Flexible Regression Models for Survival Data

Ennan Gu

Follow this and additional works at: https://scholarcommons.sc.edu/etd

Part of the Statistics and Probability Commons

Recommended Citation

This Open Access Dissertation is brought to you by Scholar Commons. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact dillarda@mailbox.sc.edu.
Flexible Regression Models for Survival Data

by

Ennan Gu

Bachelor of Arts
China Agricultural University, 2012

Master of Arts
Northeastern University, 2014

Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in
Statistics
College of Arts and Sciences
University of South Carolina
2020

Accepted by:
Lianming Wang, Major Professor
Jiajia Zhang, Committee Member/Outside Member
Xiaoyan Lin, Committee Member
Yen-Yi Ho, Committee Member
Cheryl L. Addy, Vice Provost and Dean of the Graduate School
ACKNOWLEDGMENTS

I would like to express my deepest and sincere gratitude to my advisor Dr. Lianming Wang. The solid statistical foundation of Dr. Wang’s research work in survival analysis and his innovative thinking encourage me all the time. His enthusiasm in statistics and conscientious attitude keep me engaged in my research. I will never forget the times when I felt lack of motivation and stressed, Dr. Wang offered advice and encouragement without hesitation. Without his persistent guidance, my doctoral dissertation would not be completed.

I am also very thankful to Dr. Jiajia Zhang, Dr. Xiaoyan Lin and Dr. Yen-Yi Ho for spending time serving on my committee and giving constructive suggestions on my dissertation. They have provided not only professional guidance but also life wisdom. I would especially like to thank Dr. Jiajia Zhang, who gives insightful guidance in survival analysis and many brilliant ideas on my research work. I also would like to thank all the professors and faculty members in the department of statistics for teaching me statistical knowledge and providing various resources and help during my doctoral life.

Last but not the least, I am grateful that my family and friends always support and encourage me to follow my heart and pursue my dream. Many thanks to their patience, tolerance and love.
Abstract

Survival analysis is a branch of statistics to analyze the time-to-event data or survival data. One important feature of survival data is censoring, which means that not all the subjects’ survival time are observed directly. Among all the survival data, right-censored data are the most common type and consist of some exactly observed survival times and some right-censored observations. In this dissertation, we focus on studying flexible regression models for complicated right-censored survival data when the classical proportional hazards (PH) assumption is not satisfied. Flexible semiparametric regression models can largely avoid misspecification of parametric distributions and thus provide more modeling flexibility.

Cure models are studied in this dissertation to analyze survival data, for which there is a cured group in the study population and this is evidenced by a level-off at the end of the nonparametric survival estimate. In addition, we also incorporate background mortality in the cure models to improve estimation accuracy in this research. Considering the background mortality is important based on the fact that patients dying from other causes also benefit from the treatment of the disease of interest as shown in the SEER cancer studies. In Chapter 2, a semiparametric estimation approach is proposed based on EM algorithm under the mixture cure proportional hazards model with background mortality (MCPH+BM). In Chapter 3, a promotion time cure proportional hazards model with background mortality (PTPH+BM) is proposed, and its extension to the semiparametric transformation model is under further exploration. Both models are validated via comprehensive simulation studies and real data analysis.
Another perspective on non-proportional hazards is to explore a more general model than the Cox PH model such as the generalized odds-rate (GOR) models (Dabrowska and Doksum, 1988). In Chapter 4, the identifiability problems and the estimation of parameters in the GOR models are discussed.
# Table of Contents

Acknowledgments ........................................ iii

Abstract ............................................... iv

List of Tables ........................................ viii

List of Figures ......................................... xi

Chapter 1 Introduction .................................. 1

1.1 Survival Data ...................................... 1

1.2 Motivating Data .................................... 4

1.3 Survival Models ..................................... 7

1.4 Existing Regression Approaches .................. 14

1.5 Structure of the Dissertation ..................... 16

Chapter 2 Semiparametric Estimation of the Cure Fraction in Population-based Cancer Survival Analysis .................................. 18

2.1 Introduction ....................................... 19

2.2 Mixture Cure Model with Background Mortality .......... 22

2.3 EM Algorithm ....................................... 24

2.4 Simulation Studies ................................... 30

2.5 Real Data Study .................................... 39
2.6 Conclusions and Discussions ........................................... 46

CHAPTER 3 SEMIPARAMETRIC ESTIMATION IN THE POPULATION-LEVEL
Promotion Time Cure Model ............................................. 48

3.1 Introduction ............................................................... 48
3.2 Promotion Time Cure Model with Background Mortality ............ 50
3.3 EM Algorithm ............................................................. 52
3.4 Simulation Studies ....................................................... 58
3.5 Breast Cancer Data Analysis .......................................... 61
3.6 Discussions ............................................................... 65

CHAPTER 4 REGRESSION ANALYSIS OF RIGHT-CENSORED DATA UN-
der the Generalized Odds-Rate Model ................................. 68

4.1 Introduction ............................................................... 69
4.2 Identifiability of GOR models ........................................ 71
4.3 The Proposed Method .................................................. 73
4.4 Simulation ................................................................. 79
4.5 Real Data Analysis ...................................................... 87
4.6 Discussions ............................................................... 93

BIBLIOGRAPHY .............................................................. 95

APPENDIX A THE QUANTITIES INVOLVED IN \( \text{var}(\hat{\theta}) \) ................. 102

APPENDIX B THE EM ALGORITHM FOR THE PH MODEL .................. 105
LIST OF TABLES

Table 1.1  Partial patients records in acute leukemia data  ................. 3

Table 2.1  Summary Statistics for Weibull Data with Logistic Link: by semiparametric method, flexible parametric method with monotone splines, parametric Weibull baseline function. ....................... 33

Table 2.2  Summary Statistics for Lognormal Data with Logistic Link: by semiparametric method, flexible parametric method with monotone splines, parametric Lognormal baseline function. .......... 34

Table 2.3  Summary Statistics for Loglogistic Data with Logistic Link: by semiparametric method, flexible parametric method with monotone splines, parametric Loglogistic baseline function. ............ 35

Table 2.4  Sensitivity Analysis: columns are Weibull, Lognormal and Loglogistic data distributions; rows are the parametric Weibull, Lognormal, Loglogistic methods and the proposed semiparametric method. .................................................. 36

Table 2.5  Summary Statistics for the proposed semiparametric method with logistic link (low cure rates) ...................... 37

Table 2.6  SEER Louisiana regional female breast cancer: parameters estimation in the semiparametric MCPH+BM model and the semiparametric MCPH model, race and grade are included in both the cure logistic link function and latency survival function. Model I is to fit the whole data set with 6200 observations, Model II is to fit the subgroup of age \( \leq 70 \) with 4267 observations, and Model III is to fit the subgroup of age \( > 70 \) with 1933 observations. .............................................. 42

Table 3.1  Summary Statistics for Weibull Data: by proposed semiparametric method and parametric method with Weibull baseline function. .................................................. 60
Table 3.2 Summary Statistics for Lognormal Data: by proposed semiparametric method and parametric method with Lognormal baseline function. .......................................................... 61

Table 3.3 Summary Statistics for Loglogistic Data: by proposed semiparametric method and parametric method with Loglogistic baseline function. ........................................................................ 62

Table 3.4 Sensitivity Analysis for 800 Sample Size: rows represent parametric methods and semiparametric method, and columns are the Weibull, Lognormal and Loglogistic data. .......................... 63

Table 3.5 SEER regional female breast cancer: parameters estimation in the semiparametric PCPH+BM model and the semiparametric PCPH model, race and grade are included in both the cure link function and latency survival function. Model I is to fit the whole data set with 6200 observations, Model II is to fit the subgroup data of age $\leq 70$, and model III is to fit the subgroup data of age $> 70$. .................................................................................. 64

Table 4.1 Summary Statistics of the proposed method with concave function $\Lambda_0(t) = \log(1 + t) + t^{1.5}$, $\beta = c(-1, -1)$, $\rho = 2$, $\rho = 1$ and $\rho = 0.5$ with 3 levels censoring rate (around 20% in I: Low Censored, 40% in II: Medium Censored and 75% in III: High Censored). ........................................................................... 81

Table 4.2 Summary Statistics of the proposed method with convex function $\Lambda_0(t) = 0.5t^{5/6}$, $\beta = c(-1, -1)$, $\rho = 2$, $\rho = 1$ and $\rho = 0.5$ with 3 levels censoring rate (around 20% in I: Low Censored, 40% in II: Medium Censored and 75% in III: High Censored). ... 82

Table 4.3 Summary Statistics of the proposed method with concave function $\Lambda_0(t) = \log(1 + t) + t^{1.5}$, $\beta = c(-1, -1)$ and in extreme cases of $\rho = 4$ and $\rho = 0.25$ with 3 levels censoring rate (around 20% in I: Low Censored, 40% in II: Medium Censored and 75% in III: High Censored). Larger $\rho$ provides larger frailty variance, and the setup under $\rho = 4$ needs much larger sample size to provide stable performance. .......................................................... 83
Table 4.4 Summary Statistics of the proposed method with convex function \( \Lambda_0(t) = 0.5t^{5/6} \) and in extreme cases of \( \rho = 4 \) and \( \rho = 0.25 \) with 3 levels censoring rate (around 20\% in I: Low Censored, 40\% in II: Medium Censored and 75\% in III: High Censored). Similar to the concave function \( \Lambda_0(t) \), a larger \( \rho \) provides larger frailty variance, and the setup under \( \rho = 4 \) needs much larger sample size to provide stable performance. . . . 84

Table 4.5 Summary Statistics for concave function \( \Lambda_0(t) = \log(1 + t) + t^{1.5} \), \( \beta = c(-1, -1) \), \( \rho = 2 \) with 3 levels censoring rate (around 20\% in I: Low Censored, 40\% in II: Medium Censored and 75\% in III: High Censored), the proposed EM algorithm performs much better than the direct MLE algorithm. In other settings, EM algorithm also performs better and the results are not shown. . . . 85

Table 4.6 Power Analysis of the proposed method for \( \rho = 1 \) (PO model). Note the concave function is \( \Lambda_0(t) = \log(1 + t) + t^{1.5} \), convex function is \( \Lambda_0(t) = 0.5t^{5/6} \) and \( \beta = c(-1, -1) \). Sample size is 1500. 86

Table 4.7 Summary Statistics of the proposed method with fixed parameter \( \rho \), concave function \( \Lambda_0(t) = \log(1 + t) + t^{1.5} \), convex function \( \Lambda_0(t) = 0.5t^{5/6} \), \( \beta = c(-1, -1) \) with 3 levels censoring rate (around 20\% in I: Low Censored, 40\% in II: Medium Censored and 75\% in III: High Censored), and sample size of 200. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 86

Table 4.8 SEER Louisiana Localized Lung Cancer Data: include sex, race and grade in the survival function. Model I is to fit the data by estimating \( \rho \) with regression parameters, Model II fixes \( \rho = 0 \), and Model III fits the data with fixed \( \rho = 1 \). The best number of interior knots is searched for 10, 15, 20 and 25 based on AIC and BIC. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 91

Table 4.9 SEER Louisiana Localized Lung Cancer Data: parameters estimation in the proposed EM algorithm, sex, race and grade are included in the survival function. Interior knots are selected as 15 based on AIC and BIC for the three models. . . . . . . . . . . . . . . . . . . . . . . . . . . . 92

Table 4.10 SEER Louisiana Localized Lung Cancer Data: parameters estimation in the proposed EM algorithm, sex, race, grade and also the interaction of sex and race are included in the survival function. Interior knots are selected as 15 based on AIC and BIC for the three models. . . . . . . . . . . . . . . . . . . . . . . . . . . . 92
LIST OF FIGURES

Figure 1.1 Kaplan Meier curve by race: the dark black line represents the survival probability for the black people and the light grey line is the overall survival probability for the white people. ........................................ 6

Figure 1.2 Stratified Kaplan Meier curve by Age 70: from light grey to dark black successively are the overall survival probabilities for white of age $\leq 70$, white of age $> 70$, black of age $\leq 70$ and black of age $> 70$. ................................................................. 6

Figure 2.1 Baseline Survival Probability Estimation for Weibull Data by Semiparametric MCPH+BM Model (200 Sample Size). .......................... 35

Figure 2.2 Baseline Survival Probability Estimation for Weibull Data by Semiparametric MCPH+BM Model (400 Sample Size). .......................... 36

Figure 2.3 Baseline Survival Probability Estimation for Weibull Data by Semiparametric MCPH+BM Model (800 Sample Size). .......................... 37

Figure 2.4 Baseline Survival Probability Estimation for Lognormal Data by Semiparametric MCPH+BM Model (200 Sample Size). ................. 38

Figure 2.5 Baseline Survival Probability Estimation for Lognormal Data by Semiparametric MCPH+BM Model (400 Sample Size). ................. 38

Figure 2.6 Baseline Survival Probability Estimation for Lognormal Data by Semiparametric MCPH+BM Model (800 Sample Size). ................. 39

Figure 2.7 Baseline Survival Probability Estimation for Loglogistic Data by Semiparametric MCPH+BM Model (200 Sample Size). .................. 39

Figure 2.8 Baseline Survival Probability Estimation for Loglogistic Data by Semiparametric MCPH+BM Model (400 Sample Size). .................. 40

Figure 2.9 Baseline Survival Probability Estimation for Loglogistic Data by Semiparametric MCPH+BM Model (800 Sample Size). .................. 40
Figure 2.10 Predicted Net Survival Probability for the Simulated Data of 200 Sample Size Weibull Distribution: dashed line and solid line represent the predicted survival probabilities for treatment and control arms respectively. ........................................ 41

Figure 2.11 Predicted Net Survival Probability for the Simulated Data of 200 Sample Size Lognormal Distribution: dashed line and solid line represent the predicted survival probabilities for treatment and control arms respectively. ........................................ 41

Figure 2.12 Predicted Net Survival Probability for the Simulated Data of 200 Sample Size Loglogistic Distribution: dashed line and solid line represent the predicted survival probabilities for treatment and control arms respectively. ........................................ 42

Figure 2.13 The MCPH+BM model predicts the net survival probability by race, and the MCPH model predicts the observed (net) survival probability by race. ....................................................... 44

Figure 2.14 Predicted net survival under the MCPH+BM model and the MCPH model for the subgroup of Age > 70 in the SEER regional female breast cancer. ................................................. 44

Figure 2.15 Predicted net survival under the MCPH+BM model and the MCPH+BM model for the subgroup of Age ≤ 70 in the SEER regional female breast cancer. ........................................... 45

Figure 2.16 Pseudo Data Estimation Output by the R Package “psmcur” . . 45

Figure 3.1 Predicted Survival Probability for the SEER regional female breast cancer data, the proposed PCPH+BM model estimates cure fraction higher than the standard PCPH model. ....................... 64

Figure 3.2 Predicted Survival Probability for the subgroup of age ≤ 70 in the SEER regional female breast cancer data. The proposed PCPH+BM model estimates cure fraction higher than the standard PCPH model. ....................... 65

Figure 3.3 Predicted Survival Probability for the subgroup of age > 70 in the SEER regional female breast cancer data. The proposed PCPH+BM model estimates cure fraction higher than the standard PCPH model. ....................... 65
Figure 4.1 Case i: Solid line is the estimated $\Lambda_{01}(t)$ and dashed line is the estimated $\Lambda_{02}(t)$. ................................................................. 72

Figure 4.2 Case ii: Solid line is the estimated $\Lambda_{01}(t)$ and dashed line is the estimated $\Lambda_{02}(t)$. ................................................................. 73

Figure 4.3 Estimated baseline survival probability for the concave function $\Lambda_0(t) = \log(1 + t) + t^{1.5}$, $\beta = c(-1, -1)$, $\rho = 2$ with low censored rate (sample size 200). ................................................................. 87

Figure 4.4 Estimated baseline survival probability for the concave function $\Lambda_0(t) = \log(1 + t) + t^{1.5}$, $\beta = c(-1, -1)$, $\rho = 2$ with low censored rate (sample size 500). ................................................................. 87

Figure 4.5 Estimated baseline survival probability for the concave function $\Lambda_0(t) = \log(1 + t) + t^{1.5}$, $\beta = c(-1, -1)$, $\rho = 2$ with low censored rate (sample size 1000). ................................................................. 88

Figure 4.6 Estimated baseline survival probability for the concave function $\Lambda_0(t) = \log(1 + t) + t^{1.5}$, $\beta = c(-1, -1)$, $\rho = 2$ with low censored rate (sample size 1500). ................................................................. 88

Figure 4.7 Estimated baseline survival probability for the concave function $\Lambda_0(t) = \log(1 + t) + t^{1.5}$, $\beta = c(-1, -1)$, $\rho = 2$ with medium censored rate (sample size 200). ................................................................. 89

Figure 4.8 Estimated baseline survival probability for the concave function $\Lambda_0(t) = \log(1 + t) + t^{1.5}$, $\beta = c(-1, -1)$, $\rho = 2$ with medium censored rate (sample size 500). ................................................................. 89

Figure 4.9 Estimated baseline survival probability for the concave function $\Lambda_0(t) = \log(1 + t) + t^{1.5}$, $\beta = c(-1, -1)$, $\rho = 2$ with medium censored rate (sample size 1000). ................................................................. 90

Figure 4.10 Estimated baseline survival probability for the concave function $\Lambda_0(t) = \log(1 + t) + t^{1.5}$, $\beta = c(-1, -1)$, $\rho = 2$ with medium censored rate (sample size 1500). ................................................................. 90

Figure 4.11 Predicted survival probability vs. Time (years) for White Female and White Male in the SEER Localized Lung Cancer Data under the proposed method and compared to the Kaplan Meier Curve. The proposed method predicts the survival probability higher for white female than for white male. ........................................... 93
Figure 4.12 Predicted survival probability vs. Time (years) for Black Female and Black Male in the SEER Localized Lung Cancer Data under the proposed method and compared to the Kaplan Meier Curves. The proposed method predicts the survival probability higher for black female than for black male.
Chapter 1

Introduction

1.1 Survival Data

Survival analysis is a collection of statistical procedures that study the duration of time until one particular event of interest happens and thus it is also called time-to-event analysis. The survival time can be years, months, weeks or days from the beginning of follow-up of an individual until the event of interest occurs. The event of interest can be death, disease incidence, relapse from remission or recovery from disease that may happen to a patient. For example, if the event of interest is death, then the survival time will be the time in years until a patient dies. However, the survival time may not be observed directly in some cases. In some studies especially in clinical trial studies, patients are examined only at discrete observational times, and the statuses whether the patients have experienced the event of interest or not at these observational times can be recorded. However, the event time may not be known exactly but be known to fall within some time interval. In other studies, some patients may not experience the event of interest by the end of a study, or some patients have already experienced the event of interest before being recruited in a study. All these phenomena are called censoring, where some survival time of interest can not be known exactly.

Censoring occurs when we have some information about subjects’ survival time, but we do not know what are their exact survival time. Thus censoring is a condition in which a subject’s survival time is only partially known. For example, a patient
moves to another state after being recruited to a clinical trial and is lost to be traced. The only information available on his or her survival time is the last date on which this patient was known to be alive. This date may be the last time that this patient reported to a clinic for a regular check-up.

There are various categories of censoring, such as right censoring, left censoring, and interval censoring. Right censoring occurs when a subject’s exact survival time is not observed and is known to be greater than this person’s observation time. Left censoring occurs when a subject’s exact survival time is less than or equal to this person’s observation time. Interval censoring means a subject’s exact survival is only known to lie within an interval instead of being observed directly. Below are some real life examples to illustrate different censored data.

1.1.1 Right-Censored Data

The Acute Leukemia Group (1963) has reported the results of a clinical trial of a drug 6-mercaptopurine (6-MP) versus a placebo in twenty-one pairs of patients with acute leukemia. This trial was conducted at eleven American hospitals. Patients who had a complete or partial remission of their leukemia induced by treatment with the drug prednisone were selected. A complete or partial remission means that either most or all signs of disease had disappeared from the bone marrow.

The trial was conducted by matching pairs of patients at a given hospital by remission status (complete or partial) and randomizing within the pair to either a 6-MP or placebo maintenance therapy. Patients were followed until their leukemia relapsed or until the end of the study. The event of interest is the relapse of leukemia and the survival time is the time in weeks until patients went out of remission.

Part of the patients records are shown in table 1.1. Patient A was followed from the start of the study until getting leukemia relapsed at week 5, so this person’s survival time was observed to be 5 weeks. Patient B was observed from the start
Table 1.1: Partial patients records in acute leukemia data

<table>
<thead>
<tr>
<th>Patient</th>
<th>Enter Time</th>
<th>End Time</th>
<th>Reason</th>
<th>Failed(1); censored (0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>5</td>
<td>out of remission</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>12</td>
<td>study ends</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>2.5</td>
<td>6</td>
<td>withdrawn</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>4</td>
<td>12</td>
<td>study ends</td>
<td>0</td>
</tr>
<tr>
<td>E</td>
<td>3</td>
<td>9</td>
<td>lost to follow up</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>8</td>
<td>11.5</td>
<td>out of remission</td>
<td>1</td>
</tr>
</tbody>
</table>

of the study, but did not experience the relapse until the end of the 12-week study period, so this person’s survival time was censored and at least 12 weeks. Patient C entered the study between week two and week three and was followed until he or she withdrew at week 6, so this person’s survival time was censored. Patient E entered the study at week 3 and was lost to follow at week 9, so this person’s survival time was censored and at least 6 weeks. In summary, the patient A and F experienced leukemia relapse and their survival times are known exactly. The other four patients did not experience leukemia, and thus their survival times are right-censored. This data set is an example of right-censored data in the survival literature, which contain some exactly observed survival times and some right-censored survival times.

1.1.2 LEFT-CENSORING DATA

One example is that in early childhood learning centers, research interest often focuses upon testing children to determine when a child learns to accomplish certain specified tasks such as reciting the alphabet. The age at which a child can recite the alphabet would be considered as the time-to-event. Often, some children are already able to recite the alphabet when they are recruited in the study. Such survival times are considered left-censored.

Another example is on clinical trial studies. Human immunodeficiency virus (HIV) is a chronic disease which weakens the immune system. HIV RNA or viral load (VL) measures the number of actively replicating HIV virus in a subject, which is an important biomarker for HIV disease. The suppression of VL to undetectable
levels can improve physical functioning and reduce HIV related mortality. VL is undetectable if it has less than 200 copies of HIV per milliliter of blood by the CDC guideline. In a study of the HIV, the observations start from the time points when patients’ VL reach detectable levels. However, the exact survival time that some subjects’ VL reach detectable levels are unknown and these survival times are left-censored.

1.1.3 Interval-Censored Data

De Gruttola and Lagakos (1989) discussed the study involving a cohort of hemophiliacs for whom both infection with the HIV and the onset of the acquired immunodeficiency syndrome (AIDS) or other clinical symptoms. The data were collected from 262 patients with type A or B hemophilia in France between 1978 and 1988. Twenty-five of the patients were detected to be infected with HIV on their first lab tests. By August 1988, 197 hemophiliacs had been infected and 43 of these showed the acquisition of AIDS or other clinical symptoms such as lymphadenopathy or leukopenia. These patients were believed to become infected from the infusions of contaminated blood factor they received periodically to treat their hemophilia. Blood were periodically sampled and stored to decide a time interval during which the infections occurred. Thus, the infection times were censored within a time interval with changed status from negative to positive.

1.2 Motivating Data

Right-censored data are the most common type of censored data in oncology studies and cancer statistics. In these studies, the time-to-event is usually the survival time until death. In this dissertation, we focus on analyzing right-censored data in different survival models. The data for illustration are from the Surveillance, Epidemiology, and End Results Program (SEER) (Howlader et al., 2019). The SEER cancer data
is the research data from population-based cancer registries covering approximately 34.6 % of the United States population. The SEER registries collect data on patient demographics, primary tumor site, tumor morphology, stage at diagnosis, and first course of treatment, and they follow up with patients for vital status. The vital status is recorded as dead or alive.

Breast Cancer Data

Breast cancer is fairly common in female, representing about 15.2% of all new cancer cases in the United States. Female breast cancer is the fourth leading cause of death in the United States, and it is most frequently diagnosed among women aged 55-64. Women who are diagnosed at older age may be more likely than younger women to die of the disease, and family history increases the risk of breast cancer (Howlader et al., 2019). Regional breast cancer, which is defined as cancer cells spread to regional lymph nodes, accounts for 30% of all the cases by stage. The 5-year relative survival probability for regional breast cancer is 85.5%. For illustration, we use the SEER Louisiana regional female breast cancer for patients of age > 50 with the diagnosis year between 2000 to 2012 (Howlader et al., 2019).

A total of 6200 patients were included in this study with a mean age at diagnosis of 65.7 years old, a censoring rate of 59.5% and a maximum follow-up time of 12.9 years. 1933 patients (31.2%) were older than 70 years old, 4236 patients (68.3%) are white, and 1964 patients (31.7%) are black. There are also other races in this data set, but we focus on the comparison between the white people and the black people.

Figure 1.1 shows that there is a significant difference in survival probability between the white and the black with the log-rank test p-value= 6e-08. We further illustrate the Kaplan Meier curve stratified by each race and age (> 70 and ≤ 70) group in Figure 1.2, which indicates the survival probabilities also vary by age groups. The four subgroups including white and age > 70, black and age > 70, white and
Figure 1.1: Kaplan Meier curve by race: the dark black line represents the survival probability for the black people and the light grey line is the overall survival probability for the white people.

Figure 1.2: Stratified Kaplan Meier curve by Age 70: from light grey to dark black successively are the overall survival probabilities for white of age ≤ 70, white of age > 70, black of age ≤ 70 and black of age > 70.

age ≤ 70 and black and age ≤ 70 have significant different survival probabilities (p-value=2e-16). It is worthwhile pointing out that in this data set, the event of interest is the record of vital status of either being dead or alive, so the enrolled patients may die from causes other than breast cancer, especially for elderly patients who may die from aging. Those deaths from other causes can obscure the accurate estimation of survival probability and cure fraction, because those patients dying from other causes may also benefit from the treatment of breast cancer.
Lung Cancer Data

From the SEER 1975-2016 Research Database (Howlader et al., 2019), we also select a subset of the data for the Louisiana localized lung cancer study. This data set contains subjects aged at least 50 from the state of Louisiana whose diagnosis year is between 2004 to 2016. The response variable of interest is the time to death since diagnosis, for which right-censored data are available since not all patients died by 2016. Our data set only focuses on the data from white and black races although there are many other races in the original data set.

In total, there are 9,965 observations in this data set, and the right-censored rate is 42.76%. We consider three covariates of interest: “sex” taking 1 for female and 0 for male, “race” taking 1 for black race and 0 for white race, and “grade”, which is a measurement of how closely the tumor cells resemble lung ranging from 1 to 9 representing different degrees of differentiation.

1.3 Survival Models

For the past decades, survival analysis has been adapted for application in different areas such as biology and biomedical studies. We denote by $T$ the random variable for a subject’s survival time, which is a non-negative continuous variable. We also denote by $t$ any specific value of interest for the random variable $T$. For example, if we are interested in studying whether a patient survives for more than 5 years after undergoing cancer therapy, then $t = 5$. Specifically for right-censored data, let $Y$ be the observation time of patients and $C$ denote their censoring time. Let $\delta$ denote a random variable with possible values 0 and 1 which indicates either censoring or failure. That is, $\delta = 1$ for failure if the event of interest occurs during a study, or $\delta = 0$ if the survival time is censored by the end of a study. We have $Y = T$ if a subject’s exact survival time is observed and $Y = C$ if it is censored, i.e., $Y = \min(T, C)$ and $\delta = I(T \leq C)$. 
The cumulative distribution function of $T$ is $F(t) = P(T \leq t)$ and its corresponding probability density function is denoted as $f(t)$. The survival function of $T$ is the probability that a subject survives longer than some specified time $t$, given by

$$S(t) = P(T > t) = 1 - F(t) = \int_t^{\infty} f(s)ds, \quad 0 < t < \infty.$$ 

The survival functions are nonincreasing and head-downward as $t$ increases. At time $t = 0$, $S(t) = S(0) = 1$.

The hazard function gives the instantaneous potential per unit time for the event of interest to occur given that the subject has survived up to time $t$, which is expressed as

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}.$$ 

The relationship between survival function $S(t)$ and hazard function $\lambda(t)$ is

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d\log S(t)}{dt}$$

and

$$S(t) = \exp\{- \int_0^t \lambda(s)ds\} = \exp\{-\Lambda(t)\},$$

where $\Lambda(t)$ is the cumulative hazard function until time $t$.

Below are some popular survival models which are used to estimate the effect of potential covariates on the survival time of patients.

1.3.1 The Proportional Hazards Model

One of the most important statistical models in survival analysis is the proportional hazards (PH) model proposed by Cox (1972). The PH assumption introduces covariates into the model and specifies the hazard function in the following form,

$$h(t|x) = h_0(t) \exp(\beta^T x),$$

where $h(\cdot|x)$ is the hazard function with covariates $x$, $h_0(\cdot)$ is the baseline hazard function with $x = 0$, and $\beta$ is a vector of covariate coefficients. This model has
proportional hazards with respect of the covariates \( \mathbf{x} \) such that the hazard ratio for individuals with \( \mathbf{x}_1 \) and \( \mathbf{x}_2 \) is \( \exp \{ \mathbf{\beta}^\top (\mathbf{x}_1 - \mathbf{x}_2) \} \), which is constant over time. Therefore, the Cox PH model essentially assumes that the treatment effect is constant over time. The cumulative hazard function in the PH model is the integral of baseline hazard function, defined as \( H(t|\mathbf{x}) = \int_0^\infty h(t|\mathbf{x}) \, dt \), and the logarithm of cumulative hazard function is expressed as

\[
\log \{ H(t|\mathbf{x}) \} = \mathbf{\beta}^\top \mathbf{x} + \log \{ H_0(t) \},
\]

which shows that the curves \( \log \{ H(t|\mathbf{x}) \} \) for different values of \( \mathbf{x} \) are parallel.

For estimation in the PH model, Cox (1975) proposed the partial likelihood to estimate parameters \( \mathbf{\beta} \) without involving the baseline function \( h_0(t) \), which is the most popular estimation approach in survival analysis for right-censored data. Suppose we observe \((Y_i, \delta_i, \mathbf{x}_i)\) for individual \( i = 1, \ldots, n \), where \( Y_i \) is a observed failure time random variable, \( \delta_i \) is the failure/censoring indicator (1=failure, 0=censoring), and \( \mathbf{x}_i \) represents a vector of covariates. Assume that there are \( K \) distinct failure times and there are no ties among these event times. Let \( t_{(1)} < t_{(2)} < \cdots < t_{(K)} \) represent the \( K \) ordered distinct failure times, and \( \mathbf{x}_{(i)} \) is the covariates for the subject that has the failure at time \( t_{(i)} \). Let \( R(t) = \{ i : Y_i \geq t \} \) denote the at-risk set at time \( t \), the partial likelihood for the Cox PH model is

\[
L(\mathbf{\beta}) = \prod_{i=1}^K \frac{\exp \{ \mathbf{\beta}^\top \mathbf{x}_{(i)} \}}{\sum_{j \in R(t_{(i)})} \exp (\mathbf{\beta}^\top \mathbf{x}_j)}. \tag{1.1}
\]

The partial maximum likelihood estimator of \( \mathbf{\beta} \) is found by solving the partial likelihood score equation obtained by setting the derivative of the partial likelihood (1.1) with respect to \( \mathbf{\beta} \) to 0 in the following form

\[
\sum_{i=1}^K \mathbf{x}_{(i)} - \sum_{i=1}^K \frac{\sum_{j \in R(t_{(i)})} \mathbf{x}_j \exp (\mathbf{\beta}^\top \mathbf{x}_j)}{\sum_{j \in R(t_{(i)})} \exp (\mathbf{\beta}^\top \mathbf{x}_j)} = 0.
\]
1.3.2 THE ACCELERATED FAILURE TIME MODEL

While the Cox PH model specifies that the effect of covariates is multiplicative with respect to the hazard, the accelerated failure time (AFT) models are that the effect of covariates is multiplicative or proportional with respect to survival time. The survival time $T$ in the AFT models maintain the following relationship with covariates

$$\log T = \beta^T x + \epsilon,$$

where $x$ is a vector of covariates and $\epsilon$ are independent random errors. Different parametric distributions can be assumed for $\epsilon$ to generate different survival functions. Maximum likelihood estimation can be applied in the AFT models.

1.3.3 THE PROPORTIONAL ODDS MODEL

When the effect of covariates diminishes over time in two-sample cases, one possibility is to use time-dependent covariates in the Cox PH model. Alternatively, one may use the proportional odds (PO) model (Bennett, 1983a). The PO model assumes that the odds ratio remains constant over time. The odds ratio is the ratio of failure odds of getting the event by time $t$ for different groups, and the failure odds is denoted by $F(t)/1 - F(t)$. The PO model specifies that given covariate $x$, the odds for survival at time $t$ is

$$\frac{F(t|x)}{1 - F(t|x)} = \frac{F_0(t)}{1 - F_0(t)} \exp (\beta^T x),$$

where $F_0(t)$ is an unknown baseline cumulative distribution function. The hazard function given $x$ is

$$\lambda(t|x) = \frac{\exp (\beta^T x)}{\exp (\beta^T x) R(t) + 1} \frac{dR(t)}{dt},$$

where $R(t) = F_0(t)/(1 - F_0(t))$. Hence for two different sets of covariates $x_1$ and $x_2$, the hazard ratio $\lambda(t|x_1)/\lambda(t|x_2)$ approaches unity as time $t$ increases.
The generalized odds-rate (GOR) models was proposed by Harrington and Fleming (1982) and its survival function is expressed as

\[ S(t|x) = \{1 + \rho \Lambda_0(t) \exp(\beta^\top x)\}^{-\rho^{-1}}, \quad t > 0. \]  

In the GOR survival function, \( \Lambda_0(t) \) is a strictly positive increasing function, \( \beta \) is a \( p \times 1 \) vector of regression parameters denoting the covariate effects and \( \rho \) is a positive constant. Specifically, as \( \rho \to 0 \), equation (1.2) has a limiting survival function \( S(t|x) = \exp\{-\Lambda_0(t)\} \exp(\beta^\top x) = S_0(t) \exp(\beta^\top x) \), which is the survival function under the Cox PH model. When \( \rho = 1 \), equation (1.2) becomes the PO model survival function \( S(t|x) = \{1 + \Lambda_0(t) \exp(\beta^\top x)\}^{-1} \).

Previous studies in the GOR models are mainly focused on estimating regression parameters \( \beta \) by fixing the parameter \( \rho \) and selecting model by Akaike information criterion (AIC) or Bayesian information criterion (BIC), for example, in Cheng, Wei, and Ying (1995); Cheng, Wei, and Ying (1997); Scharfstein, Tsiatis, and Gilbert (1998); Fine, Ying, and Wei (1998); Slud and Vonta (2004); Dabrowska (2006) and Yin and Zeng (2006), among others. The regression parameters \( \beta \) and the increasing function \( \Lambda_0(t) \) can be estimated by maximum likelihood estimation.

### 1.3.5 Cure Models

Although the Cox PH model and the partial likelihood estimation approach are widely used due to its fast computation and easy interpretation. In practical circumstances, the proportional hazards assumption may not be satisfied.

With the rapid development of medication, more and more diseases and even cancers are curable in the long run such as breast cancer and colon cancer, thus, there is a proportion of patients may never experience the event of interest for a long time. The Kaplan-Meier curve will level off and reach a plateau at the end of the
study to indicate the cure. The classic cure models are used to address this issue, and two commonly adopted cure models are the mixture cure model and the promotion time cure model.

**Mixture Cure Model**

The mixture cure model was first proposed by Berkson and Gage (1952) and previous research in literature include Gray and Tsiatis (1989), Kuk and Chen (1992), Sy and Taylor (2000), Peng and Dear (2000), Betensky and Schoenfeld (2001), Lambert (2007), Cai et al. (2012), Ortega et al. (2014) among others. The mixture cure model is expressed as

\[ S_{\text{pop}}(t|x, z) = \pi(z)S(t|x) + 1 - \pi(z), \]

where \(1 - \pi(z)\) is the probability of the subjects being cured depending on the covariates \(z\), and \(S(t|x)\) is referred to as a latency survival function, which is the survival probability of the uncured patients depending on the covariates \(x\). Note that \(\pi(z)\) can take different link functions such as the logit link function \(\pi(z) = \exp(\gamma^Tz)/(1 + \exp(\gamma^Tz))\), the log-log link function \(\log\{-\log\{1 - \pi(z)\}\}\), and the probit link function \(\Phi^{-1}\{\pi(z)\} = \gamma^Tz\), where \(\Phi(\cdot)\) is the cumulative distribution function of standard normal distribution. If the PH model is used to model the latency survival function \(S(t|x)\), the mixture cure model is called the mixture cure PH model.

**Promotion time cure model**

The promotion time cure model was initially motivated by a biological model to analyze the time-to-relapse in cancer studies (Andrei and Asselain, 1996; Tsodikov, 1998; Chen, Ibrahim, and Sinha, 1999; Tsodikov, Ibrahim, and Yakovlev, 2003; Yin and Ibrahim, 2005). For \(i\)th individual with covariates \(z_i\), the survival function for individual \(i\) is given by

\[ S_{\text{pop}}(t|z_i) = \exp\{-\theta(z_i)F(t)\}, \]

\(\text{where } F(t) = \int_0^t S(t)\, dt\).
where $\theta(\cdot)$ is a link function and $F(t)$ is a latency cumulative distribution function. The corresponding latency survival function is denoted by $S(t) = 1 - F(t)$, and the latency probability density function is $f(t) = dF(t)/dt$. The cure fraction is $S_{\text{pop}}(\infty|z_i) = \exp\{\theta(z_i)\}$ when time goes to infinity, and the cure fraction is always positive. The promotion time cure model is strongly motivated by biological considerations of tumour cells. For the patient $i$, let $N_i$ be the number of tumour cells which have the potential of metastasizing (metastasis-competent tumour cells), and it is assumed that $N_i$ has a Poisson distribution with mean $\theta(z_i)$. The promotion time for the $k$th ($k = 1, \ldots, N_i$) tumour cell is denoted as $\tilde{T}_k$, which is the time for the $k$th metastasis-competent tumour cell to produce a detectable tumour mass. Assume that $\tilde{T}_k$’s are independent and identically distributed with the cumulative distribution function of $F(t)$ conditional on $N_i$. Note that both $N_i$ and $\tilde{T}_k$ are unobserved latent variables which can be explored by Expectation-Maximization (EM) algorithm. The time to relapse of cancer is defined as $T_i = \min(\tilde{T}_1, \ldots, \tilde{T}_{N_i})$. The survival function can be further written as

$$S_{\text{pop}}(t|z_i) = P(T_i > t) = P(N_i = 0) + \sum_{k \geq 1} P(\tilde{T}_1 > t, \ldots, \tilde{T}_{N_i} > t|N_i = k)P(N_i = k)$$

$$= \exp\{-\theta(z_i)\} + \sum_{k=1}^{\infty} S(t)^k \frac{\theta(z_i)^k \exp\{-\theta(z_i)\}}{k!}$$

$$= \exp\{-\theta(z_i)F(t)\}. \quad (1.5)$$

The corresponding hazard rate for the subject $i$ is $\theta(z_i)f(t)$ and the link function $\theta(z_i)$ usually takes the exponential form $\exp(\gamma^T z_i)$. To be consistent with (1.3) in the mixture cure model, (1.5) is also to allow $F(t)$ to depend on a vector of covariates $x_i$, leading to

$$S_{\text{pop}}(t|z_i, x_i) = \exp\{-\theta(z_i)F(t|x_i)\}. \quad (1.6)$$

The mixture cure model and promotion time cure model are the most widely used cure fraction models in application and each has its own advantages as well as draw-
backs as discussed in Chen, Ibrahim, and Sinha (1999) and Ibrahim, Chen, and Sinha (2001). The mixture cure model is more straightforward to interpret since it involves two parts to indicate cured and uncured patients. From the Bayesian perspective, when a uniform improper prior is set for the parameters $\gamma$, the posterior distribution for $\gamma$ is proper only in the promotion time cure model but is improper in the mixture cure model. When including the covariates $z$ through the link function for $\pi$ via a standard binomial regression model, (1.3) yields improper posterior distributions for many types of noninformative improper priors, including the uniform prior.

1.4 Existing Regression Approaches

There are many new emerging approaches proposed in the past decades for analyzing right-censored data, and the ultimate goal is to estimate the covariate effects on the survival time. Parametric models work well only when the distribution of baseline functions are correctly specified, but semiparametric regression models are more preferred in application since they would avoid the problem of misspecification and gain much modeling flexibility.

The partial likelihood function in equation has eliminated the infinite dimensional baseline hazard function from the estimation of regression parameters for right-censored data. Andersen and Gill (1982) established the asymptotic properties of the maximum partial likelihood estimator. Breslow (1972) proposed the nonparametric estimator for cumulative baseline hazard function via counting process martingale theory. This type of estimator is called nonparametric maximum likelihood estimator (NPMLE). When the proportional hazards assumption is violated, Bennett (1983b) proposed a semiparametric maximum likelihood estimator (MLE) for estimating the regression parameters in the PO model. Parzen and Harrington (1993) used adaptive splines with small number of knots to estimate baseline odds of failing by time. The
NPMLE based on the profile likelihood function was studied in Murphy, Rossini, and Vaart (1997).

For the more generalized GOR models, Cheng, Wei, and Ying (1997) proposed a semiparametric approach to estimate the regression parameters. Scharfstein, Tsiatis, and Gilbert (1998) studied the NPMLE for the regression parameters conditional on \( \rho \) being fixed and known. Chen, Jin, and Ying (2002) developed simple martingale-based estimating equations to estimate regression parameters. The theoretical properties and hypothesis tests of the NPMLE for general transformation models have been studied by Bagdonavicius and Nikulin (1999), Kosorok, Lee, and Fine (2004), Zeng and Lin (2006) and Song, Kosorok, and Fine (2009).

For the cure models estimation, parametric models and flexible parametric models usually adapt Newton-Raphson algorithm to maximize the observed likelihood function. In semiparametric cure models, EM algorithm is used to estimate parameters and cure fraction. The EM algorithm was first introduced by Dempster, Laird, and Rubin (1977) (DLR). The DLR paper makes significant contributions such that it recognizes the expectation step (E-step) and the maximization step (M-step), it gives theoretical properties of the EM algorithm and it also provides a wide range of applications in statistics. The EM algorithm has become a very popular computational method in statistics for both frequentists and Bayesian approaches. The implementation of the E-step and M-step is easy for many statistical problems and even for complex models. The M-step can be performed by existing R packages such as “optim” or “nleqslv” which makes the algorithm more computationally efficient. Moreover, the EM algorithm does not require large storage space. Louis (1982) proposed the observed information matrix for the EM algorithm, and it provides the variance estimation in closed form which makes the EM algorithm even more appealing.
The most crucial deterrent in the MLE is the potentially unlimited dimension in semiparametric regression models. The Newton-Raphson method in maximizing the full likelihood function requires \(O(d^3)\) \((d\) is the number of parameters) operations to solve a system of score equations. The principal part of the \(d\) parameters in a semiparametric regression model is to specify a stepwise function \(H\) which may approach the true continuous function as \(d\) goes to infinity. A large \(d\) will make the maximization by Newton-Raphson method has high complexity and is very difficult in computation. On the contrary, the EM algorithm handles the \(H\) function in an \(O(d)\) way.

The EM algorithm can be extended to involve a profile likelihood, such that the E-step involves computation of the expectation of full likelihood function with respect to the observed data, and the M-step is to estimate covariate effects using the profile likelihood in Johansen (1983). Let \(\theta = (\beta, H)\), where \(\beta\) is a vector parameters of interest and \(H\) is the stepwise function of a nuisance vector parameters. We denote \(\mathcal{O}\) as the observed data. The profile likelihood of \(\beta\) is the likelihood function defined as

\[
\mathcal{L}_p(\beta; \mathcal{O}) = \sup_H \mathcal{L}(\beta, H; \mathcal{O}).
\]

1.5 Structure of the Dissertation

The rest of the dissertation is structured as follows. In Chapter 2, we will study the mixture cure PH model with background mortality (MCPH+BM) and its computational estimation approach. An EM algorithm with latent variables to indicate uncured status is developed to estimate the semiparametric MCPH+BM model, and a perturbation variance estimation method is also discussed. The implementation of R functions in the “psmcure” package published in Github is further described. The results of comprehensive simulation studies and a real data application are provided to evaluate the performance of the proposed method. In Chapter 3, we propose an
EM algorithm to estimate cure fraction in the promotion time cure PH model with background mortality (PCPH+BM). The simulation studies show that the proposed semiparametric method is robust to different data distributions. A real data example is provided to illustrate the proposed method and compare the estimation to the case when background mortality is ignored. A generalization of transformation model with background mortality is further considered. In Chapter 4, we discuss the identifiability problems in the GOR models and propose a novel EM algorithm to estimate all the regression parameters and the parameter $\rho$ simultaneously.
CHAPTER 2

SEMIPARAMETRIC ESTIMATION OF THE CURE FRACTION IN POPULATION-BASED CANCER SURVIVAL ANALYSIS

ABSTRACT

With rapid development in medical research, the treatment of diseases including cancer has progressed dramatically and those survivors may die from causes other than the one under study, especially among elderly patients. Motivated by the SEER female breast cancer study, background mortality is incorporated into the mixture cure proportional hazards (MCPH) model to improve the cure fraction estimation in population-based cancer studies. Here, that patients are “cured” is defined as when the mortality rate of the individuals in diseased group returns to the same level as that expected in the general population, where the population level mortality is presented by the mortality table of the United States. The semiparametric estimation method based on the EM algorithm for the MCPH model with background mortality (MCPH+BM) is further developed and validated via comprehensive simulation studies. Real data analysis shows that the proposed semiparametric MCPH+BM model may provide more accurate estimation in population-level cancer study.
2.1 Introduction

The treatment of diseases including cancer has progressed dramatically and results in a high survival probability. The 5-year relative survival probability is as high as 98.2% for thyroid cancer, 98.0% for prostate cancer, 89.9% for female breast cancer, and 64.4% for colorectal cancer from the Surveillance, Epidemiology, and End Results (SEER) Program (2009-2015) (Howlader et al., 2019). In the long follow-up time, those survivors may die from causes other than the one under study, especially in the case of many elderly patients in a study (Verheul et al., 1993), which may obscure the estimation of survival probability of interest.

Relative survival is a net survival measurement representing cancer survival in the absence of other causes of death. It helps to evaluate heath care effectiveness for the whole population or segments of the population. Relative survival was proposed by Esteve et al. (1990) such that observed survival probability is the product of corrected survival probability and expected survival probability, and observed hazard rate is the sum of corrected hazard rate and excess hazard rate. Previous studies are mainly focused on how to model the excess hazard. Giorgi et al. (2003) developed a B-splines relative survival regression model to model hazard ratios. Dickman et al. (2004) proposed a generalized linear model to estimate additive hazards. Stare, Pohar, and Henderson (2005) addressed the goodness of fit problem of relative survival model. Pohar and Stare (2006) developed a R package “re surv” to apply several relative survival regression models. Nelson et al. (2007) proposed a model restricted cubic splines on the log cumulative excess hazard scale to estimate the relative survival and excess mortality rates. Perme, Henderson, and Stare (2008) used EM algorithm by treating the cause of death as missing data which is a generalization of the Cox PH model and provides flexibility in baseline excess hazard estimation and they further proposed a new estimator of net survival probability that enables comparability between countries.
A similar situation that patients may suffer from other events is competing risk (CR), which is an event either hinders the observation of the event of interest or alters its probability of occurrence (Gooley et al., 1999). The information from CR event is needed to derive the subdistribution function and marginal function of the event of interest. When the event of interest is death, the main difference between relative survival and competing risk rests on whether the other causes of death are known. In this paper we consider the case when death is all-cause death rather than cause-specific death, therefore the relative survival model is more appropriate. We propose the cure fraction estimation with background mortality based on the structure of relative survival.

Cure models have been employed to analyze cancer studies with potentially cured patients. In the standard cure models, “cure” is defined as patients will never experience the event of interest in the future with a probability of one (Boag, 1949). The standard mixture cure model is for the case when the specific cause of death is recorded, so death is disease-related death. In competing risk, at least two causes of death are known and included in the analysis. However, when the cause of death is unknown or not recorded, the influence of mortality due to other causes should be incorporated. De Angelis et al. (1999) proposed the new definition of “cure” in a population level, that is, the mortality risk of cured patients will return to the similar mortality risk as their counterpart in the general population, which is referred to as background mortality. The mixture cure model with background mortality which incorporates the mortality of general population is more appropriate to this type of studies than the standard mixture cure model, since it accounts for the mortality due to other causes (Lambert et al., 2006). Thus, it can improve accuracy in estimating the survival probability of uncured patients and cure fraction. The background mortality can be defined via a population-level life table with matched sex and age.
De Angelis et al. (1999) incorporated the background mortality into the mixture cure proportional hazards (MCPH) model under the exponential and Weibull distributions for uncured patients. Phillips, Coldman, and McBride (2002) extended De Angelis’s model to estimate the prevalence of cancer. Sposto (2002) described three link functions for the cure fraction estimation. Lambert et al. (2006) incorporated the background mortality to the parametric Weibull cure model via Newton-Raphson algorithm which is implemented in STATA (Lambert, 2007). They (Lambert et al., 2010) also have proposed a finite mixture of Weibull distributions to add flexibility, which also adds the complexity of estimating more Weibull parameters. Royston and Lambert (2011) discussed different topics such as time-dependent and continuous covariates in relative survival. Andersson et al. (2011) proposed to use flexible parametric survival model with cubic splines to estimate cure fraction in population-based studies, which however do not allow covariates included in cure fraction function. All these previous studies on cure model with background mortality used maximum likelihood function to estimate parameters and cure fraction.

Even though there are discussions on semiparametric estimation methods of the MCPH model (Cai et al., 2012; Peng and Dear, 2000), few work is available in the literature on the semiparametric estimation method for the MCPH model with background mortality. In addition, it is well understood that there are limitations in parametric and flexible parametric estimations, because the appropriate parametric distribution is hard to specify and the number of knots and degree of splines have to be selected. The purpose of this paper is to fill the gap in the study of semiparametric estimation method in the mixture cure model with background mortality. The rest of the paper is organized as follows. Section 2.2 illustrates the MCPH model with background mortality (MCPH+BM). The semiparametric estimation method are described in Section 2.3. Section 2.4 outlines the simulation studies and Section 2.5 applies the proposed method to the real data. Section 2.6 provides some conclusions.
2.2 Mixture Cure Model with Background Mortality

Similar to the mixture cure model, the mixture cure model with background mortality has two components:

1. patients being cured, whose mortality risk will return to the similar mortality risk level as their counterpart in the general population;

2. uncured patients, who will not only have the risk of death from disease of interest, but also suffer from the similar mortality risk as their counterpart in the general population.

Let \( t \) denote failure time, \( h^*(t) \) and \( S^*(t) \) are the population-level hazard risk which is also called background mortality rate, and the corresponding background survival probability as their counterpart in the general population with matched sex and age. Denote \( h_u(t) \) and \( S_u(t) \) as the hazard rate of the disease of interest and the corresponding survival probability. For cured patients, the hazard rate is \( h^*(t) \) with survival probability of \( S^*(t) \); for uncured patients, the hazard risk is \( h^*(t) + h_u(t) \) with survival probability of \( S^*(t)S_u(t) \). Let us assume the proportion of cured patients is indicated by \( (1 - \pi) \). The mixture cure model with background mortality is expressed as

\[
S_{\text{pop}}(t) = \pi S^*(t)S_u(t) + (1 - \pi)S^*(t).
\]

The corresponding population-level hazard function is expressed as

\[
h_{\text{pop}}(t) = h^*(t) + \frac{\pi f_u(t)}{\pi S_u(t) + 1 - \pi},
\]

where \( f_u(t) \) is the probability density function associated with the survival function \( S_u(t) \). Note, the model reduces to the standard mixture cure model when \( S^*(t) = 1 \) or \( h^*(t) = 0 \), which is the mixture cure model ignoring background mortality risk. The second component in the population-level hazard function \( \pi f_u(t)/\{\pi S_u(t) + 1 - \pi\} \)
is the corresponding hazard function for the standard mixture cure model. The hazard function of the mixture cure model with background mortality is intuitively the summation of the background hazard rate and the disease-related hazard rate.

Let $z$ denote a vector of covariates, which may impact the uncured fraction, and the model turns to the mixture cure model with background mortality (Lambert et al., 2006; Lambert, 2007) such that

$$S_{\text{pop}}(t) = S^*(t)\{\pi(z)S_u(t) + 1 - \pi(z)\}, \quad (2.1)$$

where $\pi(z)$ can be modeled via the logistic link, log-log link or probit link. Let $x$ denote a vector of covariates having potential effects on the latency survival function $S_u(t)$. It is worthwhile pointing out that the same covariates are allowed for $x$ and $z$ although we use different covariate notations, which is more flexible than existing methods. When $S_u(t)$ is specified via the proportional hazards assumption of $S_u(t|x) = S_0(t)\exp(\beta^T x)$, equation (2.1) turns to the MCPH model with background mortality (MCPH+BM) (De Angelis et al., 1999; Royston and Lambert, 2011), which is written as

$$S_{\text{pop}}(t) = S^*(t)\{\pi(z)S_0(t)\exp(\beta^T x) + 1 - \pi(z)\}.$$ 

Similarly, as $S^*(t) = 1$ or $h^*(t) = 0$, equation (2.1) reduces to the standard MCPH model.

Let $O_i = (t_i, \delta_i, z_i, x_i)$ denote the observed data for subject $i = 1, 2, \ldots, n$, where $t_i$ is the observed survival time with recorded death of any cause as the event of interest, $\delta_i$ is the censoring indicator with 1 for death and 0 for right censoring, $z_i$ is a vector of covariates which have potential effects on the uncured fraction, and $x_i$ is a vector of covariates for the survival probability of uncured patients. We assume that right censoring is independent provided that it is random and non-informative such that the distribution of survival times provides no information about the distribution of censoring times. Let $\Theta = \{\gamma, \beta, S_0(\cdot), h_0(\cdot)\}$ denote the unknown parameters,
where $\gamma$ is the vector of unknown parameters in $\pi(z)$, $\beta$ are the coefficients in $S_u(t_i)$, $S_0(\cdot)$ and $h_0(\cdot)$ are the baseline survival and baseline hazard functions in $S_u(t)$, so the observed likelihood function in the MCPH+BM model is

$$L_{\text{obs}}(\Theta; O) = \prod_{i=1}^{n} \left[ \pi(z_i) S_0(t_i) \exp(\beta^T x_i) \{ h^*(t_i) + h_0(t_i) \exp(\beta^T x_i) \} + \{1 - \pi(z_i)\} h^*(t_i) \right]^{\delta_i} \times S^*(t_i) \left[ \pi(z_i) S_0(t_i) \exp(\beta^T x_i) + \{1 - \pi(z_i)\} \right]^{1-\delta_i}.$$

Note that both $h^*(t_i)$ and $S^*(t_i)$ are assumed to be sex and age matched constant values for the subject $i$. In this paper they are obtained by matching sex and age for each individual from a life table (Coleman et al., 1999), so both $h^*(t_i)$ and $S^*(t_i)$ are not affected by the modeled parameters. For application, the overall background mortality can also be used. Once $S_0(t_i)$ is fully specified as in a parametric structure, it can be estimated directly via MLE algorithm (Lambert et al., 2006) or EM algorithm. However, in practice it is often hard to find an appropriate parametric structure or such parametric structures do not even exist. In these situations, the proposed semiparametric approach in this paper can be applied. We will discuss the details of our estimation approach for the MCPH+BM model in the following section.

2.3 EM Algorithm

Similar to in the standard MCPH model (Sposto, 2002), we use a logistic link function for the uncured fraction in the MCPH+BM model. Other link functions such as log-log link and probit link can also be applied. Let $y$ be the latent uncured indicator, with $y = 1$ indicating uncured patients and $y = 0$ indicating cured patients. The uncured probability under the logistic link function is expressed as

$$P(y = 1|z) = \pi(z) = \frac{\exp(\gamma^T z)}{1 + \exp(\gamma^T z)}.$$

Note that $y$ is partially missing in the standard MCPH model because when a patient is cured with $y_i = 0$, this person only can be censored ($\delta_i = 0$) and no
longer suffers from the disease under study. However, \( y \) is completely missing in the MCPH+BM model, because patients die from other causes may also benefit from the treatment of disease and has the possibility to be cured. There are four components contributing to the complete likelihood function, \( \pi(z_i) S^*(t_i) S_0(t_i) \exp(\beta^T x_i) \{h^*(t_i) + h_0(t_i) \exp(\beta^T x_i)\} \) which indicates that a patient is uncured and dead from the disease under study, \( \{1 - \pi(z_i)\} S^*(t_i) \) which is a quantity for a cured censored patient, \( \pi(z_i) S^*(t_i) S_0(t_i) \exp(\beta^T x_i) \) which is a quantity for a uncured censored patient, and \( \{1 - \pi(z_i)\} h^*(t_i) S^*(t_i) \) which indicates a patient is cured of the disease under study but dead from background mortality.

The complete likelihood function is then written as

\[
\mathcal{L}_c(\Theta; \mathbf{O}, \mathbf{Y}) = \prod_{i=1}^{n} \left[ \{1 - \pi(z_i)\} h^*(t_i) \delta_i S^*(t_i) \right]^{(1-y_i)} \left[ \pi(z_i) S^*(t_i) S_0(t_i) \exp(\beta^T x_i) \{h^*(t_i) + h_0(t_i) \exp(\beta^T x_i)\} \delta_i \right]^{y_i},
\]

and the logarithm of the complete likelihood function is

\[
\ell_c(\Theta; \mathbf{O}, \mathbf{Y}) = \sum_{i=1}^{n} (1 - y_i) \log \{1 - \pi(z_i)\} + y_i \log \{\pi(z_i)\} + y_i \log \{S_0(t_i) \exp(\beta^T x_i)\} \\
+ y_i \log \{S^*(t_i)\} + \delta_i y_i \log \{h^*(t_i) + h_0(t_i) \exp(\beta^T x_i)\} \\
+ (1 - y_i) \log \{h^*(t_i) \delta_i S^*(t_i)\}.
\] (2.2)

The main components associated with unknown parameters \( \gamma \) and \( \beta \) in equation (2.2) can be expressed as the sum of two parts \( \ell_{c_1}(\gamma; z, \mathbf{Y}) + \ell_{c_2}\{\beta, S_0(\cdot), h_0(\cdot); t, \delta, x, \mathbf{y}\} \) with

\[
\ell_{c_1}(\gamma; z, \mathbf{Y}) = \sum_{i=1}^{n} y_i \log \{\pi(z_i)\} + (1 - y_i) \log \{1 - \pi(z_i)\}
\] (2.3)

\[
\ell_{c_2}\{\beta, S_0(\cdot), h_0(\cdot); t, \delta, x, \mathbf{y}\} = \sum_{i=1}^{n} y_i \log \{S_0(t_i) \exp(\beta^T x_i)\} \\
+ \delta_i y_i \log \{h^*(t_i) + h_0(t_i) \exp(\beta^T x_i)\}
\] (2.4)

where \( \mathbf{y} = (y_1, y_2, \ldots, y_n) \) are the vector of uncured indicators.
In the EM algorithm, the E-step is to take conditional expectations of equation (2.3) and (2.4) with respect to \( y_i \)’s given the observed data \( \mathbf{O} \) and parameters \( \Theta \). For the subject \( i \), \( y_i \) is either 1 or 0 no matter what value \( \delta_i \) is. Thus, \( y_i|\delta_i = 1 \) follows a Bernoulli distribution with the probability of success

\[
\pi(z_i)S^*(t_i)S_0(t_i)^{\exp(\beta^T x_i)} \left\{ h^*(t_i) + h_0(t_i) \exp(\beta^T x_i) \right\} \\
\{1 - \pi(z_i)\} h^*(t_i)S^*(t_i) + \pi(z_i)S^*(t_i)S_0(t_i)^{\exp(\beta^T x_i)} \left\{ h^*(t_i) + h_0(t_i) \exp(\beta^T x_i) \right\}
\]

and \( y_i|\delta_i = 0 \) follows a Bernoulli distribution with the probability of success

\[
\pi(z_i)S^*(t_i)S_0(t_i)^{\exp(\beta^T x_i)} \\
\{1 - \pi(z_i)\} S^*(t_i) + \pi(z_i)S^*(t_i)S_0(t_i)^{\exp(\beta^T x_i)}.
\]

Let \( \omega_i \) denote the conditional expectation of \( y_i \) given the observed data \( \mathbf{O} \) and parameters \( \Theta \), one has

\[
\omega_i = \mathbb{E}(y_i|\Theta, \mathbf{O}) = \delta_i \frac{\pi(z_i)S^*(t_i)S_0(t_i)^{\exp(\beta^T x_i)} \left\{ h^*(t_i) + h_0(t_i) \exp(\beta^T x_i) \right\} }{\{1 - \pi(z_i)\} h^*(t_i)S^*(t_i) + \pi(z_i)S^*(t_i)S_0(t_i)^{\exp(\beta^T x_i)} \left\{ h^*(t_i) + h_0(t_i) \exp(\beta^T x_i) \right\} } + (1 - \delta_i) \frac{\pi(z_i)S^*(t_i)S_0(t_i)^{\exp(\beta^T x_i)} }{\{1 - \pi(z_i)\} S^*(t_i) + \pi(z_i)S^*(t_i)S_0(t_i)^{\exp(\beta^T x_i)}}.
\]

The conditional expectations of equation (2.3) and (2.4) with respect to \( \omega_i \)’s given the observed data \( \mathbf{O} \) and parameters \( \Theta \) are

\[
\mathbb{E}\{\ell_{c_1} (\mathbf{y}; \mathbf{z}, \omega)\} = \sum_{i=1}^{n} \omega_i \log \{\pi(z_i)\} + (1 - \omega_i) \log \{1 - \pi(z_i)\} \tag{2.6}
\]

and

\[
\mathbb{E}\left[\ell_{c_2} (\mathbf{y}, S_0(\cdot), h_0(\cdot); \mathbf{t}, \delta, \mathbf{x}, \omega)\right] = \sum_{i=1}^{n} \omega_i \log \{S_0(t_i)^{\exp(\beta^T x_i)}\} \\
+ \delta_i \omega_i \log \{h^*(t_i) + h_0(t_i) \exp(\beta^T x_i)\}. \tag{2.7}
\]

The M-step in the EM algorithm is to maximize equation (2.6) and (2.7) via Newton-Raphson method using “optim” function in R. Note that equation (2.6) only contains the unknown parameters \( \gamma \) and equation (2.7) only involves the unknown parameters \( \beta, S_0(\cdot) \) and \( h_0(\cdot) \). Thus, the expectation of the logarithm complete
likelihood can be maximized by equations (2.6) and (2.7) separately, which improves computation efficiency. The background hazard rate $h^\ast(t)$ and background survival probability $S^\ast(t)$ are sex and age matched constant pieces in the calculation of EM algorithm.

Let $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$ be the distinct ordered uncensored failure times and $R\{t_{(j)}\}$ be the risk set at the time point $t_{(j)}$. Assume the hazard rate only jumps at these event times. Given $\beta$ and $\omega_i$, we take the first partial derivative of equation (2.7) with respect to $h_0(t_i; \beta, \omega_i)$, which is $\partial E\left[\ell_{c2}\{\beta, S_0(\cdot), h_0(\cdot); t, \delta, x, \omega}\right]/\partial h_0(t_i; \beta, \omega_i) = 0$. Then equation (2.7) can be maximized at the following baseline hazard function (Peng and Dear, 2000; Breslow, 1972)

$$h_0(t_i; \beta, \omega_i) = \frac{\delta_i \omega_i}{\sum_{l \in R(t_i)} \omega_l \exp(\beta^\top x_l)} - \frac{h^\ast(t_i)}{\exp(\beta^\top x_i)},$$

and the baseline survival function

$$S_0(t_i; \beta, \omega_i) = \exp\left(-\sum_{j: t_{(j)} \leq t_i} \left[\frac{\delta_{(j)} \omega_{(j)}}{\sum_{l \in R(t_{(j)})} \omega_l \exp(\beta^\top x_l)} - \frac{h^\ast\{t_{(j)}\}}{\exp(\beta^\top x_{(j)})}\right]\right).$$

Given estimated baseline functions in equation (2.8) and (2.9), $\beta$ can be updated by maximizing equation (2.7).

**Remark:** The equation (2.6) is exactly the same as the first expectation of the log-logarithm complete likelihood function in the standard MCPH model (Peng and Dear, 2000; Sy and Taylor, 2000), but the equation (2.7) involves the background hazard risk, which is different from the formula in the standard MCPH model. The complexity in the equation (2.7) makes the calculation of $\omega_i$ in the equation (2.5) more complex. Specifically, in the standard MCPH model, $\omega_i = 1$ when $\delta_i = 1$, since a dead patient is uncured. However, in the proposed MCPH+BM model when $\delta_i = 1$, a patient not only may die from the disease under study, but also may be cured of the disease under study and die from background mortality. Moreover, in the standard MCPH model, the baseline hazard function only takes the first quantity in the formula (2.8), but in the MCPH+BM model, patients also suffer from the background
hazard risk which is the second quantity in the formula (2.8). Thus, the corresponding baseline survival function becomes more complex as shown in (2.9).

Previous studies have discussed the identifiability issue in the standard mixture cure model. Li, Taylor, and Sy (2001) explored the mixture cure model with or without covariates in the cure fraction and the latency distribution, and theoretically showed that the mixture cure model is identifiable except for the case when the latency distribution is independent of covariates. Specifically, the mixture cure model is not identifiable when the cure fraction is not modeled by covariates. Yu et al. (2004) studied the identifiability of the mixture cure model for the grouped survival data and found the cure fraction estimates can be sensitive to the latency survival distributions. If the largest survival time is censored, the estimated survival probability at the time between the largest uncensored time and the largest time is equal to the survival probability at the largest uncensored time, which leads to an improper distribution. Consequently in the estimation of nonparametric baseline function, the model is overparameterized such that the intercept term in the logistic predictor and the improper distribution of uncured patients are not identifiable. Sy and Taylor (2000) and Peng and Dear (2000) apply the zero tail constrains in the standard MCPH model. Similarly, we impose the Taylor tail completion (zero-tail constraint) (Taylor, 1995) on the total baseline survival probability expressed as $S_0^*(t)S_0(t; \beta, \omega)$, if there are censored survival times greater than the last survival time, and the total baseline cumulative hazard function beyond the last survival time is infinity.

The EM algorithm for estimating parameters $\Theta = \{\gamma, \beta, S_0(\cdot), h_0(\cdot)\}$ is summarized as follows,

Step 0: Set initial values $\omega_i^0$ to the subject censoring indicator $\delta_i$, $(i = 1, 2, \cdots, n)$, initial values $\gamma^0$ are estimated from the logistic regression with $\omega^0$ as response variable and $z$ as covariates, and initial values $\beta^0$ are estimated from the Cox proportional hazards model with $x$ as covariates;
Step 1: In \((k+1)\) iteration, update \(\gamma^{(k+1)}\) via maximizing (2.6) using “optim” function;

Step 2: Update \(\beta^{(k+1)}\) via maximizing equation (2.7) using “optim”, where baseline hazard and baseline survival functions are replaced by (2.8) and (2.9) in equation (2.7);

Step 3: Update \(h_0^{(k+1)}\{t_i; \beta^{(k+1)}, \omega_i^{(k)}\}\) and \(S_0^{(k+1)}\{t_i; \beta^{(k+1)}, \omega_i^{(k)}\}\) using equation (2.8) and (2.9); then \(\omega_i^{(k+1)}\) is updated by equation (2.5);

Step 4: Iterate the Step 1-3 until convergence, when the differences of the updated values in two successive iterations are less than 1e-7.

For variance estimation, let \(\hat{\Theta} = \{\hat{\gamma}, \hat{\beta}, \hat{S}_0(\cdot), \hat{h}_0(\cdot)\}\) denote the converged values from the EM algorithm, and \(\hat{\omega}_i\) be the last iteration updated value of \(\omega_i\) for the subject \(i\). In the equation (2.6) and (2.7), the expectation of individual logarithm likelihood functions for the subject \(i\) are \(E\{\ell_{c_1,i}(\gamma; z_i, \omega_i)\} = \omega_i \log \{\pi(z_i)\} + (1-\omega_i) \log \{1-\pi(z_i)\}\) and \(E\{\ell_{c_2,i}(\beta, S_0(t_i), h_0(t_i); \delta_i, x_i, \omega_i)\} = \omega_i \log \{S^*(t_i)S_0(t_i)\exp(\beta^\top x_i)\} + \delta_i \omega_i \log \{h^*(t_i) + h_0(t_i)\exp(\beta^\top x_i)\} + \delta_i (1-\omega_i) \log \{h^*(t_i)\} + (1-\omega_i) \log \{S^*(t_i)\}\). Let \(D_i\) be the first derivative of these two expectation logarithm likelihood functions with respect to \(\gamma\) and \(\beta\),

\[
D_i = \left( \frac{\partial E\{\ell_{c_1,i}(\gamma; z_i, \tilde{\omega}_i)\}}{\partial \gamma}, \frac{\partial E\{\ell_{c_2,i}(\beta, S_0(t_i), h_0(t_i); \delta_i, x_i, \tilde{\omega}_i)\}}{\partial \beta} \right)_{(\hat{\beta}, \hat{S}_0(t_i), \hat{h}_0(t_i))}.
\]

The variances of \(\hat{\gamma}\) and \(\hat{\beta}\) can be obtained by inverting the empirical Fisher information matrix (Murphy and Vaart, 2000) \(\hat{I} = \sum_{i=1}^n D_iD'_i\). The first component of \(D_i\) is easily obtained because it has a closed form. The second component of \(D_i\) does not have a closed form, but it can be approximated by numerical differentiation methods such as a first-order Richardson extrapolation of the central difference (Jamshidian and Jennrich, 2000). Thus, the approximation of the second component of \(D_i\) for the
The jth coefficient in $\beta$ is expressed as
\[
\frac{\partial E\{\ell_{c2,i}(\beta, S_0(t_i), h_0(t_i))\}}{\partial \beta_j} \approx \frac{1}{12d} \left( E\left[\ell_{c2,i}(\beta - 2d v^{(j)}, S_0(t_i), h_0(t_i))\right] - 8E\left[\ell_{c2,i}(\beta - dv^{(j)}, S_0(t_i), h_0(t_i))\right] + 8E\left[\ell_{c2,i}(\beta + dv^{(j)}, S_0(t_i), h_0(t_i))\right] - E\left[\ell_{c2,i}(\beta + 2dv^{(j)}, S_0(t_i), h_0(t_i))\right]\right)_{\{\hat{\beta}, \hat{S}_0(t_i), \hat{h}_0(t_i)\}}
\]

where $d$ is a small positive value, $v^{(j)}$ is a vector with its jth component equals to 1 and all others equal to 0.

2.4 Simulation Studies

Simulation studies are conducted to evaluate the performance of the proposed semi-parametric estimation approach in the MCPH+BM model. We compare the performance of our proposed model to a flexible parametric method and fully parametric methods. The flexible parametric method used for comparison is to model the baseline hazard function by monotone splines with equally spaced knots at percentiles (Ramsay, 1988) via the EM algorithm in Section 2.3. The degree of splines is set to 3 and the number of knots is selected by Akaike’s information criterion (AIC). The fully parametric method is to model the baseline hazard rate and baseline survival probability with the corresponding parametric distributions of Weibull, Lognormal and Loglogistic via the EM algorithm. We use EM algorithm for fully parametric method in simulation studies instead of MLE algorithm since it is more reasonable to be compared under similar algorithm setting.

Note, 1) the existing flexible parametric methods for cure fraction estimation in population level such as the method proposed by Andersson et al. (2011) can be implemented by “rstm2” or “flexsurvcure” functions in R package, and “stpm2” command in STATA, but those models and packages do not have the same data
setting as the MCPH+BM model has, specifically the cure fraction cannot accommodate covariates; 2) With a simulation comparison, the MLE parametric algorithm with the implementation of Nelder and Mead method requires large number of iterations (5000) and small convergence tolerance (1e-20) to get similar simulation results as in the EM parametric algorithm (tables are not shown).

We generate the covariates $z$ from a Bernoulli distribution with probability 0.5, with 0 and 1 as the group indicator. Setting $\gamma$ to $(\log(2), -1)$ gives the cure rate about 33% and 58% for two groups for illustration. Uncured indicators are generated from a Bernoulli distribution with the probability of uncured fraction. The covariate of “age” is generated from a normal distribution with mean of 70 and standard deviation of 5, and “sex” from a Bernoulli distribution with the probability of 0.5.

For a patient being cured, the survival time is generated from the background mortality distribution. The background mortality is matched based on the estimated Weibull distribution approximated using “2015 USA Life Table” (Arias and Xu, 2018). The “2015 USA Life Table” has survival probability for male and female at the age of 0 to 110. For each sex and age combination, we approximate the life table via a Weibull distribution and those estimated parameters of Weibull distribution are used as the background mortality parameters. Then we match the sex and age in the data set with those in the background mortality table, and use the parameters to generate survival times from the background mortality.

For the uncured patients who may die from the disease under study or some other causes, the survival time is the minimum value generated from the background mortality distribution and the latency proportional hazards model. In the latency proportional hazards model, we include the covariates of group and sex, and set their coefficients $\beta$ to $(1, -1)$. Several baseline survival functions are considered in the proportional hazards model to simulate data set: 1) Weibull distribution with shape = 0.75 and scale = 1; 2) Lognormal distribution with mean = 0 and standard
deviation = 1; and 3) Loglogistic distribution with shape = 2 and scale = 1. Finally, the censoring time is generated from an exponential distribution \( \exp(c) \) where the parameter \( c \) controls the censoring rate at about 50%.

We consider a small (n=200), medium (n=400) and large (n=800) sample size with 500 replicates for each simulation setup. A sensitivity analysis of 800 sample size is further conducted to illustrate the misspecification in the fully parametric method through the performance of the proposed semiparametric approach in the MCPH+BM model and the performance of fully parametric method with correctly specified and misspecified distributions.

The simulation results including bias, average estimated standard error (StErr), empirical standard deviation (StDev), 95% coverage probability (CP) and also average running time per replicate are reported in Table 2.1-2.5. Specifically, biases are calculated as the average of differences between point estimators and their true values. StErrs are the average estimated standard errors obtained by the perturbation method with \( d = 0.1 \) in equation (2.10) for the proposed semiparametric approach, and the Hessian matrix for the flexible parametric and fully parametric methods. StDevs are the empirical standard deviations of point estimators from the 500 replicates and CPs are the average coverage probabilities of the 95% confidence intervals. The cure fraction estimation for two groups are also reported.

The results of simulation studies (Table 2.1-2.3) show that the proposed semiparametric method performs well in estimating all the parameters and cure fraction in the MCPH+BM model. The biases for all estimated parameters are small and the corresponding CPs are close to the nominal 95%, also comparable to the flexible parametric method when the best number of knots is selected according to AIC, and the fully parametric methods when correctly specifying the data distributions. Note that the performance of flexible parametric method via EM algorithm is good when the best number of knots is selected by AIC, but the computation speed is about...
Table 2.1: Summary Statistics for Weibull Data with Logistic Link: by semiparametric method, flexible parametric method with monotone splines, parametric Weibull baseline function.

<table>
<thead>
<tr>
<th></th>
<th>Semiparametric</th>
<th>Flexible Parametric</th>
<th>Parametric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Par Bias StErr StDev CP</td>
<td>Bias StErr StDev CP</td>
<td>Bias StErr StDev CP</td>
</tr>
<tr>
<td>n=200</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma_0$</td>
<td>0.073 0.247 0.266 0.938</td>
<td>0.019 0.460 0.394 0.968</td>
<td>0.047 0.293 0.306 0.950</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>-0.058 0.334 0.331 0.942</td>
<td>-0.135 0.520 0.457 0.946</td>
<td>-0.037 0.364 0.396 0.932</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.036 0.244 0.290 0.918</td>
<td>0.087 0.295 0.278 0.938</td>
<td>0.047 0.274 0.289 0.946</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.039 0.267 0.274 0.934</td>
<td>-0.089 0.280 0.265 0.954</td>
<td>-0.026 0.268 0.258 0.966</td>
</tr>
<tr>
<td>cure-1</td>
<td>-0.004 0.049 0.051 0.940</td>
<td>0.005 0.054 0.056 0.966</td>
<td>-0.003 0.051 0.054 0.950</td>
</tr>
<tr>
<td>cure-2</td>
<td>-0.027 0.052 0.054 0.942</td>
<td>-0.017 0.075 0.073 0.952</td>
<td>-0.007 0.066 0.065 0.940</td>
</tr>
<tr>
<td></td>
<td>running time: 1.433</td>
<td>running time: 17.891</td>
<td>running time: 0.726</td>
</tr>
<tr>
<td>n=400</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma_0$</td>
<td>0.057 0.170 0.178 0.932</td>
<td>0.041 0.233 0.233 0.958</td>
<td>-0.004 0.201 0.198 0.956</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>-0.046 0.235 0.231 0.941</td>
<td>-0.058 0.277 0.279 0.960</td>
<td>-0.024 0.252 0.243 0.948</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.046 0.160 0.170 0.926</td>
<td>0.053 0.192 0.191 0.946</td>
<td>-0.004 0.187 0.187 0.946</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.027 0.166 0.180 0.930</td>
<td>-0.042 0.187 0.190 0.940</td>
<td>-0.006 0.184 0.188 0.942</td>
</tr>
<tr>
<td>cure-1</td>
<td>0.004 0.036 0.036 0.932</td>
<td>0.004 0.036 0.038 0.956</td>
<td>-0.005 0.036 0.035 0.958</td>
</tr>
<tr>
<td>cure-2</td>
<td>-0.013 0.036 0.038 0.936</td>
<td>-0.007 0.050 0.050 0.958</td>
<td>0.002 0.043 0.044 0.952</td>
</tr>
<tr>
<td></td>
<td>running time: 10.646</td>
<td>running time: 36.340</td>
<td>running time: 1.903</td>
</tr>
<tr>
<td>n=800</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma_0$</td>
<td>0.011 0.128 0.124 0.952</td>
<td>0.009 0.156 0.147 0.958</td>
<td>0.014 0.142 0.137 0.958</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>-0.021 0.165 0.161 0.936</td>
<td>-0.011 0.188 0.182 0.946</td>
<td>-0.021 0.177 0.167 0.972</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.025 0.121 0.124 0.934</td>
<td>0.038 0.131 0.140 0.924</td>
<td>0.008 0.130 0.125 0.956</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.024 0.119 0.124 0.926</td>
<td>-0.045 0.129 0.128 0.944</td>
<td>-0.007 0.128 0.126 0.954</td>
</tr>
<tr>
<td>cure-1</td>
<td>0.004 0.025 0.026 0.952</td>
<td>0.000 0.027 0.027 0.956</td>
<td>0.001 0.025 0.025 0.958</td>
</tr>
<tr>
<td>cure-2</td>
<td>-0.002 0.029 0.027 0.948</td>
<td>-0.001 0.034 0.032 0.950</td>
<td>-0.003 0.031 0.030 0.962</td>
</tr>
<tr>
<td></td>
<td>running time: 12.021</td>
<td>running time: 114.612</td>
<td>running time: 2.990</td>
</tr>
</tbody>
</table>

10 times of that in the proposed semiparametric approach. The standard errors for the estimated parameters are similar in magnitude for different distributions, and become smaller when sample size increases from 200 to 400 and then to 800. Thus the accuracy of parameters estimation gets better with the increase of sample size.

The perturbation method to estimate the variance of estimators works well to provide desirable standard errors and coverage probabilities, and is also efficient in computing. We also did additional simulation to compare the perturbation method and bootstrap method for variance estimation (tables are not shown). The results from bootstrap method are similar to what we obtained from the perturbation method, but the perturbation method is more computational efficient. In each replicate, it only takes an average of 0.04 seconds for 200 sample size, 0.10 seconds for 400 sample size and 0.20 seconds for 800 sample size for variance estimation. On the contrary, the
Table 2.2: Summary Statistics for Lognormal Data with Logistic Link: by semi-parametric method, flexible parametric method with monotone splines, parametric Lognormal baseline function.

<table>
<thead>
<tr>
<th></th>
<th>Semiparametric</th>
<th>Flexible Parametric</th>
<th>Parametric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Par Bias StErr StDev CP</td>
<td>Bias StErr StDev CP</td>
<td>Bias StErr StDev CP</td>
</tr>
<tr>
<td>n=200</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma_0$</td>
<td>0.063 0.250 0.273 0.940</td>
<td>0.132 0.562 0.583 0.968</td>
<td>0.010 0.317 0.330 0.952</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>-0.080 0.334 0.330 0.948</td>
<td>-0.148 0.619 0.604 0.962</td>
<td>-0.007 0.388 0.416 0.942</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.020 0.270 0.287 0.936</td>
<td>0.104 0.302 0.313 0.929</td>
<td>0.023 0.290 0.294 0.934</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.025 0.262 0.281 0.930</td>
<td>-0.094 0.283 0.292 0.941</td>
<td>-0.020 0.276 0.297 0.934</td>
</tr>
<tr>
<td>cure-1</td>
<td>-0.004 0.050 0.050 0.942</td>
<td>0.003 0.054 0.055 0.966</td>
<td>-0.002 0.056 0.057 0.954</td>
</tr>
<tr>
<td>cure-2</td>
<td>-0.017 0.055 0.057 0.942</td>
<td>-0.016 0.081 0.082 0.960</td>
<td>0.002 0.067 0.069 0.948</td>
</tr>
<tr>
<td>running time:</td>
<td>1.909</td>
<td>running time: 14.453</td>
<td>running time: 0.795</td>
</tr>
<tr>
<td>n=400</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma_0$</td>
<td>0.041 0.174 0.199 0.934</td>
<td>0.013 0.240 0.239 0.956</td>
<td>0.004 0.217 0.219 0.954</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>0.036 0.249 0.229 0.962</td>
<td>-0.008 0.284 0.273 0.974</td>
<td>-0.001 0.267 0.265 0.950</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.024 0.185 0.195 0.926</td>
<td>0.065 0.194 0.202 0.924</td>
<td>0.011 0.198 0.191 0.960</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.042 0.193 0.195 0.932</td>
<td>-0.067 0.189 0.197 0.932</td>
<td>-0.019 0.190 0.200 0.938</td>
</tr>
<tr>
<td>cure-1</td>
<td>0.006 0.037 0.036 0.936</td>
<td>0.001 0.036 0.036 0.954</td>
<td>-0.001 0.039 0.040 0.954</td>
</tr>
<tr>
<td>cure-2</td>
<td>-0.007 0.043 0.043 0.938</td>
<td>0.005 0.051 0.052 0.948</td>
<td>0.001 0.048 0.048 0.950</td>
</tr>
<tr>
<td>running time:</td>
<td>4.579</td>
<td>running time: 32.318</td>
<td>running time: 0.933</td>
</tr>
<tr>
<td>n=800</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma_0$</td>
<td>0.023 0.125 0.123 0.944</td>
<td>0.019 0.158 0.151 0.941</td>
<td>0.002 0.153 0.158 0.938</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>0.017 0.167 0.165 0.922</td>
<td>-0.005 0.191 0.187 0.959</td>
<td>-0.003 0.187 0.187 0.950</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.025 0.135 0.134 0.952</td>
<td>0.011 0.130 0.142 0.919</td>
<td>0.010 0.138 0.141 0.944</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.016 0.130 0.133 0.936</td>
<td>-0.024 0.139 0.141 0.922</td>
<td>-0.012 0.133 0.133 0.944</td>
</tr>
<tr>
<td>cure-1</td>
<td>0.009 0.027 0.027 0.944</td>
<td>0.003 0.027 0.027 0.940</td>
<td>-0.000 0.026 0.026 0.938</td>
</tr>
<tr>
<td>cure-2</td>
<td>0.006 0.028 0.028 0.940</td>
<td>0.012 0.033 0.034 0.946</td>
<td>0.000 0.035 0.035 0.942</td>
</tr>
<tr>
<td>running time:</td>
<td>15.513</td>
<td>running time: 97.027</td>
<td>running time: 1.741</td>
</tr>
</tbody>
</table>

Bootstrap method with 100 resamples to estimate variance takes about 1.5 seconds for a small sample size of 200, which is 37 times slower.

Figures 2.1-2.9 display the average baseline survival probability estimation with their empirical 95% confidence intervals for different data distributions. The estimated baseline survival probability is close to the true baseline survival function and the 95% confidence interval becomes narrower with the increase of sample size, which indicates the proposed method provides a good estimation of the baseline function under the proposed semiparametric MCPH+BM model. In addition, Figure 2.10-2.12 are the predicted net survival probabilities for two groups when incorporating the covariate of group in both cure function and latency survival function which indicates a long term plateau.

The sensitivity analysis in Table 2.4 has the same data and parameters settings as those described above with n=800, which shows that the proposed semiparametric
Table 2.3: Summary Statistics for Loglogistic Data with Logistic Link: by semi-parametric method, flexible parametric method with monotone splines, parametric Loglogistic baseline function.

<table>
<thead>
<tr>
<th></th>
<th>Semiparametric</th>
<th>Flexible Parametric</th>
<th>Parametric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Par Bias StErr StDev CP</td>
<td>Par Bias StErr StDev CP</td>
<td>Par Bias StErr StDev CP</td>
</tr>
<tr>
<td>n=200</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma_0$</td>
<td>0.070 0.325 0.313 0.934</td>
<td>0.010 0.394 0.376 0.968</td>
<td>0.025 0.303 0.319 0.954</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>-0.049 0.333 0.371 0.934</td>
<td>-0.029 0.456 0.443 0.952</td>
<td>-0.006 0.374 0.380 0.946</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.007 0.290 0.287 0.928</td>
<td>0.079 0.295 0.305 0.952</td>
<td>0.018 0.287 0.302 0.940</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.003 0.277 0.281 0.944</td>
<td>-0.084 0.285 0.297 0.926</td>
<td>-0.023 0.271 0.272 0.956</td>
</tr>
<tr>
<td>cure-1</td>
<td>-0.006 0.050 0.051 0.934</td>
<td>0.004 0.054 0.056 0.966</td>
<td>-0.005 0.053 0.053 0.954</td>
</tr>
<tr>
<td>cure-2</td>
<td>-0.012 0.058 0.050 0.932</td>
<td>0.003 0.075 0.074 0.952</td>
<td>-0.002 0.068 0.068 0.948</td>
</tr>
<tr>
<td>running time:</td>
<td>1.356</td>
<td>running time: 33.389</td>
<td>running time: 0.709</td>
</tr>
<tr>
<td>n=400</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma_0$</td>
<td>0.030 0.188 0.185 0.944</td>
<td>0.080 0.224 0.220 0.968</td>
<td>0.016 0.211 0.211 0.958</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>-0.057 0.229 0.234 0.946</td>
<td>-0.066 0.271 0.267 0.938</td>
<td>-0.012 0.261 0.249 0.962</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.056 0.166 0.188 0.918</td>
<td>0.031 0.193 0.191 0.950</td>
<td>0.026 0.196 0.193 0.938</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.016 0.196 0.192 0.940</td>
<td>-0.030 0.190 0.205 0.930</td>
<td>-0.008 0.187 0.183 0.948</td>
</tr>
<tr>
<td>cure-1</td>
<td>0.006 0.034 0.034 0.944</td>
<td>0.003 0.037 0.038 0.966</td>
<td>-0.001 0.038 0.038 0.960</td>
</tr>
<tr>
<td>cure-2</td>
<td>-0.005 0.041 0.040 0.940</td>
<td>0.020 0.049 0.050 0.940</td>
<td>-0.002 0.046 0.046 0.956</td>
</tr>
<tr>
<td>running time:</td>
<td>3.865</td>
<td>running time: 106.963</td>
<td>running time: 1.198</td>
</tr>
<tr>
<td>n=800</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma_0$</td>
<td>0.030 0.131 0.129 0.934</td>
<td>0.024 0.147 0.150 0.940</td>
<td>0.009 0.148 0.144 0.966</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>-0.032 0.161 0.171 0.929</td>
<td>-0.010 0.182 0.178 0.964</td>
<td>-0.019 0.183 0.168 0.976</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.023 0.128 0.125 0.940</td>
<td>0.009 0.129 0.137 0.930</td>
<td>0.013 0.135 0.142 0.928</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.002 0.154 0.146 0.958</td>
<td>-0.020 0.129 0.130 0.948</td>
<td>-0.007 0.131 0.130 0.948</td>
</tr>
<tr>
<td>cure-1</td>
<td>0.007 0.025 0.025 0.934</td>
<td>0.003 0.025 0.026 0.942</td>
<td>0.002 0.025 0.025 0.964</td>
</tr>
<tr>
<td>cure-2</td>
<td>0.010 0.028 0.029 0.932</td>
<td>0.029 0.034 0.034 0.944</td>
<td>-0.001 0.031 0.032 0.964</td>
</tr>
<tr>
<td>running time:</td>
<td>12.647</td>
<td>running time: 137.027</td>
<td>running time: 3.295</td>
</tr>
</tbody>
</table>

Figure 2.1: Baseline Survival Probability Estimation for Weibull Data by Semiparametric MCPH+BM Model (200 Sample Size).

method performs better than fully parametric methods. The columns in Table 2.4 denote the true data distributions (Weibull, Lognormal and Loglogistic) while rows
Table 2.4: Sensitivity Analysis: columns are Weibull, Lognormal and Loglogistic data distributions; rows are the parametric Weibull, Lognormal, Loglogistic methods and the proposed semiparametric method.

<table>
<thead>
<tr>
<th></th>
<th>Weibull</th>
<th>Lognormal</th>
<th>Loglogistic</th>
<th>Semiparametric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Par</td>
<td>Bias</td>
<td>StErr</td>
<td>StDev</td>
</tr>
<tr>
<td><strong>Weibull</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma_0$</td>
<td>0.014</td>
<td>0.142</td>
<td>0.137</td>
<td>0.958</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>-0.021</td>
<td>0.177</td>
<td>0.167</td>
<td>0.972</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.008</td>
<td>0.130</td>
<td>0.125</td>
<td>0.956</td>
</tr>
<tr>
<td>shape</td>
<td>0.005</td>
<td>0.034</td>
<td>0.034</td>
<td>0.954</td>
</tr>
<tr>
<td>scale</td>
<td>0.012</td>
<td>0.133</td>
<td>0.131</td>
<td>0.958</td>
</tr>
<tr>
<td>cure-1</td>
<td>0.001</td>
<td>0.025</td>
<td>0.025</td>
<td>0.958</td>
</tr>
<tr>
<td>cure-2</td>
<td>-0.003</td>
<td>0.031</td>
<td>0.030</td>
<td>0.962</td>
</tr>
<tr>
<td><strong>Lognormal</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma_0$</td>
<td>0.450</td>
<td>0.202</td>
<td>0.153</td>
<td>0.332</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>-0.428</td>
<td>0.229</td>
<td>0.183</td>
<td>0.562</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.171</td>
<td>0.162</td>
<td>0.085</td>
<td>0.940</td>
</tr>
<tr>
<td>shape</td>
<td>-0.750</td>
<td>0.244</td>
<td>0.004</td>
<td>0.000</td>
</tr>
<tr>
<td>scale</td>
<td>1.067</td>
<td>0.091</td>
<td>0.121</td>
<td>0.000</td>
</tr>
<tr>
<td>cure-1</td>
<td>-0.006</td>
<td>0.032</td>
<td>0.026</td>
<td>0.330</td>
</tr>
<tr>
<td>cure-2</td>
<td>-0.091</td>
<td>0.033</td>
<td>0.028</td>
<td>0.328</td>
</tr>
<tr>
<td><strong>Loglogistic</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma_0$</td>
<td>0.367</td>
<td>0.159</td>
<td>0.152</td>
<td>0.348</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>-0.348</td>
<td>0.191</td>
<td>0.183</td>
<td>0.574</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.090</td>
<td>0.140</td>
<td>0.115</td>
<td>0.952</td>
</tr>
<tr>
<td>shape</td>
<td>0.206</td>
<td>0.046</td>
<td>0.051</td>
<td>0.002</td>
</tr>
<tr>
<td>scale</td>
<td>-0.268</td>
<td>0.118</td>
<td>0.103</td>
<td>0.380</td>
</tr>
<tr>
<td>cure-1</td>
<td>-0.005</td>
<td>0.030</td>
<td>0.026</td>
<td>0.316</td>
</tr>
<tr>
<td>cure-2</td>
<td>-0.075</td>
<td>0.030</td>
<td>0.029</td>
<td>0.344</td>
</tr>
<tr>
<td><strong>Semiparametric</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma_0$</td>
<td>0.011</td>
<td>0.128</td>
<td>0.124</td>
<td>0.952</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>-0.021</td>
<td>0.016</td>
<td>0.061</td>
<td>0.936</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.025</td>
<td>0.121</td>
<td>0.124</td>
<td>0.934</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.024</td>
<td>0.119</td>
<td>0.124</td>
<td>0.926</td>
</tr>
<tr>
<td>cure-1</td>
<td>0.004</td>
<td>0.025</td>
<td>0.026</td>
<td>0.952</td>
</tr>
<tr>
<td>cure-2</td>
<td>-0.002</td>
<td>0.029</td>
<td>0.027</td>
<td>0.918</td>
</tr>
</tbody>
</table>

Figure 2.2: Baseline Survival Probability Estimation for Weibull Data by Semiparametric MCPH+BM Model (400 Sample Size).
Table 2.5: Summary Statistics for the proposed semiparametric method with logistic link (low cure rates)

<table>
<thead>
<tr>
<th>n=200</th>
<th>Weibull Data</th>
<th>Lognormal Data</th>
<th>Loglogistic Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Par</td>
<td>Bias</td>
<td>StErr</td>
<td>StDev</td>
</tr>
<tr>
<td>(\gamma_0)</td>
<td>0.058</td>
<td>0.256</td>
<td>0.242</td>
</tr>
<tr>
<td>(\gamma_1)</td>
<td>-0.046</td>
<td>0.302</td>
<td>0.313</td>
</tr>
<tr>
<td>(\beta_1)</td>
<td>0.002</td>
<td>0.198</td>
<td>0.202</td>
</tr>
<tr>
<td>(\beta_2)</td>
<td>-0.007</td>
<td>0.150</td>
<td>0.168</td>
</tr>
<tr>
<td>cure-1</td>
<td>-0.004</td>
<td>0.038</td>
<td>0.036</td>
</tr>
<tr>
<td>cure-2</td>
<td>0.008</td>
<td>0.035</td>
<td>0.036</td>
</tr>
<tr>
<td>n=400</td>
<td>Weibull Data</td>
<td>Lognormal Data</td>
<td>Loglogistic Data</td>
</tr>
<tr>
<td>Par</td>
<td>Bias</td>
<td>StErr</td>
<td>StDev</td>
</tr>
<tr>
<td>(\gamma_0)</td>
<td>0.021</td>
<td>0.182</td>
<td>0.175</td>
</tr>
<tr>
<td>(\gamma_1)</td>
<td>-0.032</td>
<td>0.208</td>
<td>0.189</td>
</tr>
<tr>
<td>(\beta_1)</td>
<td>0.004</td>
<td>0.130</td>
<td>0.138</td>
</tr>
<tr>
<td>(\beta_2)</td>
<td>-0.005</td>
<td>0.091</td>
<td>0.084</td>
</tr>
<tr>
<td>cure-1</td>
<td>0.003</td>
<td>0.028</td>
<td>0.029</td>
</tr>
<tr>
<td>cure-2</td>
<td>0.005</td>
<td>0.022</td>
<td>0.022</td>
</tr>
<tr>
<td>n=800</td>
<td>Weibull Data</td>
<td>Lognormal Data</td>
<td>Loglogistic Data</td>
</tr>
<tr>
<td>Par</td>
<td>Bias</td>
<td>StErr</td>
<td>StDev</td>
</tr>
<tr>
<td>(\gamma_0)</td>
<td>0.023</td>
<td>0.118</td>
<td>0.116</td>
</tr>
<tr>
<td>(\gamma_1)</td>
<td>-0.028</td>
<td>0.141</td>
<td>0.136</td>
</tr>
<tr>
<td>(\beta_1)</td>
<td>0.003</td>
<td>0.082</td>
<td>0.092</td>
</tr>
<tr>
<td>(\beta_2)</td>
<td>-0.003</td>
<td>0.048</td>
<td>0.052</td>
</tr>
<tr>
<td>cure-1</td>
<td>0.002</td>
<td>0.015</td>
<td>0.014</td>
</tr>
<tr>
<td>cure-2</td>
<td>-0.001</td>
<td>0.014</td>
<td>0.015</td>
</tr>
</tbody>
</table>

Figure 2.3: Baseline Survival Probability Estimation for Weibull Data by Semiparametric MCPH+BM Model (800 Sample Size).

present the methods used to fit the data including the parametric MCPH+BM model by using Weibull, Lognormal, Loglogistic baseline distributions and our proposed semiparametric method. For the parametric methods, the diagonal estimation, where the distribution is correctly specified, provides a good estimation with small biases, small standard errors and good coverage probabilities. The proposed semiparametric
MCPH+BM model performs well in parameters and cure fraction estimation for different data distributions with small biases and the coverage probabilities close to the nominal 95%. Therefore, the semiparametric method for MCPH+BM model works well for different parametric assumptions.
2.5 Real Data Study

2.5.1 SEER Regional Female Breast Cancer

We applied the proposed semiparametric MCPH+BM model to the SEER Louisiana regional female breast cancer data (Howlader et al., 2019), and compare its performance to that under the standard MCPH model via semiparametric approach. The
Figure 2.8: Baseline Survival Probability Estimation for Loglogistic Data by Semi-parametric MCPH+BM Model (400 Sample Size).

Figure 2.9: Baseline Survival Probability Estimation for Loglogistic Data by Semi-parametric MCPH+BM Model (800 Sample Size).

The event of interest is the status of death or not. We fit the data with the covariates of race and grade in both cure rate function and latency survival function such that

\[ S_{\text{pop}}(t|\text{age, gender, race, grade}) = S^*(t|\text{age, gender}) \left\{ \pi(\text{race, grade})S_0(t)^{\exp[\beta(\text{race, grade})]} + [1 - \pi(\text{race, grade})] \right\} , \]

where race = 1 for the white people and = 2 for the black people, and grade is a measurement of how closely the tumor cells resemble normal breast cells with 1 to
Figure 2.10: Predicted Net Survival Probability for the Simulated Data of 200 Sample Size Weibull Distribution: dashed line and solid line represent the predicted survival probabilities for treatment and control arms respectively.

Figure 2.11: Predicted Net Survival Probability for the Simulated Data of 200 Sample Size Lognormal Distribution: dashed line and solid line represent the predicted survival probabilities for treatment and control arms respectively.

9 representing different degrees of differentiation. Similar to the simulation studies, the background mortality distribution is approximated by the Weibull distribution based on “2015 USA Life Table” with the matched sex and age. We also compare the performance of the MCPH+BM model to that under the standard MCPH model for the subgroup of age $\leq 70$ and age $> 70$. 
Figure 2.12: Predicted Net Survival Probability for the Simulated Data of 200 Sample Size Loglogistic Distribution: dashed line and solid line represent the predicted survival probabilities for treatment and control arms respectively.

Table 2.6: SEER Louisiana regional female breast cancer: parameters estimation in the semiparametric MCPH+BM model and the semiparametric MCPH model, race and grade are included in both the cure logistic link function and latency survival function. Model I is to fit the whole data set with 6200 observations, Model II is to fit the subgroup of age $\leq 70$ with 4267 observations, and Model III is to fit the subgroup of age $> 70$ with 1933 observations.

<table>
<thead>
<tr>
<th></th>
<th>MCPH+BM</th>
<th></th>
<th>MCPH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est StErr CI</td>
<td></td>
<td>Est StErr CI</td>
</tr>
<tr>
<td><strong>Model I: whole data</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma_0$</td>
<td>-1.140 0.111 (-1.358, -0.922)</td>
<td>-0.140 0.132 (-0.399, 0.119)</td>
<td></td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>0.476 0.070  (0.339, 0.613)</td>
<td>0.151 0.100  (-0.045, 0.347)</td>
<td></td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>0.081 0.016  (0.050, 0.112)</td>
<td>0.039 0.021  (-0.002, 0.080)</td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.382 0.039  (0.306, 0.458)</td>
<td>0.219 0.068  (0.086, 0.352)</td>
<td></td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.089 0.013  (0.064, 0.111)</td>
<td>0.060 0.015  (0.031, 0.089)</td>
<td></td>
</tr>
<tr>
<td>cure-W</td>
<td>0.605 0.020  (0.566, 0.644)</td>
<td>0.468 -      -</td>
<td></td>
</tr>
<tr>
<td>cure-B</td>
<td>0.487 0.012  (0.463, 0.511)</td>
<td>0.431 -      -</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>MCPH+BM</th>
<th></th>
<th>MCPH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est StErr CI</td>
<td></td>
<td>Est StErr CI</td>
</tr>
<tr>
<td><strong>Model II: subgroup of age $\leq 70$</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma_0$</td>
<td>-1.419 0.125 (-1.664, -1.174)</td>
<td>-0.932 0.130 (-1.187, -0.678)</td>
<td></td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>0.506 0.078  (0.353, 0.659)</td>
<td>0.351 0.083  (0.188, 0.514)</td>
<td></td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>0.071 0.018  (0.036, 0.106)</td>
<td>0.053 0.025  (0.004, 0.102)</td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.113 0.067  (-0.018, 0.244)</td>
<td>0.257 0.087  (0.086, 0.428)</td>
<td></td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.051 0.027  (-0.002, 0.104)</td>
<td>0.056 0.021  (0.015, 0.097)</td>
<td></td>
</tr>
<tr>
<td>cure-W</td>
<td>0.669 0.020  (0.639, 0.708)</td>
<td>0.604 -      -</td>
<td></td>
</tr>
<tr>
<td>cure-B</td>
<td>0.549 0.012  (0.525, 0.573)</td>
<td>0.518 -      -</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>MCPH+BM</th>
<th></th>
<th>MCPH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est StErr CI</td>
<td></td>
<td>Est StErr CI</td>
</tr>
<tr>
<td><strong>Model III: subgroup of age $&gt; 70$</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma_0$</td>
<td>-0.831 0.217 (-1.26, 0.41)</td>
<td>0.954 0.294 (0.378, 1.530)</td>
<td></td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>0.461 0.149  (0.169, 0.753)</td>
<td>0.273 0.254  (-0.225, 0.771)</td>
<td></td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>0.090 0.032  (0.027, 0.153)</td>
<td>0.017 0.046  (-0.073, 0.107)</td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.508 0.046  (0.418, 0.598)</td>
<td>0.212 0.095  (0.026, 0.398)</td>
<td></td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.038 0.015  (0.009, 0.067)</td>
<td>0.074 0.020  (0.035, 0.113)</td>
<td></td>
</tr>
<tr>
<td>cure-W</td>
<td>0.526 0.045  (0.438, 0.614)</td>
<td>0.218 -      -</td>
<td></td>
</tr>
<tr>
<td>cure-B</td>
<td>0.411 0.023  (0.366, 0.456)</td>
<td>0.175 -      -</td>
<td></td>
</tr>
</tbody>
</table>
In Table 2.6, Model I shows the estimation for the whole SEER Louisiana regional female breast cancer data set. The estimated cure rates are 60.46% and 48.73% for white people and black people in the semiparametric MCPH+BM model, compared to 46.84% and 43.11% in the standard MCPH model via semiparametric method. The race and grade are both significant in the proposed MCPH+BM model, while they are not significant in the cure function under the standard MCPH model. Model II in Table 2.6 shows that for the subgroup of age \( \leq 70 \), the estimated cure rates are 66.90% and 54.93% for white people and black in the MCPH+BM model, and are 60.41% and 51.78% in the standard MCPH model. Both race and grade have significant impact on the cure fraction in the proposed MCPH+BM model. Model III indicates that for the subgroup of age \( > 70 \), the estimated cure rates are 52.56% and 41.12%, which are much higher than the cure fraction estimation 21.81% and 17.51% under the standard MCPH model. Figure 2.13 shows the predicted observed survival probability under the standard MCPH model and predicted net survival probability under the proposed MCPH+BM model by race, which indicate higher cure fraction estimation from the proposed semiparametric MCPH+BM model compared to the standard MCPH model, and it is consistent with the estimation in Table 2.6. The similar conclusions are also found in the subset of age \( > 70 \) and \( \leq 70 \) shown in Figure 2.14 and 2.15.

2.5.2 Software

We develop a R package named “psmcure” based on the proposed MCPH+BM model, which can be downloaded from the Github repository

\[ \text{install\_github("gygygy1989/psmcure")}. \] The main function is \text{bmcure()} with arguments including \text{formula} which incidates the covariates in the latency survival function, \text{cureform} includes the covariates in cure link function, \text{link} is the type of link function which can be “logistic”, “probit”, and “cloglog”, \text{data} is the dataset.
Figure 2.13: The MCPH+BM model predicts the net survival probability by race, and the MCPH model predicts the observed (net) survival probability by race.

Figure 2.14: Predicted net survival under the MCPH+BM model and the MCPH model for the subgroup of Age $>70$ in the SEER regional female breast cancer.

used, table 1 and table 2 are the USA Life Table for female and male, na.action is the action to deal with missing data, method can be either “semiparametric”, “flexible” or “parametric”, and dis for distribution must be specified when the method is “parametric”. A pseudo real data was generated and fitted shown in Figure 2.16.
Figure 2.15: Predicted net survival under the MCPH+BM model and the MCPH+BM model for the subgroup of Age ≤ 70 in the SEER regional female breast cancer.

```r
> fit <- bmcure(formula = Surv(Time, Status) ~ race + grade,
  
cureform = ~ race + grade, link = "logistic", data = data, table1 = table1,
  
table2 = table2, na.action = na.omit, method = "semiparametric", dis = NULL)
```

---

Figure 2.16: Pseudo Data Estimation Output by the R Package “psmcur”

---

45
2.6 Conclusions and Discussions

“Cure” in a population level is defined as that the mortality risk of “cured patients” will return to the similar mortality risk as their counterpart in the general population (De Angelis et al., 1999). Many survival studies show that the MCPH+BM model is a useful tool in analyzing such data and provides insight on population-based survival study. The proposed semiparametric method in the MCPH+BM model is more appealing in practice due to its great modeling flexibility and computation efficiency. The simulation studies show that the semiparametric method in the MCPH+BM model via EM algorithm is comparable to the fully parametric method when the parametric baseline function is correctly specified and also comparable to the flexible method when the best number of knots is selected by information criterion such as AIC. Moreover, the computation of EM algorithm with perturbation variance estimation is efficient compared to bootstrap method. The real data analysis illustrates the usage of proposed semiparametric MCPH+BM model to the SEER data, and it may provide more accurate estimation of cure fraction in population level.

The proposed semiparametric method is expected to promote the application of the MCPH+BM model in practice. The MCPH+BM model works well in the situations when the cause of death is unknown or not recorded, so that death not only contributes to the disease under study, but also results from the background mortality of the whole population. Specifically, when there is a plateau observed from Kaplan-Meier curve and the death is a disease-related death, the standard MCPH model is suggested. In the case that the event at the last survival time is death indicating no plateau in Kaplan-Meier curve, there is no visual evidence suggesting the cure fraction and the advice from domain experts may help to choose the best model. If the death is all-cause death and the cause of death is unknown, the proposed MCPH+BM model is recommended to estimate cure rate in population level. Note, the appropriate background mortality should be used in practice. We used the “2015
USA Life Table” since the SEER data is population representative, while the samples in the clinical trial may not be a random sample from the whole population and the population level mortality are not applicable. In terms of variable selection in the MCPH+BM model, AIC can not be used in semiparametric method directly, but a forward or backward selection based on the Wald test combining with suggestions from domain experts can be used to select variables in cure link function and latency survival function.

Currently, there are efforts working toward adjusting the post-cancer mortality from general population background mortality through drug reimbursement, and the adjusted background mortality may be applied in the future.
Chapter 3

Semiparametric Estimation in the Population-level Promotion Time Cure Model

Abstract

Similar to the mixture cure proportional hazards model with background mortality (MCPH+BM) model, another widely used cure model, i.e., the promotion time cure model also can be extended to the population-level cancer study. The “cured” in population level is defined as when the mortality rate of the individuals in diseased group returns to the same level as that expected in the general population, where the population level mortality is presented by the mortality table of the United States. We develop a promotion time cure model with background mortality (PCPH+BM) model and propose an EM algorithm in semiparametric regression. Simulation studies and real data analysis show that the proposed semiparametric approach provides good performance in estimation of parameters, cure fraction and survival functions in population-level cancer study, and it also largely avoid the problem of misspecification.

3.1 Introduction

The cure models have been proposed and studied, considering a proportion of subjects will never experience the event of interest in the time-to-event data. Boag (1949) proposed the classic two-component mixture cure model, which treats the whole population as a mixture of cured patients and uncured patients. Specifically,
the two-component mixture cure model is a mixture of a cure fraction and the latency survival function for the uncured patients, which is more advantageous than standard survival models such as the Cox PH model when a plateau occurs on the tail of survival curves. This model has been extensively discussed by many authors, including Farewell (1982); Gray and Tsiatis (1989); Kuk and Chen (1992); Sposto, Sather, and Baker (1992); Taylor (1995); and Stangl and Greenhouse (1998) among others.

From the perspective of both frequentist and Bayesian, the mixture cure model has some limits pointed out by Chen, Ibrahim, and Sinha (1999) and Ibrahim, Chen, and Sinha (2001). Firstly, it does not have the proportional hazards structure in the presence of covariates, since many asymptotic and computational results require a proportional hazards structure for survival models. Second, the mixture cure model in the equation (1.3) yields improper posterior distributions for many noninformative improper priors including the uniform prior for the regression coefficients when covaraites are included in the link function. This is a main problem in Bayesian inference. Yakovlev et al. (1993) proposed a new formulation for the standard two-component mixture cure model to study the promotion time of cancer metastases (1.4). The non-mixture cure model or promotion time cure model (PTCM) has been studied by Andrei and Asselain (1996); Tsodikov (1998); Ibrahim, Chen, and Sinha (2001) and so on. The PTCM has its advantages such that it can inherit the proportional hazards structure for the whole population which makes the interpretation of regression parameters more acceptable. In addition, the model provides a biological derivation which involves the number of tumor cells that have potential to metastasize. Zeng, Yin, and Ibrahim (2006) proposed the nonparametric maximum likelihood estimator (NPMLE) for the PTCM. Ibrahim, Chen, and Sinha (2014) discussed Bayesian approaches to fit the PTCM. Chen and Du (2018) proposed the estimation with smoothing splines in penalized profile likelihood.
Similar to in Chapter 2, inspired by the relative survival that observed survival probability is actually the product of net survival probability and expected survival probability, we also incorporate the background mortality into the PTCM and explore the cure fraction estimation in a population level. The PTCM has been extended to incorporate background mortality rates in parametric regression models by Lambert et al. (2006). They (Lambert, 2007) also developed the STATA commands. Andersson et al. (2011) developed a flexible parametric survival model which is a special case of the PTCM. The population-level PTCM incorporating background mortality via semiparametric method has not been studied yet. The purpose of this chapter is to fill the gap in the study of semiparametric estimation approach in the promotion time cure model with background mortality. The rest of the paper is organized as follows. Section 3.2 illustrates the proposed model with background mortality (PTCM+BM). The semiparametric estimation methods and variance estimation are described in Section 3.3 based on the EM algorithm. Section 3.4 outlines the simulation studies and Section 3.5 applies the proposed method to the real data. Section 3.6 provides some discussions.

3.2 Promotion Time Cure Model with Background Mortality

The standard promotion time cure model (PTCM) has been discussed in Chapter 1. By incorporating background mortality, the promotion time cure model with background mortality (PTCM+BM) has the overall survival function for the subject \( i = 1, \ldots, n \), expressed as

\[
S_{\text{pop}}(t|z_i, x_i) = S^*_i(t) \exp\{-\pi(z_i)F(t|x_i)\}
\]

and the corresponding population-level hazard function is

\[
h_{\text{pop}}(t|z_i, x_i) = h^*_i(t) + \pi(z_i)f(t|x_i),
\]
where $S_i^*(t)$ is the background survival probability and $h_i^*(t)$ is the background mortality rate at time $t$ for the subject $i = 1, \ldots, n$. Note that $\pi(z_i)$ is a link function for the subject $i$ with respect to the covariate $z_i$. For the latency part, $F(t|x_i)$ is the latency cumulative distribution function with respect to the covariate $x_i$ for the subject $i$, and $f(t|x_i)$ is the corresponding probability density function. Considering the proportional hazards structure for the latency survival function, the PTCM+BM becomes the promotion time cure proportional hazards model with background mortality (PTPH+BM).

Suppose there are $n$ iid right-censored observations, 

$O_i = \{y_i = \min(t_i, C_i), z_i, x_i), \delta_i = I(t_i \leq C_i); i = 1, \ldots, n\}$ where $\delta_i$ is the indicator function with 1 for death and 0 for right censoring, $z_i$ is a vector of covariates which have potential effects on the link function, and $x_i$ is a vector of covariates for the latency survival function. We assume that the right-censored time $C_i$ is independent provided that it is random and non-informative such that the distribution of survival times provides no information about the distribution of censoring times. Let the unknown parameters denoted by $\Theta = \{\gamma, \beta, h_0(t), S_0(t)\}$, then the observed likelihood function is given by

$$
L_{obs}(\Theta; O) = \prod_{i=1}^{n} \left\{ h_i^*(y_i) + \pi(z_i) f(y_i|x_i) \right\}^\delta_i S_i^*(y_i) \exp\{-\pi(z_i) F(y_i|x_i)\}
$$

$$
= \prod_{i=1}^{n} \left\{ h_i^*(y_i) + \pi(z_i) h_0(y_i) \exp (\beta^\top x_i) S_0(y_i) \exp(\beta^\top x_i) \right\}^\delta_i S_i^*(y_i)
\exp \left[ -\pi(z_i) \{1 - S_0(y_i) \exp(\beta^\top x_i)\} \right]
$$

and the logarithm observed likelihood function is

$$
L_{obs}(\Theta; O) = \sum_{i=1}^{n} \left[ \delta_i \log \left\{ h_i^*(y_i) + \pi(z_i) h_0(y_i) \exp (\beta^\top x_i) S_0(y_i) \exp(\beta^\top x_i) \right\} \right.
$$

$$
+ \log S_i^*(y_i) - \pi(z_i) \{1 - S_0(y_i) \exp(\beta^\top x_i)\} \right].
$$
3.3 EM Algorithm

The latent variables $N_i$ is the number of metastasis-competent tumour cells for the subject $i$, which has a Poisson distribution with mean of $\pi(z_i)$. Similar to the equation (1.5), the survival function in the PCPH+BM model can be further derived as

$$S_{\text{pop}}(t|\mathbf{z}_i, \mathbf{x}_i) = S^*(t) \left[ \exp\{ -\pi(z_i) \} + \sum_{k=1}^{\infty} \{ 1 - F(t|\mathbf{x}_i) \} \frac{\pi(z_i)^k \exp\{ -\pi(z_i) \}}{k!} \right]$$

The promotion time for the $k$th ($k = 1, \ldots, N_i$) tumour cell is denoted as $\tilde{T}_k$, which is the time for the $k$th metastasis-competent tumour cell to produce a detectable tumour mass. The survival function for the $k$th tumour cell can be written as

$$\{ 1 - F(t|\mathbf{x}_i) \} \frac{\pi(z_i)^k \exp\{ -\pi(z_i) \}}{k!},$$

and its corresponding probability density function is derived as

$$k \{ 1 - F(t|\mathbf{x}_i) \}^{k-1} f(t_i|\mathbf{x}_i) \frac{\pi(z_i)^k \exp\{ -\pi(z_i) \}}{k!}.$$

Thus, its corresponding hazard function is $k f(t_i|\mathbf{x}_i)/\{ 1 - F(t_i|\mathbf{x}_i) \}$.

The complete likelihood function in the PCPH+BM with the latent variables $\mathbf{N}$ is

$$L_c(\Theta; \mathbf{O}, \mathbf{N}) = \prod_{i=1}^{n} \left\{ h^*(y_i) + N_i h_0(y_i) \exp (\beta^T \mathbf{x}_i) \right\}^{\delta_i} S^*(y_i) \left\{ S_0(y_i)^{\exp(\beta^T \mathbf{x}_i)} \right\}^{N_i} \frac{\pi(\mathbf{z}_i)^{N_i} \exp\{ -\pi(\mathbf{z}_i) \}}{N_i!},$$

where we suppose the link function $\pi(\mathbf{z}_i)$ takes the exponential form of $\pi(\mathbf{z}_i) = \exp(\gamma^T \mathbf{z}_i)$.

3.3.1 E-step

In EM algorithm, E-step is to take expectation of the complete likelihood function with respect to $N_i$ given the observed data $\mathbf{O}$. When $\delta_i = 1$, $N_i = 0, 1, 2, \ldots, \infty$
in the PCPH+BM, which is different from the case when the background mortality is ignored and a subject dies only from the cancer under study, then \( N_i \) takes on values of \( 1, \ldots, \infty \). The conditional complete likelihood function in the PCPH+BM, therefore becomes more complex. We will derive the algorithm in E-step in detail on two cases separately (\( \delta_i = 1 \) and \( \delta_i = 0 \)).

The equation (3.1) conditional on \( \delta_i = 1 \) is written as

\[
\mathcal{L}_c(\Theta; \mathbf{O}, \mathbf{N}|\delta_i = 1) = \prod_{i=1}^{n} \left( h^*(y_i) S^*(y_i) \exp \left[ - \{1 - S_0(y_i)^{\exp(\beta^T x_i)}\}\pi(z_i) \right] \right) \times \frac{\{S_0(y_i)^{\exp(\beta^T x_i)}\pi(z_i)\}^{N_i} \exp\{-S_0(y_i)^{\exp(\beta^T x_i)}\pi(z_i)\}}{N_i!} \\
+ S^*(y_i) N_i h_0(y_i) \exp (\beta^T x_i) \exp \left[ - \{1 - S_0(y_i)^{\exp(\beta^T x_i)}\}\pi(z_i) \right] \times \frac{\{S_0(y_i)^{\exp(\beta^T x_i)}\pi(z_i)\}^{N_i} \exp\{-S_0(y_i)^{\exp(\beta^T x_i)}\pi(z_i)\}}{N_i!}
\]

(3.2)

Taking integral of the equation (3.2) over the latent variable \( \mathbf{N} \), the marginal complete likelihood function given \( \delta_i = 1 \) is

\[
\mathcal{L}_c(\Theta; \mathbf{O}|\delta_i = 1) = \prod_{i=1}^{n} \left( \sum_{N_i=0}^{\infty} \frac{\{S_0(y_i)^{\exp(\beta^T x_i)}\pi(z_i)\}^{N_i} \exp\{-S_0(y_i)^{\exp(\beta^T x_i)}\pi(z_i)\}}{N_i!} h^*(y_i) \right) \\
\times S^*(y_i) \exp \left[ - \{1 - S_0(y_i)^{\exp(\beta^T x_i)}\}\pi(z_i) \right] \\
\times \sum_{N_i=0}^{\infty} N_i \frac{\{S_0(y_i)^{\exp(\beta^T x_i)}\pi(z_i)\}^{N_i} \exp\{-S_0(y_i)^{\exp(\beta^T x_i)}\pi(z_i)\}}{N_i!} \\
\times h_0(y_i) \exp (\beta^T x_i) S^*(y_i) \exp \left[ - \{1 - S_0(y_i)^{\exp(\beta^T x_i)}\}\pi(z_i) \right] \\
= \prod_{i=1}^{n} \left( h^*(y_i) S^*(y_i) \exp \left[ - \{1 - S_0(y_i)^{\exp(\beta^T x_i)}\}\pi(z_i) \right] \right) \\
+ S_0(y_i)^{\exp(\beta^T x_i)}\pi(z_i) h_0(y_i) \exp (\beta^T x_i) S^*(y_i) \\
\times \exp \left[ - \{1 - S_0(y_i)^{\exp(\beta^T x_i)}\}\pi(z_i) \right].
\]

To simplify expressions in the following algorithm, let \( A_i = h^*(y_i) S^*(y_i) \exp \left[ - \{1 - S_0(y_i)^{\exp(\beta^T x_i)}\}\pi(z_i) \right] \) and \( B_i = S_0(y_i)^{\exp(\beta^T x_i)}\pi(z_i) h_0(y_i) \exp (\beta^T x_i) S^*(y_i) \exp \left[ - \{1 - S_0(y_i)^{\exp(\beta^T x_i)}\}\pi(z_i) \right] \) for the subject \( i \). The conditional likelihood function of \( \mathbf{N} \)
given the parameters $\Theta$, the observed data $O$ and $\delta_i = 1$ is

$$L_c(N|\Theta, O; \delta_i = 1) = \prod_{i=1}^{n} \left( \frac{A_i}{A_i + B_i} \left\{ \frac{S_0(y_i)^{\exp(\beta^{x_i})}\pi(z_i)}{N_i} \exp\{-S_0(y_i)^{\exp(\beta^{x_i})}\pi(z_i)\} \right\} N_i! \right)$$

$$+ \left( \frac{B_i}{A_i + B_i} \right) \frac{S_0(y_i)^{\exp(\beta^{x_i})}\pi(z_i)}{N_i!} \times \left\{ \frac{S_0(y_i)^{\exp(\beta^{x_i})}\pi(z_i)}{N_i} \exp\{-S_0(y_i)^{\exp(\beta^{x_i})}\pi(z_i)\} N_i \right\},$$

and then the expectation of $N_i$ given $\Theta$, $O$ and $\delta_i = 1$ is

$$E(N_i|\Theta, O; \delta_i = 1) = \frac{A_i}{A_i + B_i} S_0(y_i)^{\exp(\beta^{x_i})}\pi(z_i) + \frac{B_i}{A_i + B_i} \left\{ 1 + S_0(y_i)^{\exp(\beta^{x_i})}\pi(z_i) \right\},$$

(3.3)

which is a weighted mean of $S_0(y_i)^{\exp(\beta^{x_i})}\pi(z_i)$ and $1 + S_0(y_i)^{\exp(\beta^{x_i})}\pi(z_i)$.

When $\delta_i = 0$, $N_i = 0, 1, 2, \ldots, \infty$, the equation (3.1) given $\delta_i = 0$ is

$$L_c(\Theta; O, N|\delta_i = 0) = \prod_{i=1}^{n} S^*(y_i) \exp\left[-\left\{ 1 - S_0(y_i)^{\exp(\beta^{x_i})} \right\}\pi(z_i) \right]$$

$$\times \left\{ \frac{S_0(y_i)^{\exp(\beta^{x_i})}\pi(z_i)}{N_i} \exp\{-S_0(y_i)^{\exp(\beta^{x_i})}\pi(z_i)\} \right\} N_i!$$

and the marginal complete likelihood function over $N$ given $\delta_i = 0$ is

$$L_c(\Theta; O|\delta_i = 0) = S^*(y_i) \exp\left[-\left\{ 1 - S_0(y_i)^{\exp(\beta^{x_i})} \right\}\pi(z_i) \right].$$

The conditional likelihood function of $N$ given $\Theta$, $O$ and $\delta_i = 0$ is

$$L_c(N|\Theta, O; \delta_i = 0) = \frac{S_0(y_i)^{\exp(\beta^{x_i})}\pi(z_i)}{N_i!} \exp\{-S_0(y_i)^{\exp(\beta^{x_i})}\pi(z_i)\},$$

and the expectation of $N_i$ given $\Theta$, $O$ and $\delta_i = 0$ is

$$E(N_i|\Theta, O; \delta_i = 0) = S_0(y_i)^{\exp(\beta^{x_i})}\pi(z_i).$$

(3.4)

In summary, the equation (3.3) and (3.4) can be written in one single formula as

$$\omega_{1i} = E(N_i|\Theta, O)$$

$$= \delta_i \left[ \frac{A_i}{A_i + B_i} S_0(y_i)^{\exp(\beta^{x_i})}\pi(z_i) + \frac{B_i}{A_i + B_i} \left\{ 1 + S_0(y_i)^{\exp(\beta^{x_i})}\pi(z_i) \right\} \right]$$

$$+ (1 - \delta_i) S_0(y_i)^{\exp(\beta^{x_i})}\pi(z_i).$$

(3.5)
The expectation of the logarithm complete likelihood function is expressed as

\[ Q(N; \Theta, O) = \sum_{i=1}^{n} \left( \delta_i E \left[ \log \{ h^*(y_i) + N_i h_0(y_i) \exp (\beta^T x_i) \} \right] + E(N_i) \log \{ S_0(y_i) \exp (\beta^T x_i) \} \right) \]

\[ = \sum_{i=1}^{n} \delta_i \omega_{2i} + \omega_{1i} \log \{ S_0(y_i) \exp (\beta^T x_i) \} + \omega_{1i} \log \{ \pi(z_i) \} - \pi(z_i), \]

where \( E \left[ \log \{ h^*(y_i) + N_i h_0(y_i) \exp (\beta^T x_i) \} \right] \) is denoted as \( \omega_{2i} \) for simplification in expression.

The equation (3.6) can be further written as the summation of two separate functions

\[ Q_1(\gamma; z, \omega_1) = \sum_{i=1}^{n} \omega_{1i} \log \{ \pi(z_i) \} - \pi(z_i) \]  \hspace{1cm} (3.7)

and

\[ Q_2(\beta, h_0(y), S_0(y); x, \omega_2) = \sum_{i=1}^{n} \delta_i \omega_{2i} + \omega_{1i} \exp (\beta^T x_i) \log \{ S_0(y_i) \}, \]  \hspace{1cm} (3.8)

where function (3.7) only involves the parameter \( \gamma \) in cure function \( \pi(z_i) = \exp(\gamma^T z_i) \) and function (3.8) involves baseline functions and the parameter \( \beta \).

### 3.3.2 M-step

In M-step, we maximize the equation (3.7) and (3.8) via Newton-Raphson method using “optim” function in R to update \( \gamma \), baseline functions and \( \beta \), respectively, which improves computation efficiency. The background mortality \( h^*(t) \) in the expectation term \( \omega_{2i} \) are sex and age matched constant pieces in the calculation.

For baseline estimation, \( t_1 < t_2 < \cdots < t_p \) are the distinct ordered uncensored failure times, and let \( \lambda_{0j} (j = 1, \cdots, p) \) be the jump size at each death point. Take first derivative of the equation (3.8) and set it to 0, we have

\[ \frac{\partial Q_2(\beta, h_0(t), S_0(t))}{\partial \lambda_{0j}} = \delta_j \frac{\partial E \log \{ h^*(t_j) + N_j \lambda_{0j} \exp (\beta^T x_j) \}}{\partial \lambda_{0j}} - \sum_{l=j}^{p} \omega_{1l} \exp (\beta^T x_i) \]

\[ = \delta_j E \left\{ \frac{N_j \exp (\beta^T x_j)}{h^*(t_j) + N_j \lambda_{0j} \exp (\beta^T x_j)} \right\} - \sum_{l=j}^{p} \omega_{1l} \exp (\beta^T x_i) = 0. \]  \hspace{1cm} (3.9)
Note that the expectation term in the equation (3.9) can be estimated by iteration until convergence such that

\[
E\left\{ \frac{N_j \exp(\beta^\top x_j)}{h^*(t_j) + N_j \lambda_{0j} \exp(\beta^\top x_j)} \right\} = E\left\{ \frac{1 - h^*(t_j)/\lambda_{0j}}{1 - \lambda_{0j} E\{1/h^*(t_j) + N_j \lambda_{0j} \exp(\beta^\top x_j)\}} \right\} = \frac{1}{\lambda_{0j}} \frac{1}{\lambda_{0j}} E\left\{ \frac{1}{h^*(t_j) + N_j \lambda_{0j} \exp(\beta^\top x_j)} \right\}.
\]

(3.10)

Update \( \lambda_{0j}^m \) in the \( m \) iteration by plugging the equation (3.10) into equation (3.9), the converged \( \lambda_{0j}^m \) would be the value when the difference between \( \lambda_{0j}^m \) and \( \lambda_{0j}^{m-1} \) is less than a small value such as 1e-7, which is written as

\[
\lambda_{0j}^m = \frac{\delta_j \left[ 1 - h^*(t_j) E\{1/h^*(t_j) + N_j \lambda_{0j}^{m-1} \exp(\beta^\top x_j)\} \right]}{\sum_{l=j}^p \omega_{1l} \exp(\beta^\top x_l)}.
\]

Thus, the baseline hazard function and baseline survival function for the subject \( i \) are

\[
h_0^m(y_i; \beta, x_i, \omega_{1i}) = \frac{\delta_i \left[ 1 - h^*(y_i) E\{1/h^*(y_i) + N_i \lambda_{0i}^{m-1} \exp(\beta^\top x_i)\} \right]}{\sum_{l:y_l \leq y_i} \omega_{1l} \exp(\beta^\top x_l)},
\]

(3.11)

and

\[
S_0^m(y_i; \beta, x_i, \omega_{1i}, \omega_{3i}) = \exp\{- \sum_{l:y_l \leq y_i} \lambda_{0i}^m (y_l)\}.
\]

(3.12)

The following conditional expectation terms for the subject \( i \) are also needed in the M-step algorithm,

\[
\omega_{2i} = E\left[ \log \left\{ h^*(y_i) + N_i h_0(y_i) \exp(\beta^\top x_i) \right\} | \Theta, O \right] = \sum_{k=0}^{\infty} \left( \log \left\{ h^*(y_i) + k h_0(y_i) \exp(\beta^\top x_i) \right\} \times \mathcal{L}_c(k|\Theta, O) \right)
\]

\[
= \delta_i \sum_{k=0}^{\infty} \left( \log \left\{ h^*(y_i) + k h_0(y_i) \exp(\beta^\top x_i) \right\} \times \mathcal{L}_c(k|\Theta, O; \delta_i = 1) \right) + \left( 1 - \delta_i \right) \sum_{j=0}^{\infty} \left( \log \left\{ h^*(y_i) + k h_0(y_i) \exp(\beta^\top x_i) \right\} \times \mathcal{L}_c(k|\Theta, O; \delta_i = 0) \right),
\]

(3.13)
\[
\omega_{3i} = E\{\frac{1}{h^*(y_i) + N_i h_0(y_i) \exp (\beta^T x_i)}|\Theta, O\}
\]
\[
= \sum_{k=0}^{\infty} \left( \frac{1}{h^*(y_i) + kh_0(y_i) \exp (\beta^T x_i)} \times L_c(k|\Theta, O) \right)
\]
\[
= \delta_i \sum_{k=0}^{\infty} \left( \frac{1}{h^*(y_i) + kh_0(y_i) \exp (\beta^T x_i)} \times L_c(k|\Theta, O; \delta_i = 1) \right)
\]
\[\quad + (1 - \delta_i) \sum_{k=0}^{\infty} \left( \frac{1}{h^*(y_i) + kh_0(y_i) \exp (\beta^T x_i)} \times L_c(k|\Theta, O; \delta_i = 0) \right). \tag{3.14}\]

The EM algorithm for estimating the parameters \( \Theta = \{\gamma, \beta, h_0(y), S_0(y)\} \) is summarized as follows,

**Step 0:** Set initial values \( \omega_{1i}^0, \omega_{2i}^0 \) and \( \omega_{3i}^0 \) as the subject censoring indicators \( \delta_i, (i = 1, 2, \cdots, n) \), the initial values \( \gamma^0 \) are estimated from the logistic regression with \( \omega_i^0 \) as the response variable and \( z \) as the covariates, and the initial values \( \beta^0 \) are estimated from the cox proportional hazards model with \( x \) as the covariates;

**Step 1:** In \((d+1)\) iteration, update \( \gamma^{(d+1)} \) via maximizing \((3.7)\) using “optim”;

**Step 2:** Update \( \beta^{(d+1)} \) via maximizing equation \((3.8)\) using “optim”, where the baseline hazard and survival functions are replaced by \((3.11)\) and \((3.12)\);

**Step 3:** Update \( h_0^{(d+1)}(y_i; \beta^{(d+1)}, x_i, \omega_{1i}^{(d)}, \omega_{3i}^{(d)}) \) and \( S_0^{(d+1)}(y_i; \beta^{(d+1)}, x_i, \omega_{1i}^{(d)}, \omega_{3i}^{(d)}) \) using equation \((3.11)\) and \((3.12)\); then update \( \omega_{1i}^{(d+1)}, \omega_{2i}^{(d+1)} \) and \( \omega_{3i}^{(d+1)} \) by the equation \((3.5),(3.13)\) and \((3.14)\), respectively;

**Step 4:** Iterate steps 1-3 till convergence when the difference of two successive iterations is less than \(1e^{-7}\).

For variance estimation, the resampling method proposed by Jin, Ying, and Wei (2001) is adapted. We introduce a random variable \( V \) which follows the distribution Gamma\((1,1)\), and the likelihood functions in the equations \((3.7)\) and \((3.8)\) become,

\[
V_1(\gamma; z, \omega) = \sum_{i=1}^{n} V_i \left[ \omega_{1i} \log \{ \pi(z_i) \} - \pi(z_i) \right], \tag{3.15}
\]
The maximization of equations (3.15) and (3.16) can be implemented in the same way as in the proposed EM algorithm to find the estimators \( \tilde{\Theta} \)'s, and the resampling procedure is started with \( M = 500 \) samples of \( V_i \), as recommended in the Jin, Ying, and Wei (2001). The variance estimation of parameters is then the empirical variance of the \( M \) estimators \( \tilde{\Theta} \)'s.

### 3.4 Simulation Studies

Simulation studies are conducted to evaluate the performance of the proposed semi-parametric estimation method in the PCPH+BM model. We compare the performance of our proposed model to fully parametric methods. The fully parametric method is to model the baseline hazard rate and baseline survival probability with the corresponding parametric distributions of Weibull, Lognormal and Loglogistic via the EM algorithm.

Similar to in Chapter 2, we generate covariates \( z \) from a Bernoulli distribution with probability 0.5 with 0 and 1 for the group indicator. Setting \( \gamma \) to \((\log(2), -1)\) gives the cure rate about 33% and 58% for two groups for illustration. Uncured indicators are generated from a Bernoulli distribution with the probability of uncured rate. The covariate of “age” is generated from a normal distribution with mean of 70 and standard deviation of 5, and “sex” from a Bernoulli distribution with the probability of 0.5.

For a patient being cured, the survival time is generated from the background mortality distribution. The background mortality is matched based on the estimated Weibull distribution approximated using “2015 USA Life Table” (Arias and Xu, 2018). The “2015 USA Life Table” has survival probability for male and female at the age...
of 0 to 110. For each sex and age combination, we approximate the life table via a Weibull distribution and those estimated parameters of Weibull distribution are used as the background mortality parameters. Then we match the sex and age in the data set with those in the background mortality table, and use the parameters to generate survival times from the background mortality.

For the uncured patients who may die from the disease under study or some other causes, the survival time is the minimum value generated from the background mortality distribution and the latency proportional hazards model. In the latency proportional hazards model, we include the covariates of group and sex, and set their coefficients $\beta$ to $(1, -1)$. Several baseline survival functions, Weibull distribution with shape=0 and scale=1, Lognormal distribution with mean=0 and standard deviation=1, and Loglogistic distribution with shape=2 and scale=1 are considered in the proportional hazards model to simulate data set. The censoring time is generated from an exponential distribution $\exp(c)$ where $c$ controls the censoring rate at about 50%.

A small (n=200), medium (n=400) and large (n=800) sample size with 500 replicates are considered for each simulation setup. A sensitivity analysis of 800 sample size is further done to illustrate the problem of misspecification in the fully parametric method. The simulation results include bias which are calculated as the average of differences between point estimators and their true values, average estimated standard error (StErr), empirical standard deviation (StDev), CPs which are the average coverage probabilities of the 95% confidence intervals.

Table 3.1-3.3 show that the proposed semiparametric PCPH+BM model estimate parameters and cure fraction well and the performance is comparable to the fully parametric methods. The biases for all parameters and cure fraction are small and the corresponding CPs are close to the nominal 95%. The estimated standard errors for parameters are similar in magnitude for different data distributions, and become
Table 3.1: Summary Statistics for Weibull Data: by proposed semiparametric method and parametric method with Weibull baseline function.

<table>
<thead>
<tr>
<th></th>
<th>Semiparametric Method</th>
<th>Parametric Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Par Bias StErr StDev CP</td>
<td>Bias StErr StDev CP</td>
</tr>
<tr>
<td>n=200</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$b_0$</td>
<td>0.031 0.205 0.170</td>
<td>0.009 0.140 0.149</td>
</tr>
<tr>
<td>$b_1$</td>
<td>-0.021 0.184 0.184</td>
<td>-0.017 0.194 0.204</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.022 0.239 0.235</td>
<td>0.023 0.257 0.251</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.011 0.134 0.138</td>
<td>-0.029 0.148 0.148</td>
</tr>
<tr>
<td>shape</td>
<td>- - -</td>
<td>0.016 0.062 0.062</td>
</tr>
<tr>
<td>scale</td>
<td>- - -</td>
<td>0.052 0.344 0.351</td>
</tr>
<tr>
<td>cure-1</td>
<td>-0.005 0.058 0.056</td>
<td>0.001 0.057 0.058</td>
</tr>
<tr>
<td>cure-2</td>
<td>-0.004 0.047 0.045</td>
<td>0.001 0.038 0.039</td>
</tr>
<tr>
<td>n=400</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$b_0$</td>
<td>0.024 0.147 0.118</td>
<td>0.020 0.097 0.101</td>
</tr>
<tr>
<td>$b_1$</td>
<td>-0.024 0.127 0.130</td>
<td>-0.010 0.134 0.133</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.017 0.175 0.170</td>
<td>0.011 0.177 0.176</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.012 0.098 0.100</td>
<td>-0.014 0.100 0.098</td>
</tr>
<tr>
<td>shape</td>
<td>- - -</td>
<td>0.008 0.042 0.041</td>
</tr>
<tr>
<td>scale</td>
<td>- - -</td>
<td>0.045 0.233 0.232</td>
</tr>
<tr>
<td>cure-1</td>
<td>-0.001 0.041 0.039</td>
<td>-0.004 0.039 0.039</td>
</tr>
<tr>
<td>cure-2</td>
<td>-0.004 0.033 0.034</td>
<td>-0.004 0.026 0.027</td>
</tr>
<tr>
<td>n=800</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$b_0$</td>
<td>0.015 0.102 0.083</td>
<td>0.008 0.070 0.069</td>
</tr>
<tr>
<td>$b_1$</td>
<td>-0.022 0.090 0.092</td>
<td>-0.006 0.098 0.099</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.014 0.128 0.126</td>
<td>0.018 0.128 0.117</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.017 0.062 0.066</td>
<td>-0.011 0.072 0.067</td>
</tr>
<tr>
<td>shape</td>
<td>- - -</td>
<td>0.007 0.030 0.028</td>
</tr>
<tr>
<td>scale</td>
<td>- - -</td>
<td>0.012 0.162 0.150</td>
</tr>
<tr>
<td>cure-1</td>
<td>0.002 0.028 0.026</td>
<td>-0.001 0.029 0.029</td>
</tr>
<tr>
<td>cure-2</td>
<td>-0.003 0.025 0.023</td>
<td>-0.001 0.020 0.019</td>
</tr>
</tbody>
</table>

The accuracy of parameters estimation gets better with the increase of sample size. The sensitivity analysis in Table 3.4 uses sample size of 800 and indicates that the proposed semiparametric method performs better than fully parameter methods. For the parametric estimation methods, the diagonal estimation provides good estimation with small biases, small standard errors and good coverage probabilities, since the data distribution is correctly specified. The proposed semiparametric PCPH+BM model performs well in estimating all parameters and cure fraction for different data
Table 3.2: Summary Statistics for Lognormal Data: by proposed semiparametric method and parametric method with Lognormal baseline function.

<table>
<thead>
<tr>
<th></th>
<th>Semiparametric Method</th>
<th>Parametric Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Par</td>
<td>Bias</td>
</tr>
<tr>
<td>n=200</td>
<td>b₀</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>b₁</td>
<td>-0.002</td>
</tr>
<tr>
<td></td>
<td>β₁</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>β₂</td>
<td>0.032</td>
</tr>
<tr>
<td></td>
<td>shape</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>scale</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>cure-1</td>
<td>-0.003</td>
</tr>
<tr>
<td></td>
<td>cure-2</td>
<td>0.004</td>
</tr>
<tr>
<td>n=400</td>
<td>b₀</td>
<td>-0.001</td>
</tr>
<tr>
<td></td>
<td>b₁</td>
<td>-0.003</td>
</tr>
<tr>
<td></td>
<td>β₁</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>β₂</td>
<td>-0.015</td>
</tr>
<tr>
<td></td>
<td>shape</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>scale</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>cure-1</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>cure-2</td>
<td>0.003</td>
</tr>
<tr>
<td>n=800</td>
<td>b₀</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>b₁</td>
<td>-0.017</td>
</tr>
<tr>
<td></td>
<td>β₁</td>
<td>0.022</td>
</tr>
<tr>
<td></td>
<td>β₂</td>
<td>-0.007</td>
</tr>
<tr>
<td></td>
<td>shape</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>scale</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>cure-1</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>cure-2</td>
<td>0.001</td>
</tr>
</tbody>
</table>

distributions. Therefore, the semiparametric method for PCPH+BM model would largely avoid the misspecification problem in fully parametric methods.

3.5 Breast Cancer Data Analysis

We applied the proposed semiparametric PCPH+BM model to the SEER Louisiana regional female breast cancer data described in Chapter 1.2 (Howlader et al., 2019) and compare its performance to the standard PCPH model via semiparametric method. We fit the data with the covariates of race and grade in both the link function and the latency survival function such that

\[ S_{\text{pop}}(t|\text{age,sex,race,grade}) = S^*(t|\text{age,sex}) \exp\{-\pi(\text{race,grade})F(t|\text{race,grade})\} \]

Similar to the simulation studies, the background mortality distribution is approximated by the Weibull distribution based on “2015 USA Life Table” with the
Table 3.3: Summary Statistics for Loglogistic Data: by proposed semiparametric method and parametric method with Loglogistic baseline function.

<table>
<thead>
<tr>
<th></th>
<th>Semiparametric Method</th>
<th>Parametric Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Par Bias StErr StDev CP</td>
<td>Bias StErr StDev CP</td>
</tr>
<tr>
<td>n=200</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(b_0)</td>
<td>-0.004 0.187 0.189 0.956</td>
<td>0.012 0.141 0.143 0.948</td>
</tr>
<tr>
<td>(b_1)</td>
<td>0.011 0.180 0.173 0.942</td>
<td>-0.021 0.203 0.197 0.954</td>
</tr>
<tr>
<td>(\beta_1)</td>
<td>-0.023 0.260 0.272 0.932</td>
<td>0.031 0.266 0.247 0.968</td>
</tr>
<tr>
<td>(\beta_2)</td>
<td>0.012 0.153 0.142 0.930</td>
<td>-0.022 0.157 0.149 0.956</td>
</tr>
<tr>
<td>shape</td>
<td>- - - -</td>
<td>0.056 0.179 0.175 0.944</td>
</tr>
<tr>
<td>scale</td>
<td>- - - -</td>
<td>0.013 0.146 0.143 0.960</td>
</tr>
<tr>
<td>cure-1</td>
<td>-0.004 0.053 0.054 0.952</td>
<td>0.002 0.057 0.059 0.946</td>
</tr>
<tr>
<td>cure-2</td>
<td>0.006 0.047 0.049 0.940</td>
<td>-0.001 0.036 0.037 0.950</td>
</tr>
<tr>
<td>n=400</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(b_0)</td>
<td>-0.004 0.151 0.135 0.942</td>
<td>-0.002 0.098 0.095 0.956</td>
</tr>
<tr>
<td>(b_1)</td>
<td>-0.001 0.138 0.125 0.928</td>
<td>-0.003 0.142 0.139 0.960</td>
</tr>
<tr>
<td>(\beta_1)</td>
<td>0.002 0.176 0.185 0.932</td>
<td>0.010 0.185 0.176 0.962</td>
</tr>
<tr>
<td>(\beta_2)</td>
<td>-0.002 0.115 0.103 0.942</td>
<td>0.000 0.107 0.109 0.950</td>
</tr>
<tr>
<td>shape</td>
<td>- - - -</td>
<td>0.009 0.120 0.113 0.964</td>
</tr>
<tr>
<td>scale</td>
<td>- - - -</td>
<td>0.007 0.103 0.102 0.958</td>
</tr>
<tr>
<td>cure-1</td>
<td>0.001 0.042 0.039 0.948</td>
<td>0.001 0.040 0.040 0.954</td>
</tr>
<tr>
<td>cure-2</td>
<td>0.004 0.036 0.035 0.932</td>
<td>0.002 0.025 0.026 0.952</td>
</tr>
<tr>
<td>n=800</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(b_0)</td>
<td>0.012 0.124 0.100 0.932</td>
<td>0.009 0.069 0.072 0.936</td>
</tr>
<tr>
<td>(b_1)</td>
<td>-0.003 0.011 0.006 0.928</td>
<td>-0.007 0.100 0.100 0.948</td>
</tr>
<tr>
<td>(\beta_1)</td>
<td>0.007 0.118 0.130 0.930</td>
<td>0.007 0.129 0.126 0.964</td>
</tr>
<tr>
<td>(\beta_2)</td>
<td>-0.011 0.071 0.066 0.958</td>
<td>-0.003 0.074 0.074 0.954</td>
</tr>
<tr>
<td>shape</td>
<td>- - - -</td>
<td>0.010 0.084 0.084 0.960</td>
</tr>
<tr>
<td>scale</td>
<td>- - - -</td>
<td>0.007 0.072 0.072 0.948</td>
</tr>
<tr>
<td>cure-1</td>
<td>0.005 0.030 0.027 0.936</td>
<td>-0.001 0.029 0.029 0.936</td>
</tr>
<tr>
<td>cure-2</td>
<td>0.003 0.026 0.025 0.930</td>
<td>-0.002 0.018 0.019 0.942</td>
</tr>
</tbody>
</table>

matched sex and age. We also compare the performance of PCPH+BM model to that of PCPH model for the subgroup of age > 70 and ≤ 70.

In Table 3.5, first part shows the estimation for the whole SEER Louisiana regional female breast cancer data set. The estimated cure rates are 62.3% and 45.6% for the white people and the black people in the semiparametric PCPH+BM model compared to 47.0% and 42.9% in the standard PCPH model via semiparametric method. The second part in Table 3.5 shows that for the subgroup of age ≤ 70, the estimated cure rates are 69.1% and 54.2% for the white and black people in the PCPH+BM model, and are 60.6% and 51.6% in the PCPH model. For the subgroup of age > 70, the estimated cure rates are 55.1% and 34.4% for the white and black people under the semiparametric PCPH+BM model, while the cure rates are estimated to be 22.1% and 17.4% under the semiparametric PCPH model. Parameters estimation are all
Table 3.4: Sensitivity Analysis for 800 Sample Size: rows represent parametric methods and semiparametric method, and columns are the Weibull, Lognormal and Loglogistic data.

<table>
<thead>
<tr>
<th>Par</th>
<th>Weibull</th>
<th>Lognormal</th>
<th>Loglogistic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>b₀</td>
<td>b₁</td>
<td>β₁</td>
</tr>
<tr>
<td>Weibull</td>
<td>0.008</td>
<td>-0.066</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td>0.070</td>
<td>0.098</td>
<td>0.128</td>
</tr>
<tr>
<td></td>
<td>0.069</td>
<td>0.099</td>
<td>0.117</td>
</tr>
<tr>
<td></td>
<td>0.942</td>
<td>0.952</td>
<td>0.968</td>
</tr>
<tr>
<td></td>
<td>-0.092</td>
<td>-0.002</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td>0.055</td>
<td>0.081</td>
<td>0.088</td>
</tr>
<tr>
<td></td>
<td>0.070</td>
<td>0.103</td>
<td>0.147</td>
</tr>
<tr>
<td></td>
<td>0.592</td>
<td>0.734</td>
<td>0.670</td>
</tr>
<tr>
<td></td>
<td>-0.012</td>
<td>0.006</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>0.188</td>
<td>0.012</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>0.011</td>
<td>0.003</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>-0.014</td>
<td>0.007</td>
<td>-0.020</td>
</tr>
<tr>
<td></td>
<td>0.128</td>
<td>0.124</td>
<td>0.124</td>
</tr>
<tr>
<td></td>
<td>0.126</td>
<td>0.100</td>
<td>0.101</td>
</tr>
<tr>
<td></td>
<td>0.962</td>
<td>0.962</td>
<td>0.962</td>
</tr>
<tr>
<td></td>
<td>0.103</td>
<td>0.910</td>
<td>0.906</td>
</tr>
<tr>
<td></td>
<td>0.362</td>
<td>0.936</td>
<td>0.936</td>
</tr>
</tbody>
</table>

in the same magnitudes for the three data sets under both models, and race and grade both have significant impact on the cure fraction. When race changes from white to black, the net survival probability becomes smaller. With the level of cancer cell grade increases, the net survival probability decreases. The effect of background mortality on the predicted survival probability for younger patients group is smaller than that for elder patients group. Figure 3.1 - 3.3 show that the predicted survival probabilities by race are higher by the proposed semiparametric PCPH+BM model and the white people have higher net survival probability than the black people, which are consistent with the estimation in Table 3.5.
Table 3.5: SEER regional female breast cancer: parameters estimation in the semiparametric PCPH+BM model and the semiparametric PCPH model, race and grade are included in both the cure link function and latency survival function. Model I is to fit the whole data set with 6200 observations, Model II is to fit the subgroup data of age $\leq 70$, and model III is to fit the subgroup data of age $> 70$.

<table>
<thead>
<tr>
<th>Par</th>
<th>Est</th>
<th>StErr</th>
<th>CI</th>
<th>Est</th>
<th>StErr</th>
<th>CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PCPH+BM</td>
<td>PCPH</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>γ₀</td>
<td>-1.477</td>
<td>0.111 (-1.695,-1.259)</td>
<td>-0.477</td>
<td>0.065</td>
<td>(-0.605,-0.349)</td>
</tr>
<tr>
<td>γ₁</td>
<td>0.505</td>
<td>0.070 (0.369,0.642)</td>
<td>0.113</td>
<td>0.047 (0.022,0.205)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>γ₂</td>
<td>0.076</td>
<td>0.016 (0.045,0.106)</td>
<td>0.028</td>
<td>0.010 (0.008,0.048)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>β₁</td>
<td>0.318</td>
<td>0.039 (0.243,0.394)</td>
<td>0.195</td>
<td>0.041 (0.114,0.276)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>β₂</td>
<td>0.063</td>
<td>0.014 (0.036,0.090)</td>
<td>0.058</td>
<td>0.012 (0.035,0.081)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cure-W</td>
<td>0.623</td>
<td>0.013 (0.598,0.648)</td>
<td>0.470</td>
<td>0.009 (0.452,0.488)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cure-B</td>
<td>0.456</td>
<td>0.020 (0.417,0.495)</td>
<td>0.429</td>
<td>0.017 (0.396,0.462)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Par</th>
<th>Est</th>
<th>StErr</th>
<th>CI</th>
<th>Est</th>
<th>StErr</th>
<th>CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>γ₀</td>
<td>-1.740</td>
<td>0.126 (1.946,-1.493)</td>
<td>-1.083</td>
<td>0.091 (-1.262,-0.904)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>γ₁</td>
<td>0.506</td>
<td>0.077 (0.355,0.657)</td>
<td>0.277</td>
<td>0.061 (0.158,0.397)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>γ₂</td>
<td>0.081</td>
<td>0.017 (0.046,0.115)</td>
<td>0.039</td>
<td>0.014 (0.012,0.066)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>β₁</td>
<td>0.193</td>
<td>0.055 (0.085,0.301)</td>
<td>0.201</td>
<td>0.053 (0.097,0.305)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>β₂</td>
<td>0.047</td>
<td>0.022 (0.004,0.090)</td>
<td>0.052</td>
<td>0.016 (0.022,0.083)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cure-W</td>
<td>0.691</td>
<td>0.013 (0.666,0.716)</td>
<td>0.606</td>
<td>0.012 (0.582,0.630)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cure-B</td>
<td>0.542</td>
<td>0.020 (0.503,0.581)</td>
<td>0.516</td>
<td>0.020 (0.477,0.555)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Par</th>
<th>Est</th>
<th>StErr</th>
<th>CI</th>
<th>Est</th>
<th>StErr</th>
<th>CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>γ₀</td>
<td>-1.466</td>
<td>0.239 (1.933,-0.998)</td>
<td>0.225</td>
<td>0.096 (0.037,0.413)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>γ₁</td>
<td>0.584</td>
<td>0.158 (0.274,0.894)</td>
<td>0.145</td>
<td>0.082 (-0.015,0.305)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>γ₂</td>
<td>0.123</td>
<td>0.033 (0.057,0.188)</td>
<td>0.015</td>
<td>0.017 (-0.019,0.048)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>β₁</td>
<td>0.646</td>
<td>0.050 (0.549,0.743)</td>
<td>0.160</td>
<td>0.075 (0.012,0.307)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>β₂</td>
<td>0.070</td>
<td>0.015 (0.039,0.100)</td>
<td>0.078</td>
<td>0.020 (0.039,0.118)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cure-W</td>
<td>0.481</td>
<td>0.029 (0.424,0.538)</td>
<td>0.221</td>
<td>0.015 (0.192,0.250)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cure-B</td>
<td>0.312</td>
<td>0.050 (0.214,0.410)</td>
<td>0.174</td>
<td>0.028 (0.119,0.229)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.1: Predicted Survival Probability for the SEER regional female breast cancer data, the proposed PCPH+BM model estimates cure fraction higher than the standard PCPH model.
3.6 Discussions

The semiparametric PCPH+BM model has desirable performance in estimation of parameters and cure fraction, and is comparable to fully parametric methods only when the baseline distribution can be correctly specified. Thus, the proposed semi-

Figure 3.2: Predicted Survival Probability for the subgroup of age $\leq 70$ in the SEER regional female breast cancer data. The proposed PCPH+BM model estimates cure fraction higher than the standard PCPH model.

Figure 3.3: Predicted Survival Probability for the subgroup of age $> 70$ in the SEER regional female breast cancer data. The proposed PCPH+BM model estimates cure fraction higher than the standard PCPH model.
parametric PCPH+BM model is more flexible in estimation and largely avoid the problem of misspecification.

There are different properties for the proposed PCPH+BM model, compared to the MCPH+BM model in Chapter 2. First, in terms of estimation shown in Table 2.6 and 3.5, all the regression parameters are significant with the corresponding confidence interval not containing 0. The point estimators are all in the same magnitude. The cure rates estimation are pretty close from these two models. In Model I for the whole data set, the PCPH+BM model estimates the cure rates for the white people and black people as 62.3% and 45.6%, respectively, compared to 60.5% and 48.7% under the MCPH+BM model. However, the MCPH+BM model is more efficient in computation compared to the PCPH+BM model. The former uses a perturbation variance estimation and the latter adapts resampling variance estimation. Previous studies on the semiparametric EM algorithm in analyzing survival data with a cure fraction have shown the estimators are asymptotic normality and there is no potential issue in simulation studies (Peng and Dear, 2000; Peng and Carriere, 2002; Peng, 2003b; Peng, 2003a; Peng and Zhang, 2008; Cai et al., 2012). The proposed PCPH+BM model and the MCPH+BM model in Chapter 2 are both to estimate parameters via semiparametric EM algorithm, which has background survival probabilities and background hazard rates involved in calculation as constant values, and the asymptotic normality of parameters estimation still hold. The simulation studies in Section 3.4 also support this claim.

Secondly, the AIC can not be used in semiparametric estimation directly. However, the model comparison can be conducted by using the difference of AIC values from the PCPH+BM model and the MCPH+BM model (-2ll) since both models have the same number of parameters. In the breast cancer data analysis, Model I, Model II and Model III give $-2ll_{\text{Model I}} = 784.976$, $-2ll_{\text{Model II}} = 498.288$ and $-2ll_{\text{Model III}} = 742.842$, respectively. Specifically, for the SEER female regional breast
cancer data, regardless of for young people or elder people, the PCPH+BM model always estimates the AIC smaller than the MCPH+BM model. Thus, the PCPH+BM model is selected.

Thirdly, in terms of model structure, the MCPH+BM model is a mixture of two components, uncured patients and cured patients. The subjects in the MCPH+BM model have been labeled into two categories. The PCPH+BM model has biological deviation involving the number of metastasizing-competent tumor cells. From this perspective, the model selection is also based on the suggestions from domain experts and research preference.

We also consider to extend the PCPH+BM model to the semiparametric transformation cure models (Zeng, Yin, and Ibrahim, 2006). In the PCPH+BM model, the assumption that the promotion time \((\tilde{T}_1, \ldots, \tilde{T}_k)\) are mutually independent, may not be realistic since they are unobserved random variables taken on the same subject. Therefore, a subject-specific frailty \(\xi_i\) is introduced such that conditional on both \(N_i = k\) and \(\xi_i\), the promotion time \((\tilde{T}_1, \ldots, \tilde{T}_k)\) are mutually independent with distribution function \(F(t)\). Also, conditional on \(z_i\) and \(\xi_i\), \(N_i\) has a Poisson distribution with rate \(\xi_i \theta(z_i)\), thus \(\xi_i\) represents the heterogeneity of the Poisson rates in the \(N_i\)’s.
CHAPTER 4
REGRESSION ANALYSIS OF RIGHT-CENSORED DATA
UNDER THE GENERALIZED ODDS-RATE MODEL

ABSTRACT

Generalized odds-rate (GOR) models, also referred to as $G^\rho$ family in the literature, are a general class of semiparametric regression models taking the classic proportional hazards model as a limiting case and the proportional odds model as a special case. Although many approaches have been proposed for analyzing right-censored data using the GOR models in the literature, most of these studies have assumed $\rho$ to be known and reported that estimating $\rho$ together with the regression parameters is problematic. This article investigates the identifiability issues associated with the GOR models when treating $\rho$ as unknown and proves that all parameters are identifiable in the conventional regression settings. An novel estimation approach based on the EM algorithm is proposed for analyzing right-censored data using the GOR models treating $\rho$ as unknown. The proposed approach provides variance estimation in closed-form. A comprehensive simulation study has shown that the proposed approach always has an excellent performance in estimating the regression parameters and the baseline survival function, but can estimate $\rho$ accurately only when sample size is very large. The proposed approach is also illustrated by a real data set from the SEER lung cancer study.
4.1 Introduction

The proportional hazards (PH) model is the most popular survival models for analyzing censored data in the literature. However, the PH model assumption that the hazards are proportional for different subgroups implies that the corresponding survival functions are parallel for these subgroups. This is also violated in real life data analysis. One possible remedy of this problem is to apply some more flexible regression models. The generalized odds-rate (GOR) models are a class of more general semiparametric regression models. The survival function in the GOR models is expressed as

\[ S(t\|x) = \{1 + \rho \Lambda_0(t) \exp(x^T \beta)\}^{-\rho^{-1}}, \] (4.1)

where \( \Lambda_0(t) \) is a strictly positive increasing function, \( \beta \) is a \( p \times 1 \) vector of regression parameters denoting covariate effects, and \( \rho \) is a positive constant. Specifically, as \( \rho \to 0 \), equation (4.1) leads to a limiting survival function \( S(t\|x) = \exp\{-\Lambda_0(t)\exp(x^T \beta)\} \) under the Cox PH model. When \( \rho = 1 \), equation (4.1) becomes the PO model survival function \( S(t\|x) = \{1 + \Lambda_0(t) \exp(x^T \beta)\}^{-1} \). The GOR models are also well recognized as a special class of linear transformation models in the form of \( g_{\rho}\{S(t\|x)\} = \alpha(t) + x^T \beta \) with \( g_{\rho}(s) = \log\{(s^{-\rho} - 1)/\rho\} \) for \( \rho > 0 \). Note that \( \alpha(t) = \log\{\Lambda_0(t)\} \) or in the form of \( \alpha(T) = -x^T \beta + \epsilon \), where \( \exp(\epsilon) \) follows a Pareto distribution with \( \rho > 0 \). The term of \( (s^{-\rho} - 1)/\rho = \Lambda_0(t) \exp(x^T \beta) \) is the Box-Cox transformation. When \( x = 0 \), the function \( \Lambda_0(t) \) can be interpreted as the Box-Cox transformed baseline survival function, and \( 1/\{1 + \Lambda_0(t)\} \) is the frailty PH model survival function. It is worth noting that the linear transformation models in the literature do not contain the additional parameter \( \rho \) in the transformation function \( g \). The popular gamma frailty PH models, which are used for modeling multivariate or clustered failure times, have the GOR models as the marginal distributions of the failure times.
Many approaches have been proposed for analyzing right-censored data by using the GOR models. Harrington and Fleming (1982) proposed the $G^\rho$ statistic for testing the regression parameters in the two-sample setting. Cheng, Wei, and Ying (1995) and Cheng, Wei, and Ying (1997) proposed a semiparametric approach to estimate the regression parameters. Scharfstein, Tsiatis, and Gilbert (1998) studied the NPMLE for the regression parameters when the time-to-event outcome is subject to right censoring conditional on $\rho$ being fixed and known. Chen, Jin, and Ying (2002) developed simple martingale-based estimating equations to estimate regression parameters. The theoretical properties and hypothesis tests of the NPMLE for general transformation models have been studied by Bagdonavicius and Nikulin (1999); Kosorok, Lee, and Fine (2004); Zeng and Lin (2006) and Song, Kosorok, and Fine (2009). Some Bayesian approaches have also been developed based on the GOR models. For example, Hanson and Yang (2007); Hanson and Jara (2013) used a mixture of finite Polya trees prior for the baseline survival function, and Banerjee et al. (2007) adopted the piece-wise exponential function for the baseline cumulative hazard function.

Previous studies on the GOR models are mainly focused on the regression parameters estimation by treating $\rho$ as fixed and known. However, such strategy reduces the flexibility of the GOR models. The simulation study in Zeng, Yin, and Ibrahim (2006) indicates that the performance of the NPMLEs is poor and the convergence is problematic with a sample size of 400 considering the estimation of $\rho$, since the likelihood function tends to be flat when $\rho$ varies around the true value. In general it is reported that $\rho$ can not be estimated accurately in the literature. A natural question is whether the parameters in the GOR models are identifiable.

In this article, we investigate the identifiability problem of the GOR models. Our investigation has found that the GOR models are non-identifiable in the case when there are no covariates, but are indeed identifiable in the usual regression settings
including the commonly seen two-sample or \(k\)-sample settings. We further propose a joint estimating approach of the parameter \(\rho\), the regression parameters, and the baseline parameters for right-censored data. The proposed approach allows one to conduct model selection among the GOR models by making inference on \(\rho\) in addition to the conventional interests of identifying significant risk factors and estimating their effects. The rest of the paper is organized as follows. Section 4.2 discusses the identifiability issues of the GOR models. Section 4.3 gives the details of the proposed approach for the regression analysis of right-censored data. Section 4.4 outlines the simulation studies and Section 4.5 applies the proposed method to the real data. Section 4.6 provides some conclusions and discussions.

4.2 Identifiability of GOR models

For the GOR models in equation (4.1), previous studies have theoretically proved the following results for the GOR models without or with covariates (Yao, 2016).

**Fact:** The GOR models with \(x^\top \beta = 0\) are non-identifiable.

For the regression settings, the first case is to include a binary covariate, which is called two-sample setting. The second case is the general setting when there are a mixture of continuous and binary covariates. Assuming the \(p\) covariates are not linearly correlated, which means that not any single covariate can not be written as a linear combination of the others, we have the following results.

**Theorem 1:** The GOR models defined in (4.1) are identifiable in the case of one binary covariate.

**Theorem 2:** The GOR models defined in (4.1) are identifiable in general regression settings.

To explore the identifiability of \(\rho\) by real number examples, we consider the following cases:
(i) $X$ is generated from bernoulli(0.5) with possible values 0 and 1. Let $t$ is a sequence of value changing from 0 to 50 with step of 0.1. Let $\Lambda_{01}(t) = t$, $\beta = -1$ and $\rho_1 = 1$. Suppose $\Lambda_{02}(t) = a\Lambda_{01}(t) + b\Lambda_{01}^2(t) + c\Lambda_{01}^3(t)$ which is a transformation of the function $\Lambda_{01}(t)$, and $\beta_2$ is unknown and to be estimated. The estimators $\rho_2 = 1.43$, $\hat{a} = 1.092$, $\hat{b} = 0.096$, $\hat{c} = -0.001$ and $\hat{\beta}_2 = -1.152$ are found to make the maximum of $|S_1(t|X) - S_2(t|X)| = 0.00987 < 0.01$. Figure 4.1 shows the difference between the function $\Lambda_{01}(t)$ and $\Lambda_{02}(t)$ in case i.

(ii) Generate $X$ from bernoulli(0.5) with possible values 0 and 1. Let $t$ be a sequence of values changing from 0 to 50 with step of 0.1. Let $\Lambda_{01}(t) = \log(1 + t) + t^{1.5}$ which is a concave function, $\beta = -1$ and $\rho_1 = 1$. Suppose $\Lambda_{02}(t) = a\Lambda_{01}(t) + b\Lambda_{01}^2(t) + c\Lambda_{01}^3(t)$, and $\beta_2$ is unknown and to be estimated. A set of parameters $\rho_2 = 1.250$, $\hat{a} = 1.086$, $\hat{b} = 0.033$, $\hat{c} = -8e - 05$ and $\hat{\beta}_2 = -1.105$ make the maximum of $|S_1(t|X) - S_2(t|X)| = 0.00978 < 0.01$. Figure 4.2 shows the difference between function $\Lambda_{01}(t)$ and $\Lambda_{02}(t)$ in case ii.
4.3 The Proposed Method

4.3.1 The GOR model

For the seek of easy computation, we reparameterize \( \tau = 1/\rho \) and write the survival function of the GOR model as

\[
S(t|x) = \{1 + \tau^{-1} \Lambda_0(t) \exp(x^\top \beta)\}^{-\tau}.
\]

Under this specification, the corresponding density function and hazard function are

\[
f(t|x) = \lambda_0(t) \exp(x^\top \beta) \{1 + \tau^{-1} \Lambda_0(t) \exp(x^\top \beta)\}^{-\tau-1}
\]

and

\[
h(t|x) = \lambda_0(t) \exp(x^\top \beta) \{1 + \tau^{-1} \Lambda_0(t) \exp(x^\top \beta)\}^{-1},
\]

respectively, where \( \lambda_0(t) \) is the first derivative of \( \Lambda_0(t) \).
4.3.2 The observed data and likelihood

Let $O_i = (y_i, \delta_i, x_i)$ denote the observed data for subject $i$ ($i = 1, 2, \ldots, n$), where $y_i$ is the observation time as the minimizer of failure times $T_i$ and censoring time $C_i$, $\delta_i = I(T_i \leq C_i)$ is the censoring indicator taking 1 for exactly observed and 0 for right-censored failure time, and $x_i$ is a vector of covariates for subject $i$. Under the independence assumption between the failure time and censoring time, the observed likelihood function based on observed data $O$ takes the following form

$$L_{obs} = \prod_{i=1}^{n} \left[ \lambda_0(y_i) \exp(x_i^T \beta) \{1 + \tau^{-1} \Lambda_0(y_i) \exp(x_i^T \beta)\}^{-\delta_i} \{1 + \tau^{-1} \Lambda_0(y_i) \exp(x_i^T \beta)\}^{-\tau} \right] \delta_i \{1 + \tau^{-1} \Lambda_0(y_i) \exp(x_i^T \beta)\}^{-\tau}$$

and the logarithm observed likelihood function is

$$\ell_{obs} = \sum_{i=1}^{n} \left( \delta_i \log \{\lambda_0(y_i)\} + \delta_i x_i^T \beta - \delta_i \log \{1 + \tau^{-1} \Lambda_0(y_i) \exp(x_i^T \beta)\} \right) - \tau \log \{1 + \tau^{-1} \Lambda_0(y_i) \exp(x_i^T \beta)\}.$$ (4.2)

The main research interests in the GOR models are to assess the covariate effects and to estimate the survival functions for different subgroups for prediction purpose. These require one to estimate the unknown parameters $\theta = \{\rho, \beta, \lambda_0(\cdot)\}$. Estimating $\rho$ under the GOR models has been regarded as a tough issue in the literature for right-censored data as mentioned in the introduction section.

4.3.3 Monotone Splines

The cumulative baseline hazard function $\Lambda_0(\cdot)$ is an unspecified increasing function, and it has infinite-dimensional parameters. Using splines is popular to model non-parametric functions, and it leads to a finite number of parameters to estimate while maintaining adequate modeling flexibility by not assuming a specific shape for the unknown functions. We propose to model $\Lambda_0(\cdot)$ with the monotone splines (Ramsay, 1988)

$$\Lambda_0(t) = \sum_{l=1}^{k} \gamma_l b_l(t),$$
where $b_l$’s are nondecreasing integrated spline basis functions ranging from 0 to 1, and $\gamma = (\gamma_1, \ldots, \gamma_L)$ are nonnegative spline coefficients which can ensure $\Lambda_0(\cdot)$ is nondecreasing. For spline basis functions, the degree and knots need to be specified within a time range (Ramsay, 1988). The degree of basis functions controls the overall smoothness of the basis functions; For instance, the degree $d = 1, 2$ and $3$ correspond to linear, quadratic and cubic basis functions, respectively, and the specifying of degree of 2 or 3 would attain adequate smoothness. The placement of knots controls the overall modeling flexibility such that more knots in a region provides greater modeling flexibility. The choice of knots in the spline functions can be either equally spaced or at quartiles (Ramsay, 1988). The number of basis functions $k$ is determined by $k = d + m - 2$, which is the sum of the degree and the number of interior knots, then the spline basis functions will be fully determined with degree and placement of knots. Ramsay (1988) suggested using a small number of strategically placed interior knots at median or quartiles. The number of interior knots is further recommended using approximately 10 to 30 (Wang and Dunson, 2011; Wang and Lin, 2011). The Akaike’s information criterion (AIC) can be used in selecting the number of knots in real data application (McMahan, Wang, and Tebbs, 2013).

4.3.4 EM Algorithm

The novel EM algorithm we propose is to estimate $\theta = (\Lambda_0(\cdot), \beta, \tau)$ with data augmentation. In the first stage, by introducing a gamma frailty variable $\phi_i \sim Ga(\tau, \tau)$ for subject $i = 1, \ldots, n$ where $\tau$ is the shape and rate of gamma distribution, the equation (4.2) involves the survival function

$$S(t|x) = \int_0^\infty e^{-\Lambda_0(t)\exp(x^t\beta)\phi} \frac{\tau^{\tau-1}e^{-\tau\phi}}{\Gamma(\tau)} d\phi.$$  \hspace{1cm} (4.3)
We further introduce a second-stage latent variable $\psi_i \sim Ga(1, \tau)$ only for those exactly observed subjects with $\delta_i = 1$ such that

$$\{1 + \tau^{-1} \Lambda_0(t) \exp(x_i^T \beta)\}^{-1} = \int_0^{\infty} e^{-\Lambda_0(t) \exp(x_i^T \beta) \psi_i} \tau e^{-\tau \psi_i} d\psi_i.$$

The third-stage latent variable is a $Z_i \sim \text{multinomial}[1, (\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k})]$ only for those exactly observed subjects with $\delta_i = 1$. Hence, the complete likelihood function with the three sets of latent variables is written as

$$L_c = \prod_{i=1}^n \left[ \prod_{l=1}^{k} \{\gamma_l M_l(y_i)\} Z_{il} \exp(x_i^T \beta) \exp\{-\Lambda_0(y_i) \exp(x_i^T \beta) \psi_i - \tau \psi_i\} \right]^{\delta_i}$$

$$\times \exp\{-\Lambda_0(y_i) \exp(x_i^T \beta) \phi_i\} \frac{\tau^{\phi_i - 1} e^{-\tau \phi_i}}{\Gamma(\tau)}.$$

In the EM algorithm, the E-step is to take the conditional expectation of the logarithm complete likelihood function (4.4) with respect to the latent variables $\phi_i$'s, $\psi_i$'s, and $Z_{il}$'s, given the observed data $O$ and the current parameters $\theta^{(d)}$ at $d$th iteration, which is denoted as $Q(\theta, \theta^{(d)}) = E[\log\{L_c(\theta)\}|O; \theta^{(d)}]$. The expected logarithm complete likelihood functions can be written as the summation of two parts,

$$Q_1(\beta, \gamma, \theta^{(d)}) = \sum_{i=1}^n \delta_i \left[ \sum_{l=1}^{k} E(Z_{il}|O; \theta^{(d)}) \log\{\gamma_l M_l(y_i)\} \right] + \delta_i x_i^T \beta$$

$$- \delta_i \sum_{l=1}^{k} \gamma_l b_l(y_i) \exp(x_i^T \beta) E(\psi_i|O; \theta^{(d)})$$

$$- \sum_{l=1}^{k} \gamma_l b_l(y_i) \exp(x_i^T \beta) E(\phi_i|O; \theta^{(d)}),$$

and

$$Q_2(\tau, \theta^{(d)}) = \sum_{i=1}^n -\delta_i \tau E(\psi_i|O; \theta^{(d)}) + \delta_i \log(\tau) - \tau E(\phi_i|O; \theta^{(d)}) + \tau \log(\tau)$$

$$+ (\tau - 1) E(\log \phi_i|O; \theta^{(d)}) - \log\{\Gamma(\tau)\}.$$
and \( \lambda_0(y_i) = \sum_{l=1}^{k} \gamma_l M_l(y_i) \). The conditional expectations involved in equation (4.5) and (4.6) all have explicit forms as follows,

\[
E(\psi_i|O; \theta^{(d)}) = \frac{\delta_i}{\Lambda_0^{(d)}(y_i) \exp(x_i^T \beta^{(d)}) + \tau^{(d)}},
\]

\[
E(\phi_i|O; \theta^{(d)}) = \frac{\tau}{\Lambda_0^{(d)}(y_i) \exp(x_i^T \beta^{(d)}) + \tau^{(d)}},
\]

\[
E(Z_{ij}|O; \theta^{(d)}) = \frac{\gamma_j^{(d)} M_j(y_i)}{\sum_{l=1}^{k} \gamma_l^{(d)} M_l(y_i)} \delta_{ij},
\]

and

\[
E\{\log(\phi)|O; \theta^{(d)}\} = \frac{\Gamma'(\tau)}{\Gamma(\tau)} - \log\{\Lambda_0^{(d)}(y_i) \exp(x_i^T \beta^{(d)}) + \tau^{(d)}\}.
\]

The M-step is to maximize the equations (4.5) and (4.6) with respect to the unknown parameters \( \theta \) by taking partial derivatives and setting them to zero such that,

\[
\frac{\partial Q(\theta, \theta^{(d)})}{\partial \gamma_j} = \sum_{i=1}^{n} \delta_i E(Z_{ij}|O; \theta^{(d)}) \frac{\delta_i \exp(x_i^T \beta) E(\psi_i|O; \theta^{(d)}) b_j(y_i)}{\gamma_j} - \delta_i \exp(x_i^T \beta) E(\phi_i|O; \theta^{(d)}) b_j(y_i) = 0
\]

(4.7)

\[
\frac{\partial Q(\theta, \theta^{(d)})}{\partial \beta} = \sum_{i=1}^{n} \delta_i x_i - \delta_i \left\{ \sum_{j=1}^{k} \gamma_j b_j(y_i) \right\} E(\psi_i|O; \theta^{(d)}) \exp(x_i^T \beta) x_i
\]

(4.8)

\[
\frac{\partial Q(\theta, \theta^{(d)})}{\partial \tau} = \sum_{i=1}^{n} -\delta_i E(\psi_i|O; \theta^{(d)}) + \delta_i \tau^{-1} - E(\phi_i|O; \theta^{(d)}) + \log(\tau) + 1
\]

(4.9)

\[
+ E(\log(\phi_i)|O; \theta^{(d)}) \frac{\Gamma'(\tau)}{\Gamma(\tau)} = 0.
\]

Note that the equation (4.7) leads to a closed form for \( \gamma_j \) as a function of \( \beta \),

\[
\gamma_j^{*}(\beta) = \frac{\sum_{i=1}^{n} \delta_i E(Z_{ij}|O; \theta^{(d)}) \exp(x_i^T \beta) E(\psi_i|O; \theta^{(d)}) b_j(y_i) + \exp(x_i^T \beta) E(\phi_i|O; \theta^{(d)}) b_j(y_i)}{\sum_{i=1}^{n} \delta_i \exp(x_i^T \beta) E(\psi_i|O; \theta^{(d)}) b_j(y_i) + \exp(x_i^T \beta) E(\phi_i|O; \theta^{(d)}) b_j(y_i)},
\]

(4.9)

\( j = 1, \ldots, k \).

Thus, one can replace \( \gamma_j \) with \( \gamma_j^{*}(\beta) \) for each \( j \) in the equation (4.8) and solve for \( \beta^{(d+1)} \). Then \( \gamma_j^{(d+1)} \) is obtained as \( \gamma_j^{*}(\beta^{(d+1)}) \), for \( j = 1, \ldots, k \). The proposed EM algorithm to estimate parameters \( \theta = (\gamma_{j=1,2,\ldots,k}, \beta, \tau) \) can be summarized as follows,
Step 0: Set initial values \( \beta^0 \) as 0.1, \( \tau^0 \) as 1, and the spline parameters \( r^{(0)}_{j=1,2,...,k} \) as 0.1;

Step 1: Obtain \( \beta^{(d+1)} \) by solving the \( p \) equations in (4.8), where

\[
\gamma_{j}^{(d)}(\beta) = \frac{\sum_{i=1}^{n} \delta_i E(Z_{ij}|O; \theta^{(d)})}{\sum_{i=1}^{n} \delta_i \exp(x_i^T \beta) E(\psi_i|O; \theta^{(d)}) b_j(y_i) + \exp(x_i^T \beta) E(\phi_i|O; \theta^{(d)}) b_j(y_i)},
\]

\( j = 1, \ldots, k \).

Step 2: Update \( \gamma_{j}^{(d+1)} = \gamma_{j}^{(d)}(\beta^{(d+1)}) \) for \( j = 1, \ldots, k \) and increase \( d \) by 1.

Step 3: Obtain \( \tau^{(d+1)} \) by solving the equation

\[
\sum_{i=1}^{n} -\delta_i E(\psi_i|O; \theta^{(d)}) + \delta_i \tau^{-1} - E(\phi_i|O; \theta^{(d)}) + \log(\tau) + 1 + E(\log \phi_i|O; \theta^{(d)}) \Gamma'(\tau) / \Gamma(\tau) = 0.
\]

Repeat Step 1 - Step 3 until convergence. The convergence of the EM algorithm is claimed when the maximum change of all elements of \( \beta \), \( \gamma_j \)'s and \( \tau \) between successive iterations is less than a prespecified small value \( \epsilon \), say \( 10^{-7} \). The proposed EM algorithm has two likelihood functions involving separate parameters which is appealing in computation. Moreover, the updated \( \gamma_{j}^{(d+1)} \) can be guaranteed to be nonnegative from equation (4.9), so there is no constrain needed when estimate \( \gamma_j \). The estimator for \( \rho \) is transformed back as \( \hat{\rho} = \hat{\tau}^{-1} \).

For variance estimation, let \( \hat{\theta} \) denote the converged values of \( \theta^{(d)} = (\gamma^{(d)\prime}, \beta^{(d)\prime}, \tau^{(d)\prime}) \) and \( \hat{\theta} \) is a maximum likelihood estimator of \( \theta \). Variance can be estimated by Louis’s method (Louis, 1982) which provides closed-form expressions for all parameters. Specifically, \( \text{var}(\hat{\theta}) \) is the inverse of the observed information matrix \( I(\hat{\theta}) \) and \( I(\theta) \) is obtained using the missing information principle as follows,

\[
I(\theta) = -\frac{\partial^2 Q(\theta, \hat{\theta})}{\partial \theta \partial \theta^\prime} - \text{var}
\left( \frac{\partial \log L_c(\theta)}{\partial \theta} | O, \hat{\theta} \right).
\] (4.10)

All the quantities involved to calculate these two terms in equation (4.10) have closed-form expressions, and the details are shown in the Appendix A. Those closed-form expressions make the variance estimation easy to compute, which is another appealing point in the proposed approach.
The proposed approach allows one to make inference on \( \rho \). This is appealing since it provides plausible values of \( \rho \) based on the 95\% confidence interval for a specific data set using the GOR model. Also, we can validate the PO model assumption by conducting a formal hypothesis test about \( \rho = 1 \).

As the PH model is the limiting case of the GOR models when \( \rho \to 0 \), a simplified version of the proposed EM algorithm can be used for the PH model. We provide some details of the EM algorithm and the variance estimation under the PH model in Appendix B.

4.4 Simulation

A comprehensive simulation study was conducted to assess the performance of the proposed approach. The failure time was generated from

\[
S(t|x_i) = \{1 + \rho \Lambda_0(t) \exp(x_i^\top \beta)\}^{-\rho^{-1}},
\]

where \( x_i = (x_{i1}, x_{i2}) \) are two representative continuous and discrete covariates with \( x_{i1} \sim N(0,1) \) and \( x_{i2} \sim \text{Bernoulli}(0.5) \). The true baseline cumulative function was taken to be a concave function \( \Lambda_0(t) = \log(1 + t) + t^{1.5} \) and a convex function \( \Lambda_0(t) = 0.5 t^{5/6} \) in two simulation scenarios. The regression parameters were set as \( \beta = c(-1,-1) \), and \( \rho \) took on 0.5, 1, 2 and as well as two relatively extreme values of 0.25 and 4. The right-censoring time \( C_i \) was generated from an exponential distribution with parameter \( c \) to control the censoring rate. For example, in the concave function \( \Lambda_0(t) = \log(1 + t) + t^{1.5} \) and \( \rho = 1 \), taking \( c \) to be 0.1, 0.4, and 3 corresponds to a right-censoring rate about 20\%, 40\%, and 75\%, respectively. We generated 500 independent data sets, each with a sample size of \( n = 200, 500, 800, 1000 \) and 1500 for each parameter configuration.

For each data set, we ran the proposed method described in Section 4.3 under the GOR model. We used interior knots of 10 and degree of 3 for adequate smoothness for
the monotone spline specification. The initial values for \( \beta_1 \) and \( \beta_2 \) were both 0.1, the initial value for \( \tau \) was 1, and the initial values for \( \gamma_\ell s \) were set as 0.1. The convergence was claimed when the maximum changes of all parameters for two successive iterations was smaller than \( 10^{-7} \).

The summary statistics include the difference between the average of 500 point estimates and true values (bias), the average of 500 estimated standard errors (StErr), the empirical standard deviations of 500 point estimates (StDev) and the 95% coverage probability (CP). Tables 4.1 and 4.2 illustrate that the proposed method gives good estimation in large sample size of 1000 or 1500 in both concave and convex baseline cumulative baseline functions. As seen, biases are small for all parameters; StErrs are close to StDevs indicating the variance estimation based on Louis’s method are accurate; all the 95% CPs are close to the nominal value 0.95 which indicates the asymptotic normality of the parameter estimation is valid. It is also observed that the StErrs tend to become smaller as \( \rho \) changes from 2 to 0.5, since larger \( \rho \) results in less information about the failure time in data. High censored rate data tends to have larger variances estimation than low censored rate data. Figures 4.3 and 4.10 show that the estimated baseline survival probability is very close to the true baseline survival function for low censored rate and medium censored rate data, and the corresponding confidence interval becomes narrower when sample size increases. The weak identifiability issue in the GOR models makes the estimation of the regression parameters \( \beta \) and the parameter \( \rho \) all less precise since all the parameters react with each other in the EM algorithm.

In Tables 4.3 and 4.4, we explored the performance of the proposed EM algorithm in extreme cases of \( \rho = 4 \) when frailty variance is large and \( \rho = 0.25 \) when frailty variance is small in the equation (4.3). In the setup of \( \rho = 4 \) for low censored rate data, larger sample size such as 2500 is needed to provide good estimation, while on the contrary, the performance of the proposed EM algorithm is good enough in the
Table 4.1: Summary Statistics of the proposed method with concave function $\Lambda_0(t) = \log(1 + t) + r^{1.5}$, $\beta = c(-1, -1)$, $\rho = 2$, $\rho = 1$ and $\rho = 0.5$ with 3 levels censoring rate (around 20% in I: Low Censored, 40% in II: Medium Censored and 75% in III: High Censored).

<table>
<thead>
<tr>
<th></th>
<th>$\rho = 2$</th>
<th>$\rho = 1$</th>
<th>$\rho = 0.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Par</td>
<td>Bias</td>
<td>StErr</td>
</tr>
<tr>
<td>I: Low Censored</td>
<td>n=200</td>
<td>$\beta_1$</td>
<td>-0.074</td>
</tr>
<tr>
<td></td>
<td>n=500</td>
<td>$\beta_2$</td>
<td>-0.031</td>
</tr>
<tr>
<td></td>
<td>n=800</td>
<td>$\rho$</td>
<td>0.046</td>
</tr>
<tr>
<td>II: Medium Censored</td>
<td>n=200</td>
<td>$\beta_1$</td>
<td>-0.013</td>
</tr>
<tr>
<td></td>
<td>n=500</td>
<td>$\beta_2$</td>
<td>-0.016</td>
</tr>
<tr>
<td></td>
<td>n=800</td>
<td>$\rho$</td>
<td>0.048</td>
</tr>
<tr>
<td>III: High Censored</td>
<td>n=200</td>
<td>$\beta_1$</td>
<td>-0.003</td>
</tr>
<tr>
<td></td>
<td>n=500</td>
<td>$\beta_2$</td>
<td>-0.008</td>
</tr>
<tr>
<td></td>
<td>n=800</td>
<td>$\rho$</td>
<td>0.027</td>
</tr>
<tr>
<td></td>
<td>n=1000</td>
<td>$\beta_1$</td>
<td>-0.013</td>
</tr>
<tr>
<td></td>
<td>n=1500</td>
<td>$\rho$</td>
<td>0.046</td>
</tr>
</tbody>
</table>

The proposed EM algorithm was compared to the direct MLE algorithm in Table 4.5 for different censored rates data, and the former algorithm always performs better

The sample size of 500 for $\rho = 0.25$. This makes sense because the larger frailty variance is, the less information we can get from the data.
Table 4.2: Summary Statistics of the proposed method with convex function $\Lambda_t(t) = 0.5^{t/6}$, $\beta = c(1,-1)$, $\rho = 2$, $\rho = 1$ and $\rho = 0.5$ with 3 levels censoring rate (around 20% in I: Low Censored, 40% in II: Medium Censored and 75% in III: High Censored).

<table>
<thead>
<tr>
<th>Par</th>
<th>$\rho = 2$</th>
<th>$\rho = 1$</th>
<th>$\rho = 0.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>-0.124</td>
<td>-0.124</td>
<td>-0.026</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.126</td>
<td>-0.126</td>
<td>-0.200</td>
</tr>
<tr>
<td>$\rho$</td>
<td>0.366</td>
<td>0.197</td>
<td>0.013</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>-0.117</td>
<td>-0.099</td>
<td>0.0345</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.170</td>
<td>0.251</td>
<td>0.0965</td>
</tr>
<tr>
<td>$\rho$</td>
<td>0.160</td>
<td>0.351</td>
<td>0.108</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.203</td>
<td>0.196</td>
<td>0.079</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.156</td>
<td>-0.581</td>
<td>-0.021</td>
</tr>
<tr>
<td>$\rho$</td>
<td>-0.019</td>
<td>-0.072</td>
<td>-0.077</td>
</tr>
</tbody>
</table>

exists in other settings (not shown).

I: Low Censored

| n=200 | $\beta_1$ | -0.0124 | 0.263 | 0.426 | 0.812 |
| n=500 | $\beta_1$ | -0.0288 | 0.169 | 0.188 | 0.862 |
| n=800 | $\beta_1$ | -0.0204 | 0.249 | 0.229 | 0.918 |
| n=1000 | $\beta_1$ | -0.0210 | 0.197 | 0.192 | 0.922 |
| n=1500 | $\beta_1$ | -0.0138 | 0.098 | 0.092 | 0.950 |

II: Medium Censored

| n=200 | $\beta_1$ | -0.0965 | 0.270 | 0.509 | 0.868 |
| n=500 | $\beta_1$ | -0.0677 | 0.400 | 0.517 | 0.896 |
| n=800 | $\beta_1$ | -0.0233 | 0.170 | 0.187 | 0.904 |
| n=1000 | $\beta_1$ | -0.0345 | 0.251 | 0.258 | 0.918 |
| n=1500 | $\beta_1$ | -0.0011 | 0.147 | 0.112 | 0.946 |

III: High Censored

| n=200 | $\beta_1$ | -0.3500 | 0.396 | 1.480 | 0.806 |
| n=500 | $\beta_1$ | -0.0386 | 0.226 | 0.245 | 0.830 |
| n=800 | $\beta_1$ | -0.0147 | 0.229 | 0.343 | 0.944 |
| n=1000 | $\beta_1$ | -0.0165 | 0.175 | 0.172 | 0.954 |
| n=1500 | $\beta_1$ | -0.0019 | 0.155 | 0.159 | 0.950 |

than the latter for different censored rates and sample sizes. This superiority also exists in other settings (not shown).

We provide illustration on model testing in terms of $\rho$. For example, for testing the PO model ($\rho = 1$), Table 4.6 presents the power analysis for rejecting $\rho = 1$. The sample size $n = 1500$ was used. Under the PO model, the power is very close to the nominal value 0.05 when the true value $\rho = 1$, and the power for other values of $\rho$ is
higher when \( \rho \) is farther away from \( \rho = 1 \). High censored rate data tends to provide lower power than low censored rate data since the likelihood function is less sharply
Table 4.4: Summary Statistics of the proposed method with convex function $\Lambda_0(t) = 0.5t^{5/6}$, $\beta = c(-1, -1)$ and in extreme cases of $\rho = 4$ and $\rho = 0.25$ with 3 levels censoring rate (around 20% in I: Low Censored, 40% in II: Medium Censored and 75% in III: High Censored). Similar to the concave function $\Lambda_0(t)$, a larger $\rho$ provides larger frailty variance, and the setup under $\rho = 4$ needs much larger sample size to provide stable performance.

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$\rho = 4$</th>
<th>$\rho = 0.25$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>-0.200</td>
<td>-0.278</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.168</td>
<td>0.276</td>
</tr>
<tr>
<td>$\rho$</td>
<td>1.572</td>
<td>7.755</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>-0.071</td>
<td>0.173</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.050</td>
<td>0.296</td>
</tr>
<tr>
<td>$\rho$</td>
<td>0.677</td>
<td>7.739</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>-0.041</td>
<td>0.148</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.047</td>
<td>0.243</td>
</tr>
<tr>
<td>$\rho$</td>
<td>0.327</td>
<td>0.814</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>-0.022</td>
<td>0.133</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.021</td>
<td>0.212</td>
</tr>
<tr>
<td>$\rho$</td>
<td>0.303</td>
<td>0.759</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>-0.006</td>
<td>0.116</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.021</td>
<td>0.179</td>
</tr>
<tr>
<td>$\rho$</td>
<td>0.186</td>
<td>0.627</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>-0.003</td>
<td>0.096</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.004</td>
<td>0.144</td>
</tr>
<tr>
<td>$\rho$</td>
<td>0.106</td>
<td>0.627</td>
</tr>
</tbody>
</table>

peaked. For both the concave and convex cumulative baseline hazard functions, the power analysis are similar.
Since the existing approaches all estimate regression parameters by treating \( \rho \) as known in the literature, and we implemented a simplified version of our approach with known \( \rho \) for our simulated data. Table 4.7 provides the summarized results of regression parameters estimation \( \beta \) in this case. As seen in Table 4.7, the estimation
Table 4.6: Power Analysis of the proposed method for $\rho = 1$ (PO model). Note the concave function is $\Lambda_0(t) = \log(1 + t) + t^{1.5}$, convex function is $\Lambda_0(t) = 0.5t^{5/6}$ and $\beta = c(-1, -1)$. Sample size is 1500.

<table>
<thead>
<tr>
<th></th>
<th>concave</th>
<th></th>
<th>convex</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho = 0.25$</td>
<td>1.000</td>
<td>0.986</td>
<td>0.046</td>
<td>0.926</td>
<td>0.094</td>
<td>1.000</td>
<td>0.968</td>
</tr>
<tr>
<td>$\rho = 0.5$</td>
<td>0.068</td>
<td>0.884</td>
<td>0.960</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho = 1$</td>
<td>0.872</td>
<td>0.960</td>
<td>0.978</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho = 4$</td>
<td>0.780</td>
<td>0.324</td>
<td>0.068</td>
<td>0.174</td>
<td>0.884</td>
<td>0.960</td>
<td>0.174</td>
</tr>
</tbody>
</table>

Table 4.7: Summary Statistics of the proposed method with fixed parameter $\rho$, concave function $\Lambda_0(t) = \log(1 + t) + t^{1.5}$, convex function $\Lambda_0(t) = 0.5t^{5/6}$, $\beta = c(-1, -1)$ with 3 levels censoring rate (around 20% in I: Low Censored, 40% in II: Medium Censored and 75% in III: High Censored), and sample size of 200.

<table>
<thead>
<tr>
<th></th>
<th>Par</th>
<th>Bias</th>
<th>StErr</th>
<th>StDev</th>
<th>CP</th>
<th></th>
<th>Par</th>
<th>Bias</th>
<th>StErr</th>
<th>StDev</th>
<th>CP</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I: Low Censored</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho = 0.25$</td>
<td>$\beta_1$</td>
<td>-0.021</td>
<td>0.116</td>
<td>0.122</td>
<td>0.948</td>
<td>-0.017</td>
<td>0.114</td>
<td>0.117</td>
<td>0.948</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho = 0.5$</td>
<td>$\beta_2$</td>
<td>-0.024</td>
<td>0.200</td>
<td>0.209</td>
<td>0.946</td>
<td>-0.011</td>
<td>0.197</td>
<td>0.192</td>
<td>0.956</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho = 1$</td>
<td>$\beta_2$</td>
<td>0.004</td>
<td>0.222</td>
<td>0.223</td>
<td>0.952</td>
<td>-0.000</td>
<td>0.223</td>
<td>0.234</td>
<td>0.942</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho = 4$</td>
<td>$\beta_2$</td>
<td>0.004</td>
<td>0.229</td>
<td>0.237</td>
<td>0.937</td>
<td>-0.012</td>
<td>0.330</td>
<td>0.308</td>
<td>0.956</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| II: Medium Censored |     |      |       |       |    |          |     |      |       |       |    |          |
| $\rho = 0.25$ | $\beta_1$ | -0.012 | 0.127 | 0.131 | 0.950 | -0.018 | 0.127 | 0.128 | 0.954 |
| $\rho = 0.5$  | $\beta_2$ | -0.008 | 0.222 | 0.232 | 0.934 | -0.007 | 0.221 | 0.231 | 0.936 |
| $\rho = 1$    | $\beta_2$ | 0.006  | 0.216 | 0.227 | 0.960 | -0.029 | 0.242 | 0.246 | 0.948 |
| $\rho = 4$    | $\beta_2$ | -0.031 | 0.236 | 0.255 | 0.932 | -0.023 | 0.233 | 0.248 | 0.928 |

| III: High Censored |     |      |       |       |    |          |     |      |       |       |    |          |
| $\rho = 0.25$ | $\beta_1$ | -0.041 | 0.183 | 0.180 | 0.962 | -0.029 | 0.184 | 0.192 | 0.952 |
| $\rho = 0.5$  | $\beta_2$ | -0.033 | 0.326 | 0.343 | 0.942 | -0.041 | 0.337 | 0.353 | 0.938 |
| $\rho = 1$    | $\beta_2$ | 0.015  | 0.209 | 0.212 | 0.952 | -0.038 | 0.215 | 0.227 | 0.926 |
| $\rho = 4$    | $\beta_2$ | -0.014 | 0.277 | 0.279 | 0.960 | -0.036 | 0.474 | 0.484 | 0.944 |

of the regression parameters is very good for all cases even when sample size is only 200.
Figure 4.3: Estimated baseline survival probability for the concave function $\Lambda_0(t) = \log(1 + t) + t^{1.5}$, $\beta = c(-1, -1)$, $\rho = 2$ with low censored rate (sample size 200).

Figure 4.4: Estimated baseline survival probability for the concave function $\Lambda_0(t) = \log(1 + t) + t^{1.5}$, $\beta = c(-1, -1)$, $\rho = 2$ with low censored rate (sample size 500).

4.5 Real Data Analysis

The Surveillance, Epidemiology, and End Results (SEER) Program (Howlader et al., 2019) is a source of epidemiologic information on the incidence and survival rates of cancer in the United States, supported by the Surveillance Research Program (SRP) in NCI’s Division of Cancer Control and Population Sciences (DCCPS). SEER col-
Figure 4.5: Estimated baseline survival probability for the concave function \( \Lambda_0(t) = \log(1 + t) + t^{1.5} \), \( \beta = c(-1, -1) \), \( \rho = 2 \) with low censored rate (sample size 1000).

Figure 4.6: Estimated baseline survival probability for the concave function \( \Lambda_0(t) = \log(1 + t) + t^{1.5} \), \( \beta = c(-1, -1) \), \( \rho = 2 \) with low censored rate (sample size 1500).

The SEER Program registries routinely collect data on patient demographics, primary tumor site, tumor morphology and stage at diagnosis, first course of treatment, and follow-up for vital status of dead or alive.
Figure 4.7: Estimated baseline survival probability for the concave function $\Lambda_0(t) = \log(1 + t) + t^{1.5}$, $\beta = c(-1, -1)$, $\rho = 2$ with medium censored rate (sample size 200).

Figure 4.8: Estimated baseline survival probability for the concave function $\Lambda_0(t) = \log(1 + t) + t^{1.5}$, $\beta = c(-1, -1)$, $\rho = 2$ with medium censored rate (sample size 500).

Our data set comes from the SEER 1975-2016 Research Database (Howlader et al., 2019). Specifically, we select a subset of the data for the Louisiana localized lung cancer study. This data set contains subjects aged at least 50 from the state of Louisiana whose diagnosis year is between 2004 to 2016. The response variable of interest is the time to death since diagnosis, for which right-censored data are available since not all patients died by 2016. Our data set only focuses on the data
Figure 4.9: Estimated baseline survival probability for the concave function $\Lambda_0(t) = \log(1 + t) + t^{1.5}$, $\beta = c(-1, -1)$, $\rho = 2$ with medium censored rate (sample size 1000).

Figure 4.10: Estimated baseline survival probability for the concave function $\Lambda_0(t) = \log(1 + t) + t^{1.5}$, $\beta = c(-1, -1)$, $\rho = 2$ with medium censored rate (sample size 1500).

from white and black races although there are many other races in the original data set. In total, there are 9,965 observations in this data set, and the right-censored rate is 42.76%. We consider three covariates in this data analysis: “sex” taking 1 for female and 0 for male, “race” taking 1 for black race and 0 for white race, and “grade”, which is a measurement of how closely the tumor cells resemble lung ranging from 1 to 9 representing different degrees of differentiation.
Table 4.9 presents the parameters estimation by the proposed approach under the GOR model (Model I), the PH model (Model II) and the PO model (Model III). The interior knots selected by the AIC and BIC are both 15 for the three models shown in Table 4.8. Of the three models, Model I provides the smallest AIC of 29079.62. The regression estimators are (-0.244, 0.070, 0.135), which means that as sex changes from male to female, the ratio of survival probability compared to the placebo group would increase by \( \exp(-0.244) = 0.783 \), and as race changes from black to white, the ratio of survival probability increases by \( \exp(0.070) = 1.073 \), and the ratio of survival probability will rise by \( \exp(0.135) = 1.145 \) as grade increases by 1. The estimator of \( \rho \) is 1.493 in Model I, indicating the Model III has a slightly higher AIC of 29083.56, and Model II provides the largest AIC of 29153.82. For all these models, the estimated effects of sex and grade are significant with their corresponding 95% confidence interval excluding 0, but race is not significant with the 95% confidence interval containing 0. The estimator of the parameter \( \rho \) is 1.493 with the 95% confidence interval of (0.983, 2.003). Since this confidence interval contains 1, it suggests that the PO model is also reasonable to fit this data, although the GOR model with \( \rho = 1.493 \) is preferred. We also conducted the analysis with an additional interaction term of sex and race in Table 4.10, which shows the interaction term is not significant with the 95% confidence interval containing 0.

Table 4.8: SEER Louisiana Localized Lung Cancer Data: include sex, race and grade in the survival function. Model I is to fit the data by estimating \( \rho \) with regression parameters, Model II fixes \( \rho = 0 \), and Model III fits the data with fixed \( \rho = 1 \). The best number of interior knots is searched for 10, 15, 20 and 25 based on AIC and BIC.

<table>
<thead>
<tr>
<th>Model:</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model I: proposed approach</td>
<td>AIC</td>
<td>BIC</td>
<td>AIC</td>
<td>BIC</td>
</tr>
<tr>
<td></td>
<td>29136.66</td>
<td>29259.17</td>
<td>29079.62</td>
<td>29317.24</td>
</tr>
<tr>
<td></td>
<td>29085.21</td>
<td>29315.82</td>
<td>29122.66</td>
<td>29317.24</td>
</tr>
<tr>
<td>Model II: Fix ( \rho ) at 0 (PH Model)</td>
<td>AIC</td>
<td>BIC</td>
<td>AIC</td>
<td>BIC</td>
</tr>
<tr>
<td></td>
<td>29205.09</td>
<td>29230.40</td>
<td>29153.82</td>
<td>29380.91</td>
</tr>
<tr>
<td></td>
<td>29157.50</td>
<td>29380.91</td>
<td>29194.27</td>
<td>29380.91</td>
</tr>
<tr>
<td>Model III: Fix ( \rho ) at 1 (PO Model)</td>
<td>AIC</td>
<td>BIC</td>
<td>AIC</td>
<td>BIC</td>
</tr>
<tr>
<td></td>
<td>29138.65</td>
<td>29253.96</td>
<td>29083.56</td>
<td>29313.12</td>
</tr>
<tr>
<td></td>
<td>29088.36</td>
<td>29311.77</td>
<td>29125.74</td>
<td>29313.12</td>
</tr>
</tbody>
</table>
Table 4.9: SEER Louisiana Localized Lung Cancer Data: parameters estimation in the proposed EM algorithm, sex, race and grade are included in the survival function. Interior knots are selected as 15 based on AIC and BIC for the three models.

<table>
<thead>
<tr>
<th>Par</th>
<th>Est</th>
<th>StErr</th>
<th>95% CI</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Model I: Proposed Approach</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>-0.244</td>
<td>0.045</td>
<td>(-0.332, -0.154)</td>
<td></td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.070</td>
<td>0.051</td>
<td>(-0.030, 0.170)</td>
<td></td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>0.135</td>
<td>0.010</td>
<td>(0.115, 0.154)</td>
<td></td>
</tr>
<tr>
<td>$\rho$</td>
<td>1.493</td>
<td>0.260</td>
<td>(0.983, 2.003)</td>
<td></td>
</tr>
<tr>
<td>$AIC=$</td>
<td>29079.62</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model II: Fix $\rho$ at 0 (PH Model)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>-0.155</td>
<td>0.027</td>
<td>(-0.209, -0.102)</td>
<td></td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.033</td>
<td>0.031</td>
<td>(-0.028, 0.094)</td>
<td></td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>0.077</td>
<td>0.004</td>
<td>(0.069, 0.084)</td>
<td></td>
</tr>
<tr>
<td>$AIC=$</td>
<td>29153.82</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model III: Fix $\rho$ at 1 (PO Model)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>-0.218</td>
<td>0.039</td>
<td>(-0.294, -0.142)</td>
<td></td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.056</td>
<td>0.045</td>
<td>(-0.032, 0.144)</td>
<td></td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>0.119</td>
<td>0.006</td>
<td>(0.108, 0.130)</td>
<td></td>
</tr>
<tr>
<td>$AIC=$</td>
<td>29083.56</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.10: SEER Louisiana Localized Lung Cancer Data: parameters estimation in the proposed EM algorithm, sex, race, grade and also the interaction of sex and race are included in the survival function. Interior knots are selected as 15 based on AIC and BIC for the three models.

<table>
<thead>
<tr>
<th>Par</th>
<th>Est</th>
<th>StErr</th>
<th>95% CI</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Model I: Proposed Approach</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>-0.261</td>
<td>0.051</td>
<td>(-0.362, -0.160)</td>
<td></td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.041</td>
<td>0.064</td>
<td>(-0.084, 0.166)</td>
<td></td>
</tr>
<tr>
<td>$\beta_{12}$</td>
<td>0.080</td>
<td>0.105</td>
<td>(-0.126, 0.285)</td>
<td></td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>0.135</td>
<td>0.010</td>
<td>(0.115, 0.154)</td>
<td></td>
</tr>
<tr>
<td>$\rho$</td>
<td>1.495</td>
<td>0.260</td>
<td>(0.985, 2.006)</td>
<td></td>
</tr>
<tr>
<td>$AIC=$</td>
<td>29081.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model II: Fix $\rho$ at 0 (PH Model)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>-0.166</td>
<td>0.031</td>
<td>(-0.226, -0.105)</td>
<td></td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.017</td>
<td>0.039</td>
<td>(-0.059, 0.093)</td>
<td></td>
</tr>
<tr>
<td>$\beta_{12}$</td>
<td>0.046</td>
<td>0.065</td>
<td>(-0.081, 0.174)</td>
<td></td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>0.077</td>
<td>0.004</td>
<td>(0.069, 0.084)</td>
<td></td>
</tr>
<tr>
<td>$AIC=$</td>
<td>29155.31</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model III: Fix $\rho$ at 1 (PO Model)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>-0.233</td>
<td>0.044</td>
<td>(-0.319, -0.147)</td>
<td></td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.031</td>
<td>0.056</td>
<td>(-0.079, 0.141)</td>
<td></td>
</tr>
<tr>
<td>$\beta_{12}$</td>
<td>0.069</td>
<td>0.093</td>
<td>(-0.113, 0.251)</td>
<td></td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>0.119</td>
<td>0.006</td>
<td>(0.108, 0.130)</td>
<td></td>
</tr>
<tr>
<td>$AIC=$</td>
<td>29085.00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 4.11 and 4.12 show the predicted survival probability for different combination of sex and race, and grade is set at the mean value of the data set. The predicted survival probabilities are comparable to the Kaplan Meier curve. For both
Figure 4.11: Predicted survival probability vs. Time (years) for White Female and White Male in the SEER Localized Lung Cancer Data under the proposed method and compared to the Kaplan Meier Curve. The proposed method predicts the survival probability higher for white female than for white male.

the white and black people, female tends to have higher survival probability than male, which is consistent with the estimation results in Table 4.9.

4.6 DISCUSSIONS

In this paper, we investigate the identifiability issues of the GOR models and prove that all the parameters including the parameter $\rho$ are identifiable in regression settings. However, estimating $\rho$ in the GOR models can be problematic as reported in many research articles. Our simulation results suggest that the estimation performance of the propose approach is satisfactory if sample size is large enough. However, the performance is heavily influenced by the right-censoring rate and the true value of $\rho$. For example, when the true value of $\rho$ is 0.25 and the right-censoring rate is about 20%, the estimation of all parameters can be accurate if the sample size is 200; however, when the true value of $\rho$ is 4 with the same censoring rate, the sample size needed for a satisfactory performance of the proposed method is about 2500.
Figure 4.12: Predicted survival probability vs. Time (years) for Black Female and Black Male in the SEER Localized Lung Cancer Data under the proposed method and compared to the Kaplan Meier Curves. The proposed method predicts the survival probability higher for black female than for black male.

The proposed approach based on the EM algorithm has several appealing properties such as being robust to initial values, easy to implement, fast to converge, and providing variance estimates in closed form. Our approach allows to estimate the regression parameters, the spline coefficients, and $\rho$ jointly. Thus, it allows one to make inference on $\rho$ and to conduct model comparison and testing in addition to the estimation of the covariate effects and survival functions.

In terms of model selection, the proposed approach for the GOR models provides more flexibility since it can estimate the parameter $\rho$ with the regression parameters simultaneously. AIC or BIC can be used to select the best model. In real life, however, the PO model may have better interpretation and easy to understand in odds ratio compared to the transformed odds ratio in the GOR models. If the confidence interval of $\rho$ estimator includes 1, then the PO model is suggested.


Hanson, TIMOTHY E and ALEJANDRO Jara (2013). “Surviving fully Bayesian nonparametric regression models”. In: *Bayesian Theory and Applications*, pp. 593–615.


Yao, Bin (2016). “Semiparametric regression analysis of panel count data and interval-censored failure time data”. In:


Appendix A

The quantities involved in var(\( \hat{\theta} \))

For notation simplicity, we use \( E(Y) \) to denote the conditional expectation of a quantity \( Y \) conditional on the observed data \( O \), i.e., \( E(Y) = E(Y|O, \theta(d)) \). Similarly, the covariances and variances below refer to the conditional covariances and variances given the observed data \( O \). In the Louis’s method, the first part of quantities are the second derivative of \( Q(\hat{\theta}) \) with respective to \( \theta \),

\[
\frac{\partial^2 Q(\theta, \theta(d))}{\partial \beta \partial \beta'} = \sum_{i=1}^{n} \left\{ -\delta_{i} E(\psi_{i}) - E(\phi_{i}) \right\} \sum_{l=1}^{k} \gamma_{l} b_{l}(y_{i}) \exp(x_{i}^{\top} \beta) x_{i} x_{i}'
\]

\[
\frac{\partial^2 Q(\theta, \theta(d))}{\partial \gamma^2} = \sum_{i=1}^{n} -\gamma_{j}^{-2} \delta_{i} E(Z_{ij})
\]

\[
\frac{\partial^2 Q(\theta, \theta(d))}{\partial \beta \partial \gamma_{j}} = \sum_{i=1}^{n} \left\{ -\delta_{i} E(\psi_{i}) - E(\phi_{i}) \right\} b_{j}(y_{i}) \exp(x_{i}^{\top} \beta) x_{i}
\]

\[
\frac{\partial^2 Q(\theta, \theta(d))}{\partial \tau^2} = \sum_{i=1}^{n} \left\{ -\delta_{i} \tau^{-2} + \tau^{-1} - \frac{\partial^2 \ln(\Gamma(\tau))}{\partial \tau^2} \right\}.
\]

Note that, \( \frac{\partial^2 Q(\theta, \theta(d))}{\partial \gamma_{j} \partial \gamma_{j'}} = 0 \) for \( j \neq j' \), and also, \( \frac{\partial^2 Q(\theta, \theta(d))}{\partial \beta \partial \gamma_{j}} = 0 \) and \( \frac{\partial^2 Q(\theta, \theta(d))}{\partial \beta \partial \tau} = 0 \).

The second part of quantities in Louis’s method is the covariance of the first derivative of log \( L_{c}(\theta) \) with respect of \( \theta \),

\[
\frac{\partial \log L_{c}(\theta)}{\partial \beta} = \sum_{i=1}^{n} \left\{ \delta_{i} x_{i} - \delta_{i} \sum_{j=1}^{k} \gamma_{j} b_{j}(y_{i}) \exp(x_{i}^{\top} \beta) \psi_{i} x_{i} - \sum_{j=1}^{k} \gamma_{j} b_{j}(y_{i}) \exp(x_{i}^{\top} \beta) \phi_{i} x_{i} \right\}
\]

\[
\frac{\partial \log L_{c}(\theta)}{\partial \gamma_{j}} = \sum_{i=1}^{n} \left\{ \delta_{i} Z_{ij} - \delta_{i} b_{j}(y_{i}) \exp(x_{i}^{\top} \beta) \psi_{i} - b_{j}(y_{i}) \exp(x_{i}^{\top} \beta) \phi_{i} \right\}
\]

\[
\frac{\partial \log L_{c}(\theta)}{\partial \tau} = \sum_{i=1}^{n} \left\{ -\delta_{i} \psi_{i} + \delta_{i} \tau^{-1} - \phi_{i} + \log(\tau) + 1 + \log(\phi_{i}) - \frac{\Gamma'(\tau)}{\Gamma(\tau)} \right\}.
\]
Then

\[
\text{cov}\left(\frac{\partial \log \mathcal{L}_c(\theta)}{\partial \beta}, \frac{\partial \log \mathcal{L}_c(\theta)}{\partial \beta}\right) = \sum_{i=1}^{n} \left[ \delta_i \left\{ \sum_{j=1}^{k} \gamma_j b_j(y_i) \right\}^2 \exp(2x_i^T \beta) \text{var}(\psi_i)x_i x_i' \right. \\
+ \left\{ \sum_{j=1}^{k} \gamma_j b_j(y_i) \right\}^2 \exp(2x_i^T \beta) \text{var}(\phi_i)x_i x_i' \right]
\]

\[
\text{cov}\left(\frac{\partial \log \mathcal{L}_c(\theta)}{\partial \gamma_j}, \frac{\partial \log \mathcal{L}_c(\theta)}{\partial \gamma_j}\right) = \sum_{i=1}^{n} \left\{ \delta_i \text{var}(Z_{ij}) \gamma_j^2 + \delta_i b_j^2(y_i) \exp(2x_i^T \beta) \text{var}(\psi_i) \right. \\
+ \delta_i b_j(y_i) b_j(y_i) \exp(2x_i^T \beta) \text{var}(\phi_i) \right\} \\
\text{cov}\left(\frac{\partial \log \mathcal{L}_c(\theta)}{\partial \gamma_j}, \frac{\partial \log \mathcal{L}_c(\theta)}{\partial \gamma_{j'}}\right) = \sum_{i=1}^{n} \left\{ \delta_i \text{cov}(Z_{ij}, Z_{ij'}) \gamma_j \gamma_{j'} + \delta_i b_j(y_i) b_{j'}(y_i) \exp(2x_i^T \beta) \text{var}(\psi_i) \right. \\
+ \delta_i b_j(y_i) b_{j'}(y_i) \exp(2x_i^T \beta) \text{var}(\phi_i) \right\} \\
\text{cov}\left(\frac{\partial \log \mathcal{L}_c(\theta)}{\partial \beta}, \frac{\partial \log \mathcal{L}_c(\theta)}{\partial \gamma_j}\right) = \sum_{i=1}^{n} \left\{ \delta_i \sum_{j=1}^{k} \gamma_j b_j(y_i) \exp(2x_i^T \beta) b_j(y_i) \text{var}(\psi_i)x_i \right. \\
+ \sum_{j=1}^{k} \gamma_j b_j(y_i) \exp(2x_i^T \beta) b_j(y_i) \text{var}(\phi_i)x_i \right\} \\
\text{cov}\left(\frac{\partial \log \mathcal{L}_c(\theta)}{\partial \beta}, \frac{\partial \log \mathcal{L}_c(\theta)}{\partial \tau}\right) = \sum_{i=1}^{n} \left\{ \delta_i \text{var}(\psi_i) + \text{var}(\phi_i) + \text{var}(\log \phi_i) - 2\text{cov}(\phi_i, \log \phi_i) \right\} \\
\text{cov}\left(\frac{\partial \log \mathcal{L}_c(\theta)}{\partial \tau}, \frac{\partial \log \mathcal{L}_c(\theta)}{\partial \tau}\right) = \sum_{i=1}^{n} \left\{ \delta_i \text{var}(\psi_i) + \text{var}(\phi_i) + \text{var}(\log \phi_i) - 2\text{cov}(\phi_i, \log \phi_i) \right\} \\
\text{cov}\left(\frac{\partial \log \mathcal{L}_c(\theta)}{\partial \gamma_j}, \frac{\partial \log \mathcal{L}_c(\theta)}{\partial \tau}\right) = \sum_{i=1}^{n} \left\{ \delta_i b_j(y_i) \exp(x_i^T \beta) \text{var}(\psi_i) + b_j(y_i) \exp(x_i^T \beta) \text{var}(\phi_i) \right. \\
- \left. \delta_i b_j(y_i) \exp(x_i^T \beta) \text{cov}(\phi_i, \log(\phi_i))x_i \right\} \\
\text{cov}(\psi_i, \phi_i) = \text{cov}(Z_{ij}, \psi_i) = \text{cov}(Z_{ij}, \phi_i) = \text{cov}(\psi_i, \log \phi_i) = 0, \text{ and} \\
\text{cov}(\phi_i, \log \phi_i) = E(\phi_i \log \phi_i) - E(\phi_i)E(\log \phi_i)
\]

\[
= \frac{\Gamma'(\tau + 1)}{\Gamma(\tau)} \left\{ \Lambda_0(y_i) \exp(x_i^T \beta) + \tau \right\} - \frac{\tau \log \{ \Lambda_0(y_i) \exp(x_i^T \beta) + \tau \}}{\Lambda_0(y_i) \exp(x_i^T \beta) + \tau} - \frac{\tau}{\Lambda_0(y_i) \exp(x_i^T \beta) + \tau} \left[ \Gamma'(\tau) - \log \{ \Lambda_0(y_i) \exp(x_i^T \beta) + \tau \} \right].
\]
The conditional covariance and variance terms have the following forms,

\[
\text{var}(\psi_i|O; \theta^{(d)}) = \frac{\delta_i}{\{\Lambda_0^{(d)}(y_i) \exp(x_i^T \beta^{(d)}) + \tau^{(d)}\}^2}
\]

\[
\text{var}(\phi_i|O; \theta^{(d)}) = \frac{\tau}{\{\Lambda_0^{(d)}(y_i)e^{\beta^{(d)}x_i} + \tau^{(d)}\}^2}
\]

\[
\text{var}(Z_{ij}|O; \theta^{(d)}) = \frac{\gamma_j^{(d)} M_j(y_i)}{\sum_{k=1}^k \gamma_k^{(d)} M_k(y_i)} \{1 - \frac{\gamma_j^{(d)} M_j(y_i)}{\sum_{k=1}^k \gamma_k^{(d)} M_k(y_i)}\} \delta_i
\]

\[
\text{cov}(Z_{ij}, Z_{ij'}|O; \theta^{(d)}) = -\frac{\gamma_j^{(d)} M_j(y_i)}{\sum_{k=1}^k \gamma_k^{(d)} M_k(y_i)} \frac{\gamma_{j'} M_{j'}(y_i)}{\sum_{k=1}^k \gamma_k^{(d)} M_k(y_i)} \delta_i
\]

\[
\text{var}(\log(\phi)|O; \theta^{(d)}) = \frac{\Gamma''(\tau)}{\Gamma(\tau)} - \left\{ \frac{\Gamma'(\tau)}{\Gamma(\tau)} \right\}^2.
\]

These closed-form expressions for variance estimation make computation more efficient. The variance estimator for \( \rho \) is transformed via \( \text{var}(\rho) = \tau^{-4} \text{var}(\tau) \).
Appendix B

The EM algorithm for the PH model

For the PH model ($\rho = 0$), the observed likelihood function with observed data $O$ and parameters $\theta = (\Lambda_0(\cdot), \beta, \tau)$ is

$$
\mathcal{L}_{\text{obs}}(\theta; O) = \prod_{i=1}^{n} h(y_i|x_i)^{\delta_i} S(y_i|x_i)
= \prod_{i=1}^{n} \{\lambda_0(y_i) \exp(x_i^T \beta)\}^{\delta_i} \exp\{-\Lambda_0(y_i)\}\exp(x_i^T \beta),
$$

and the log observed likelihood function is

$$
\ell_{\text{obs}}(\theta; O) = \sum_{i=1}^{n} \delta_i \log\{\lambda_0(y_i)\} + \delta_i x_i^T \beta - \exp(x_i^T \beta).
$$

The complete likelihood function with latent variable $Z_i \sim \text{multinomial}[1, (\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k})]$ for $\delta_i = 1$ is

$$
\mathcal{L}_c(\theta, \theta^{(d)}; O) = \prod_{i=1}^{n} \left[ \prod_{l=1}^{k} \{\gamma_l M_l(y_i)\}^{Z_{il}} \exp(x_i^T \beta) \right]^{\delta_i} \exp\{-\Lambda_0(y_i)\}\exp(x_i^T \beta),
$$

and then the expectation of logarithm complete likelihood function is expressed as

$$
Q(\theta, \theta^{(d)}; O) = \sum_{i=1}^{n} \delta_i \left[ \sum_{l=1}^{k} E(Z_{il}|O; \theta^{(d)}) \log\{\gamma_l M_l(y_i)\} \right] + \delta_i x_i^T \beta - \exp(x_i^T \beta) \sum_{l=1}^{k} \gamma_l b_l(y_i).
$$

By taking first partial derivative of equation (B.1) with $\gamma_j$ and setting it to zero, the updated quantity has closed form

$$
\gamma_j^*(\beta) = \frac{\sum_{i=1}^{n} \delta_i E(Z_{ij}|O; \theta^{(d)})}{\sum_{i=1}^{n} \delta_i \exp(x_i^T \beta) b_j(y_i)}, j = 1, \ldots, k.
$$

$\beta$ is updated by solving

$$
\sum_{i=1}^{n} \delta_i x_i - \sum_{j=1}^{k} \gamma_j^{*(d)} b_j(y_i) \exp(x_i^T \beta) x_i = 0.
$$
For variance estimation via the Louis’s method, the quantities involved are

\[
\frac{\partial^2 Q(\theta, \theta^{(d)})}{\partial \beta \partial \beta'} = \sum_{i=1}^{n} - \exp(x_i^T \beta)x_i x_i' \sum_{l=1}^{k} \gamma_l b_l(y_i)
\]

\[
\frac{\partial^2 Q(\theta, \theta^{(d)})}{\partial \gamma_j^2} = \sum_{i=1}^{n} -\gamma_j^{-2} \delta_i E(Z_{ij})
\]

\[
\frac{\partial^2 Q(\theta, \theta^{(d)})}{\partial \beta \partial \gamma_j} = \sum_{i=1}^{n} -b_j(y_i) \exp(x_i^T \beta)x_i
\]

and

\[
\text{cov}\left(\frac{\partial \log \mathcal{L}_c(\theta)}{\partial \gamma_j}, \frac{\partial \log \mathcal{L}_c(\theta)}{\partial \gamma_j}\right) = \sum_{i=1}^{n} \delta_i^2 \text{var}(Z_{ij}) \gamma_j^2
\]

\[
\text{cov}\left(\frac{\partial \log \mathcal{L}_c(\theta)}{\partial \gamma_j}, \frac{\partial \log \mathcal{L}_c(\theta)}{\partial \gamma_{j'}}\right) = \sum_{i=1}^{n} \delta_i^2 \text{cov}(Z_{ij}, Z_{ij'}) \gamma_j \gamma_{j'}
\]