. Hence, a subsequent draw from the DP can be expressed as:

$$\zeta_n = \begin{cases} \zeta_i^* & \text{with prob } \frac{n_i}{n-1+\varphi} \\ \zeta, \zeta \sim J_0 & \text{with prob } \frac{\varphi}{n-1+\varphi} \end{cases} \quad (4.18)$$

Notice that n_i represents the number of occurrences of ζ_i^* in $\zeta_{1:n-1}$.

Here, we can represent the DP as a nonparametric prior within the context of a hierarchical Bayes model (Antoniak, 1974). Let us simulate the data according to the following scheme

$$\begin{aligned}
J &\sim DP(\varphi, J_0(\cdot|\kappa)) \\
\zeta_n &\sim J \\
X_n &\sim p(\cdot|\zeta_n, \iota) \\
\varpi = (\iota, \kappa, \varphi) &\sim \pi
\end{aligned} \tag{4.19}$$

which is known as the Dirichlet process mixture model. Notice that ι is a vector parameter associated with \mathbf{x} , $J_0(\cdot|\kappa)$ is a specified base probability measure which is dependent on κ , a vector of unknown parameters, and φ adheres to a parametric distribution π .

Basu and Chib (2003) obtain the marginal likelihood according the following calculation:

$$\begin{aligned}
m(\mathbf{x}) &= \int L(\mathbf{x}|\iota, \kappa, \varphi, J_0) \pi(\iota, \kappa, \varphi) d\iota d\kappa d\varphi \\
&= \int \int p(\mathbf{x}|\iota, J) d\mathcal{P}(J|\varphi, J_0, \kappa) \pi(\iota, \kappa, \varphi) d\iota d\kappa d\varphi \\
&= \int \int \left\{ \prod_{i=1}^n \int p(\mathbf{x}_i|\zeta_i, \iota) dJ(\zeta_i) \right\} \\
&\quad \times d\mathcal{P}(J|\varphi, J_0, \kappa) \pi(\iota, \kappa, \varphi) d\iota d\kappa d\varphi,
\end{aligned} \tag{4.20}$$

where $\mathcal{P}(\cdot|\varphi, J_0, \kappa)$ represents the DP measure. Notice that the marginal likelihood acts as a constant to normalize the posterior distribution. Hence, the above equation can be expressed as

$$m(\mathbf{x}) = \frac{L(\mathbf{x}|\iota^*, \kappa^*, \varphi^*, J_0) \pi(\iota^*, \kappa^*, \varphi^*)}{\pi(\iota^*, \kappa^*, \varphi^*|\mathbf{x})}, \tag{4.21}$$

where a single point within in parameter space is represented by $(\iota^*, \kappa^*, \varphi^*)$. The prior and posterior densities at this particular point are $\pi(\iota^*, \kappa^*, \varphi^*)$ and $\pi(\iota^*, \kappa^*, \varphi^*|\mathbf{x})$, respectively. Next, Basu and Chib (2003) outline a method for calculating the estimates of the likelihood $\hat{L}(\mathbf{x}|\iota^*, \kappa^*, \varphi^*, J_0)$ and posterior ordinate $\hat{\pi}(\iota^*, \kappa^*, \varphi^*|\mathbf{x}, J_0)$.

To estimate the posterior ordinate, Basu and Chib (2003) implement Markov Chain sampling and use the previous framework of Chib (1995). The posterior ordinate is decom-

posed as the following:

$$\begin{aligned} \log \pi(\iota^*, \kappa^*, \varphi^* | \mathbf{x}) &= \log \pi(\iota^* | \mathbf{x}) \\ &+ \log \pi(\varphi^* | \mathbf{x}, \iota^*) + \log \pi(\kappa^* | \mathbf{x}, \iota^*, \varphi^*) \end{aligned} \quad (4.22)$$

The first term on the right hand side of the equation can be estimated by MCMC sampling iterated for $j = 1, \dots, J_1$ cycles; that is,

$$\hat{\pi}(\iota^* | \mathbf{x}) = \frac{1}{J_1} \sum_{j=1}^{J_1} \pi(\iota^* | \zeta^{(j)}, \kappa^{(j)}, \varphi^{(j)}, \mathbf{x}), \quad (4.23)$$

where the superscript (j) represents numbers drawn at iteration j . The other ordinate is estimated after additional iterations, denoted as J_2 ; that is,

$$\hat{\pi}(\varphi^* | \mathbf{x}, \iota^*) = \frac{1}{J_2} \sum_{j=J_1+1}^{J_1+J_2} \pi(\varphi^* | \zeta^{(j)}, \iota^*, \kappa^{(j)}, \mathbf{x}). \quad (4.24)$$

To estimate $\hat{\pi}(\kappa^* | \mathbf{x}, \iota^*, \varphi^*)$, Basu and Chib (2003) execute the chain another J_3 iterations after setting ι and φ at ι^* and φ^* , respectively.

Next, we notice that the likelihood ordinate is written as

$$L(\mathbf{x} | \iota^*, \kappa^*, \varphi^*) = \int \left\{ \prod_{i=1}^n \int p(\mathbf{x}_i | \zeta_i, \iota^*) dJ(\zeta_i) \right\} d\mathcal{P}(J | \varphi^*, J_0, \kappa^*). \quad (4.25)$$

Unfortunately, this expression is nonanalytic and has no closed form. To calculate the above expression, Basu and Chib (2003) show that $L(\mathbf{x} | \iota^*, \kappa^*, \varphi^*)$ can be computed via sequential importance sampling (SIS). Here, ζ is simulated sequentially according to the following density

$$\pi^*(\zeta_1, \dots, \zeta_n | \mathbf{x}, \varpi^*) = \prod_{i=1}^n \pi(\zeta_i | \mathbf{x}_{(i)}, \zeta_{(i-1)}, \varpi^*) \quad (4.26)$$

The importance weight, as computed by Kong *et al.* (1994), equals

$$\frac{\pi(\zeta_1, \dots, \zeta_n | \mathbf{x}, \varpi^*)}{\pi^*(\zeta_1, \dots, \zeta_n | \mathbf{x}, \varpi^*)} = \frac{w}{L(\mathbf{x} | \varpi^*)} \quad (4.27)$$

with the weights w expressed as

$$w(\zeta_1, \dots, \zeta_n) = p(\mathbf{x}_1 | \varpi^*) \prod_{i=2}^n p(\mathbf{x}_i | \mathbf{x}_{(i-1)}, \zeta_{(i-1)}, \varpi^*) \quad (4.28)$$

Because the likelihood term in equation (4.27) does not depend on the vector ζ , the quantity given by equation (4.27) can create an estimate of the likelihood $L(\mathbf{x}|\varpi^*)$. If we replicate this procedure M times, the mean of the weights $\bar{w} = \frac{1}{M} \sum_{j=1}^M w^{(j)}$ is a consistent estimate of the desired likelihood ordinate when conducting Monte Carlo simulation.

A potential drawback to sequential importance sampling (SIS) is that the weights tend to be quite volatile. As a result, Basu and Chib (2003) implement an alternative method for computing the weights called collapsed SIS. This updated approach seeks to eliminate ζ_i by integration. The modified sequential importance sampling has less variability due to a reduction in space where the sequential importance sampling method works. This idea was also developed in the context of general weighted Chinese restaurant processes to estimate the marginal density by Ishwaran *et al.* (2001). In summary, ζ_i is no longer sampled but instead, a sampled cluster is marginalized over ζ_i (see Basu and Chib (2003) for details).

4.2 Variational Methods

We will now implement variational methods for easier computation of the marginal distribution which will enable us to perform model selection. Mean-field variational inference approximates likelihoods/posteriors for a seemingly intractable probability distribution. Throughout the statistical literature, many of these variational methods have been implemented for parametric models, such as those by Jordan *et al* (1999), Ueda and Ghahramani (2002), and Wainwright and Jordan (2003). The lower bound of a model with observed variables \mathbf{x} and unobserved, hidden variables \mathbf{U} can be obtained by Jensen's inequality; that is,

$$\begin{aligned} \log p(\mathbf{x}) &= \log \int_{\mathbf{u}} p(\mathbf{x}, \mathbf{u}) d\mathbf{u} \\ &= \log \int_{\mathbf{u}} \frac{q(\mathbf{u})p(\mathbf{x}, \mathbf{u})}{q(\mathbf{u})} d\mathbf{u} \\ &\geq \int_{\mathbf{u}} q(\mathbf{u}) \log p(\mathbf{x}, \mathbf{u}) - \int_{\mathbf{u}} q(\mathbf{u}) \log q(\mathbf{u}) \\ &= \text{E}[\log p(\mathbf{x}, \mathbf{U})] - \text{E}[\log q(\mathbf{U})], \end{aligned} \tag{4.29}$$

for an arbitrary density $q(\mathbf{u})$. Note that $p(\cdot)$ represents a probability density.

The main concept regarding variational methods is to restrict the function $q(\mathbf{u})$ to a parametric family in a way that optimizing the lower bound in the above equation is computationally feasible. This optimal function q is the distribution which is closest in Kullback-Liebler distance to the true posterior in the particular, chosen parametric family. In the upcoming discussion, we will focus more on the variational approach as applied to the DP, which was developed by Blei and Jordan (2004). This algorithm is based on the Dirichlet process mixture model and is applicable to data even when the Gibbs sampling algorithm converges slowly.

We can compute the marginal distribution along with the corresponding Bayes Factor by iteratively minimizing all variational parameters via a normal approximation to the likelihood; that is,

$$BF = \frac{p(x|M_1)}{p(x|M_2)}, \quad (4.30)$$

where $p(x|M_i)$ denotes the marginal density for Model i . Notice that we are averaging over the parameters rather than maximizing the likelihood as in the frequentist approach. If we parameterize by vectors of parameters denoted as ψ_1 and ψ_2 the Bayes factor can be written as:

$$BF = \frac{\int p(\psi_1|M_1)p(x|\psi_1, M_1)d\psi_1}{\int p(\psi_2|M_2)p(x|\psi_2, M_2)d\psi_2} \quad (4.31)$$

If $BF > 1$, there is stronger evidence contained in the data to support M_1 versus M_2 . Jeffreys (1961) has given an empirical scale regarding the interpretation of the Bayes factor.

4.2.1 Variational methods for Dirichlet process mixture models

The procedure for constructing a Dirichlet process mixture model as described above in Section 4.1 can be thought of as an “infinite” mixture model; that is, the current data only contains a finite number of components, yet additional data can uncover previously unobserved, hidden components (Neal, 2000). In the following construction of J (Sethuraman, 1994), let us consider the nonfinite collection of independent random variables denoted as $V_i \sim \text{Beta}(1, \varphi)$ and $\zeta_i^* \sim J_0$ for $i = 1, 2, 3, \dots$. It follows that J can now

be expressed as:

$$\begin{aligned}\theta_i &= V_i \prod_{j=1}^{i-1} (1 - V_j), \\ J(\zeta) &= \sum_{i=1}^{\infty} \theta_i \delta(\zeta, \zeta_i^*).\end{aligned}\tag{4.32}$$

In the above equation, θ represents an infinite vector of various mixing proportions while $\zeta_{1:\infty}^*$ denotes variables regarding an infinite collection of mixture components.

We can deduce that the lower bound on equation (4.29) is:

$$\begin{aligned}\log p(\mathbf{x}) &= \log \int p(\mathbf{x}, \mathbf{V}, \zeta^*) dv d\zeta^* \\ &= \log \int p(\mathbf{x}|\mathbf{V}, \zeta^*) p(\mathbf{V}, \zeta^*) dv d\zeta^* \\ &= \log \int \left\{ \frac{p(\mathbf{x}|\mathbf{V}, \zeta^*) p(\mathbf{V}, \zeta^*)}{q(\mathbf{V}, \zeta^*)} \right\} q(\mathbf{V}, \zeta^*) dv d\zeta^* \\ &\geq \int \{\log p(\mathbf{x}|\mathbf{V}, \zeta^*)\} q(\mathbf{V}, \zeta^*) dv d\zeta^* \\ &\quad + \int \{\log p(\mathbf{V}, \zeta^*)\} q(\mathbf{V}, \zeta^*) dv d\zeta^* \\ &\quad - \int \{\log q(\mathbf{V}, \zeta^*)\} q(\mathbf{V}, \zeta^*) dv d\zeta^*,\end{aligned}\tag{4.33}$$

where \mathbf{V} and η^* are independent; that is,

$$p(\mathbf{V}, \zeta^*) = p(\mathbf{V})p(\zeta^*).\tag{4.34}$$

We will impose the following priors:

$$\begin{aligned}V_j &\stackrel{iid}{\sim} Be(1, \varphi), \quad j = 1, \dots, K-1, V_K \equiv 1 \\ \zeta_j &\stackrel{ind}{\sim} N(\mu_0, \sigma_0^2), \quad j = 1, \dots, K \\ \beta_k &\stackrel{ind}{\sim} N(\xi_k, \tau_k^2), \quad k = 1, \dots, m \\ \gamma_k &\stackrel{ind}{\sim} N(\vartheta_k, \varrho_k^2), \quad k = 1, \dots, m \\ \frac{1}{h^2} &\stackrel{ind}{\sim} Ga(r_0, s_0)\end{aligned}\tag{4.35}$$

The lower bound can now be expressed as the following decomposition:

$$\log p(\mathbf{x}|\varphi) \geq \mathbb{E}_q[\log p(\mathbf{V}|\varphi)] \quad (4.36)$$

$$+ \sum_{i=1}^N \mathbb{E}_q[\log p(x_i|\zeta^*)] \quad (4.37)$$

$$+ \mathbb{E}_q[\log p(\zeta^*)] \quad (4.38)$$

$$- \mathbb{E}_q[\log q(\mathbf{V}, \zeta^*)] \quad (4.39)$$

To calculate the above expression we will first impose variational parameters with the following distributions:

$$\begin{aligned} V_j &\overset{ind}{\sim} Be(a_j, b_j), j = 1, \dots, K-1, V_K \equiv 1 \\ \zeta_j &\overset{ind}{\sim} N(\mu_j^*, \sigma_j^{*2}), j = 1, \dots, K \\ \beta_k &\overset{ind}{\sim} N(\xi_k^*, \tau_k^{*2}), k = 1, \dots, m \\ \gamma_k &\overset{ind}{\sim} N(\vartheta_k^*, \varrho_k^{*2}), k = 1, \dots, m \\ \frac{1}{h^2} &\sim Ga(r_0^*, s_0^*) \end{aligned} \quad (4.40)$$

To make this probability distribution computationally feasible, we will cut off the variational distribution at a set value K by letting $q(V_k = 1) = 1$. As a result, we can disregard ζ_k^* for $k > K$, since θ_k (the mixture proportions) equals to 0. In our case, we will take $K = 4$ and $m = 3$.

Now, we will define $w = (w_1, w_2, \dots, w_k)$ according the following scheme:

$$\begin{aligned} w_1 &= v_1 \\ w_2 &= (1 - v_1)v_2 \\ &\dots \\ w_{k-1} &= (1 - v_1) \dots (1 - v_{k-2})v_{k-1} \\ w_k &= (1 - v_1) \dots (1 - v_{k-1}), \end{aligned} \quad (4.41)$$

where $v_1, v_2, \dots, v_k \stackrel{iid}{\sim} Be(1, \varphi)$. We represent $p(x_i|\zeta^*, \beta, \gamma)$ as the following:

$$\prod_{i=1}^N \left[\frac{e^{\beta Z_i} \sum_{j=1}^K \frac{1}{h} \phi \left(\frac{x - \zeta_j^*}{h} \right) w_j}{1 + (e^{\gamma Z_i} - 1) \sum_{j=1}^K \Phi \left(\frac{x - \zeta_j^*}{h} \right) w_j} \right]^{\Delta_i} \left[1 + e^{-\gamma Z_i} \frac{\left(1 - \sum_{j=1}^K \Phi \left(\frac{x - \zeta_j^*}{h} \right) w_j \right)}{\sum_{j=1}^K \Phi \left(\frac{x - \zeta_j^*}{h} \right) w_j} \right]^{-e^{\beta Z_i}}, \quad (4.42)$$

where h represents the bandwidth, ϕ and Φ represent the pdf and cdf of a normal distribution, respectively. To compute (4.33), we will need to iteratively minimize all variational parameters and first compute the estimated information of the MLE via the normal approximation to the likelihood; that is,

$$\hat{I} = \hat{I}(\hat{V}_1, \dots, \hat{V}_K, \hat{\zeta}_1, \dots, \hat{\zeta}_K, \hat{\beta}_1, \dots, \hat{\beta}_m, \hat{\gamma}_1, \dots, \hat{\gamma}_m, \frac{1}{\hat{h}^2}). \quad (4.43)$$

The next step is to minimize the following function of thirty variables based on the decomposition expressed by (4.36), (4.37), (4.38), and (4.39) by minimizing one parameter while fixing the others at an initial guess of 0.5. (4.36) can be represented by:

$$(\varphi - 1) \sum_{j=1}^K E_{a_j, b_j} [\log V], \quad (4.44)$$

(4.37) can be written as:

$$\begin{aligned} & \frac{1}{2} \left(\frac{a_1}{a_1+b_1} - \hat{V}_1, \dots, \frac{a_k}{a_k+b_k} - \hat{V}_k, \mu_1^* - \hat{\zeta}_1, \dots, \mu_k^* - \hat{\zeta}_k, \xi_1^* - \hat{\beta}_1, \dots, \xi_m^* - \hat{\beta}_m, \vartheta_1^* - \hat{\gamma}_1, \dots, \vartheta_m^* - \hat{\gamma}_m \right)^T \\ & \times \hat{I} \times \left(\frac{a_1}{a_1+b_1} - \hat{V}_1, \dots, \frac{a_k}{a_k+b_k} - \hat{V}_k, \mu_1^* - \hat{\zeta}_1, \dots, \mu_k^* - \hat{\zeta}_k, \xi_1^* - \hat{\beta}_1, \dots, \xi_m^* - \hat{\beta}_m, \vartheta_1^* - \hat{\gamma}_1, \dots, \vartheta_m^* - \hat{\gamma}_m \right) \times \frac{r_0^*}{s_0^*} \\ & + \frac{1}{2} \text{tr} \left(\hat{I} \times \text{diag} \left(\frac{a_1 b_1}{(a_1+b_1)^2 (a_1+b_1+1)}, \dots, \frac{a_k b_k}{(a_k+b_k)^2 (a_k+b_k+1)}, \sigma_1^{*2}, \dots, \sigma_k^{*2} \right) \right) \times \frac{r_0^*}{s_0^*} \\ & - \frac{N}{2} E_q(\log H) - \log |I| + \frac{p_d}{2} \log(2\pi), \end{aligned} \quad (4.45)$$

where $p_d = 2(m+K)$, and $2(m+K) \times 2(m+K)$ represents the dimension of the information matrix. The terms in (4.38) are:

$$\begin{aligned} & \frac{1}{2\sigma_0^2} \sum_{j=1}^K (\mu_j^* - \mu_0)^2 + \frac{1}{2\sigma_0^2} \sum_{j=1}^K \sigma_j^{*2} \\ & \sum_{k=1}^m \left\{ \frac{(\xi_k^* - \xi_k)^2}{2\tau_k^2} + \frac{\tau_k^{*2}}{2\tau_k^2} \right\} + \sum_{k=1}^m \left\{ \frac{(\varphi_k^* - \varphi_k)^2}{2\varrho_k^2} + \frac{\varrho_k^{*2}}{2\varrho_k^2} \right\}, \end{aligned} \quad (4.46)$$

and (4.39) can be expressed as:

$$\begin{aligned}
& -\sum_{j=1}^K \left\{ (a_j - 1)E_{a_j, b_j}[\log V] + (b_j - 1)E_{a_j, b_j}[\log(1 - V)] - \log B(a_j, b_j) \right\} \\
& \qquad \qquad \qquad -\sum_{j=1}^K \log \sigma_j^{*2} - \sum_{k=1}^m \log \tau_k^{*2} - \sum_{k=1}^m \log \varrho_k^{*2}, \quad (4.47)
\end{aligned}$$

where $E_q(\log H) = \frac{N}{2}(\Psi(r_0^*) - \log s_0^*)$, $H \sim Ga(r_0^*, s_0^*)$, $E[\log V] = \Psi(a) - \Psi(a + b)$, $E[\log(1 - V)] = \Psi(b) - \Psi(a + b)$, B represents the beta function, and Ψ represents the digamma function. After implementing the above functions to find the minimum of each of the thirty parameters, we will repeat the procedure until a tolerance of 0.001 is reached. The result is that we will find the variational distribution Q which is closest to the true posterior distribution P , within the restrictions of the function's parameters. This method can be categorized as a full Bayesian approach.

4.2.2 A Variational Method Based on Bayesian Bootstrap

Next, we will explore another variational method based on Bayesian bootstrap. The variational method based on Bayesian bootstrap simplifies the amount of expressions needed. The lower bound now becomes:

$$\begin{aligned}
\log p(\mathbf{x}) &= \log \int p(\mathbf{x}, \mathbf{V}, \beta, \gamma) dv d\beta d\gamma \quad (4.48) \\
&= \log \int p(\mathbf{x}|\mathbf{V}, \beta, \gamma)p(\mathbf{V}, \beta, \gamma) dv d\beta d\gamma \\
&= \log \int \left\{ \frac{p(\mathbf{x}|\mathbf{V}, \beta, \gamma)p(\mathbf{V}, \beta, \gamma)}{q^*(\mathbf{V}, \beta, \gamma)} \right\} q^*(\mathbf{V}, \beta, \gamma) dv d\beta d\gamma \\
&\geq \int \{\log p(\mathbf{x}|\mathbf{V}, \beta, \gamma)\} q^*(\mathbf{V}, \beta, \gamma) dv d\beta d\gamma \\
&+ \int \{\log p(\mathbf{V}, \beta, \gamma)\} q^*(\mathbf{V}, \beta, \gamma) dv d\beta d\gamma \\
&- \int \{\log q^*(\mathbf{V}, \beta, \gamma)\} q^*(\mathbf{V}, \beta, \gamma) dv d\beta d\gamma.
\end{aligned}$$

The lower bound further simplifies to:

$$\int \{\log p(\mathbf{x}|\mathbf{V}, \beta, \gamma)\} q^*(\mathbf{V}, \beta, \gamma) dv d\beta d\gamma \quad (4.49)$$

$$+ \sum_{i=1}^N \int \{\log p(V_i)\} q^*(V_i) dv_i \quad (4.50)$$

$$\begin{aligned} &+ \int \{\log p(\beta)\} q^*(\beta) d\beta \\ &+ \int \{\log p(\gamma)\} q^*(\gamma) d\gamma \\ &- \sum_{i=1}^N \int \{\log q^*(V_i)\} q^*(V_i) dv_i \quad (4.51) \\ &- \int \{\log q^*(\beta)\} q^*(\beta) d\beta \\ &- \int \{\log q^*(\gamma)\} q^*(\gamma) d\gamma \end{aligned}$$

To define $p(\mathbf{x}|\mathbf{V}, \beta, \gamma)$, let us recall the equation based on Poisson form given earlier in Chapter 3; that is,

$$\begin{aligned} &\prod_{j=1}^{k_n} \left[\prod_{i \in D(t_{(j)})} \left\{ 1 - \left(1 - \sum_{j=1}^k \frac{w_j}{1 + (e^{\gamma^T Z_i} - 1) \prod_{j' \leq j} (1 - w_{j'})} \right)^{e^{\beta^T Z_i}} \right\} \right] \\ &\times \exp \left(- \sum_{i \in R(t_{(j)})} e^{\beta^T Z_i} \sum_{j=1}^k \frac{w_j}{1 + (e^{\gamma^T Z_i} - 1) \prod_{j' \leq j} (1 - w_{j'})} \right), \quad (4.52) \end{aligned}$$

where $R(t_{(j)}) = \{i : Y_i(t_{(j)}) = 1\}$ and $D(t_{(j)}) = \{i : \Delta N_i(t_{(j)}) = 1\}$. Note that the set $D(t_{(j)})$ represents the set of observations which fail at time $t_{(j)}$ and $R(t_{(j)})$ represents the set of observations still at risk at time $t_{(j)}$. Next, by the normal approximation, we can write:

$$\begin{aligned} -2 \log p(\mathbf{x}|\mathbf{V}, \beta, \gamma) &\approx \left[\left((V, \beta, \gamma) - (\hat{V}, \hat{\beta}, \hat{\gamma}) \right)^T \hat{I} \left((V, \beta, \gamma) - (\hat{V}, \hat{\beta}, \hat{\gamma}) \right) \right] \\ &- \log |\hat{I}| + \frac{p_d}{2} \log(2\pi), \quad (4.53) \end{aligned}$$

where $p_d = (k_n - 1) + 2m$ and $(k_n - 1) + 2m \times (k_n - 1) + 2m$ represents the dimension of the information matrix. After taking the expectation of this expression, we realize that

(4.53) reduces to the same calculation given in (4.36)-(4.39) excluding the variable ζ and the bandwidth h ; that is, we can simplify the lower bound given above. First, our priors, which are data dependent, can be specified as:

$$\begin{aligned} V_j &\stackrel{iid}{\sim} Be(a_j, b_j), \quad j = 1, \dots, k_n - 1, V_{k_n} \equiv 1 \\ \beta_k &\stackrel{iid}{\sim} N(\xi_l, \tau_l^2), \quad l = 1, \dots, m \\ \gamma_l &\stackrel{iid}{\sim} N(\vartheta_l, \varrho_l^2), \quad l = 1, \dots, m \end{aligned} \quad (4.54)$$

(4.49) can be written as:

$$\begin{aligned} &\frac{1}{2} \left(\frac{a_1}{a_1+b_1} - \hat{V}_1, \dots, \frac{a_k}{a_k+b_k} - \hat{V}_k, \xi_1^* - \hat{\beta}_1, \dots, \xi_m^* - \hat{\beta}_m, \vartheta_1^* - \hat{\gamma}_1, \dots, \vartheta_m^* - \hat{\gamma}_m \right)^T \\ &\quad \times \hat{I} \times \left(\frac{a_1}{a_1+b_1} - \hat{V}_1, \dots, \frac{a_k}{a_k+b_k} - \hat{V}_k, \xi_1^* - \hat{\beta}_1, \dots, \xi_m^* - \hat{\beta}_m, \vartheta_1^* - \hat{\gamma}_1, \dots, \vartheta_m^* - \hat{\gamma}_m \right) \\ &\quad + \frac{1}{2} \text{tr} \left(\hat{I} \times \text{diag} \left(\frac{a_1 b_1}{(a_1+b_1)^2 (a_1+b_1+1)}, \dots, \frac{a_k b_k}{(a_k+b_k)^2 (a_k+b_k+1)} \right) \right) \\ &\quad - \log |\hat{I}| + \frac{pd}{2} \log(2\pi) \end{aligned} \quad (4.55)$$

(4.50) can be denoted by:

$$\begin{aligned} &(\varphi - 1) \sum_{j=1}^K E_{a_j, b_j} [\log V] \\ &\quad + \sum_{k=1}^m \left\{ \frac{(\xi_k^* - \xi_k)^2}{2\tau_k^2} + \frac{\tau_k^{*2}}{2\tau_k^2} \right\} + \sum_{k=1}^m \left\{ \frac{(\varphi_k^* - \varphi_k)^2}{2\varrho_k^2} + \frac{\varrho_k^{*2}}{2\varrho_k^2} \right\}, \end{aligned} \quad (4.56)$$

and (4.51) can be expressed as:

$$\begin{aligned} &-\sum_{j=1}^K \left\{ (a_j - 1) E_{a_j, b_j} [\log V] + (b_j - 1) E_{a_j, b_j} [\log(1 - V)] - \log B(a_j, b_j) \right\} \\ &\quad - \sum_{k=1}^m \log \tau_k^{*2} - \sum_{k=1}^m \log \varrho_k^{*2}, \end{aligned} \quad (4.57)$$

Our variational parameters are defined as:

$$\begin{aligned} V_j &\stackrel{iid}{\sim} Be(1, n - j), \quad j = 1, \dots, k_n - 1, V_{k_n} \equiv 1 \\ \beta_l &\stackrel{iid}{\sim} N(\xi_l^*, \tau_l^{*2}), \quad l = 1, \dots, m \\ \gamma_l &\stackrel{iid}{\sim} N(\vartheta_l^*, \varrho_l^{*2}), \quad l = 1, \dots, m \end{aligned} \quad (4.58)$$

We will next write the “ V_j ’s” in terms of the “ w_j ’s” by expressing w according to the stick-breaking construction. In this instance, we subtract each particular observation j from the number of non-censored observations; that is,

$$\begin{aligned} w_j &= \prod_{l=1}^{j-1} (1 - V_l) V_j \\ V_j &\stackrel{\text{ind}}{\sim} Be(1, k_n - j) \end{aligned} \quad (4.59)$$

Notice that

$$\begin{aligned} \left(1 - \sum_{l=1}^j w_l\right) &= \prod_{l=1}^j (1 - V_l) \\ V_j &= 1 - \frac{1 - \sum_{l=1}^j w_l}{1 - \sum_{l=1}^{j-1} w_l} \\ &= \frac{w_j}{\sum_{l=1}^{j-1} w_l} \end{aligned} \quad (4.60)$$

To handle censoring, let us define a distribution on the V ’s that will serve the role as the prior. Using the likelihood defined above, we will update the variational parameters based on the beta distribution. If censoring is present, we will start at the last observation (assuming that it is non-censored), move backwards, and subtract the number of consecutive censored observations; that is,

$$V_j \stackrel{\text{ind}}{\sim} Be(1, n - j - c_j), \quad j = 1, \dots, k_n - 1, V_{k_n} \equiv 1, \quad (4.61)$$

where c_j represents the number of consecutive censored observations that occur after the j th failure time.

4.3 Simulation Studies

Next, we will design a simulation study using both the full Bayesian and Bayesian bootstrap variational methods outlined above to determine whether the Yang-Prentice ($\beta \neq 0, \gamma \neq 0$), proportional hazards ($\gamma = 0$), proportional odds ($\beta = 0$), or null model ($\beta = \gamma =$

0) is more appropriate. To start, we will simulate data according to each of the following four cases: 1. $\beta \neq 0, \gamma \neq 0$; 2. $\gamma = 0, \beta \neq 0$; 3. $\gamma \neq 0, \beta = 0$; and 4. $\beta = \gamma = 0$, and calculate the required marginal density after implementing variational methods. Note that when we specify $\beta \neq 0$, we are setting $\beta = (0, 0, 0.1)$. Likewise, when we specify $\gamma \neq 0$, we are setting $\gamma = (0.7, 0.4, 0.3)$. We will fit the data and calculate the marginal density under any misspecified model represented by the three other cases; e.g., we simulate data from the Yang-Prentice model, but fit the null model. We will denote the values of the marginal density by p_{1l}, p_{2l}, p_{3l} , and p_{4l} , which originate from the YP, PH, PO, and NULL models respectively, and l represents the number of replicates. To calculate the marginal probabilities, we will compute the maximum among these four models; that is, $\max(p_{jl}, j = 1, \dots, 4) = m_l$. Next, we set $r_{jl} = \frac{p_{jl}}{m_l}$, and calculate the marginal probabilities as $\frac{1}{100} \sum_{l=1}^{100} I(r_{jl} = 1)$, $j = 1, \dots, 4$. After calculating the marginal density values for both the true and misspecified models, we will perform model selection by computing the log Bayes factor; that is, $\log p_{1l} - \log p_{jl}$, where $j = 2, \dots, 4$, and calculating the 25th, 50th, and 75th percentiles. According to Jeffrey's scale, the ranges, $(0, 1.15)$, $(1.15, 3.45)$, $(3.45, 4.60)$, and $(4.60, \infty)$, correspond to the decisions against our particular model of interest: "not worth mentioning", "substantial", "strong", and "very strong." In our setup, we will simulate 100 replicates with sample size $n = 100$, and 30% censoring. The results for the full Bayesian variational method are given by Tables 4.1 and 4.2:

Table 4.1: Results corresponding to the probability of selecting a particular model when simulating data from the assumptions of four different cases under the full Bayesian variational approach: $\beta \neq 0, \gamma \neq 0$; $\gamma = 0, \beta \neq 0$; $\gamma \neq 0, \beta = 0$; $\beta = \gamma = 0$. Prob. = probability of selecting the specified model, Assump. = assumption, Sim. = simulated, Mar. = marginal, MC s.e. = Monte-Carlo standard error.

Data Assump. (Sim. Mar. Prob., MC s.e.)	Model	Fitted Mar. Prob., MC s.e.
YP: $\gamma \neq 0, \beta \neq 0$ (0.8200, 0.0384)	PH: $\gamma = 0, \beta \neq 0$	0.0200, 0.0142
	PO: $\beta = 0, \gamma \neq 0$	0.1500, 0.0357
	NULL: $\beta = \gamma = 0$	0.0100, 0.0099
PO: $\beta = 0, \gamma \neq 0$ (0.7800, 0.0414)	YP: $\beta \neq 0, \gamma \neq 0$	0.1400, 0.0347
	PH: $\gamma = 0, \beta \neq 0$	0.0700, 0.0255
	NULL: $\beta = \gamma = 0$	0.0100, 0.0099
PH: $\gamma = 0, \beta \neq 0$ (0.7800, 0.0414)	YP: $\beta \neq 0, \gamma \neq 0$	0.0100, 0.0099
	PO: $\beta = 0, \gamma \neq 0$	0.0100, 0.0099
	NULL: $\beta = \gamma = 0$	0.2000, 0.0400
NULL: $\gamma = \beta = 0$ (0.8300, 0.0376)	YP: $\beta \neq 0, \gamma \neq 0$	0.0400, 0.0196
	PO: $\beta = 0, \gamma \neq 0$	0.0100, 0.0099
	PH: $\gamma = 0, \beta \neq 0$	0.1200, 0.0325

Table 4.2: Results corresponding to the probability of selecting a particular model conducting pairwise comparisons when simulating data from the assumptions of four different cases under the full Bayesian variational approach: $\beta \neq 0, \gamma \neq 0$; $\gamma = 0, \beta \neq 0$; $\gamma \neq 0, \beta = 0$; $\beta = \gamma = 0$. Assump. = assumption, med. = median, BF = Bayes Factor = Sim. Marginal Prob. / Fitted Marginal Prob., % correct = percentage that the correct model is identified (log BF ≥ 0). Range refers to distance between the 25th and 75th percentiles.

Data Assump.	Model	log BF med. (range)	% correct
YP: $\gamma \neq 0, \beta \neq 0$	PH: $\gamma = 0, \beta \neq 0$	4.2937 (3.5879, 6.8672)	96%
	PO: $\beta = 0, \gamma \neq 0$	2.0687 (0.4842, 3.4115)	84%
	NULL: $\beta = \gamma = 0$	4.6826 (3.9776, 6.2000)	98%
PO: $\beta = 0, \gamma \neq 0$	YP: $\beta \neq 0, \gamma \neq 0$	2.4505 (1.1835, 3.8809)	85%
	PH: $\gamma = 0, \beta \neq 0$	3.3631 (1.7045, 5.4243)	88%
	NULL: $\beta = \gamma = 0$	4.5077 (1.8932, 6.7346)	95%
PH: $\gamma = 0, \beta \neq 0$	YP: $\beta \neq 0, \gamma \neq 0$	3.9615 (2.9218, 4.9257)	99%
	PO: $\beta = 0, \gamma \neq 0$	4.2384 (3.0283, 5.2953)	98%
	NULL: $\beta = \gamma = 0$	1.0110 (0.1354, 1.6932)	80%
NULL: $\gamma = \beta = 0$	YP: $\beta \neq 0, \gamma \neq 0$	1.6297 (0.8323, 2.4788)	95%
	PO: $\beta = 0, \gamma \neq 0$	2.2620 (1.6127, 3.0611)	99%
	PH: $\gamma = 0, \beta \neq 0$	1.1018 (0.4600, 1.8674)	86%

From these results, we can notice that the median log Bayes Factor (BF) are all larger than zero when comparing the true model versus the misspecified model. The procedure usually chooses the correct model ($\geq 80\%$), even in the cases where the only difference between the true and fitted models is the fact that $\beta_3 = 0.1$ versus $\beta_3 = 0$. If we choose a model where the true $\gamma = 0$, and fit a misspecified model with $\gamma \neq 0$, then the true model has higher posterior probability in the simulation examples. Also, notice that we have chosen true values for γ that are further away from zero than β . The Bayes Factor

tends to be higher when comparing the true model to those where we have misspecified γ versus those where we have only misspecified β . For example, when γ only is misspecified, the evidence against the incorrect model ranges from “substantial” to “strong”. However, when β only is misspecified, the evidence against the incorrect model ranges from “not worth mentioning” to “substantial.” Table 4.3 provides estimates of the biases of β in the cases when we have fit a model (either misspecified or not) with $\beta \neq 0$.

Table 4.3: Biases of $\beta = (\beta_0, \beta_1, \beta_2)$ under both the true and misspecified models using the full Bayesian variational approach, s.e. = standard error

Data Assumption	Model	Bias of β , s.e.
YP: $\gamma \neq 0, \beta \neq 0$	YP: $\gamma \neq 0, \beta \neq 0$	(-0.0229, -0.0304, 0.0158), (0.0413, 0.0191, 0.0343)
	PH: $\gamma = 0, \beta \neq 0$	(-0.2184, -0.1842, 0.2045), (0.0186, 0.0277, 0.0141)
PO: $\beta = 0, \gamma \neq 0$	YP: $\gamma \neq 0, \beta \neq 0$	(-0.1671, -0.1659, 0.1172), (0.0241, 0.0476, 0.0472)
	PH: $\gamma = 0, \beta \neq 0$	(-0.5004, -0.4466, 0.2225), (0.0532, 0.0467, 0.0364)
PH: $\gamma = 0, \beta \neq 0$	PH: $\gamma = 0, \beta \neq 0$	(0.0203, 0.0242, 0.0231), (0.0275, 0.0283, 0.0535)
	YP: $\gamma \neq 0, \beta \neq 0$	(0.2561, 0.3660, -0.2353), (0.0849, 0.0459, 0.0481)
NULL: $\gamma = \beta = 0$	YP: $\beta \neq 0, \gamma \neq 0$	(-0.2611, -0.1123, 0.2203), (0.0246, 0.0210, 0.0252)
	PH: $\gamma = 0, \beta \neq 0$	(-0.2635, -0.0381, -0.0752), (0.0375, 0.0235, 0.0364)

From these results, we can notice that the biases are large in all cases where the model is incorrectly specified, but remain small when the correct model is fitted. This shows the usefulness of the YP model which contains both the proportional hazards and proportional odds models. We will next conduct the same analyses by implementing the Bayesian bootstrap variational method whose results are given by Tables 4.4-4.6:

Table 4.4: Results corresponding to the probability of selecting a particular model when simulating data from the assumptions of four different cases under the Bayesian bootstrap variational approach: $\beta \neq 0, \gamma \neq 0$; $\gamma = 0, \beta \neq 0$; $\gamma \neq 0, \beta = 0$; $\beta = \gamma = 0$. Prob. = probability of selecting the specified model, Assump. = assumption, Sim. = simulated, Mar. = marginal, MC s.e. = Monte-Carlo standard error.

Data Assump. (Sim. Mar. Prob., MC s.e.)	Model	Fitted Mar. Prob., MC s.e.
YP: $\gamma \neq 0, \beta \neq 0$ (0.7200, 0.0449)	PH: $\gamma = 0, \beta \neq 0$	0.0900, 0.0286
	PO: $\beta = 0, \gamma \neq 0$	0.1400, 0.0347
	NULL: $\beta = \gamma = 0$	0.0500, 0.0218
PO: $\beta = 0, \gamma \neq 0$ (0.7500, 0.0433)	YP: $\beta \neq 0, \gamma \neq 0$	0.1500, 0.0357
	PH: $\gamma = 0, \beta \neq 0$	0.0700, 0.0255
	NULL: $\beta = \gamma = 0$	0.0300, 0.0171
PH: $\gamma = 0, \beta \neq 0$ (0.8500, 0.0357)	YP: $\beta \neq 0, \gamma \neq 0$	0.0200, 0.0140
	PO: $\beta = 0, \gamma \neq 0$	0.0100, 0.0099
	NULL: $\beta = \gamma = 0$	0.1200, 0.0325
NULL: $\gamma = \beta = 0$ (0.7000, 0.0438)	YP: $\beta \neq 0, \gamma \neq 0$	0.0700, 0.0255
	PO: $\beta = 0, \gamma \neq 0$	0.0700, 0.0255
	PH: $\gamma = 0, \beta \neq 0$	0.1600, 0.0367

Table 4.5: Results corresponding to the probability of selecting a particular model conducting pairwise comparisons when simulating data from the assumptions of four different cases under the Bayesian bootstrap variational approach: $\beta \neq 0, \gamma \neq 0$; $\gamma = 0, \beta \neq 0$; $\gamma \neq 0, \beta = 0$; $\beta = \gamma = 0$. Assump. = assumption, Med. = median, BF = Bayes Factor = Sim. Marginal Prob. / Fitted Marginal Prob., % correct = percentage that the correct model is identified (log BF ≥ 0). Range refers to distance between the 25th and 75th percentiles.

Data Assump.	Model	log BF med. (range)	% correct
YP: $\gamma \neq 0, \beta \neq 0$	PH: $\gamma = 0, \beta \neq 0$	3.6448 (1.9182, 5.8133)	91%
	PO: $\beta = 0, \gamma \neq 0$	1.9731 (0.3472, 3.1938)	80%
	NULL: $\beta = \gamma = 0$	4.1568 (2.2883, 6.1624)	93%
PO: $\beta = 0, \gamma \neq 0$	YP: $\beta \neq 0, \gamma \neq 0$	2.4372 (0.9052, 4.4732)	80%
	PH: $\gamma = 0, \beta \neq 0$	3.2612 (1.7333, 5.5454)	91%
	NULL: $\beta = \gamma = 0$	3.9719 (1.8117, 5.7376)	91%
PH: $\gamma = 0, \beta \neq 0$	YP: $\beta \neq 0, \gamma \neq 0$	2.8591 (2.0102, 3.6673)	98%
	PO: $\beta = 0, \gamma \neq 0$	2.9142 (2.2537, 3.7193)	99%
	NULL: $\beta = \gamma = 0$	1.4247 (0.4759, 2.4322)	87%
NULL: $\gamma = \beta = 0$	YP: $\beta \neq 0, \gamma \neq 0$	1.7818 (0.9651, 2.6626)	91%
	PO: $\beta = 0, \gamma \neq 0$	2.4633 (1.3474, 3.2255)	90%
	PH: $\gamma = 0, \beta \neq 0$	1.0142 (0.0976, 1.9840)	80%

Table 4.6: Biases of $\beta = (\beta_0, \beta_1, \beta_2)$ under both the true and misspecified models using the Bayesian bootstrap variational approach, s.e. = standard error

Data Assumption	Model	Bias of β , s.e.
YP: $\gamma \neq 0, \beta \neq 0$	YP: $\gamma \neq 0, \beta \neq 0$	(-0.0212, 0.0638, 0.0338), (0.0156, 0.0264, 0.0261)
	PH: $\gamma = 0, \beta \neq 0$	(-0.2975, -0.1672, 0.1264), (0.0249, 0.0163, 0.1324)
PO: $\beta = 0, \gamma \neq 0$	YP: $\gamma \neq 0, \beta \neq 0$	(-0.2517, 0.1208, 0.0907), (0.0521, 0.0479, 0.0394)
	PH: $\gamma = 0, \beta \neq 0$	(-0.1839, -0.0375, 0.1176), (0.0570, 0.0114, 0.0147)
PH: $\gamma = 0, \beta \neq 0$	PH: $\gamma = 0, \beta \neq 0$	(-0.0276, 0.0194, 0.0632), (0.0201, 0.0146, 0.0615)
	YP: $\gamma \neq 0, \beta \neq 0$	(-0.3993, -0.2885, 0.1156), (0.0458, 0.1689, 0.0178)
NULL: $\gamma = \beta = 0$	YP: $\beta \neq 0, \gamma \neq 0$	(-0.3656, -0.2824, 0.1104), (0.0330, 0.0224, 0.0178)
	PH: $\gamma = 0, \beta \neq 0$	(-0.1214, -0.1494, -0.0443), (0.0335, 0.0104, 0.0156)

These results are very similar to those using the full Bayesian variational approach. In the majority of cases, the Bayes Factor is slightly more extreme under the full Bayesian variational approach. However, both approaches can effectively identify the correct model, even under cases where there is only a slight model misspecification. Note that in either approach, across all cases, the correct model is chosen at least 80% of the time. Overall, both methods perform model selection well.

Bibliography

- [1] Andersen, P.K., Borgan O., Gill R.D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.
- [2] Basu, S. and Chib, S. (2003). Marginal Likelihood and Bayes Factors for Dirichlet Process Mixture Models. *Journal of the American Statistical Association*, **98**, 224-35.
- [3] Belisle, C. (1992). Convergence theorems for a class of simulated annealing algorithms. *Journal of Applied Probability*, **29**, 885-95.
- [4] Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine*, **2**, 273-77.
- [5] Blei, D., and Jordan, M. (2004). Variational Methods for the Dirichlet Process. *Proceedings of the 21st International Conference on Machine Learning*. Banff, Canada.
- [6] Bowman, A.W. (1984). An alternative method of cross-validation in the smoothing of density estimates. *Biometrika*, **71**, 353-60.
- [7] Chib, S. (1995). Marginal likelihood from the Gibbs Output *Journal of the American Statistical Association*, **90**, 1313-21.
- [8] Chib, S., and Jeliazkov, I. (2001). Marginal Likelihood From the Metropolis-Hastings Output *Journal of the American Statistical Association*, **96**, 270-81.
- [9] Cox, D.R. (1972). Regression models and life-tables (with Discussion). *Journal of the Royal Statistical Society, Series B*, **30**, 248-75

- [10] Deheuvels, P. (1977). Estimation nonparamétrique de la densité par Histogrammes Généralisés, *Revue Statistique Appliquée*, **25**, 5-42.
- [11] Durbin, R., Eddy, S.R., Krogh, A., and Mitchison, G.J. (1998). Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge, UK.
- [12] Fleming, T., and Harrington, D. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.
- [13] Gastrointestinal Tumor Study Group (1982). A comparison of combination and combined modality therapy for locally advanced gastric carcinoma. *Cancer*, **49**, 1771-77.
- [14] Gilks, W.R. (1992). Derivative-free adaptive rejection sampling for Gibbs sampling. In *Bayesian Statistics 4*, Oxford: Clarendon.
- [15] Gilks, W.R., Best, N.G., and Tan, K.K.C. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Applied Statistics*, **44**, 455-72.
- [16] Han, C., and Carlin, B.P. (2001). MCMC Methods for Computing Bayes Factors: A Comparative Review. *Journal of the American Statistical Association*, **96**, 1122-32.
- [17] Ishwaran, H., James, L.F., and Sun, J. (2001). Bayesian Model Selection in Finite Mixtures by Marginal Density Decomposition. *Journal of the American Statistical Association*, **96**, 1316-32.
- [18] Jeffreys, H. (1961) *The Theory of Probability*, **3e**, 432.
- [19] Jones, M.C., Marron, J.S., and Sheather, S.J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, **91**, 401-07.
- [20] Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. (1999). Introduction to variational methods for graphical models. *Machine Learning*, **37**, 183-33.

- [21] Kalbfleisch, J.D. and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley.
- [22] Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457-81.
- [23] Kim, Y., and Lee, J. (2003). Bayesian Bootstrap for Proportional Hazards Models. *The Annals of Statistics*, **31**, 1905-22.
- [24] Klein, J., and Moeschberger, M. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer.
- [25] Kong, A., Liu, J.S., and Wong, W.H. (1994). Sequential Imputations and Bayesian Missing Data Problems. *Journal of the American Statistical Association*, **89**, 278-88.
- [26] Lo, A. (1993). A Bayesian bootstrap for censored data. *The Annals of Statistics*, **21**, 100-23.
- [27] Neal, R. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, **9**, 249-65.
- [28] Owen, A. (1990). Empirical Likelihood Ratio Confidence Regions. *The Annals of Statistics*, **18**, 90-120.
- [29] Park, B.-U. and Marron, J.S. (1990). Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, **85**, 66-72.
- [30] Price, G., Crooks, G., Green, R., and Brenner, S. (2005). Statistical evaluation of pairwise protein sequence comparison with the Bayesian bootstrap. *Bioinformatics*, **21**, 3824-31.
- [31] Ritter, C., and Tanner, M.A. (1992). Facilitating the Gibbs Sampler: The Biggs Stopper and the Griddy-Gibbs Sampler. *Journal of the American Statistical Association*, **87**, 861-68.

- [32] Rubin, D.B. (1981). The Bayesian bootstrap *The Annals of Statistics*, **9**, 130-34.
- [33] Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, **9**, 65-78.
- [34] Scott, D.W. and Terrell, G.R. (1987). Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, **82**, 1131-46.
- [35] Sethuraman, J. (1994) A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639-50.
- [36] Susarla, V. and Van Ryzin, J. (1976). Nonparametric Bayesian elimination of survival curves from incomplete observations. *Journal of the American Statistical Association*, **71**, 897-02.
- [37] Ueda, N., and Ghahramani, Z. (2002). Bayesian model search for mixture models based on optimizing variational bounds. *Neural Networks*, **15**, 1223-41.
- [38] Wainwright, M. and Jordan, M. (2003). Graphical models, exponential families, and variational inference (Technical Report 649). U.C. Berkeley, Department of Statistics.
- [39] Whittemore, A.S. and Keller, J.B. (1986). Survival estimation using splines. *Biometrika*, **42**, 495-06.
- [40] Yang, S. and Prentice R. (1999). Semiparametric inference in the proportional odds regression model, *Journal of the American Statistical Association*, **94**, 125-36.
- [41] Yang, S. and Prentice R. (2005). Semiparametric analysis of short-term and long-term hazard ratios with two-sample survival data. *Biometrika*, **92**, 1-17.

Appendices

APPENDIX A

Theorem 1: $S_1(t) = [1 + \frac{\theta_1}{\theta_2} K(t)]^{-\theta_2}$

Proof:

Let $u = \theta_1 + (\theta_2 - \theta_1)S_2(t)$.

This implies that $S_1(t) = \exp[-\int_0^m \frac{\theta_1 \theta_2}{\theta_1 + (\theta_2 - \theta_1)S_2(t)} h_2(t) dt] = \exp[-\int_a^b \frac{\theta_1 \theta_2}{(u^2 - \theta_1 u)} du]$,
where $a = \theta_1 + \frac{\theta_2 - \theta_1}{1 + K(m)}$ and $b = \theta_2$.

Next, we use the fact that $\int \frac{\theta_1}{u(u - \theta_1)} du = \int (\frac{1}{u - \theta_1} - \frac{1}{u}) du = \log \frac{u - \theta_1}{u}$
It follows that $\int \frac{\theta_1 \theta_2}{u(u - \theta_1)} du = \theta_2 \log \frac{u - \theta_1}{u}$

$$\begin{aligned} \text{Hence, } S_1(t) &= \exp[-\int_a^b \frac{\theta_1 \theta_2}{(u^2 - \theta_1 u)} du] = \exp(-\theta_2 \log \left[\left(\frac{\theta_2 - \theta_1}{\theta_2} \right) \left(\frac{\theta_1 + \frac{\theta_2 - \theta_1}{1 + K(m)}}{\frac{\theta_2 - \theta_1}{1 + K(m)}} \right) \right]) \\ &= \exp(-\theta_2 \log \left[\left(\frac{\theta_1(1 + K(m)) + (\theta_2 - \theta_1)}{\theta_2} \right) \left(\frac{1 + K(m)}{1 + K(m)} \right) \right]) = \exp(-\theta_2 \log \left[\frac{\theta_2 + \theta_1 K(m)}{\theta_2} \right]) \\ &= \exp(\log \left[1 + \frac{\theta_1}{\theta_2} K(m) \right]^{-\theta_2}) = \left[1 + \frac{\theta_1}{\theta_2} K(m) \right]^{-\theta_2}. \end{aligned}$$

It also follows trivially that $S_1(t) = \left\{ 1 - \frac{\theta_1}{\theta_2} + \frac{\theta_1}{\theta_2} e^{H_2(t)} \right\}^{-\theta_2}$.

Theorem 2: $\int_0^t \frac{a}{1 + bS_0(s)} dH_0(s) = a \log \frac{1 + bS_0(t)}{S_0(t)(1 + b)}$

Proof:

Let us first apply a change of variable where $u = H_0(s)$ to obtain that:

$$\int_0^t \frac{a}{1 + bS_0(s)} dH_0(s) = \int_0^{H_0(t)} \frac{a}{1 + be^{-u}} du.$$

$$\begin{aligned} \text{It follows that } \int_0^{H_0(t)} \frac{a}{1 + be^{-u}} du &= \int_0^{H_0(t)} \frac{ae^u}{e^u + b} du = \int_{1+b}^{b+e^{H_0(t)}} \frac{a}{v} dv \\ &= a \log \frac{b + e^{H_0(t)}}{1 + b} = a \log \frac{1 + bS_0(t)}{S_0(t)(1 + b)}. \end{aligned}$$

Notice that we have made the substitutions: $S_0(s) = e^{-H_0(s)}$ and $v = e^u + b$.

APPENDIX B

Likelihood Derivations

The Proportional Hazards Model

$$L(\theta) = \prod_{j=1}^{n_1} (h_1(x_{1j}, \theta))^{\Delta_{1j}} [S_1(x_{1j}, \theta)] = \prod_{j=1}^{n_1} (\theta \hat{h}_2(x_{1j}))^{\Delta_{1j}} (\hat{S}_2(x_{1j}))^\theta$$

It follows that $\prod_{j=1}^{n_1} (\theta \hat{h}_2(x_{1j}))^{\Delta_{1j}} (\hat{S}_2(x_{1j}))^\theta \propto \theta^{\sum_{j=1}^{n_1} \Delta_{1j}} (\prod_{j=1}^{n_1} S_2(x_{1j}))^\theta$

$$\text{and } \theta^{\sum_{j=1}^{n_1} \Delta_{1j}} (\prod_{j=1}^{n_1} S_2(x_{1j}))^\theta = \theta^{\sum_{j=1}^{n_1} \Delta_{1j}} e^{-\theta \sum_{j=1}^{n_1} (-\log \hat{S}_2(x_{1j}))}$$

$$\log L(\theta) = \sum_{j=1}^{n_1} \Delta_{1j} \log \theta - \theta \sum_{j=1}^{n_1} (-\log \hat{S}_2(x_{1j}))$$

$$\frac{d \log L(\theta)}{d\theta} = \frac{\sum_{j=1}^{n_1} \Delta_{1j}}{\theta} + \sum_{j=1}^{n_1} \log \hat{S}_2(x_{1j})$$

$$\text{Hence, we can obtain } \hat{\theta} = \frac{\sum_{j=1}^{n_1} \Delta_{1j}}{-\sum_{j=1}^{n_1} \log \hat{S}_2(x_{1j})}$$

$$\frac{d^2 \log L(\theta)}{d\theta^2} = -\frac{\sum_{j=1}^{n_1} \Delta_{1j}}{\theta^2}$$

$$\text{This implies that } \frac{d \log L(\theta)}{d\theta} \Big|_{\theta=\hat{\theta}} = -\frac{(\sum_{j=1}^{n_1} \log \hat{S}_2(x_{1j}))^2}{\sum_{j=1}^{n_1} \Delta_{1j}} < 0.$$

The AFT Model

$$L(\theta) = \prod_{j=1}^{n_1} (h_1(x_{1j}, \theta))^{\Delta_{1j}} [S_1(x_{1j}, \theta)] = \prod_{j=1}^{n_1} [\theta \hat{h}_2(\theta x_{1j})]^{\Delta_{1j}} (\hat{S}_2(\theta x_{1j}))$$

It follows that

$$\prod_{j=1}^{n_1} [\theta \hat{h}_2(\theta x_{1j})]^{\Delta_{1j}} (\hat{S}_2(\theta x_{1j})) \propto \theta^{\sum_{j=1}^{n_1} \Delta_{1j}} \prod_{j=1}^{n_1} h_2(\theta x_{1j})^{\Delta_{1j}} (\prod_{j=1}^{n_1} \hat{S}_2(\theta x_{1j}))$$

$$\log L(\theta) = \sum_{j=1}^{n_1} \Delta_{1j} \log \theta + \sum_{j=1}^{n_1} \Delta_{1j} \log h_2(\theta x_{1j}) + \sum_{j=1}^{n_1} \log S_2(\theta x_{1j})$$

$$\frac{d \log L(\theta)}{d\theta} = \frac{\sum_{j=1}^{n_1} \Delta_{1j}}{\theta} + \frac{\sum_{j=1}^{n_1} \Delta_{1j} x_{1j} h_2'(\theta x_{1j})}{h_2(\theta x_{1j})} - \sum_{j=1}^{n_1} x_{1j} h_2(\theta x_{1j})$$

$$\begin{aligned} \frac{d^2 \log L(\theta)}{d\theta^2} &= -\frac{\sum_{j=1}^{n_1} \Delta_{1j}}{\theta^2} - \frac{\sum_{j=1}^{n_1} \Delta_{1j} x_{1j} [h_2'(\theta x_{1j})]^2 x_{1j}}{[h_2(\theta x_{1j})]^2} + \frac{\sum_{j=1}^{n_1} \Delta_{1j} x_{1j} h_2''(\theta x_{1j}) x_{1j}}{h_2(\theta x_{1j})} \\ &\quad - \sum_{j=1}^{n_1} x_{1j} h_2'(\theta x_{1j}) x_{1j} \end{aligned}$$

Hence, $\frac{d \log L(\theta)}{d\theta} \Big|_{\theta=\hat{\theta}} < 0$.

Proportional Odds Model (POM)

$$\begin{aligned} L(\theta) &= \prod_{j=1}^{n_1} (h_1(x_{1j}, \theta))^{\Delta_{1j}} [S_1(x_i, \theta)] \\ &= \prod_{j=1}^{n_1} \left[\frac{\hat{h}_2(x_{1j})\theta}{\hat{S}_2(x_{1j}) + \theta(1 - \hat{S}_2(x_{1j}))} \right]^{\Delta_{1j}} \prod_{j=1}^{n_1} \left[\frac{\hat{S}_2(x_{1j})}{\hat{S}_2(x_{1j}) + \theta(1 - \hat{S}_2(x_{1j}))} \right] \end{aligned}$$

$$\log L(\theta) \propto \sum_{j=1}^{n_1} \Delta_{1j} \log \theta - \sum_{j=1}^{n_1} (\Delta_{1j} + 1) \log (\hat{S}_2(x_{1j}) + \theta(1 - \hat{S}_2(x_{1j})))$$

$$\frac{d \log L(\theta)}{d\theta} = \frac{\sum_{j=1}^{n_1} \Delta_{1j}}{\theta} - \sum_{j=1}^{n_1} \left[\frac{\hat{F}_2(x_{1j})}{\hat{S}_2(x_{1j}) + \theta \hat{F}_2(x_{1j})} \right] (\Delta_{1j} + 1)$$

$$\frac{d^2 \log L(\theta)}{d\theta^2} = -\frac{\sum_{j=1}^{n_1} \Delta_{1j}}{\theta^2} + \sum_{j=1}^{n_1} \left[\frac{\hat{F}_2(x_{1j})^2 (\Delta_{1j} + 1)}{(\hat{S}_2(x_{1j}) + \theta \hat{F}_2(x_{1j}))^2} \right] < \sum_{j=1}^{n_1} \frac{\hat{F}_2(x_{1j}) (\Delta_{1j} + 1)}{\hat{S}_2(x_{1j}) + \theta \hat{F}_2(x_{1j})}$$

This implies that $\frac{d \log L(\theta)}{d\theta} \Big|_{\theta=\hat{\theta}} < 0$.

The Yang-Prentice Model (YP)

$$\begin{aligned} L(\theta) &= \prod_{j=1}^{n_1} (h_1(x_{1j}, \theta))^{\Delta_{1j}} [S_1(x_i, \theta)] \\ &= \prod_{j=1}^{n_1} \left(\frac{\theta_1 \theta_2 \hat{h}_2(x_{1j})}{\theta_1 + (\theta_2 - \theta_1) \hat{S}_2(x_{1j})} \right)^{\Delta_{1j}} \prod_{j=1}^{n_1} \left[1 + \frac{\theta_1 \hat{F}_2(x_{1j})}{\theta_2 \hat{S}_2(x_{1j})} \right]^{-\theta_2} \end{aligned}$$

$$\log L(\theta) = \sum_{j=1}^{n_1} \Delta_{1j} \log \left[\frac{\theta_1 \theta_2 \hat{h}_2(x_{1j})}{\theta_1 + (\theta_2 - \theta_1) \hat{S}_2(x_{1j})} \right] - \theta_2 \sum_{j=1}^{n_1} \log \left[1 + \frac{\theta_1 \hat{F}_2(x_{1j})}{\theta_2 \hat{S}_2(x_{1j})} \right]$$

$$\begin{aligned} \frac{d \log L(\theta)}{d\theta_1} &= \sum_{j=1}^{n_1} \Delta_{1j} \frac{[\theta_2^2 \hat{S}_2(x_{1j}) \hat{h}_2(x_{1j}) - \theta_1 \theta_2 \hat{h}_2(x_{1j}) (1 - \hat{S}_2(x_{1j}))]}{[\theta_1 + (\theta_2 - \theta_1) \hat{S}_2(x_{1j})]^2} \\ &\quad - \theta_2 \sum_{j=1}^{n_1} \frac{\hat{F}_2(x_{1j})}{\theta_2 \hat{S}_2(x_{1j}) + \theta_1 \hat{F}_2(x_{1j})} \end{aligned}$$

$$\begin{aligned} \frac{d \log L(\theta)}{d\theta_2} &= \sum_{j=1}^{n_1} \Delta_{1j} \frac{\theta_1 + (\theta_2 - \theta_1) \hat{S}_2(x_{1j})}{\theta_1 \theta_2 \hat{h}_2(x_{1j})} \frac{(\theta_1 + (\theta_2 - \theta_1) \hat{S}_2(x_{1j})) \theta_1 \hat{h}_2(x_{1j}) - \theta_1 \theta_2 \hat{h}_2(x_{1j}) \hat{S}_2(x_{1j})}{[\theta_1 + (\theta_2 - \theta_1) \hat{S}_2(x_{1j})]^2} \\ &\quad - \sum_{j=1}^{n_1} \log \left[1 + \frac{\theta_1 \hat{F}_2(x_{1j})}{\theta_2 \hat{S}_2(x_{1j})} \right] + \theta_2 \sum_{j=1}^{n_1} \frac{\theta_1 \hat{F}_2(x_{1j})}{\theta_2^2 \hat{S}_2(x_{1j})} \frac{\theta_2 \hat{S}_2(x_{1j})}{\theta_2 \hat{S}_2(x_{1j}) + \theta_1 \hat{F}_2(x_{1j})} \end{aligned}$$

$$\frac{d^2 \log L(\theta)}{d\theta_1^2} = \frac{\sum_{j=1}^{n_1} \Delta_{1j} [\theta_1(\hat{S}_2(x_{1j})-1) + 3\theta_2 \hat{S}_2(x_{1j})] \hat{h}_2(x_{1j}) \theta_2 (\hat{S}_2(x_{1j})-1)}{[\theta_1(\hat{S}_2(x_{1j})-1) - \theta_2 \hat{S}_2(x_{1j})]^3} + \frac{\theta_2 \sum_{j=1}^{n_1} \hat{F}_2^2(x_{1j})}{[\theta_2 \hat{S}_2(x_{1j}) + \theta_1 \hat{F}_2(x_{1j})]^2}$$

$$\begin{aligned} \frac{d^2 \log L(\theta)}{d\theta_2^2} &= \frac{\sum_{j=1}^{n_1} \Delta_{1j} [2\theta_2 \hat{S}_2(x_{1j}) - \theta_1(\hat{S}_2(x_{1j})-1)] \theta_1 (\hat{S}_2(x_{1j})-1)}{\theta_2^2 [(\theta_2 \hat{S}_2(x_{1j}) - \theta_1(\hat{S}_2(x_{1j})-1))]^2} + \frac{\sum_{j=1}^{n_1} \theta_1 \hat{F}_2(x_{1j})}{\theta_2 [\theta_2 \hat{S}_2(x_{1j}) + \theta_1 \hat{F}_2(x_{1j})]} \\ &\quad - \frac{\sum_{j=1}^{n_1} \hat{F}_2(x_{1j}) \theta_1 \hat{S}_2(x_{1j})}{[\theta_2 \hat{S}_2(x_{1j}) + \theta_1 \hat{F}_2(x_{1j})]^2} \end{aligned}$$

$$\frac{d^2 \log L(\theta)}{d\theta_1 d\theta_2} = \frac{\sum_{j=1}^{n_1} \Delta_{1j} [3\theta_2 \hat{S}_2(x_{1j}) + \theta_1(\hat{S}_2(x_{1j})-1)] \hat{h}_2(x_{1j}) \theta_1 (\hat{S}_2(x_{1j})-1)}{[\theta_2 \hat{S}_2(x_{1j}) - \theta_1(\hat{S}_2(x_{1j})-1)]^3} - \frac{\sum_{j=1}^{n_1} \hat{F}_2^2(x_{1j}) \theta_1}{[\theta_2 \hat{S}_2(x_{1j}) + \theta_1 \hat{F}_2(x_{1j})]^2}$$