ON DIFFERENT COMPUTATIONAL ASPECTS FOR BOX-COX TRANSFORMATION CURE RATE MODEL

by

PEI WANG

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON
August 2021

To my husband Dr. Dezhi Dai and our parents.

ACKNOWLEDGEMENTS

My deepest gratitude goes first and foremost to my supervisor, Dr. Suvra Pal, and also to Dr. Shan Sun-Mitchell who always helped me and led me into the world of statistics. Without Dr. Pal's inspiration, encouragement and invaluable guidance in both academics and life throughout my three and half years of study, I would not be the better and more confident person that I am right now. I appreciate the countless hours that he spent on discussing my research work and the generous help that he provided on revising journal papers and this dissertation. It has been a great honor for me to work with him.

I would like to express my deep appreciation to Dr. Shan Sun-Mitchell. The fantasitc illustrations in her class are so great that I couldn't imagine how she can make the obscure theorems and theories accessible and understandable like magic. She made me to dive into the statistics world with joy and enjoyment starting from the class MATH 4313 when I was a rookie. I will cherish the moments she consoled me when I was upset or felt disappointed.

I am extremely thankful to Dr. Andrzej Korzeniowski and Dr. Jonghyun Yun. Dr. Korzeniowski's real analysis and applied probability classes benifited me a lot. I always marveled at his ingenious angles to attack the problems. Dr. Yun's data mining class introduced me to machine learning field. He provided quite detailed materials for the class and these materials helped me a lot. I would like to extend my gratitude to Dr. Souvik Roy for his valuable and insightful advice in algorithm comparison.

I express my sincere gratitude to Dr. Hristo V. Kojouharov. He was the first professor I meet in our math department. He was always so patient and ready to help when I inqured about things such as class enrollment.

I wish to thank the College of Science and the Department of Mathematics for the financial support and the scholarships during my master and doctoral programs. I would also like to acknowledge Jodi Treszoks, Erol Ozkan, Jeremy Valdez, Hrishabh Khakurel, Craig Gore, Jiashan Ji, Madhu Gupta, Daniel Le, Julio Enciso Alva and Yifei Yu. I learned a lot from them and our friendship made my time in UTA so memorable.

My sincere thanks goes to Dr. Sandip Barui from the Indian Institue of Management Kerala for helping me out with chapter 3 of this dissertation.

Finally, I am extremely grateful to my parents and parents-in-law for their love, sacrifice and encouragement. My husband Dr. Dezhi Dai helped me a lot not only in improving my coding skills but also in supporting and taking care of me in my daily life. My special thank you goes to my cats Li and Milk, my furry friends could always melt my heart and let me get ride of fatigue easily.

August 3, 2021

ABSTRACT

ON DIFFERENT COMPUTATIONAL ASPECTS FOR BOX-COX
TRANSFORMATION CURE RATE MODEL

Pei Wang, Ph.D.

The University of Texas at Arlington, 2021

Supervising Professor: Dr. Suvra Pal

Cure rate modeling is an emerging area of research not only in biomedical science but also in other disciplines such as sociology, criminal justice, economics and engineering reliability. In the first part of this thesis, use of the wider class of generalized gamma distributions is proposed as the distribution of the lifetime for a particular transformation cure rate model, known as the Box-Cox transformation cure rate model. The maximum likelihood estimation of the Box-Cox transformation cure model parameters is studied through the calculated bias, mean square error and coverage probabilities of the asymptotic confidence intervals. The flexibilities of both generalized gamma distribution and Box-Cox model are utilized to carry out power studies to demonstrate the power of the likelihood ratio test in rejecting mis-specified models. Furthermore, the bias and efficiency of the estimators of the cure rates are studied when a wrong lifetime distribution is specified for a given cure rate model as well as when a wrong cure rate model is specified for a given lifetime distribution. The studies strongly suggest the importance of selecting a correct lifetime distribution and a correct cure rate model, which can be achieved through the proposed Box-Cox model with generalized gamma lifetime distribution. An illustration of this two-way flexibility is provided using data on breast cancer study.

In the second part of this thesis, the mixture cure rate model is considered as a special case of the Box-Cox transformation cure rate model. Instead of modeling the incidence part by using the traditional logistic or sigmoid link function, a new modeling approach based on the support vector machine (SVM) is proposed under the assumption of interval censored data. The proposed approach inherits the features of the SVM and provides flexibility to capture non-linearity in the data. A new estimation procedure based on the expectation maximization algorithm, that makes use of the sequential minimal optimization technique and Platt's scaling method, is developed to estimate the model parameters. The results of an extensive simulation study show that the proposed approach performs better in capturing complex classification boundaries when compared to the existing logistic regression-based approach. It is also verified that the ability to capture complex classification boundaries improve the estimation results corresponding to the latency parameters. For illustration, the proposed approach is applied to an interval censored smoking cessation data.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Figure Page

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Cure rate model

Models for time-to-event data (also called lifetime or survival data) that incorporates the possibility of cure are called cure rate models or long-term survival models. In these models, the population under consideration is modeled as a mixture of two types of patient groups. The group that responds well to the treatment and is no more susceptible to the event is called the “immuned” or the “cured” group. The other group for which the treatment is not effective and the event is either observed before the end of study or takes place after the study is called the “susceptible” or the “non-cured” group. Note that from a given time-to-event data, the cured status of subjects under study cannot be observed due to censoring, that is, the cured status is a latent variable. To observe the proportion of subjects who are cured of the event, also called the “cure rate”, all subjects under study should be followed for a sufficiently long period of time, see Sy and Taylor [1]. If a time-to-event data actually has a cure component, the plot of survival function does not tend to zero but levels off to some non-zero proportion, which indicates the presence of cure.

“Cured” subjects arise not only in cancer clinical trials but also in other disciplines. Examples are employees who never lose their jobs and married couples who never break in sociology studies, offenders who never commit crimes again in criminal justice, companies and subjects who are never in default on loans in economics, and parts in a complex system that never fail in engineering and so on. They all can be viewed as cured subjects and cure models should be considered when analyzing time-to-event data produced in these situations [2].

In such cases, traditional methods of survival analyses, including the well-known Cox's regression model, are not applicable since such methods assume that all subjects are susceptible to the event, thereby not accommodating the possibility of cure. Existing cure rate models are modified to include the probability of being cured, and based on how the cure proportion is introduced, the cure models can be roughly divided into two types: mixture cure models and non-mixture cure models. Cure rate modeling is a hot area of modern biostatistical research, and such an idea started back in 1940's when Boag [3] first proposed the mixture cure rate model. The formulation of the cured fraction has changed and improved over the years, many authors developed and improved the original mixture model further. Farewell [4] used a mixture model as a combination of logistic model and Weibull distribution to model the toxicant and stress level for laboratory animals. Kuk and Chen [5] proposed a semiparametric mixture cure model consisting of a logistic model for the probability of cure, and a proportional hazard model for the time to event of interest for uncured subjects. Goldman [6], Taylor [7], Peng and Dear [8], among others, have also investigated parametric, semiparametric, and nonparametric mixture cure rate models [2].

The non-mixture cure rate model is another type of cure models for modeling time-to-event data with a cure fraction. Non-mixture cure models were first introduced by Yakovlev [9], Ibrahim [10] and Chen [11]. These models were motivated by an underlying biological mechanism for cancer cells, which assumes that the number of cancer cells after cancer treatment follows a Poisson distribution. Most of the current investigations on the non-mixure cure models are in the Bayesian context due to its special form [2].

1.2 Literature review

The most widely used cure model is the mixture cure rate model introduced by Boag [3]. The population survival function can be represented by (Berkson and Gage [12])

$S_p(y) = p_0 + (1 - p_0)S_u(y)$, where p_0 is the cure proportion and $S_u(\cdot)$ is the survival function of the susceptible group. Usually, p_0 is linked to a set of covariates $\boldsymbol{x}$ using a logistic function $p_0 = \frac{1}{1+e^{\boldsymbol{x}'\boldsymbol{\beta}}}$, where $\boldsymbol{\beta}$ is the vector of regression coefficients. Note that covariates can also be introduced through $S_u(\cdot)$. Since the introduction of this mixture cure rate model, many researchers have studied this model by making different assumptions on $S_u(\cdot)$, by proposing different ways to include covariates, and by proposing different approaches for parameter estimation. For example, Farewell [4] proposed a logistic regression model for the cure rate p_0 and a Weibull distribution to model the lifetime of the susceptible subjects. Kuk and Chen [5] proposed a proportional hazards model for the lifetime of the susceptible subjects and employed a marginal likelihood approach for parameter estimation. Maller and Zhou [13] proposed a non-parametric approach based on Kaplan-Meier estimator to estimate the proportion of immunes. A semi-parametric approach based on the expectation maximization (EM) algorithm was proposed by Sy and Taylor [1]. Zhao et al. [14] developed a Bayesian approach for estimating the Cox proportional hazard cure rate model parameters. Pal and Balakrishnan [15] developed the likelihood inference based on the EM algorithm for a cure rate model that looks at the elimination of risk factors after an initial passage of time. For a book-length account on cure rate models, one may refer to the monograph by Maller and Zhou [16].

Under a competing risks scenario, let M be a latent random variable denoting the number of competing risks that can result in the event. Furthermore, let $W_i, i = 1, 2, \cdots, M$, denote the progression time due to the i-th competing risk. Then, assuming M to follow a Poisson distribution with mean η, Chen et al. [11] showed that the population survival function under such a competing risks scenario is given by $S_p(y) = e^{-\eta F(y)}$, where $F(\cdot)$ is the common distribution function of the progression times W_i, $i = 1, 2, \cdots, M$. This is known as the promotion time cure rate model or the Poisson cure rate model. In this case, the cure rate is given by $p_0 = \lim_{y\to\infty} S_p(y) = e^{-\eta}$. Covariates can be linked to p_0

through the parameter η using the log-linear function $\eta = e^{\boldsymbol{x}'\boldsymbol{\beta}}$. Furthermore, note that the survival function of the susceptible subjects is given by $S_u(y) = \frac{S_p(y)-p_0}{1-p_0}$. Equivalently, we can write $S_u(y) = \frac{e^{-\eta F(y)}-e^{-\eta}}{1-e^{-\eta}}$. Under a competing risks scenario, Rodrigues et al. [17] proposed the Conway-Maxwell Poisson (COM-Poisson) distribution to capture the unobserved number of competing risks. Balakrishnan and Pal [18] developed the EM algorithm for the COM-Poisson cure rate model under the assumption of Weibull lifetime for each competing risk. Balakrishnan et al. [19] proposed a piecewise linear approximation to model the hazard functions of competing risks in the context of mixture and promotion time cure rate models. Very recently, Pal and Roy [20, 21 developed a non-linear conjugate gradient type estimation algorithm for some cure rate models that look at the elimination process of competing risks.

The mixture and the promotion time cure rate models are the two most commonly used and widely explored cure rate models, where one may be looked as a competitor of the other. Yin and Ibrahim [22] first proposed a wider class of cure rate models that contain both mixture and promotion time cure rate models as special cases. The proposed wider class was indexed by a link parameter and was built using a Box-Cox transformation [23] on the population survival function. Peng and Xu [24] provided a biological interpretation for the Box-Cox cure rate model and proposed an estimation method under a proportional hazards framework. Diao and Yin [25] incorporated a frailty term in the Box-Cox cure rate model and proposed a non-parametric estimation technique using multivariate time-to-event data. Koutras and Milienos [26] proposed a flexible family of transformation cure rate models that was mainly motivated by the biological mechanism of the well studied promotion time cure rate model of Chen et al. [11] and by assuming that a metastasis-competent tumor cell would produce a detectable tumor only when a certain number of biological factors affect the cell. Zeng et al. [27] proposed a class of transformation survival model with a cure fraction that was motivated by biological considerations and includes the proportional hazards and the proportional odds cure rate models as particular cases. The authors proposed an efficient recursive algorithm for the maximum likelihood estimation of the model parameters.

Yu et al. [28] explored the use of generalized gamma distribution for cure rate estimation from mixture cure rate model for grouped survival data. The authors found the cure rate estimates from the model with generalized gamma distribution to be quite robust. Balakrishnan and Peng [29] explored generalized gamma distribution in the context of frailty survival model. In particular, the authors used the generalized gamma distribution as the distribution of the frailty instead of the commonly used gamma distribution to model the frailty. Balakrishnan and Pal [30] considered the wider class of generalized gamma distribution to model the competing risk lifetime in the context of COM-Poisson cure rate model. Pal et al. [31] used the generalized gamma distribution for right censored survival data and developed model discrimination methods.

In the mixture cure rate model, the incidence part, say $\pi(\boldsymbol{z})$, can also be referred as uncured rate, is traditionally and extensively modeled by sigmoid or logistic function $\pi(\boldsymbol{z}) = \frac{\exp(\boldsymbol{z}^{*\mathrm{T}}\boldsymbol{\beta})}{1+\exp(\boldsymbol{z}^{*\mathrm{T}}\boldsymbol{\beta})}$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_q)^{\mathrm{T}}$ and $\boldsymbol{z}^* = (1, \boldsymbol{z}^{\mathrm{T}})^{\mathrm{T}}$ [4, 5, 8]. As observed in the case of logistic regression, the logistic model works well when subjects are linearly separable into the cure or susceptible groups with respect to covariates. However, problem arises when subjects cannot be separated using a linear boundary. Other options to model the incidence include assuming a probit link function ($\Phi^{-1}(\pi(\boldsymbol{z})) = \boldsymbol{z}^{*\mathrm{T}}\boldsymbol{\beta}$) or a complementary log-log link function ($\log[-\log(1-\pi(\boldsymbol{z}))] = \boldsymbol{z}^{*\mathrm{T}}\boldsymbol{\beta}$), where Φ is the cumulative distribution function of the standard normal distribution [32–34]. However, these link functions do not offer non-linear separability and are not sufficient to capture more complex effects of $\boldsymbol{x}$ on the incidence. Cure rate is also estimated with few non-parametric strategies, e.g., generalized Kaplan-Meier estimate at maximum uncensored failure time

[24] and modified Beran-type estimator [35]. Again, these strategies fail to capture more complex effects of $\boldsymbol{x}$ on the incidence, especially, when multiple covariates are involved. Therefore, there exists necessity to identify a group of classifiers which would be able to model the incidence part more effectively by allowing non-linear separating boundaries between the cured and non-cured subjects.

To this end, support vector machine (SVM) could be a reasonable choice. SVM [36] is a supervised learning model with asscociated learning algorithms that analyze data for classificaiton and regression in machine learning, it can easily handle multiple continuous and categorical variables. SVM constructs a hyperplane in an iterative manner to minimize an error in multidimensional space to separate different classes, see Figure 5.1. Recently, Li et al. [37] studied the effect of the covariates on the incidence $\pi(\boldsymbol{z})$ by implementing the SVM. The new mixture model is seen to outperform existing cure rate models especially in the estimation of the incidence, and performs well for non-linearly separable classes and high dimensional covariates. However, Li et al. [37] have conducted the entire study considering data generated from non-informative right censoring mechanism.

1.3 Real data application

1.3.1 Breast cancer data

Breast cancer data can be downloaded from the R package “flexsurv”. The data represents the time-to-death (time-to-event) or the censoring time of 686 patients who had primary node positive breast cancer. In our application, we use the categorical variable prognostic group as the only covariate. This variable can take values 0, 1 and 2 depending on whether the prognostic group status is “poor”, “medium” and “good”, respectively. The observed time-to-event has the mean and the standard deviation as 3.08 years and

1.76 years, respectively. The total percentage of censored observations is 56%. Interested readers may look at [38] for more details about the data.

1.3.2 Smoking cessation data

Smoking cessation data set [39, 40] contains 223 subjects' observations who had enrolled for the study during November 1986 to February 1989 [41, 42]. Only those subjects who had tried to quit smoking at least once and who had identifiable Minnesota zip codes during the study period are considered in the analysis set. These subjects were all smokers at the time of enrollment, and were randomly assigned to two groups, namely, the smoking intervention (SI, treatment group) and the usual care (UC, control group). The subjects were monitored once every year for a period of 5 consecutive years. Information on whether they had relapsed or not (1:Yes and 0:No) are present in the data set. Relapse implies resumption of smoking and the event of interest for our illustration is the time to relapse. Obviously, the exact relapse time was unobserved since the relapse could have happened anytime in between two consecutive annual visits. Hence, the study falls under the scope of interval censored data analysis. Information on several additional variables are also available, e.g., gender (GEN, 1:Female and 0:Male), duration of smoking (DUR, time in years elapsed between commencement of smoking and entry to the study) and average number of cigarettes smoked per day (AVGCIG) before the study period. These variables are treated as covariates since these factors supposedly can influence the relapse.

1.4 Scope of the thesis

In this thesis, we study Box-Cox cure rate model introduced by Yin and Ibrahim [22] with lifetime following generalized gamma distribution in chapter 2. We present the model fitting, model discrimination and sensitivity studies results of GGBCT cure rate model. Moreover, we compare the performance of the proposed GGBCT cure rate model with the piecewise exponential Box-Cox transformation cure rate model that was originally proposed by Yin and Ibrahim [22]. Finally, we illustrate the flexibility of the proposed GGBCT model using a real data on breast cancer study and again compare our proposed approach with thepiecewise exponential approach of Yin and Ibrahim [22] for the considered data.

We discuss about the mixture cure rate model framework for interval-censored data and develop an estimation procedure based on the expectation maximization (EM) algorithm that employs the SVM to model the incidence part in chapter 3. Detailed simulation study is carried out to demonstrate the performance of our proposed model in terms of flexibility, accuracy and robustness. We also compare this model with the existing logistic regression based mixturecure rate models, the model performance is further examined and illustrated through an interval censored data on smoking cessation.

CHAPTER 2

A generalized gamma Box-Cox transformation cure rate model

2.1 Introduction

According to Boag [3], the population survival function can be represented by

$$S_p(y) = p_0 + (1 - p_0)S_u(y), \tag{2.1}$$

where p_0 is the cure proportion and $S_u(\cdot)$ is the survival function of the susceptible group. Usually, p_0 is linked to a set of covariates $\boldsymbol{x}$ using a logistic function

$$p_0 = \frac{1}{1 + e^{\boldsymbol{x}'\boldsymbol{\beta}}}, \tag{2.2}$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients. It is clear from eqn.(2.1) that $S_p(y)$ is an improper survival function, meaning that $\lim_{y\to\infty} S_p(y) = p_0$. If J denotes the latent cured status variable and Y denotes the time-to-event variable, then, $S_u(y) = P[Y > y|J = 1]$, where J takes the value one if the subject is susceptible and it takes the value zero if the subject is cured. It is also easy to see that $p_0 = P[J = 0]$. Since the introduction of this mixture cure rate model, many researchers have studied this model by making different assumptions on $S_u(\cdot)$, by proposing different ways to include covariates, and by proposing different approaches for parameter estimation.

As discussed before, let M be a latent random variable denoting the number of competing risks that can result in an event of interest. Furthermore, let W_i, $i = 1, 2, \cdots, M$, denote the lifetime due to the i-th competing risk. If we assume M to follow a Poisson distribution with mean η, it can be shown that the population survival function

$$S_p(y) = e^{-\eta F(y)}, \tag{2.3}$$

where $F(\cdot)$ is the common distribution function of the progression times W_i, $i = 1, 2, \cdots, M$. In this case, the cure rate is given by

$$p_0 = \lim_{y \to \infty} S_p(y) = e^{-\eta}. \tag{2.4}$$

Covariates can be introduced through the parameter η using the log-linear function $\eta = e^{\boldsymbol{x}'\boldsymbol{\beta}}$. Furthermore, note the survival function of the susceptible subjects can be expressed as

$$S_u(y) = \frac{e^{-\eta F(y)} - e^{-\eta}}{1 - e^{-\eta}}. \tag{2.5}$$

In this chapter, we consider the unified cure rate model proposed by Yin and Ibrahim [22] that contains both mixture cure rate model in eqn.(2.1) and promotion time cure rate model in eqn.(2.3) as special cases. Assuming a completely parametric framework, the main contribution is in proposing a flexible distribution for the lifetime and demonstrating its importance through model discrimination and sensitivity studies. Model fitting results for such a flexible cure rate model is also of primary interest. Furthermore, we compare our parametric approach with the piecewise exponential approach considered by Yin and Ibrahim[22].

The rest of this chapter is organized as follows. In Section 2.2, we introduce the generalized gamma Box-Cox transformation (GGBCT) cure rate model that unifies the mixture cure rate model of Boag [3] and promotion time cure rate model of Chen et al. [11] and also provides flexibility in modeling $F(\cdot)$. We also discuss different ways of carrying out model discrimination study for the proposed GGBCT cure rate model. In Section 2.3, we present the model fitting results of the GGBCT cure rate model through the calculated bias, root mean square error (RMSE) and the coverage probabilities of the asymptotic confidence intervals. We also present the model discrimination results, where we demonstrate the power of the likelihood ratio test to reject a mis-specified lifetime distribution for a given Box-Cox model as well as the power of the likelihood ratio test to reject a mis-specified Box-Cox model for a given lifetime distribution. Furthermore, we demonstrate the sensitivity of the estimators of cure rates under model mis-specification. In particular, we study the sensitivity with respect to the two quantities of interest, total relative bias and total relative efficiency, of the estimators of the cure rates. Moreover, through simulated data, we compare the performance of the proposed GGBCT cure rate model with the piecewise exponential Box-Cox transformation cure rate model that was originally proposed by Yin and Ibrahim [22]. In Section 2.4, we illustrate the flexibility of the proposed GGBCT model using a real data on breast cancer study and again compare our proposed approach with the piecewise exponential approach of Yin and Ibrahim [22] for the considered data. We show that our proposed approach results in a better model fit.

2.2 Generalized gamma Box-Cox transformation (GGBCT) cure rate model

The mixture and the promotion time cure rate models are the two most commonly used cure rate models, where one may be looked as a competitor of the other. Yin and Ibrahim [22] first proposed a wider class of cure rate models that contain both mixture and promotion time cure rate models as special cases. The proposed wider class was indexed by a link parameter and was built using a Box-Cox transformation [23] on the population survival function. The Box-Cox transformation on a variable Z, indexed by a transformation parameter ϕ, is defined as

$$Z^{(\phi)} = \begin{cases} \frac{Z^{\phi}-1}{\phi}, & \text{if } \phi \neq 0, \\ \log(Z), & \text{if } \phi = 0. \end{cases} \tag{2.6}$$

Now, if we apply the Box-Cox transformation on the population survival function that depends on a set of covariates $\boldsymbol{x}$, the Box-Cox transformation cure rate model is defined as

$$S_p^{(\phi)}(y|\boldsymbol{x}) = -\psi(\phi, \boldsymbol{x})F(y), \;\; 0 \leq \phi \leq 1, \tag{2.7}$$

where

$$\psi(\phi, \boldsymbol{x}) = \begin{cases} \frac{\exp(\boldsymbol{x}'\boldsymbol{\beta})}{1+\phi\exp(\boldsymbol{x}'\boldsymbol{\beta})}, & \text{if } 0 < \phi \leq 1, \\ \exp(\boldsymbol{x}'\boldsymbol{\beta}), & \text{if } \phi = 0 \end{cases} \tag{2.8}$$

and $F(\cdot)$ is a proper distribution function. On applying eqn.(2.6) in the left hand side of eqn.(2.7), the population survival function can be expressed as

$$S_p(y|\boldsymbol{x}) = \begin{cases} \{1 - \phi\psi(\phi, \boldsymbol{x})F(y)\}^{\frac{1}{\phi}}, & \text{if } 0 < \phi \leq 1, \\ \exp\{-\psi(0, \boldsymbol{x})F(y)\}, & \text{if } \phi = 0. \end{cases} \tag{2.9}$$

From eqns.(2.8) and (2.9), it is easy to verify that if $\phi = 1$,

$$\begin{aligned} S_p(y|\boldsymbol{x}) &= 1 - \frac{\exp(\boldsymbol{x}'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}'\boldsymbol{\beta})}F(y) \\ &= p_0 + (1 - p_0)S(y), \end{aligned} \tag{2.10}$$

which reduces to the mixture cure rate model in eqn.(2.1) with $p_0 = \frac{1}{1+\exp(\boldsymbol{x}'\boldsymbol{\beta})}$ and $S(y) = 1 - F(y)$. Similarly, if $\phi = 0$ in eqn.(2.9), we have

$$S_p(y|\boldsymbol{x}) = \exp\{-\eta F(y)\}, \tag{2.11}$$

which reduces to the promotion time cure rate model in eqn.(2.3) with $\eta = \psi(0, \boldsymbol{x}) = \exp(\boldsymbol{x}'\boldsymbol{\beta})$. Thus, the Box-Cox transformation cure rate model is an attractive and elegant way to unify the mixture and promotion time cure rate models. Furthermore, eqn.(2.8) introduces a general link function that allows us to study the effect of covariates on the cure rate. It is important to note that our main interest for ϕ is in the interval $[0, 1]$ as it results in an intermediate modeling structure between the promotion time or Poisson cure rate model ($\phi = 0$) and the mixture cure rate model ($\phi = 1$). Mathematically, ϕ can take any value in the real line.

From eqn.(2.9), the expression of the cure rate can be easily obtained as

$$p_0(\boldsymbol{x}) = \lim_{y\to\infty} S_p(y|\boldsymbol{x}) = \begin{cases} [1-\phi\psi(\phi,\boldsymbol{x})]^{\frac{1}{\phi}}, & \text{if } 0<\phi\le 1, \\ \exp\{-\psi(0,\boldsymbol{x})\}, & \text{if } \phi=0. \end{cases} \tag{2.12}$$

$$= \begin{cases} \left[\frac{1}{1+\phi\exp(\boldsymbol{x}'\boldsymbol{\beta})}\right]^{\frac{1}{\phi}}, & \text{if } 0<\phi\le 1, \\ \exp\{-\exp(\boldsymbol{x}'\boldsymbol{\beta})\}, & \text{if } \phi=0. \end{cases}$$

From eqn.(2.9), the population density function, denoted by $f_p(\cdot)$, can also be obtained as

$$f_p(y|\boldsymbol{x}) = -S'_p(y|\boldsymbol{x}) = \begin{cases} S_p(y|\boldsymbol{x})\psi(\phi,\boldsymbol{x})f(y)\{1-\phi\psi(\phi,\boldsymbol{x})F(y)\}^{-1}, & \text{if } 0<\phi\le 1, \\ S_p(y|\boldsymbol{x})\psi(0,\boldsymbol{x})f(y), & \text{if } \phi=0, \end{cases} \tag{2.13}$$

where $f(\cdot)$ is the density function corresponding to $F(\cdot)$. We now turn our attention to flexible modeling of $f(\cdot)$ or, equivalently, $F(\cdot) = 1 - S(\cdot)$ in eqns. (2.9) and (2.13). For this purpose, we consider a fully parametric setup and assume $f(\cdot)$ in eqn.(2.13) to follow a wider class of generalized gamma distribution with the density and survival functions respectively given by

$$f(y) = \begin{cases} \dfrac{q\left(q^{-2}\right)^{q^{-2}}(\lambda y)^{q^{-2}(q/\sigma)}\, e^{-q^{-2}(\lambda y)^{q/\sigma}}}{\Gamma\left(q^{-2}\right)\sigma y}, & \text{if } q>0 \\ \dfrac{e^{-(\ln(\lambda y))^2/\left(2\sigma^2\right)}}{\sqrt{2\pi}\sigma y}, & \text{if } q=0 \end{cases} \tag{2.14}$$

and

$$S(y) = \begin{cases} \dfrac{\Gamma\left(q^{-2}, q^{-2}(\lambda y)^{q/\sigma}\right)}{\Gamma\left(q^{-2}\right)}, & \text{if } q>0 \\ 1-\Phi\left(\frac{\ln(\lambda y)}{\sigma}\right), & \text{if } q=0. \end{cases} \tag{2.15}$$

Therefore, we have

$$F(y) = 1-S(y) = \begin{cases} 1-\dfrac{\Gamma\left(q^{-2}, q^{-2}(\lambda y)^{q/\sigma}\right)}{\Gamma\left(q^{-2}\right)}, & \text{if } q>0 \\ \Phi\left(\frac{\ln(\lambda y)}{\sigma}\right), & \text{if } q=0, \end{cases} \tag{2.16}$$

where $q > 0$ and $\sigma > 0$ are the shape parameters, whereas $\lambda > 0$ is the scale parameter. Also, $\Gamma(\cdot)$ represents the complete gamma function and $\Phi(\cdot)$ represents the distribution function of a standard normal distribution. Some of the commonly used lifetime distributions are included as special cases of the generalized gamma distribution. As an illustration, the generalized gamma distribution in eqn.(2.14) reduces to a Weibull distribution when $q = 1$, it reduces to a lognormal distribution when $q \to 0$, and, finally, it reduces to a gamma distribution when $q = \sigma$. Thus, the introduction of the generalized gamma distribution brings in adequate flexibility in cure rate modeling which may be easily missed if we just use its special cases.

Once we substitute $F(\cdot)$ in eqn.(2.9) with the distribution function of the generalized gamma distribution as presented in eqn.(2.16), we introduce a two-way flexible cure rate model that has not been studied before. The first flexibility is with respect to the Box-Cox cure rate model which contains the two most commonly used cure rate models in the literature, whereas the other flexibility is with respect to the generalized gamma distribution, as a distribution to model the lifetime, that contains the commonly used lifetime distributions. Such a two-way flexible model will allow us to determine a suitable cure rate model (within the Box-Cox family) and a suitable lifetime distribution (within the generalized gamma family) that will jointly provide the best fit to a given time-to-event data. We call this two-way flexible cure rate model as the generalized gamma Box-Cox transformation (GGBCT) cure rate model.

2.2.1 Likelihood function and estimation

Considering the form of the data to be right censored, let Y denote the true time-to-event variable and T denote the observed time-to-event. If C denotes the right censoring time, then, $T = \min\{Y, C\}$. Furthermore, if δ denotes the right censoring indicator, then, $\delta = I(Y \leq C)$ with $I(A) = 1$, if the event A is true, and is 0, otherwise. Now, assuming the censoring mechanism to be non-informative, the observed data log-likelihood function can be expressed as

$$l(\boldsymbol{\theta}) = \sum_{i:\delta_i=1} \log f_p(t_i|\boldsymbol{x}_i) + \sum_{i:\delta_i=0} \log S_p(t_i|\boldsymbol{x}_i), \tag{2.17}$$

where $S_p(\cdot)$ and $f_p(\cdot)$ are as in eqns. (2.9) and (2.13), respectively, and $\boldsymbol{\theta} = (\boldsymbol{\beta}', \phi, q, \sigma, \lambda)'$ denotes the vector of unknown parameters. The maximum likelihood estimates (MLEs) of the model parameters can be obtained by directly maximizing $l(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. For this purpose, we use the function "nlm()" readily available in R software. Note that under a similar parametric setup with a Weibull distribution for the lifetime, Pal and Balakrishnan [43] developed the EM algorithm for parameter estimation, where the maximization step of the EM algorithm was carried out using a one-step Newton Raphson method. The authors noted that the simultaneous maximization of the model parameters was not possible. To circumvent this issue, the Box-Cox transformation parameter ϕ was kept fixed and estimated using a profiling technique in conjunction with the EM algorithm. Although this technique performed satisfactorily, there were issues with the estimation results such as under-coverage of the model parameters. In this work, even though we are dealing with a more complicated lifetime distribution, we show that the "nlm()" [44] performs very well in simultaneously maximizing all model parameters and in retrieving the true parameter values very accurately.

Once we obtain the MLEs of the model parameters, we can calculate the asymptotic variances and covariances of the MLEs by inverting the observed information matrix and evaluating it at the MLEs. Then, we can construct the asymptotic confidence intervals of the parameters by using the asymptotic normality of the MLEs and the estimate of the asymptotic variance-covariance matrix of the MLEs. To judge the accuracy of this asymptotic method, we can study the coverage probabilities of these confidence intervals through a Monte Carlo simulation study.

The “nlm()” function requires us to provide initial values of the model parameters to start the iterative algorithm. The procedure for finding the initial values may differ depending on whether we are analyzing a simulated data or whether we are analyzing a real data. Since in a simulation data the true parameter values are known, for a given model parameter, say θ, we can first create an interval by taking 20% deviation off its true value on either side, i.e., $(0.8\theta, 1.2\theta)$. Then, we can randomly select a value from this interval which may be used as the parameter’s initial value. This simple procedure will not work for the real data analysis since we do not know what the true parameter values are. Furthermore, the procedure for finding the initial values for a real data may depend on the type of covariates as well as the number of covariates. We provide a technique for finding initial values for the breast cancer data with a categorical covariate (prognostic group status having three levels) that we analyze later in Section 2.4. For this purpose, we first plot the non-parametric Kaplan-Meier survival curves stratified by the prognostic group status variable. Then, from the leveling off tendency of each survival curve, we can guess the crude estimates of the cure rates of subjects belonging to different group status. Since we have one covariate, we introduce two regression parameters β_0 and β_1. Now, to find an initial guess of the parameters β_0, β_1 and ϕ, we can use these three crude estimates of cure rates and equate them to their corresponding theoretical expressions. This gives us three equations involving three unknown parameters (β_0, β_1 and ϕ) solving which we can obtain the initial values of β_0, β_1 and ϕ. Next, to find the initial values of the generalized gamma lifetime parameters, q, σ and λ, we can first select a set of fixed values of the generalized gamma shape parameter q, e.g., $q = \{0, 0.1, 0.2, \cdots, 2\}$. Then, for each fixed value of q, we can equate the mean and the variance of the observed time-to-event data to the theoretical mean and variance of the generalized gamma distribution. Thus, for each fixed value of q, we can solve these equations to find the values of the other two generalized gamma parameters, i.e., σ and λ. This gives us a set of values of q, σ and λ. Finally, we select the set $(q, \sigma, \lambda, \beta_0, \beta_1, \phi)$ as the set of initial values for which the observed data log-likelihood function is the maximum.

2.2.2 Model discrimination

Since we are dealing with a two-way flexible model, we propose different ways to carry out model discrimination studies. In the first case, we utilize the flexibility of the Box-Cox transformation cure rate model for a given (or fixed) lifetime distribution. For this purpose, we vary the transformation parameter ϕ and generate data from different Box-Cox models. In particular, we use different values of ϕ as $\phi = \{0, 0.25, 0.50, 0.75, 1\}$ that covers a wide range of Box-Cox models in the interval $\phi \in [0, 1]$. For each generated data (true model), we fit different models (fitted model) and evaluate the performance of the likelihood ratio test in rejecting each fitted model. Based on a Monte Carlo simulation study, we can report the observed levels as well as the observed rejection rates of the likelihood ratio test, where all tests can be carried out at, say, 5% level of significance. The likelihood ratio test statistic is defined as $\Lambda = -2(l_0 - l)$, where l_0 is the maximized log-likelihood value under the constrained model (i.e., under the null hypothesis) and l is the maximized log-likelihood value under the unconstrained model (i.e., under the full model). The asymptotic null distribution of the likelihood ratio test statistic Λ is a chi-square with one degree of freedom, under the standard likelihood theory. However, when testing is done in the boundary of a parameter space, such as testing for $H_0 : \phi = 0$ and $H_0 : \phi = 1$, the asymptotic null distribution of Λ is a mixture of chi-square distributions (Self and Liang, 1987), i.e., the asymptotic null distribution of Λ is such that $P[\Lambda \leq x] = 0.5 + 0.5P[\chi_1^2 \leq x]$.

In the second case, we utilize the flexibility of the generalized gamma lifetime for a given Box-Cox model. For this purpose, we vary the generalized gamma shape parameter q and generate data from different lifetime distributions within the generalized gamma family. In particular, we can consider different values of q as $q = \{0, 0.5, \sigma, 1\}$. For each generated data (true model), we fit different lifetimes (fitted model) belonging to the wider class of generalized gamma lifetime and once again evaluate the performance of the likelihood ratio test in rejecting each fitted model. Based on a Monte Carlo simulation study, we can report the observed levels and the observed rejection rates of the likelihood ratio test, where all tests can be carried out at, say, 5% level of significance.

Note that the aforementioned ways of carrying out model discrimination require us to fit the full model under consideration and involves actual tests of hypotheses. An alternate simpler way is to compare the candidate models using some well-known information-based criteria such as the Akaike information criterion (AIC), the corrected Akaike information criterion (AICc) and the Bayesian information criterion (BIC). The definitions of the AIC, AICc and BIC are as follows:

$$AIC = -2l + 2k, \;\; AICc = AIC + \frac{2k^2 + 2k}{n - k - 1} \;\text{ and }\; BIC = -2l + k\log(n).$$

In the above formulae, k denotes the number of parameters in the fitted model and n denotes the sample size. For a given information-based criterion, a smaller value would imply a better model fit. It must be noted here that using these information-based criteria do not give users any warning of how good or how bad the model fit is. For example, it may happen that all candidate models give poor fit to the data, however, these information-based criteria would still select a model as the best fitted model.

2.3 Simulation study

2.3.1 Model fitting with one binary covariate

In this section, we first describe the mechanism to generate data from the GGBCT model when there is one binary covariate present. We consider a scenario where subjects are randomly assigned to either the treatment group or the placebo group. In this way, we bring in the covariate x, where we assign $x = 1$ for subjects belonging to the treatment group and $x = 0$ for those belonging to the control group. The cure proportions for the treatment and control groups are respectively denoted by p_{0t} and p_{0c}. If we fix the true values of p_{0t}, p_{0c} and ϕ, the two regression parameters, β_0 and β_1, can be obtained using the following expressions:

$$\beta_0 = \begin{cases} \log\left[\frac{1}{\phi}\left\{\frac{1}{p_{0c}^{\phi}} - 1\right\}\right], & \text{if } 0 < \phi \leq 1 \\ \log\{-\log(p_{0c})\}, & \text{if } \phi = 0 \end{cases}$$

and

$$\beta_1 = \begin{cases} \log\left[\frac{1}{\phi}\left\{\frac{1}{p_{0t}^{\phi}} - 1\right\}\right] - \beta_0, & \text{if } 0 < \phi \leq 1 \\ \log\{-\log(p_{0t})\} - \beta_0, & \text{if } \phi = 0. \end{cases}$$

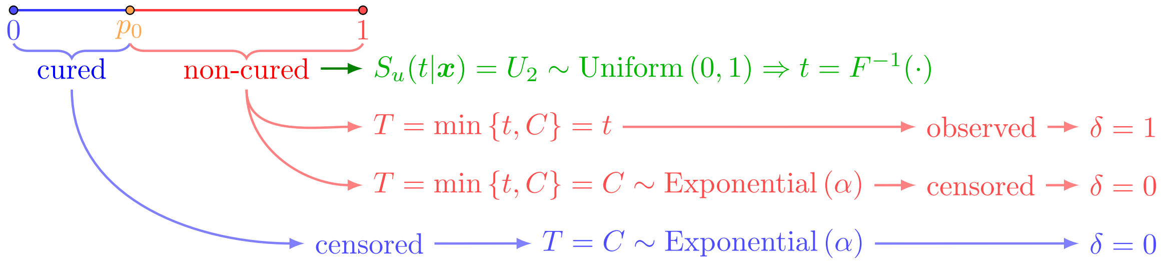
Now, to generate the observed time-to-event data T for a subject with cure rate p_0 (which can be either p_{0t} or p_{0c}) and covariate x (which can be either 1 or 0) from the GGBCT model, we first generate $U_1 \sim Uniform(0, 1)$ and censoring time $C \sim Exponential(rate = \alpha)$. If $U_1 \leq p_0$, we set the observed time as the censoring time, i.e., $T = C$. On the other hand, if $U_1 > p_0$, it implies that the subject is susceptible, and we generate $U_2 \sim Uniform(0,1)$ and perform the following:

$$\begin{cases} \frac{\{1-\phi\psi(\phi,x)F(t)\}^{\frac{1}{\phi}}-p_0}{1-p_0} = U_2, & \text{if } 0 < \phi \leq 1 \\ \\ \frac{\exp\{-\psi(0,x)F(t)\}-p_0}{1-p_0} = U_2, & \text{if } \phi = 0 \end{cases}$$

$$\Rightarrow \begin{cases} F(t) = \frac{1-\{p_0+(1-p_0)U_2\}^{\phi}}{\phi\psi(\phi,x)}, & \text{if } 0 < \phi \leq 1 \\ \\ F(t) = \frac{-\log\{p_0+(1-p_0)U_2\}}{\psi(0,x)}, & \text{if } \phi = 0 \end{cases}$$

$$\Rightarrow \begin{cases} t = F^{-1}\left[\frac{1-\{p_0+(1-p_0)U_2\}^{\phi}}{\phi\psi(\phi,x)}\right], & \text{if } 0 < \phi \leq 1 \\ \\ t = F^{-1}\left[\frac{-\log\{p_0+(1-p_0)U_2\}}{\psi(0,x)}\right], & \text{if } \phi = 0, \end{cases}$$

where $F(\cdot)$ is the distribution function of the generalized gamma distribution defined in eqn.(2.16). Finally, the observed time is given by $T = \min\{t, C\}$. In any case, if $T = C$, we set the censoring indicator $\delta = 0$. Else, we set $\delta = 1$. Let n denote the sample size which is split into two parts, n_1 and n_2, with n_1 denoting the sample size for the treatment group and n_2 denoting the sample size for the control group. We consider different values of the sample size $n(n_1, n_2)$ as $n = 200(130, 70)$ and $n = 400(230, 170)$. We also consider the true values of the cure rates for the treatment and control groups as $p_{0t} = 0.65$ and $p_{0c} = 0.35$, respectively, and the censoring rates for the treatment and control groups as $\alpha_t = 0.25$ and $\alpha_c = 0.15$, respectively. To decide on the true values of the lifetime parameters, we first equate the theoretical mean and the theoretical variance of the underlying lifetime distribution to some fixed values. For example, if the lifetime distribution is generalized gamma with parameters q, σ and λ, we equate the theoretical mean $\frac{(q^2)^{\sigma/q}}{\lambda\Gamma(1/q^2)}\Gamma(\frac{\sigma}{q} + \frac{1}{q^2})$ and the theoretical variance $\frac{q^{4\sigma/q}}{\lambda^2}\frac{1}{\Gamma(1/q^2)}\left[\Gamma(\frac{2\sigma}{q}+\frac{1}{q^2})-\frac{\Gamma^2(\frac{\sigma}{q}+\frac{1}{q^2})}{\Gamma(1/q^2)}\right]$ to 0.25 and 0.05, respectively, after fixing a true value for q. Thus, upon solving these two equations we can find the true values of σ and λ. Note that the parameter q being the shape parameter, for different fixed values of q we get different values of σ and λ for the same chosen fixed values of the mean and the variance. Figure 2.1 represents a schematic view of the data generation. All simulations are done using the R statistical software and the results are averaged over 1000 Monte Carlo runs.



0
p_0
1
cured
non-cured
$S_u(t|\boldsymbol{x}) = U_2 \sim \text{Uniform}\,(0,1) \Rightarrow t = F^{-1}(\cdot)$
$T = \min\{t, C\} = t$
observed
$\delta = 1$
$T = \min\{t, C\} = C \sim \text{Exponential}\,(\alpha)$
censored
$\delta = 0$
censored
$T = C \sim \text{Exponential}\,(\alpha)$
$\delta = 0$

Figure 2.1: Data generation: schematic diagram.

In Tables 2.1 and 2.2, we present the model fitting results, i.e., the bias, root mean square error (RMSE), standard error (SE) and coverage probabilities (CP), for the two special cases of the Box-Cox transformation cure rate model. In each of these cases, we consider the distribution of the lifetime to be generalized gamma as well as its special cases. From the tables, it is clear that the "nlm()" performs very well in estimating the true parameters values accurately. The bias in the estimators are reasonably small and both SE and RMSE decrease when the sample size is large. The coverage probabilities are also close to the nominal levels used. Note that when the lifetime distribution is generalized gamma, the SE and RMSE corresponding to the parameter λ are large when compared to the other model parameters. However, they both decrease significantly when the sample size is large.

In Tables 2.3 and 2.4, we present the model fitting results for the Box-Cox $(\phi = 0.25)$ and Box-Cox $(\phi = 0.75)$ cure rate models, respectively. Note that in these cases, the parameter ϕ is also estimated along with the other model parameters. The "nlm()" once again retrieves the true parameter values quite accurately. The SE and RMSE once again decrease with an increase in sample size. We do notice a slight over-coverage for the Box-Cox transformation parameter ϕ. The SE and RMSE of ϕ are large when compared to other model parameters and this is true for any considered lifetime distribution. In the case of generalized gamma lifetime, the SE and RMSE of λ also turns out to be large. Once again, when the sample size is large, a decrease in both SE and RMSE can be seen.

2.3.2 Model fitting with two covariates: one binary and one continuous

In this section, we first describe the mechanism to generate data from the GGBCT model when we consider one binary covariate (x_1) and one continuous covariate (x_2). We consider x_1 to be the group covariate, where we assign $x_1 = 1$ for subjects belonging to the treatment group and assign $x_1 = 0$ for subjects belonging to the control group. Similarly, we can consider x_2 to be the tumor thickness (measured in mm), where we assume the minimum tumor thickness to be 0.1mm and the maximum tumor thickness to be 20 mm. To generate the tumor thickness data, we can simply generate random numbers from Uniform(0.1,20) distribution. Once again, we consider two different values of the sample size $n(n_1, n_2)$ as 200 (130,70) and $400(230, 170)$, where n_1 denotes the sample size for the treatment group and n_2 denotes the sample size for the control group. Since we consider two covariates, we bring in three regression parameters, i.e., β_0, β_1 and β_2. In this simulation study, we consider the true values of β_0, β_1 and β_2 as 0.4, -0.5 and 0.1, respectively. We also consider two different true values of ϕ as 0.25 and 0.75. Note that due

Table 2.1: Model fitting results for the Poisson cure rate model with generalized gamma lifetime distribution and its special cases. For the lognormal and gamma lifetime, the true cure rates are $(p_{0t}, p_{0c}) = (0.65, 0.25)$, whereas for the generalized gamma and Weibull lifetimes, the true cure rates are $(p_{0t}, p_{0c}) = (0.65, 0.35)$

Lifetime	$n(n_1, n_2)$	Parameter	SE	Bias	RMSE	CP	
						0.90	0.95
Gamma	$200(130, 70)$	$\beta_0 = 0.327$	0.145	-0.003	0.150	0.880	0.944
		$\beta_1 = -1.169$	0.206	-0.002	0.218	0.878	0.932
		$\sigma = 0.447$	0.030	-0.009	0.089	0.872	0.938
		$\lambda = 2.00$	0.101	0.005	0.101	0.896	0.950
	$400(230, 170)$	$\beta_0 = 0.327$	0.093	-0.002	0.090	0.906	0.958
		$\beta_1 = -1.169$	0.144	0.003	0.142	0.906	0.960
		$\sigma = 0.447$	0.020	-0.012	0.104	0.902	0.946
		$\lambda = 2.00$	0.070	-0.001	0.068	0.916	0.960
Generalized Gamma	$200(130, 70)$	$\beta_0 = 0.049$	0.283	-0.005	0.289	0.886	0.944
		$\beta_1 = -0.891$	0.309	-0.002	0.308	0.902	0.946
		$q = 0.500$	0.306	0.031	0.311	0.938	0.966
		$\lambda = 4.632$	0.779	0.032	0.737	0.892	0.950
		$\sigma = 0.871$	0.086	-0.029	0.116	0.878	0.936
	$400(230, 170)$	$\beta_0 = 0.049$	0.098	-0.004	0.096	0.914	0.952
		$\beta_1 = -0.891$	0.148	-0.004	0.151	0.916	0.954
		$q = 0.500$	0.179	-0.007	0.175	0.912	0.970
		$\lambda = 4.632$	0.487	0.074	0.483	0.906	0.966
		$\sigma = 0.871$	0.054	-0.006	0.055	0.910	0.950
Lognormal	$200(130, 70)$	$\beta_0 = 0.327$	0.145	0.147	0.147	0.896	0.934
		$\beta_1 = -1.169$	0.206	0.213	0.213	0.902	0.952
		$\sigma = 0.241$	0.017	0.018	0.018	0.872	0.932
		$\lambda = 0.206$	0.005	0.006	0.006	0.872	0.944
	$400(230, 170)$	$\beta_0 = 0.327$	0.094	0.004	0.094	0.906	0.954
		$\beta_1 = -1.169$	0.145	-0.011	0.142	0.910	0.968
		$\sigma = 0.241$	0.012	-0.001	0.012	0.880	0.934
		$\lambda = 0.206$	0.004	0.000	0.004	0.908	0.960
Weibull	$200(130, 70)$	$\beta_0 = 0.049$	0.198	0.005	0.205	0.890	0.948
		$\beta_1 = -0.891$	0.283	-0.032	0.282	0.904	0.960
		$\sigma = 0.316$	0.033	-0.005	0.033	0.896	0.934
		$\lambda = 0.179$	0.010	0.001	0.011	0.888	0.942
	$400(230, 170)$	$\beta_0 = 0.049$	0.128	-0.012	0.129	0.910	0.954
		$\beta_1 = -0.891$	0.199	-0.001	0.205	0.910	0.950
		$\sigma = 0.316$	0.023	-0.002	0.024	0.902	0.936
		$\lambda = 0.179$	0.007	0.000	0.007	0.906	0.962

Table 2.2: Model fitting results for the Bernoulli cure rate model with generalized gamma lifetime distribution and its special cases. For the lognormal and gamma lifetime, the true cure rates are $(p_{0t}, p_{0c}) = (0.65, 0.25)$, whereas for the generalized gamma and Weibull lifetimes, the true cure rates are $(p_{0t}, p_{0c}) = (0.65, 0.35)$

Lifetime	$n(n_1, n_2)$	Parameter	SE	Bias	RMSE	CP	
						0.90	0.95
Gamma	$200(130, 70)$	$\beta_0 = 1.099$	0.279	-0.005	0.287	0.884	0.930
		$\beta_1 = -1.718$	0.335	0.009	0.342	0.894	0.944
		$\sigma = 0.447$	0.031	-0.010	0.081	0.884	0.940
		$\lambda = 2.000$	0.090	0.004	0.090	0.906	0.942
	$400(230, 170)$	$\beta_0 = 1.099$	0.179	0.012	0.183	0.900	0.952
		$\beta_1 = -1.718$	0.227	-0.011	0.225	0.908	0.952
		$\sigma = 0.447$	0.021	-0.007	0.071	0.894	0.940
		$\lambda = 2.000$	0.062	0.004	0.063	0.902	0.950
Generalized Gamma	$200(130, 70)$	$\beta_0 = 0.619$	0.253	0.019	0.258	0.920	0.948
		$\beta_1 = -1.238$	0.314	-0.028	0.329	0.896	0.954
		$q = 0.500$	0.272	0.010	0.258	0.934	0.984
		$\lambda = 4.632$	0.696	0.019	0.657	0.892	0.934
		$\sigma = 0.871$	0.073	-0.018	0.079	0.884	0.938
	$400(230, 170)$	$\beta_0 = 0.619$	0.161	-0.004	0.158	0.892	0.946
		$\beta_1 = -1.238$	0.213	-0.005	0.209	0.910	0.958
		$q = 0.500$	0.181	0.001	0.179	0.912	0.960
		$\lambda = 4.632$	0.474	0.016	0.462	0.908	0.954
		$\sigma = 0.871$	0.050	-0.012	0.090	0.904	0.952
Lognormal	$200(130, 70)$	$\beta_0 = 1.099$	0.282	0.002	0.279	0.894	0.950
		$\beta_1 = -1.718$	0.337	-0.004	0.341	0.898	0.954
		$\sigma = 0.241$	0.017	-0.001	0.018	0.882	0.920
		$\lambda = 0.206$	0.005	0.000	0.005	0.882	0.944
	$400(230, 170)$	$\beta_0 = 1.099$	0.180	0.010	0.186	0.890	0.956
		$\beta_1 = -1.718$	0.228	-0.025	0.233	0.910	0.938
		$\sigma = 0.241$	0.012	-0.001	0.011	0.906	0.948
		$\lambda = 0.206$	0.003	0.000	0.003	0.906	0.960
Weibull	$200(130, 70)$	$\beta_0 = 0.619$	0.346	0.008	0.369	0.892	0.950
		$\beta_1 = -1.238$	0.441	-0.035	0.471	0.880	0.938
		$\sigma = 0.316$	0.033	-0.009	0.034	0.866	0.926
		$\lambda = 0.179$	0.009	0.001	0.009	0.910	0.944
	$400(230, 170)$	$\beta_0 = 0.619$	0.220	0.011	0.219	0.918	0.952
		$\beta_1 = -1.238$	0.300	0.000	0.301	0.900	0.950
		$\sigma = 0.316$	0.023	-0.003	0.023	0.902	0.946
		$\lambda = 0.179$	0.006	0.000	0.006	0.928	0.966

Table 2.3: Model fitting results for the Box-Cox ($\phi = 0.25$) cure rate model with generalized gamma lifetime distribution and its special cases. For the lognormal and gamma lifetime, the true cure rates are $(p_{0t}, p_{0c}) = (0.65, 0.25)$, whereas for the generalized gamma and Weibull lifetimes, the true cure rates are $(p_{0t}, p_{0c}) = (0.65, 0.35)$

Lifetime	$n(n_1, n_2)$	Parameter	SE	Bias	RMSE	CP	
						0.90	0.95
Gamma	200(130,70)	$\beta_0 = 0.505$	0.679	0.088	0.712	0.918	0.952
		$\beta_1 = -1.293$	0.547	-0.075	0.558	0.900	0.948
		$\sigma = 0.447$	0.031	-0.007	0.073	0.884	0.940
		$\lambda = 2.000$	0.196	0.011	0.202	0.904	0.960
		$\phi = 0.250$	0.778	0.019	0.809	0.946	0.982
	400(230.170)	$\beta_0 = 0.505$	0.396	0.026	0.390	0.904	0.950
		$\beta_1 = -1.293$	0.322	-0.028	0.319	0.886	0.938
		$\sigma = 0.447$	0.022	-0.005	0.060	0.902	0.940
		$\lambda = 2.000$	0.139	0.005	0.136	0.914	0.962
		$\phi = 0.250$	0.495	-0.006	0.481	0.910	0.972
Generalized Gamma	200(130,70)	$\beta_0 = 0.183$	0.778	-0.213	0.754	0.876	0.936
		$\beta_1 = -0.971$	0.512	0.093	0.507	0.892	0.938
		$q = 0.500$	0.301	0.012	0.282	0.952	0.978
		$\lambda = 4.632$	1.602	0.249	1.501	0.904	0.940
		$\sigma = 0.871$	0.102	-0.037	0.103	0.910	0.958
		$\phi = 0.250$	1.417	-0.044	1.268	0.966	0.994
	400(230,170)	$\beta_0 = 0.183$	0.507	-0.198	0.547	0.896	0.964
		$\beta_1 = -0.971$	0.344	0.072	0.361	0.898	0.944
		$q = 0.500$	0.200	0.015	0.188	0.926	0.958
		$\lambda = 4.632$	1.088	0.083	0.992	0.908	0.938
		$\sigma = 0.871$	0.068	-0.018	0.068	0.904	0.964
		$\phi = 0.250$	0.900	-0.026	0.851	0.944	0.984
Lognormal	200(130,70)	$\beta_0 = 0.505$	0.614	-0.331	0.729	0.896	0.958
		$\beta_1 = -1.293$	0.487	0.209	0.575	0.887	0.938
		$\sigma = 0.241$	0.018	-0.002	0.018	0.900	0.936
		$\lambda = 0.206$	0.011	0.000	0.011	0.900	0.950
		$\phi = 0.250$	0.782	-0.037	0.813	0.934	0.978
	400(230,170)	$\beta_0 = 0.505$	0.402	-0.327	0.537	0.902	0.946
		$\beta_1 = -1.293$	0.326	0.198	0.407	0.893	0.951
		$\sigma = 0.241$	0.012	-0.001	0.013	0.876	0.938
		$\lambda = 0.206$	0.008	0.000	0.008	0.906	0.952
		$\phi = 0.250$	0.503	-0.011	0.492	0.924	0.966
Weibull	200(130,70)	$\beta_0 = 0.183$	1.188	0.082	1.307	0.926	0.976
		$\beta_1 = -0.971$	0.832	-0.136	0.928	0.886	0.934
		$\sigma = 0.316$	0.047	-0.003	0.046	0.902	0.946
		$\lambda = 0.179$	0.027	0.007	0.029	0.938	0.976
		$\phi = 0.250$	1.979	0.342	2.045	0.978	0.998
	400(230,170)	$\beta_0 = 0.183$	0.666	-0.142	0.679	0.894	0.948
		$\beta_1 = -0.971$	0.495	0.043	0.507	0.868	0.910
		$\sigma = 0.316$	0.032	-0.002	0.031	0.912	0.954
		$\lambda = 0.179$	0.017	0.003	0.017	0.930	0.970
		$\phi = 0.250$	1.127	0.122	1.088	0.946	0.976

Table 2.4: Model fitting results for the Box-Cox ($\phi = 0.75$) cure rate model with generalized gamma lifetime distribution and its special cases. For the lognormal and gamma lifetime, the true cure rates are $(p_{0t}, p_{0c}) = (0.65, 0.25)$, whereas for the generalized gamma and Weibull lifetimes, the true cure rates are $(p_{0t}, p_{0c}) = (0.65, 0.35)$

Lifetime	$n(n_1, n_2)$	Parameter	SE	Bias	RMSE	CP	
						0.90	0.95
Gamma	$200(130, 70)$	$\beta_0 = 0.891$	0.725	0.197	0.977	0.894	0.944
		$\beta_1 = -1.567$	0.574	-0.148	0.681	0.878	0.928
		$\sigma = 0.447$	0.031	-0.012	0.088	0.898	0.946
		$\lambda = 2.000$	0.202	0.037	0.237	0.894	0.952
		$\phi = 0.750$	0.819	0.134	1.081	0.910	0.968
	$400(230, 170)$	$\beta_0 = 0.891$	0.450	0.066	0.455	0.916	0.962
		$\beta_1 = -1.567$	0.371	-0.058	0.373	0.902	0.956
		$\sigma = 0.447$	0.021	-0.011	0.081	0.874	0.932
		$\lambda = 2.000$	0.139	0.013	0.146	0.902	0.950
		$\phi = 0.750$	0.506	0.032	0.502	0.940	0.970
Generalized Gamma	$200(130, 70)$	$\beta_0 = 0.468$	0.878	0.121	0.856	0.906	0.958
		$\beta_1 = -1.144$	0.577	-0.094	0.569	0.888	0.944
		$q = 0.500$	0.307	0.021	0.314	0.950	0.988
		$\lambda = 4.632$	1.663	0.263	1.601	0.886	0.924
		$\sigma = 0.871$	0.095	-0.041	0.103	0.892	0.936
		$\phi = 0.750$	1.497	0.076	1.439	0.954	0.990
	$400(230, 170)$	$\beta_0 = 0.468$	0.547	0.043	0.536	0.910	0.956
		$\beta_1 = -1.144$	0.380	-0.042	0.382	0.904	0.952
		$q = 0.500$	0.202	0.026	0.205	0.908	0.950
		$\lambda = 4.632$	1.077	0.083	1.054	0.882	0.936
		$\sigma = 0.871$	0.063	-0.023	0.068	0.916	0.956
		$\phi = 0.750$	0.897	0.012	0.874	0.918	0.972
Lognormal	$200(130, 70)$	$\beta_0 = 0.891$	0.547	0.056	0.732	0.882	0.934
		$\beta_1 = -1.567$	0.018	-0.052	0.589	0.864	0.920
		$\sigma = 0.241$	0.011	-0.004	0.019	0.868	0.916
		$\lambda = 0.206$	0.797	0.000	0.011	0.896	0.944
		$\phi = 0.750$	0.732	-0.013	0.846	0.916	0.976
	$400(230, 170)$	$\beta_0 = 0.891$	0.457	0.049	0.449	0.914	0.948
		$\beta_1 = -1.567$	0.374	-0.045	0.375	0.900	0.942
		$\sigma = 0.241$	0.012	-0.001	0.012	0.912	0.960
		$\lambda = 0.206$	0.008	0.001	0.008	0.890	0.966
		$\phi = 0.750$	0.522	0.035	0.524	0.928	0.958
Weibull	$200(130, 70)$	$\beta_0 = 0.468$	1.311	0.412	1.630	0.938	0.988
		$\beta_1 = -1.144$	0.923	-0.343	1.181	0.910	0.946
		$\sigma = 0.316$	0.046	-0.007	0.046	0.884	0.942
		$\lambda = 0.179$	0.026	0.007	0.030	0.928	0.976
		$\phi = 0.750$	2.057	0.333	2.187	0.982	0.998
	$400(230, 170)$	$\beta_0 = 0.468$	0.755	0.140	0.768	0.950	0.986
		$\beta_1 = -1.144$	0.557	-0.106	0.566	0.916	0.964
		$\sigma = 0.316$	0.031	-0.003	0.031	0.914	0.948
		$\lambda = 0.179$	0.017	0.004	0.018	0.928	0.962
		$\phi = 0.750$	1.205	0.141	1.220	0.954	0.990

to the presence of a continuous covariate, the cure rate for each subject will be different. In particular, the cure rate can be calculated by

$$p_0(x_1, x_2) \quad = \quad \begin{cases} \left[\frac{1}{1+\phi\exp(\beta_0+\beta_1 x_1+\beta_2 x_2)}\right]^{\frac{1}{\phi}}, & \text{if } 0 < \phi \leq 1, \\ \exp\{-\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)\}, & \text{if } \phi = 0. \end{cases}$$

We assume the lifetimes to follow a generalized gamma distribution and follow the same technique as described in Section 2.3.1 to find the true values of the generalized gamma lifetime parameters. To generate the observed time-to-event data T for a subject with cure rate $p_0(x_1, x_2)$, we first generate $U_1 \sim Uniform(0, 1)$ and censoring time $C \sim Exponential(rate = \alpha)$. If $U_1 \leq p_0(x_1, x_2)$, we set the observed time as the censoring time, i.e., $T = C$. On the other hand, if $U_1 > p_0(x_1, x_2)$, it implies that the subject is susceptible. In this case, we generate $U_2 \sim Uniform(0, 1)$ and calculate

$$\begin{cases} t = F^{-1}\left[\frac{1-\{p_0(x_1,x_2)+(1-p_0(x_1,x_2))U_2\}^{\phi}}{\phi\psi(\phi,x_1,x_2)}\right], & \text{if } 0 < \phi \leq 1 \\ \\ t = F^{-1}\left[\frac{-\log\{p_0(x_1,x_2)+(1-p_0(x_1,x_2))U_2\}}{\psi(0,x_1,x_2)}\right], & \text{if } \phi = 0, \end{cases}$$

where F is the distribution function of the generalized gamma distribution defined in eqn.(2.16). Finally, we calculate the observed time by $T = \min\{t, C\}$. Once again, if $T = C$, we set the censoring indicator $\delta = 0$, otherwise, we set $\delta = 1$. In Table 2.5, we present the model fitting results. It is easy to see that the "nlm()" once again performs very well in retrieving the true parameter values quite accurately. Note that the bias, standard error and RMSE all decreases with an increase in the sample size. The coverage probabilities are also close to the true nominal levels. Thus, even with an additional covariate (continuous), the performance of the "nlm()" is still very good.

Table 2.5: Model fitting results for the Box-Cox cure rate model with generalized gamma lifetime in the presence of two covariates

Lifetime	$n(n_1, n_2)$	Parameter	SE	Bias	RMSE	CP	
						0.90	0.95
Generalized Gamma	200(130,70)	$\beta_0 = 0.400$	0.366	0.042	0.394	0.910	0.954
		$\beta_1 = -0.500$	0.359	-0.077	0.381	0.904	0.942
		$\beta_2 = 0.100$	0.045	0.009	0.045	0.924	0.966
		$q = 0.500$	0.229	0.009	0.220	0.928	0.977
		$\lambda = 4.632$	1.135	0.163	1.137	0.904	0.946
		$\sigma = 0.871$	0.087	-0.019	0.086	0.908	0.947
		$\phi = 0.250$	0.294	0.013	0.280	0.940	0.979
	400(230,170)	$\beta_0 = 0.400$	0.214	0.003	0.227	0.901	0.948
		$\beta_1 = -0.500$	0.221	-0.035	0.230	0.908	0.950
		$\beta_2 = 0.100$	0.030	0.006	0.031	0.903	0.960
		$q = 0.500$	0.155	-0.004	0.154	0.900	0.949
		$\lambda = 4.632$	0.733	0.067	0.752	0.897	0.947
		$\sigma = 0.871$	0.059	-0.008	0.058	0.911	0.960
		$\phi = 0.250$	0.181	-0.005	0.181	0.913	0.966
	$200(130, 70)$	$\beta_0 = 0.400$	0.666	0.106	0.698	0.916	0.947
		$\beta_1 = -0.500$	0.487	-0.065	0.520	0.909	0.955
		$\beta_2 = 0.100$	0.058	-0.016	0.062	0.906	0.960
		$q = 0.500$	0.252	-0.008	0.242	0.915	0.974
		$\lambda = 4.632$	1.821	0.516	1.920	0.905	0.942
		$\sigma = 0.871$	0.082	-0.024	0.080	0.916	0.958
		$\phi = 0.750$	0.757	0.154	0.788	0.944	0.980
	$400(230, 170)$	$\beta_0 = 0.400$	0.369	0.042	0.424	0.890	0.940
		$\beta_1 = -0.500$	0.291	-0.040	0.316	0.887	0.932
		$\beta_2 = 0.100$	0.036	-0.006	0.039	0.901	0.941
		$q = 0.500$	0.170	0.001	0.172	0.891	0.946
		$\lambda = 4.632$	1.081	0.183	1.191	0.886	0.932
		$\sigma = 0.871$	0.053	-0.014	0.054	0.914	0.957
		$\phi = 0.750$	0.443	0.043	0.488	0.923	0.955

2.3.3 Model discrimination results using likelihood ratio test

In Table 2.6, we present the model discrimination results corresponding to utilizing the flexibility of the Box-Cox transformation cure rate model for a given lifetime distribution. Based on 1000 Monte Carlo runs, we report the observed levels (in bold) as well as the observed rejection rates of the likelihood ratio test. Overall, from Table 2.6, it is easy to see that for any given lifetime distribution the observed levels are close to the significance level. Thus, we can say that the chi-square and the mixture chi-square distributions both provide good approximations to the null distribution of the likelihood ratio test statistic. In some cases, however, the observed levels are found to be a little conservative. More interestingly, the observed rejection rates (can also be termed as the power of the likelihood ratio test to reject a wrong or mis-specified model) vary and depends on both true and fitted model. For example, under the assumption of generalized gamma lifetime, when the true model is $\phi = 0$, the rejection rates of the likelihood ratio test keep increasing as ϕ deviates from 0. In this regard, note that the likelihood ratio test rejects $\phi = 1$ with a high power, i.e., 82.5%. Similarly, when the true model is $\phi = 1$, the rejection rate or the power of the likelihood ratio test to reject $\phi = 0$ is 77% and such a rejection rate decreases as the fitted model gets closer to the true model. These findings clearly indicate that the likelihood ratio test can distinctly discriminate between $\phi = 0$ and $\phi = 1$ models. Now, when the true model is $\phi = 0.25$, the power values of the likelihood ratio test to reject $\phi = 0$, $\phi = 0.5$, and $\phi = 0.75$ are low to moderate only. However, in this case, the power to reject $\phi = 1$ is better and turns out to be quite high for other lifetime distributions. Thus, we can say that the likelihood ratio test can discriminate between $\phi = 0.25$ and $\phi = 1$ models. When the true model is $\phi = 0.5$, the power values of the likelihood ratio test to reject $\phi = 0.25$ and $\phi = 0.75$ are low. In this case, the likelihood ratio test only has moderate power to reject $\phi = 0$ and $\phi = 1$. Finally, when the true model is $\phi = 0.75$, the likelihood ratio test has low power to reject $\phi = 0.5$ and $\phi = 1$. In this case, the power to reject $\phi = 0$ is still high (much better for Weibull and lognormal lifetimes, and only moderate for gamma lifetimes), whereas the power to reject $\phi = 0.25$ is only moderate.

Table 2.6: Observed levels and observed rejection rates of the likelihood ratio test with respect to the Box-Cox model for a given lifetime distribution and with $n = 400$

	True Box-Cox cure rate model				
	Generalized Gamma Lifetime				
Fitted Model	$\phi = 0$	$\phi = 0.25$	$\phi = 0.5$	$\phi = 0.75$	$\phi = 1$
$\phi = 0$	**0.039**	0.394	0.559	0.659	0.769
$\phi = 0.25$	0.042	**0.061**	0.198	0.416	0.452
$\phi = 0.5$	0.234	0.112	**0.042**	0.110	0.245
$\phi = 0.75$	0.562	0.272	0.108	**0.029**	0.142
$\phi = 1$	0.825	0.630	0.480	0.273	**0.045**
	Weibull Lifetime				
$\phi = 0$	**0.041**	0.080	0.777	0.932	0.958
$\phi = 0.25$	0.064	**0.059**	0.041	0.572	0.758
$\phi = 0.5$	0.341	0.112	**0.034**	0.075	0.187
$\phi = 0.75$	0.655	0.512	0.107	**0.039**	0.068
$\phi = 1$	0.979	0.881	0.651	0.206	**0.047**
	Lognormal Lifetime				
$\phi = 0$	**0.041**	0.509	0.738	0.906	0.933
$\phi = 0.25$	0.187	**0.025**	0.249	0.487	0.627
$\phi = 0.5$	0.464	0.271	**0.037**	0.152	0.283
$\phi = 0.75$	0.843	0.733	0.232	**0.034**	0.102
$\phi = 1$	0.973	0.958	0.791	0.201	**0.039**
	Gamma Lifetime				
$\phi = 0$	**0.052**	0.094	0.183	0.437	0.872
$\phi = 0.25$	0.067	**0.038**	0.047	0.105	0.518
$\phi = 0.5$	0.625	0.059	**0.033**	0.034	0.099
$\phi = 0.75$	0.834	0.735	0.134	**0.026**	0.082
$\phi = 1$	0.939	0.931	0.788	0.111	**0.069**

In Figure 2.2, we plot the rejection rates of the likelihood ratio test for different true values of ϕ. For this purpose, we assume the lifetime distribution to be generalized gamma and the sample size to be 400. For each true value of ϕ, we fit models with different values of ϕ as $\phi = \{0, 0.1, 0.2, \cdots, 1\}$. The plot clearly shows that as the fitted model deviates from the true model, the power of the likelihood ratio test to reject the fitted model increases.

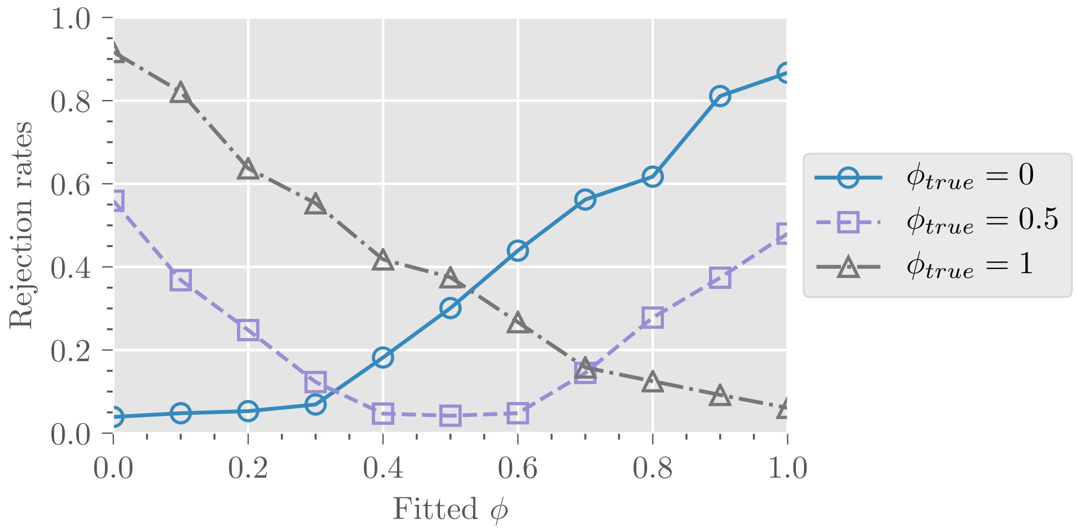


Figure 2.2: Plot showing the rejection rates of the likelihood ratio test

In Table 2.7, we present the model discrimination results corresponding to utilizing the flexibility of the generalized gamma lifetime for a given Box-Cox model. Based on 1000 Monte Carlo runs, we report the observed levels (in bold) and the observed rejection rates of the likelihood ratio test. First, it is clear that the observed levels are very close to the true level of significance. Next, for any fixed Box-Cox model, we can see that the likelihood ratio test has very high power to reject the Weibull (lognormal) lifetime when the true lifetime is lognormal (Weibull). Thus, the likelihood ratio test can distinctly discriminate between the Weibull and lognormal lifetimes. Similarly, the likelihood ratio test can also discriminate between the lognormal and generalized gamma $q = 0.5$ models as well as between the gamma and Weibull models with adequate powers. When the true lifetime is lognormal, the likelihood ratio test do not possess much power to reject the gamma lifetime, however, when the true lifetime is gamma, the likelihood ratio test do have a high power ($>$ 80%) to reject the lognormal lifetime. Finally, note that the likelihood ratio test lacks power to discriminate between the generalized gamma ($q = 0.5$) and gamma lifetimes although it does have moderate power to discriminate between the generalized gamma ($q = 0.5$) and Weibull lifetimes.

2.3.4 Sensitivity analysis

In this section, we study how sensitive the bias and the mean square error (MSE) of the estimators of the cure rates are when (i) a wrong cure rate model within the Box-Cox family is fitted assuming the lifetime to follow a generalized gamma distribution and (ii) a wrong lifetime distribution within the generalized gamma family is fitted for a given Box-Cox model. Since in our simulation study setup we have considered two groups, a combined measure of bias involved in the estimation of cure rates from both groups over 1000 Monte Carlo runs can be termed as the total relative bias (TRB) and can be defined as

$$\text{TRB} = \frac{1}{1000}\sum_{j=1}^{1000}\left\{\left[\frac{|\widehat{p_{0t,\, j}} - p_{0t(true\ value)}|}{p_{0t(true\ value)}}\right] + \left[\frac{|\widehat{p_{0c,\, j}} - p_{0c(true\ value)}|}{p_{0c(true\ value)}}\right]\right\}. \tag{2.18}$$

Similarly, we can use the MSE to define the total relative efficiency (TRE) under wrongly specified model (i.e., the fitted model) over 1000 Monte Carlo runs as

$$\text{TRE} = \frac{1}{1000}\sum_{j=1}^{1000}\frac{\text{MSE}(p_{0t(true\ model,\ j)}) + \text{MSE}(p_{0c(true\ model,\ j)})}{\text{MSE}(\widehat{p_{0t,\, j}}) + \text{MSE}(\widehat{p_{0c,\, j}})}. \tag{2.19}$$

In Table 2.8, we present the TRB values and the TRE values (in parenthesis) in the estimation of the cure rates when mis-specification is done with respect to the Box-Cox family of cure rate models. It is clear that when the true Box-Cox model and the fitted Box-Cox model are the same, the TRB is always the lowest, as one would expect. However, when the fitted Box-Cox model deviates from the true Box-Cox model, i.e., when the model is mis-specified, the TRB is seen to increase, whereas the TRE is seen to decrease. In this regard, note that when the true model is $\phi = 0$ ($\phi = 1$) and the fitted model is $\phi = 1$ ($\phi = 0$),

Table 2.7: Observed levels and observed rejection rates of the likelihood ratio test with respect to the generalized gamma distribution for a given Box-Cox model and with $n = 400$

	True lifetime distribution within generalized gamma family			
	Poisson ($\phi = 0$)			
Fitted Lifetime	Lognormal	Generalized Gamma ($q = 0.5$)	Gamma	Weibull
Lognormal	**0.058**	0.837	0.806	0.981
Generalized Gamma ($q = 0.5$)	0.725	**0.049**	0.057	0.556
Gamma	0.107	0.370	**0.044**	0.691
Weibull	0.999	0.758	0.869	**0.059**
	Bernoulli ($\phi = 1$)			
Lognormal	**0.034**	0.853	0.803	0.991
Generalized Gamma ($q = 0.5$)	0.744	**0.046**	0.182	0.588
Gamma	0.358	0.397	**0.040**	0.713
Weibull	0.998	0.774	0.841	**0.052**
	Box-Cox ($\phi = 0.75$)			
Lognormal	**0.043**	0.849	0.816	0.986
Generalized Gamma ($q = 0.5$)	0.729	**0.047**	0.253	0.627
Gamma	0.324	0.392	**0.051**	0.724
Weibull	0.999	0.764	0.857	**0.049**

the TRE is very low and varies between $15\% - 33\%$ only. The findings clearly suggest that for a given real data one should be very careful in choosing the correct cure rate model. Otherwise, the estimators of the cure rates will be biased and less efficient. Also, note that with an increase in the sample size, the TRB decreases whereas the TRE increases.

In Table 2.9, we present the TRB values and the TRE values (in parenthesis) in the estimation of the cure rates when mis-specification is done with respect to the lifetime distribution belonging to the wider class of the generalized gamma distribution. For any given cure rate model, i.e., either the Poisson model, or the Bernoulli model, or the Box-Cox $(\phi = 0.75)$ model, it is easy to see that when the lifetime distribution is correctly specified, the TRB is always the lowest. On the other hand, when the true lifetime distribution is lognormal (Weibull) and the fitted lifetime distribution is Weibull (lognormal), the TRB is the highest and the TRE is the lowest. This finding is in line with the findings from Table 2.7, where the likelihood ratio test was able to discriminate between the lognormal and Weibull lifetimes with the highest power. In general, for any considered scenario, when a wrong model is specified for the lifetime distribution, the TRB is much higher compared to the case when there is no mis-specification in the lifetime distribution. In this case of mis-specification of lifetime distribution, the TRE is also considerably less than one. These findings suggest that for a given real data, where we do not know the true lifetime distribution, one should be very careful in coming up with a proper choice of the lifetime distribution. If there is any mis-specification in the lifetime distribution, the resulting inference on the cure rates will be biased and less efficient. In section 2.4, through a real breast cancer survival data, we illustrate how the Box-Cox family of cure rate models together with the generalized gamma family of lifetime distributions can aid us in selecting a parsimonious cure rate model together with a parsimonious lifetime distribution that jointly provides the best fit.

Table 2.8: TRB and TRE (in parenthesis) in the estimation of cure rates when fitting different Box-Cox models for a given true Box-Cox model and assuming generalized gamma lifetime

	True Box-Cox Model				
Fitted Model	$\phi = 0$	$\phi = 0.25$	$\phi = 0.5$	$\phi = 0.75$	$\phi = 1$
			$n = 400$		
$\phi = 0$	0.083(-)	0.109(0.946)	0.135(0.524)	0.155(0.361)	0.173(0.259)
$\phi = 0.25$	0.107(0.943)	0.072(-)	0.109(0.890)	0.139(0.493)	0.156(0.299)
$\phi = 0.5$	0.143(0.646)	0.103(0.920)	0.073(-)	0.105(0.819)	0.143(0.455)
$\phi = 0.75$	0.216(0.417)	0.144(0.638)	0.101(0.935)	0.068(-)	0.101(0.751)
$\phi = 1$	0.221(0.334)	0.183(0.458)	0.146(0.580)	0.099(0.872)	0.061(-)
			$n = 200$		
Fitted Model	$\phi = 0$	$\phi = 0.25$	$\phi = 0.5$	$\phi = 0.75$	$\phi = 1$
$\phi = 0$	0.121(-)	0.144(0.631)	0.168(0.374)	0.188(0.244)	0.216(0.147)
$\phi = 0.25$	0.141(0.730)	0.107(-)	0.140(0.659)	0.178(0.365)	0.211(0.241)
$\phi = 0.5$	0.175(0.458)	0.135(0.730)	0.100(-)	0.131(0.624)	0.177(0.291)
$\phi = 0.75$	0.240(0.266)	0.172(0.417)	0.126(0.706)	0.095(-)	0.127(0.587)
$\phi = 1$	0.262(0.230)	0.219(0.262)	0.174(0.400)	0.123(0.702)	0.074(-)

2.3.5 Comparison with piecewise exponential approximation

Assuming the lifetimes of patients to be non-homogeneous, and based on a proportional hazards model, we model $F(\cdot)$ in eqn.(2.9) as

$$F(t|x) = 1 - S_0(t)^{\exp(\gamma x)}, \tag{2.20}$$

where $S_0(t)$ is the baseline survival function, x is a binary covariate and γ is the corresponding regression parameter. Note that we only considered one covariate (binary) for the sake of simplicity. The approach can be easily extended to multiple covariates. As done in Yin and Ibrahim [45], we approximate $S_0(t)$ using a piecewise exponential (PE) function. For this purpose, we divide the entire time axis into J partitions, i.e., we have $0 < s_1 < \cdots < s_J$, where s_J is greater than the maximum of the observed lifetimes. We denote the constant hazard corresponding to the interval $(s_{j-1}, s_j]$ by λ_j, $j = 1, 2, \cdots, J$. Note that $J = 1$ reduces to a parametric exponential assumption for $S_0(t)$. In Tables 2.10 and 2.11, we compare the GGBCT fit (our proposed model) with the PE fit (approach pro-

Table 2.9: TRB and TRE (in parenthesis) in the estimation of cure rates when fitting different lifetime distributions for a given true lifetime distribution and assuming a given Box-Cox model $(n = 400)$

		True Lifetime Distribution		
		Poisson ($\phi = 0$)		
Fitted Distribution	Lognormal	Generalized Gamma ($q = 0.5$)	Gamma	Weibull
Lognormal	0.022(-)	0.125(0.596)	0.113(0.613)	0.142(0.199)
Generalized Gamma ($q = 0.5$)	0.108(0.610)	0.101(-)	0.095(0.864)	0.092(0.398)
Gamma	0.089(0.678)	0.108(0.735)	0.082(-)	0.096(0.383)
Weibull	0.113(0.533)	0.125(0.593)	0.108(0.648)	0.048(-)
		Bernoulli ($\phi = 1$)		
Lognormal	0.043(-)	0.131(0.565)	0.123(0.757)	0.183(0.295)
Generalized Gamma ($q = 0.5$)	0.110(0.629)	0.123(-)	0.115(0.893)	0.102(0.738)
Gamma	0.045(0.888)	0.128(0.657)	0.078(-)	0.158(0.423)
Weibull	0.116(0.557)	0.136(0.479)	0.128(0.635)	0.063(-)
		Box-Cox ($\phi = 0.75$)		
Lognormal	0.048(-)	0.124(0.626)	0.122(0.806)	0.228(0.125)
Generalized Gamma ($q = 0.5$)	0.060(0.853)	0.102(-)	0.103(0.836)	0.107(0.573)
Gamma	0.083(0.723)	0.117(0.705)	0.074(-)	0.134(0.368)
Weibull	0.089(0.668)	0.127(0.539)	0.127(0.726)	0.080(-)

posed by Yin and Ibrahim [45]) when data are generated from a Box-Cox transformation cure rate model with lifetime distribution as exponential and generalized gamma, respectively. Based on the AIC, AICc and BIC values, it is clear that in both cases the proposed GGBCT model provides a better fit.

Table 2.10: Comparison of GGBCT fit with the PE fit when data is generated from a Box-Cox ($\phi = 0.5$) model with exponential lifetime distribution having rate parameter λ=4.632. Note that $(n_1, n_2) = (230, 170)$ and $(p_{0t}, p_{0c}) = (0.65, 0.35)$

Model	l	AIC	AICc	BIC	Parameters	MLE	SE
PE $(J = 1)$	-131.643	273.286	273.438	275.269	β_0	0.265	0.474
					β_1	-1.129	0.351
					γ	-0.216	0.210
					ϕ	0.514	0.808
					λ_1	5.197	1.267
PE $(J = 2)$	-122.863	257.725	257.939	257.708	β_0	0.351	0.471
					β_1	-1.085	0.336
					γ	-0.171	0.210
					ϕ	0.527	0.591
					λ_1	5.001	1.281
					λ_2	7.382	2.344
PE $(J = 3)$	-141.64	297.28	297.566	295.263	β_0	0.329	0.410
					β_1	-1.056	0.548
					γ	-0.222	0.507
					ϕ	0.495	0.860
					λ_1	1.861	0.377
					λ_2	0.743	0.096
					λ_3	0.058	0.091
GGBCT	-121.982	255.964	256.178	255.947	β_0	0.338	0.533
					β_1	-1.083	0.378
					ϕ	0.499	0.882
					q	1.018	0.232
					λ	4.679	1.342
					σ	0.982	0.120

2.4 Illustration with breast cancer data

In this section, we apply the GGBCT model to the data on breast cancer survival patients introduced in section 1.3.1.

In Figure 2.3, we present the non-parametric Kaplan-Meier survival curves stratified by the prognostic group status variable. As can be seen from the plot, subjects in the "poor" group status has the lowest survival probability. Furthermore, the survival probability increases with an improvement in the prognostic group status. Moreover, from the leveling off tendency of each survival curve, we can say that the crude estimates of the cure rates of subjects belonging to the "poor", "medium" and "good" groups are 0.07, 0.32 and 0.66, respectively. In the real data analysis, we first fit the general GGBCT model. The MLEs of the parameters of the full model, i.e., the GGBCT model, together with their standard errors are presented in Table 2.12. Since the estimate of β_1 is negative, it implies that with an improvement in the prognostic group status the cure rate also increases. In particular, the estimates of the cure rates turn out to be $\hat{p_0}(\text{poor}) = 0.019$, $\hat{p_0}(\text{medium}) = 0.174$ and $\hat{p_0}(\text{good}) = 0.463$.

Now, from Table 2.12, we can see that the estimate of the Box-Cox transformation parameter ϕ is close to zero. This clearly suggests that the promotion time (or Poisson) cure rate model might be appropriate for this dataset. So, we test for the suitability of the promotion time model, i.e., we test for $H_0 : \phi = 0$. Using the likelihood ratio test, the p-value turns out to be 0.611. As such, we fail to reject the promotion time cure rate model. The estimate of the generalized gamma shape parameter q also encourages us to test for the suitability of the lognormal lifetime distribution, i.e., $H_0 : q = 0$. In this case, the p-value corresponding to the likelihood ratio test turns out to be 0.007. Clearly, the assumption of the lognormal lifetime distribution gets rejected at both 1% and 5% nominal levels. We can also test for the joint suitability of the promotion time model with the lognormal lifetime, i.e., $H_0 : \phi = 0, q = 0$. Interestingly, the null hypothesis gets rejected at 5% level of significance, however, we fail to reject the null hypothesis at 1% level of significance. In Table 2.13, we present the likelihood ratio test results for different tests of hypotheses that one may be interested in carrying out in the given context. Note that in Table 2.13, l_0 and l represents the restricted (i.e., under H_0) and the unrestricted maximized log-likelihood function values, respectively. It is clear that at 5% level of significance only the model with $\phi = 0$ doesn't get rejected and, as such, can be considered to be suitable. Note that none of the commonly used special cases of the generalized gamma lifetime distribution turns out to be suitable, which indicates that our working model is the promotion time (or Poisson) cure rate model with generalized gamma lifetime distribution.

In Table 2.14, we present the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the corrected AIC (AICc) values to discriminate between different models. It is once again clear that the model with $\phi = 0$ together with the generalized gamma lifetime provides the best fit to the data since all the three AIC, BIC and AICc turns out to be the lowest among all candidate models. Thus, the conclusion drawn from the likelihood ratio test is in line with the conclusion drawn the information-based criteria. In Table 2.15, we present the MLEs and the standard errors (SE) of the parameters corresponding to the working model, i.e., promotion time (Poisson) cure rate model with generalized gamma lifetime. In this case, the estimates of cure rates turn out to be $\hat{p_0}(\text{poor}) = 0.037$, $\hat{p_0}(\text{medium}) = 0.223$ and $\hat{p_0}(\text{good}) = 0.507$.

In Figure 2.4, for the working model, we present the estimated parametric survival curves which are superimposed on the non-parametric Kaplan-Meier curves. As can be seen, both the parametric and the non-parametric survival curves show close concordance. Finally, we are interested in checking for the goodness-of-fit or model adequacy. For this purpose, we calculate the randomized quantile residuals (Dunn and Smyth, 1996). In Figure 2.5, we present the quantile-quantile (QQ) plot, where each point in the plot corresponds to tPiise he median of five sets of ordered residuals. The plot clearly suggests that the promotion time cure rate model with the generalized gamma lifetime provides a very good fit to the considered breast cancer data. Using the Kolmogorov-Smirnov's test, we also test for the normality of the residuals. The p-value, in this case, turns out to be 0.918, which provides a very strong evidence for the normality of residuals.

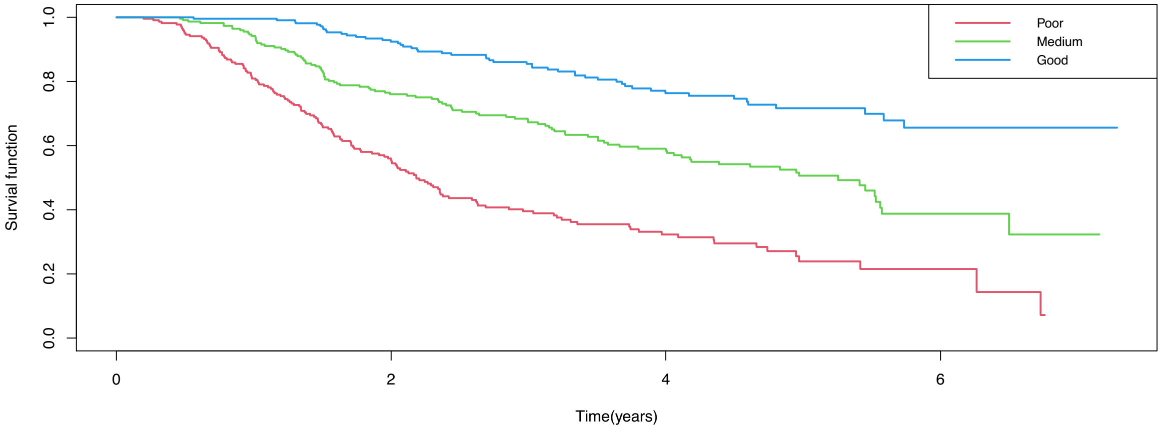


Figure 2.3: Kaplan-Meier survival curves for different prognostic group status

In Table 2.16, we present the results of PE approximation for different values of J. When compared to our proposed parametric approach with generalized gamma lifetime, it is clear that the parametric approach results in better model fit based on the AIC, BIC and AICc values (see Table 2.14 for comparison). Note that Yin and Ibrahim (2005) also concluded that a parametric approach (i.e., $J = 1$, which reduces to an exponential model) results in a better fit. Hence, our findings are in line with the findings of Yin and Ibrahim [22], who employed a Bayesian framework to develop the inference.

Table 2.11: Comparison of GGBCT fit with the PE fit when data is generated from a Box-Cox ($\phi = 0.5$) model with generalized gamma lifetime distribution having parameters $(q, \lambda, \sigma) = (0.5, 4.632, 0.871)$. Note that $(n_1, n_2) = (230, 170)$ and $(p_{0t}, p_{0c}) = (0.65, 0.35)$

Model	l	AIC	AICc	BIC	Parameters	MLE	SE
PE $(J = 1)$	-159.47	328.939	329.091	330.922	β_0	0.450	0.443
					β_1	-1.155	0.328
					γ	-0.258	0.196
					ϕ	0.835	0.730
					λ_1	4.629	1.029
PE $(J = 2)$	-155.151	322.302	322.515	322.285	β_0	0.179	0.406
					β_1	-1.034	0.297
					γ	-0.099	0.219
					ϕ	0.459	0.535
					λ_1	4.045	0.984
					λ_2	9.762	4.344
PE $(J = 3)$	-175.210	364.420	364.706	362.403	β_0	0.335	0.414
					β_1	-1.044	0.550
					γ	-0.220	0.468
					ϕ	0.517	0.764
					λ_1	1.899	0.317
					λ_2	0.742	0.092
					λ_3	0.027	0.083
GGBCT	-141.890	295.780	295.993	295.762	β_0	0.340	0.529
					β_1	-1.094	0.378
					ϕ	0.485	0.868
					q	0.486	0.213
					λ	4.669	1.102
					σ	0.861	0.068

Table 2.12: MLEs and standard errors (SE) of GGBCT model parameters corresponding to the breast cancer data

Parameter	MLE	SE
β_0	1.385	1.078
β_1	-0.822	0.426
ϕ	0.005	0.292
q	0.084	0.407
λ	0.152	0.079
σ	1.049	0.335

Table 2.13: Likelihood ratio test results for different tests of hypotheses

H_0	l_0	l	Λ	p$-$value
$\phi = 0\,, q = 0$	-799.155	-795.416	7.476	0.024
$\phi = 0\,, q = \sigma$	-799.797	-795.416	8.762	0.013
$\phi = 0\,, q = 1$	-803.745	-795.416	16.658	$\approx$ 0.000
$\phi = 1\,, q = 0$	-806.593	-795.416	22.354	$\approx$ 0.000
$\phi = 1\,, q = \sigma$	-811.859	-795.416	32.886	$\approx$ 0.000
$\phi = 1\,, q = 1$	-816.916	-795.416	43.000	$\approx$ 0.000
$q = 0$	-799.050	-795.416	7.268	0.007
$\phi = 1$	-805.441	-795.416	20.050	$\approx$ 0.000
$\phi = 0$	-795.546	-795.416	0.260	**0.611**

Table 2.14: AIC, BIC and AICc for different fitted models

Cure Rate Model	Lifetime	AIC	BIC	AICc
($\phi = 0$) Poisson	Generalized Gamma	**1601.092**	**1623.746**	**1601.180**
	Lognoramal	1606.309	1624.433	1606.368
	Gamma	1607.594	1625.718	1607.653
	Weibull	1615.489	1633.613	1615.548
($\phi = 1$) Bernoulli	Generalized Gamma	1620.882	1643.536	1620.970
	Lognormal	1621.185	1639.309	1621.244
	Gamma	1631.719	1649.842	1631.778
	Weibull	1641.831	1659.955	1641.890
Box-Cox (ϕ)	Generalized Gamma	1602.833	1630.018	1602.957

Table 2.15: MLEs and standard errors (SE) of the working model corresponding to the breast cancer data

Parameter	MLE	SE
β_0	1.196	0.227
β_1	-0.791	0.079
q	0.084	0.279
λ	0.173	0.037
σ	1.002	0.164

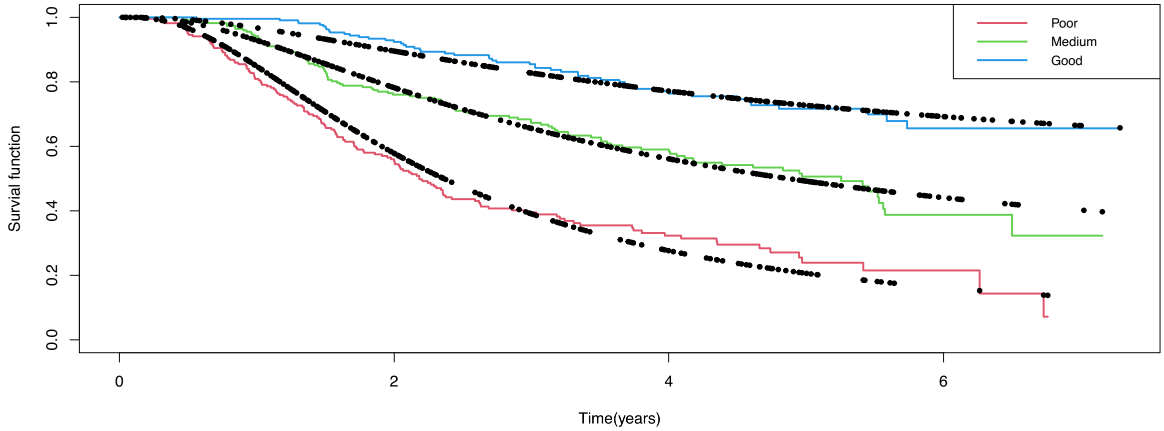


Figure 2.4: Fitted survival curves stratified by prognostic group status (for the working model)

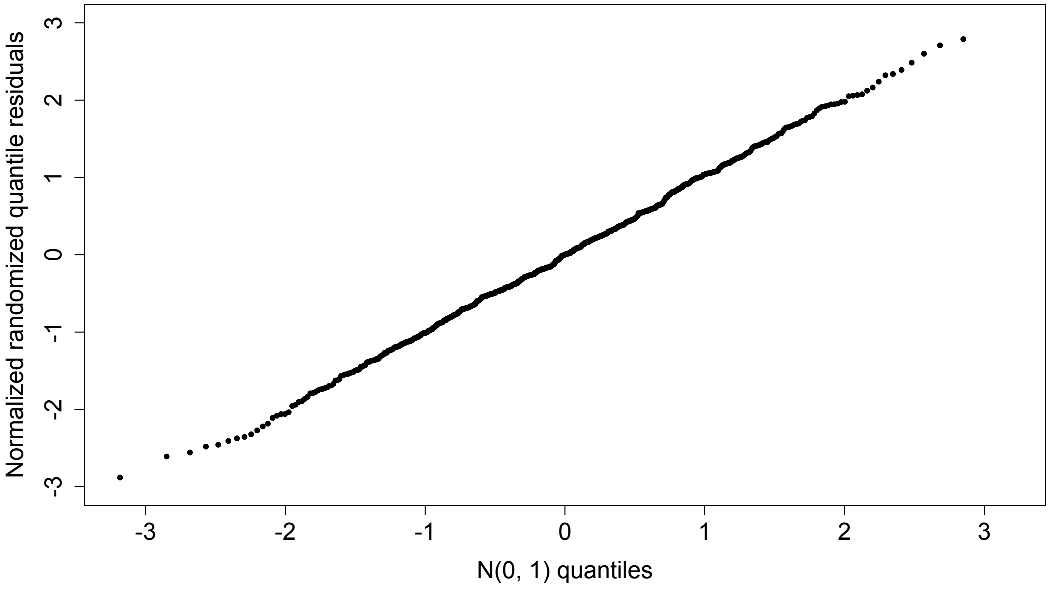


Figure 2.5: QQ-plot of the normalized randomized quantile residuals for the breast cancer survival data

Table 2.16: Piecewise exponential approximation results corresponding to the breast cancer data

J	l	AIC	AICc	BIC	Parameters	MLE	SE
1	-835.876	1681.753	1681.841	1704.407	β_0	1.279	0.311
					β_1	-0.890	0.209
					ϕ	0.204	0.142
					λ_1	0.216	0.173
					γ	-0.157	0.069
2	-830.342	1672.685	1672.808	1699.870	β_0	1.860	0.209
					β_1	-0.332	0.199
					ϕ	0.009	0.115
					λ_1	0.051	0.032
					λ_2	0.068	0.046
					γ	-0.597	0.145
3	-827.713	1669.425	1669.590	1701.141	β_0	1.573	0.527
					β_1	-0.948	0.268
					ϕ	0.014	0.263
					λ_1	0.316	0.188
					λ_2	0.409	0.262
					λ_3	0.411	0.275
					γ	-0.140	0.038
4	-827.650	1671.299	1671.512	1707.546	β_0	1.680	0.920
					β_1	-0.628	0.537
					ϕ	0.596	0.317
					λ_1	0.210	0.426
					λ_2	0.301	0.616
					λ_3	0.322	0.658
					λ_4	0.465	0.627
					γ	-0.612	0.466
5	-818.892	1655.784	1656.050	1696.561	β_0	1.698	0.356
					β_1	-0.726	0.499
					ϕ	0.169	0.274
					λ_1	0.179	0.487
					λ_2	0.390	0.309
					λ_3	0.309	0.736
					λ_4	0.307	0.356
					λ_5	0.481	0.248
					γ	-0.568	0.551
6	-814.801	1649.603	1649.929	1694.912	β_0	1.449	0.366
					β_1	-0.689	0.368
					ϕ	0.065	0.106
					λ_1	0.165	0.140
					λ_2	0.407	0.128
					λ_3	0.321	0.134
					λ_4	0.384	0.123
					λ_5	0.381	0.100
					λ_6	0.503	0.182
					γ	-0.524	0.046
7	-810.672	1643.344	1643.736	1693.184	β_0	1.526	3.708
					β_1	-0.614	1.384
					ϕ	0.498	2.766
					λ_1	0.134	0.345
					λ_2	0.361	0.881
					λ_3	0.301	0.680
					λ_4	0.279	0.588
					λ_5	0.290	0.587
					λ_6	0.291	0.559
					λ_7	0.729	1.408
					γ	-0.599	0.474

CHAPTER 3

Inference for mixture cure model with support vector machine under interval censored data

3.1 Introduction

Introduced by Boag [3] and exclusively studied by Berkson and Gage [12], the mixture cure rate model can also be expressed in the following form. If T^* denotes the lifetime of a susceptible (not cured) subject, then, the actual lifetime T for any subject can be modeled by

$$T = JT^* + (1 - J)\infty, \tag{3.1}$$

where J is a cure indicator denoting if an individual is cured ($J = 0$) or not ($J = 1$). Further, considering $S_p(t) = P(T > t)$ and $S_u(t) = P(T^* > t)$ as the respective survival functions corresponding to T and T^*, we can express

$$S_p(t) = (1 - \pi) + \pi S_u(t), \tag{3.2}$$

where $\pi = P(J = 1)$. The latency part $S_u(t) = S_u(t|\boldsymbol{x})$ and the incidence part $\pi = \pi(\boldsymbol{z})$ are generally modeled to incorporate the effects of covriates $\boldsymbol{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ and $\boldsymbol{z} = (z_1, \ldots, z_q)^{\mathrm{T}}$ for any integers p and q. Note here that $\boldsymbol{x}$ and $\boldsymbol{z}$ may have overlap.

The incidence part $\pi(\boldsymbol{z})$ is traditionally and extensively modeled by sigmoid or logistic function

$$\pi(\boldsymbol{z}) = \frac{\exp(\boldsymbol{z}^{*\mathrm{T}}\boldsymbol{\beta})}{1 + \exp(\boldsymbol{z}^{*\mathrm{T}}\boldsymbol{\beta})}, \tag{3.3}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_q)^{\mathrm{T}}$ and $\boldsymbol{z}^* = (1, \boldsymbol{z}^{\mathrm{T}})^{\mathrm{T}}$ [4, 5, 8]. As observed in the case of logistic regression, the logistic model works well when subjects are linearly separable into the cure or susceptible groups with respect to covariates. However, problem arises when subjects can not be separated using a linear boundary. To this end, support vector machine (SVM) could be a reasonable choice. Motivated by Li et al. [37] we propose to employ the SVM based modeling to study the effects of covariates on the incidence part of the mixture cure rate model by considering the form of the data to be interval-censored.

Unlike right-censored data, interval-censored data occur for a study where subjects are inspected at regular intervals, and not continuously over time. If a subject experiences the event of interest, the exact survival time is not observed. However, it is only known that the event has occurred between two consecutive inspections. Interval-censored data marked by cure prospect are often observed in follow-up clinical studies (cancer biochemical recurrence or AIDS drug resistance) dealing with events having low fatality and patients monitored at regular intervals [46, 47]. As in the case of right-censored data, some subjects may never encounter the event of interest, and are considered as cured. Mixture cure rate models with interval censored data are examined based on several estimation techniques for both semi-parametric and non-parametric set-ups [42, 48–51].

The rest of the chapter is arranged as follows. In Section 3.2, we discuss about the mixture cure rate model framework for interval-censored data and develop an estimation procedure based on the expectation maximization (EM) algorithm that employs the SVM to model the incidence part. In Section 3.3, a detailed simulation study is carried out to demonstrate the performance of our proposed model in terms of flexibility, accuracy and robustness. Comparisons of our model with the existing logistic regression based mixture cure rate models are made in this section. The model performance is further examined and illustrated in Section 3.4 through an interval censored data on smoking cessation.

3.2 SVM based mixture cure rate model with interval censoring

3.2.1 Censoring scheme and modeling lifetimes

The data we observe in situations with interval censoring are of the form $(L_i, R_i, \delta_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ for $i = 1, \ldots, n$, where n denotes the sample size. For a subject, L_i denotes the last inspection time before the event and R_i denotes the first subsequent inspection time just after the event. Note that $L_i < R_i$. The censoring indicator is denoted by $\delta_i = I(R_i < \infty)$, which takes the value 0 if $R_i = \infty$, meaning that the event is not observed for a subject before the last inspection time, and takes the value 1 if $R_i < \infty$, meaning that the event took place but its exact time is not known and is only known to belong to the interval $[L_i, R_i]$. Now, $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are the respective p dimensional and q dimensional covariate vectors affecting the incidence and latency parts, respectively, of the mixture cure rate model. To demonstrate the effect of covariates on the latency part, we consider a proportional hazards structure to model the lifetime distribution of the susceptible or non-cured subjects. That is, for the susceptible subjects, we model the hazard function by

$$h(t_i|\boldsymbol{x}_i) = h_0(t_i) \exp\left\{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\gamma}\right\}, \tag{3.4}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^{\mathrm{T}}$ is the p dimensional regression parameter vector measuring the effects of $\boldsymbol{x}$ and $h_0(\cdot)$ is the unspecified baseline hazard function. To facilitate our discussion, we assume the baseline hazard to follow a parametric form given by $h_0(t_i) = \alpha t_i^{\alpha-1}$, where $\alpha > 0$. One is of course free to use other non-parametric or semi-parametric forms for the baseline hazard. Therefore, we have

$$h(t_i|\boldsymbol{x}_i) = \alpha t_i^{\alpha-1} \exp\left\{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\gamma}\right\}. \tag{3.5}$$

Note that eqn.(3.5) turns out to be the hazard function of a Weibull distribution with shape parameter α and scale parameter $\{e^{\boldsymbol{x}_i^{\mathrm{T}}\gamma}\}^{-1/\alpha}$. Consequently, the susceptible lifetime follows a Weibull distribution with the aforementioned scale and shape parameters. Weibull distribution is a popular and flexible choice for modeling lifetimes or failure times in survival analysis. It is closed under proportional hazards family when the shape parameter remains constant, and it accommodates decreasing ($\alpha < 1$), constant ($\alpha = 1$) and increasing ($\alpha > 1$) failure rates [4, 52, 53]. From eqn.(3.2), the resulting survival function and density function of any subject in the study (irrespective of the cured status) are respectively given by

$$S_p(t_i|\boldsymbol{x}_i, \boldsymbol{z}_i) = 1 - \pi(\boldsymbol{z}_i) + \pi(\boldsymbol{z}_i)\exp\left\{-\left(t_i/m_i\right)^{\alpha}\right\} \tag{3.6}$$

and

$$f_p(t_i|\boldsymbol{x}_i, \boldsymbol{z}_i) = \pi(\boldsymbol{z}_i)\frac{\alpha t_i^{\alpha-1}}{m_i^{\alpha}} \times \exp\left\{-\left(t_i/m_i\right)^{\alpha}\right\}, \tag{3.7}$$

where $m_i = \{e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\gamma}}\}^{-1/\alpha}$.

3.2.2 Form of the likelihood function

As missing observations are inherent to the problem set-up and model framework, we propose to employ the EM algorithm to estimate the unknown parameters [1, 8, 54]. For implementing the EM algorithm, we need the form of the complete data likelihood function. Let us define $\Delta_0 = \{i : \delta_i = 0\}$ and $\Delta_1 = \{i : \delta_i = 1\}$. Missing observations that appear in this context are in terms of the cure indicator variable J, where J is as defined in eqn. (3.1). Note that J_i's are all known to take the value 1 if $i \in \Delta_1$. However, if $i \in \Delta_0$, J_i can either take 0 or 1, and is thus unknown or missing. Using these J_i's as the missing data, we can define the complete data as $(L_i, R_i, \delta_i, J_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$, for $i = 1, \ldots, n$, which contains both observed and missing data. Under the interval censoring mechanism, we can now express the complete data likelihood function and log-likelihood function as:

$$L_c = \prod_{i\in\Delta_1}\left[\pi(\boldsymbol{z}_i)\left\{S_u(L_i|\boldsymbol{x}_i) - S_u(R_i|\boldsymbol{x}_i)\right\}\right]^{J_i} \times \prod_{i\in\Delta_0}(1-\pi(\boldsymbol{z}_i))^{1-J_i}\left\{\pi(\boldsymbol{z}_i)S_u(L_i|\boldsymbol{x}_i)\right\}^{J_i} \tag{3.8}$$

and

$$
\begin{aligned}
l_c = &\sum_{i \in \Delta_1} J_i \left[\log \pi(\boldsymbol{z}_i) + \log \left\{S_u(L_i|\boldsymbol{x}_i) - S_u(R_i|\boldsymbol{x}_i)\right\}\right] \\
&+ \sum_{i \in \Delta_0} (1 - J_i) \log(1 - \pi(\boldsymbol{z}_i)) + J_i \left\{\log \pi(\boldsymbol{z}_i) + \log S_u(L_i|\boldsymbol{x}_i)\right\}, \qquad (3.9)
\end{aligned}
$$

where $S_u(t_i|\boldsymbol{x}_i) = \exp\{-(t_i/m_i)^{\alpha}\}$, J_i is unobserved for $i \in \Delta_0$ and $J_i = 1$ for $i \in \Delta_1$ [55]. It can be further noted that

$$
l_c = l_{c1} + l_{c2}, \qquad (3.10)
$$

where

$$
l_{c1} = \sum_{i=1}^{n} \left[J_i \log \pi(\boldsymbol{z}_i) + (1 - J_i) \log(1 - \pi(\boldsymbol{z}_i))\right] \qquad (3.11)
$$

is a function that depends on the incidence part only and

$$
l_{c2} = \sum_{i=1}^{n} \left[\delta_i \log \left\{S_u(L_i|\boldsymbol{x}_i) - S_u(R_i|\boldsymbol{x}_i)\right\} + (1 - \delta_i) J_i \log S_u(L_i|\boldsymbol{x}_i)\right] \qquad (3.12)
$$

is a function that depends on the latency part only. The steps of the EM algorithm by incorporating the SVM to model the incidence part $\pi(\boldsymbol{z}_i)$ are discussed henceforth.

3.2.3 Modeling the incidence part with support vector machine

Let us assume that J_i for $i \in \Delta_0$ are observed by some mechanism to assist our theory. Support vector machine algorithm maximizes the linear or non-linear margin between the two closest points belonging to the opposite classification groups (cured and susceptible). That is, SVM solves the following optimization problem for $d_i; i = 1, \ldots, n$:

$$
\max_{d_1,\ldots,d_n} \left[-\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} d_i d_j (2J_i - 1)(2J_j - 1) \Phi_k(\boldsymbol{z}_i, \boldsymbol{z}_j) + \sum_{i=1}^{n} d_i\right] \qquad (3.13)
$$

subject to the constraint $\sum_{i=1}^{n} (2J_i - 1) d_i = 0$ and $0 \leq d_i \leq C$, for $i = 1, \ldots, n$, where C is a parameter that trades off between the margin width and misclassification proportion.

Smaller values of C cause optimizer to look for a larger margin width allowing higher misclassification. $\Phi_k(.,.)$ is a symmetric positive semi definite kernel function, which we consider to be the radial basis function (RBF) given by $\Phi_k(\boldsymbol{z}_i, \boldsymbol{z}_j) = \exp\left\{-\frac{(\boldsymbol{z}_i - \boldsymbol{z}_j)^{\mathrm{T}}(\boldsymbol{z}_i - \boldsymbol{z}_j)}{\sigma^2}\right\}$. RBF is a popular choice of the kernel function owing to its robustness in transforming observations to higher dimension. The parameter σ^2 determines the kernel-width. Both hyper-parameters C and σ^2 are to be tuned to obtain the highest classification accuracy using cross-validation methods [56]. Grid search can be implemented to determine C and σ^2. Low values of σ^2 result in overfitting and jagged separator, while high values of σ^2 result in more linear and smoother decision boundaries. Also, it is recommended to standardize the covariate vector $\boldsymbol{z}$.

The mapping J_i to $2J_i - 1$ converts the respective 0 and 1s to -1 and +1s, which aids in formulation of the optimization problem under the SVM framework. Once d_i's are obtained, we can derive a threshold b as $b = \sum_{i=1}^{n}(2J_i - 1)d_i\Phi_k(\boldsymbol{z}_i, \boldsymbol{z}_j) - (2J_j - 1)$, for some $d_j > 0$. For any new covariate vector $\boldsymbol{z}_{new}$, the optimal decision or classification rule is given by

$$\psi(\boldsymbol{z}_{new}) = \sum_{i=1}^{n} d_i(2J_i - 1)\Phi_k(\boldsymbol{z}_i, \boldsymbol{z}_{new}) - b. \tag{3.14}$$

As suggested by Li et al. [37], the sequential minimal optimization method (SMO), introduced by Platt [57], can be applied to solve eqn.(3.13). As opposed to solving large quadratic optimization problems to train a SVM model, SMO solves a series of smallest possible quadratic problems. Thus, SMO is relatively time inexpensive algorithm. Any subject with covariate $\boldsymbol{z}_{new}$ is assigned to the susceptible group if $\psi(\boldsymbol{z}_{new}) > 0$ and to the cured group if $\psi(\boldsymbol{z}_{new}) < 0$.

In the given context, note that it is not enough to just classify subjects as being cured or susceptible. It is also of our interest to obtain the estimates of uncured probabilities $\pi(\boldsymbol{z}_i)$ or equivalently the cured probabilities $1 - \pi(\boldsymbol{z}_i)$. For this purpose, we use the Platt scaling method to obtain an estimate of $\pi(\boldsymbol{z}_i)$ from the classification rule $\psi(.)$ [58]. The estimate of $\pi(\boldsymbol{z}_i)$ by Platt scaling method is given by

$$\hat{\pi}(\boldsymbol{z}_i) = \frac{1}{1 + \exp\{A\psi(\boldsymbol{z}_i) + B\}}, \tag{3.15}$$

where A and B are obtained by maximizing the following function:

$$\sum_{i=1}^{n} (1 - \zeta_i)[A\psi(\boldsymbol{z}_i) + B] - \log[1 + \exp\{A\psi(\boldsymbol{z}_i) + B\}]. \tag{3.16}$$

Here,

$$\zeta_i = \begin{cases} \frac{n^{(1)}+1}{n^{(1)}+2}, & \text{if } J_i = 1 \\ \\ \frac{1}{n^{(0)}+2}, & \text{if } J_i = 0, \end{cases} \tag{3.17}$$

and $n^{(1)}$ and $n^{(0)}$ represents the number of subjects in the susceptible and cured groups, respectively.

Now, we started our discussion on the SVM based modeling of the incidence part with the assumption that J_is are observed and available for training purpose. However, in practice, the cure status J_i is not known for $i \in \Delta_0$. Multiple imputation based approach can be applied here to obtain $\hat{\pi}(\boldsymbol{z}_i)$ with imputed values of J_i for $i = 1, \ldots, n$. The steps are as follows:

1. For a pre-defined integer N^* and $n^* = 1, 2, \ldots, N^*$, generate $\{J_i^{(n^*)} : i = 1, \ldots, n\}$, where $J_i^{(n^*)}$ is a Bernoulli random variable with success probability $p_i^{(n^*)}$. The discussion on deriving $p_i^{(n^*)}$ is provided in Section 3.2.4.
2. For the imputed data $\{J_i^{(n^*)} : i = 1, \ldots, n\}$, obtain $\hat{\pi}^{(n^*)}(\boldsymbol{z}_i)$ as the estimate of $\pi(\boldsymbol{z}_i)$ by the Platt scaling method given in eqn.(3.15) for $n^* = 1, 2, \ldots, N^*$.
3. $\hat{\pi}(\boldsymbol{z}_i) = (1/N^*) \sum_{n^*=1}^{N^*} \hat{\pi}^{(n^*)}(\boldsymbol{z}_i)$ is the final estimate of $\pi(\boldsymbol{z}_i)$.

3.2.4 Development of the EM algorithm

Since the EM algorithm involves finding the conditional expectation of the complete data log-likelihood function given the current estimates (say, at the $(r+1)$-th iteration step) and the observed data, we begin our discussion by deriving the conditional expectation of J_i given the observed data, $\pi(\boldsymbol{z}_i)$ and $(\alpha, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$, as

$$w_i^{(r+1)} = \delta_i + (1-\delta_i)\frac{\pi^{(r)}(\boldsymbol{z}_i)S_u^{(r)}(L_i|\boldsymbol{x}_i)}{1-\pi^{(r)}(\boldsymbol{z}_i)+\pi^{(r)}(\boldsymbol{z}_i)S_u^{(r)}(L_i|\boldsymbol{x}_i)}, \quad i = 1, \ldots, n, \quad (3.18)$$

where $S_u^{(r)}(L_i|\boldsymbol{x}_i) = \exp\left\{-\left(L_i/m_i^{(r)}\right)^{\alpha^{(r)}}\right\}$ with $m_i^{(r)} = \{e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\gamma}^{(r)}}\}^{-1/\alpha^{(r)}}$. Note that eqn. (3.18) implies that $w_i^{(r+1)} = 1$ for all $i \in \Delta_1$. Hence, we obtain the conditional expectation of l_c by simply replacing J_i's with $w_i^{(r+1)}$ in eqn.(3.9). We denote the aforementioned conditional expectation by

$$Q_c = Q_{c1} + Q_{c2}, \quad (3.19)$$

where

$$Q_{c1} = \sum_{i=1}^{n}\left[w_i^{(r+1)}\log\pi(\boldsymbol{z}_i) + (1-w_i^{(r+1)})\log(1-\pi(\boldsymbol{z}_i))\right] \quad (3.20)$$

and

$$Q_{c2} = \sum_{i=1}^{n}\left[\delta_i\log\{S_u(L_i|\boldsymbol{x}_i) - S_u(R_i|\boldsymbol{x}_i)\} + (1-\delta_i)w_i^{(r+1)}\log S_u(L_i|\boldsymbol{x}_i)\right]. \quad (3.21)$$

For $r = 0, 1, \ldots$, the procedure below is given for the $(r+1)$-th iteration step of the EM algorithm.

1. Carry out the multiple imputation technique, as described in Section 3.2.3, by considering $p_i^{(n^*)} = w_i^{(r+1)}$, for $n^* = 1, \ldots, N^*$ and $i = 1, \ldots, n$. Obtain $\hat{\pi}^{(r+1)}(\boldsymbol{z}_i) = (1/N^*)\sum_{n^*=1}^{N^*}\hat{\pi}^{(n^*)}(\boldsymbol{z}_i)$ by applying the Platt scaling method with the classification rule $\psi(\cdot)$ defined in eqn. (3.14). Recall that the classification rule is built based on the imputed data $\{J_i^{(n^*)} : i = 1, \ldots, n\}$, where $J_i^{(n^*)}$ is a Bernoulli random variable with success probability $p_i^{(n^*)}$.
2. Obtain $(\alpha^{(r+1)}, \boldsymbol{\gamma}^{(r+1)\mathrm{T}})$ by maximizing the function Q_{c2}, as defined in eqn. (3.21), with respect to α and $\boldsymbol{\gamma}$. That is, find

$$(\alpha^{(r+1)}, \boldsymbol{\gamma}^{(r+1)\mathrm{T}})^{\mathrm{T}} = \arg\max_{\alpha, \boldsymbol{\gamma}} Q_{c2}. \tag{3.22}$$

3. Check for the convergence as follows:

$$||\boldsymbol{\theta}^{(r+1)} - \boldsymbol{\theta}^{(r)}||_2^2 < \epsilon,$$

where $\boldsymbol{\theta}^{(k)} = (\overline{\pi^{(k)}}(\boldsymbol{z}), \alpha^{(k)}, \boldsymbol{\gamma}^{(k)\mathrm{T}})^{\mathrm{T}}$, with $\overline{\pi^{(k)}}(\boldsymbol{z}) = \frac{1}{n}\sum_{i=1}^{n} \pi^{(k)}(\boldsymbol{z}_i)$, $\epsilon > 0$ is some pre-determined and sufficiently small tolerance and $|| \cdot ||_2$ is the L_2-norm. If the above criterion is satisfied, then, stop the algorithm. In this case, $\hat{\pi}^{(r+1)}(\boldsymbol{z}_i)$, for $i = 1, \ldots, n$, and $(\alpha^{(r+1)}, \boldsymbol{\gamma}^{(r+1)\mathrm{T}})^{\mathrm{T}}$ are the final pointwise estimates. On the other hand, if the above criterion is not met, continue to Step 4.

4. Update $w_i^{(r+1)}$ in eqn. (3.18) to

$$w_i^{(r+2)} = \delta_i + (1 - \delta_i)\frac{\hat{\pi}^{(r+1)}(\boldsymbol{z}_i) S_u^{(r+1)}(L_i|\boldsymbol{x}_i)}{1 - \hat{\pi}^{(r+1)}(\boldsymbol{z}_i) + \hat{\pi}^{(r+1)}(\boldsymbol{z}_i) S_u^{(r+1)}(L_i|\boldsymbol{x}_i)}, \tag{3.23}$$

where $S_u^{(r+1)}(t_i|\boldsymbol{x}_i) = \exp\left\{-\left(t_i/m_i^{(r+1)}\right)^{\alpha^{(r+1)}}\right\}$ and $m_i^{(r+1)} = \{e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\gamma}^{(r+1)}}\}^{-1/\alpha^{(r+1)}}$.

5. Repeat steps 1-4 until convergence is achieved.

3.2.5 Calculating the standard errors

The standard errors are estimated by non-parametric bootstrapping. For $b' = 1, \ldots, B$, b'-th bootstrapped data set is obtained by resampling with replacement from the original data. The sample size of the b'-th bootstrapped data is the same as the original data. Then, we carry out steps 1-5 of the EM algorithm as detailed in Section 3.2.4 to obtain the estimates of model parameters for each bootstrapped data. This gives us B estimates for each model parameter. For each parameter, the standard deviation of these B estimates provide an estimate of the standard error of the parameter.

3.2.6 Finding the initial values

To start the EM algorithm, we need to provide initial values of $\pi(\boldsymbol{z}_i)$, for $i = 1, \ldots, n$, along with α and $\boldsymbol{\gamma}$. To come up with an initial guess of $\pi(\boldsymbol{z}_i)$, first, we can consider the censoring indicator $\delta_i, i = 1, \ldots, n$, as the cure indicator (i.e., $\delta_i = 0$ would imply $J_i = 0$ and $\delta_i = 1$ would imply $J_i = 1$). Then, we can apply the SVM to come up with the classification rule, as given in eqn.(3.14), and, finally, we apply the Platt scaling method, as given in eqn.(3.15), to obtain $\pi(\boldsymbol{z}_i)$. Now, to obtain an initial guess of the latency parameters α and $\boldsymbol{\gamma}$, we make use of the form of the survival function of the susceptible subjects, i.e., $S(t_i) = \exp\left\{-\left(t_i/m_i\right)^{\alpha}\right\}$, where $m_i = \{e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\gamma}}\}^{-1/\alpha}$. Note that this form implies that

$$\log\{-\log S(t_i)\} = \alpha \log t_i + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\gamma}, \ \ i = 1, \ldots, n.$$

Hence, we can fit a linear regression model using $\log\{-\log S(t_i)\}$ as the response to obtain estimates of α and $\boldsymbol{\gamma}$, which can be used as the initial guesses. Here, $S(t_i)$ can be the estimated using the non-parametric Kaplan-Meier estimates. Since the form of the data is interval censored, we can take $t_i = \frac{L_i+R_i}{2}$, if $R_i < \infty$, and take $t_i = L_i$, if $R_i = \infty$, for all $i = 1, \ldots, n$. Note that this procedure may result in negative estimates of α. As such, we can take the initial guess of α as 0.05 or 0.1 if the estimate of α turns out to be negative.

3.3 Simulation study

In this section, we assess the performance of the proposed SVM based EM algorithm to estimate the model parameters of the mixture cure rate model for interval censored data. We generate two random values x_1 and x_2 independently from standard normal distribution and assume $\boldsymbol{x} = \boldsymbol{z}$ with $\boldsymbol{x} = (x_1, x_2)^{\mathrm{T}}$. We consider two different sample sizes: $n = 300$

and $n = 400$ and use the following links to generate uncured probabilities $\pi(\boldsymbol{z})$:

$$\begin{aligned}\text{Scenario 1: } \pi(\boldsymbol{z}) &= \frac{e^{0.3-5z_1-3z_2}}{1+e^{0.3-5z_1-3z_2}};\\ \text{Scenario 2: } \pi(\boldsymbol{z}) &= \frac{e^{0.3+10z_1^2-5z_2^2}}{1+e^{0.3+10z_1^2-5z_2^2}};\\ \text{Scenario 3: } \pi(\boldsymbol{z}) &= \exp\{-\exp(0.3-4\cos z_1-5\sin z_2)\}.\end{aligned}$$

Note that Scenario 1 represents the standard logistic regression model which captures a linear classification boundary. On the other hand, Scenarios 2 and 3 capture non-linear or more complex classification boundaries; see Figure 3.1. In Figure 3.2, we present the plots of simulated uncured probabilities and how they vary with respect to the covariates z_1 and z_2.

We assume lifetimes of the susceptible subjects follow proportional hazards structure with the hazard function

$$h(t) = h_0(t)\exp(\gamma_1 x_1 + \gamma_2 x_2)$$

where $h_0(t) = \alpha t^{\alpha-1}$. As discussed before, the above hazard function implies that the susceptible lifetime follows a Weibull distribution with shape parameter α and scale parameter $\{\exp(\gamma_1 x_1 + \gamma_2 x_2)\}^{-\frac{1}{\alpha}}$. We consider the true values of $(\alpha, \gamma_1, \gamma_2)$ as $(0.5, 1, 0.5)$. The censoring time is generated from a Uniform $(0, 20)$ distribution. Using these, the cure probabilities range from $50\% - 65\%$, whereas the overall censoring proportions range from $60\% - 75\%$. To generate interval censored lifetime data $(L_i, R_i, \delta_i), i = 1, 2, \cdots, n$, we carry out the following steps:

Step 1: Generate a Uniform (0,1) random variable U_i and a censoring time C_i;

Step 2: If $U_i \leq 1 - \pi(\boldsymbol{z}_i)$, set $L_i = C_i$, $R_i = \infty$, and $\delta_i = 0$;

Step 3: If $U_i > 1 - \pi(\boldsymbol{z}_i)$, generate T_i from a Weibull distribution with shape parameter α and scale parameter $\{\exp(\gamma_1 x_{1i} + \gamma_2 x_{2i})\}^{-\frac{1}{\alpha}}$;

Step 4:

a. If $\min\{T_i, C_i\} = C_i$, set $L_i = C_i$, $R_i = \infty$, and $\delta_i = 0$;

b. If $\min\{T_i, C_i\} = T_i$, set $\delta_i = 1$, and generate L_{1i} from Uniform $(0.2, 0.7)$ distribution and L_{2i} from Uniform $(0, 1)$ distribution. Next, create intervals $(0, L_{2i}], (L_{2i}, L_{2i} + L_{1i}], \cdots, (L_{2i} + k \times L_{1i}, \infty], k = 1, 2, \cdots,$ and select (L_i, R_i) that satisfies $L_i < T_i \leq R_i$.

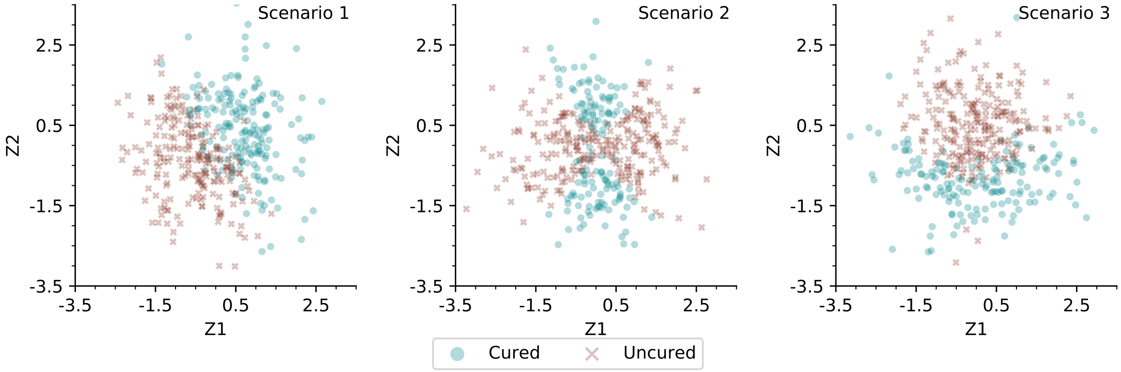


Figure 3.1: Simulated cured and uncured observations for the three considered scenarios

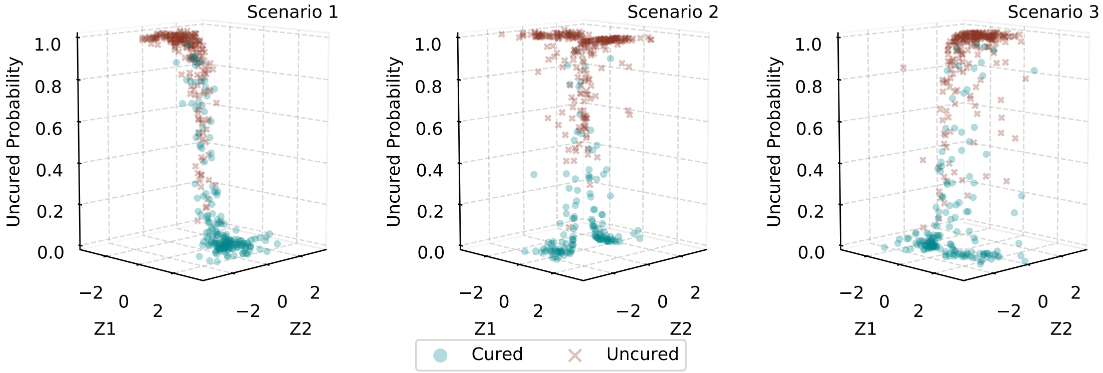


Figure 3.2: Simulated uncured probabilities and their behavior with respect to the covariates for the three considered scenarios

All simulations are done using the R statistical software (version 4.0.4) and all results are based on $M = 500$ Monte Carlo runs. To employ our proposed methodology, we consider number of imputations in the multiple imputation technique to be 5, which is in line with [37]; see also [59]. In Table 3.1, we report the bias and mean squared error (MSE) of the estimated uncured probability $\hat{\pi}(\boldsymbol{z})$ and the overall survival probability $\hat{S}_p = \hat{S}_p(.,.;\boldsymbol{x},\boldsymbol{z})$. These are calculated as:

$$\text{Bias}(\hat{\pi}(\boldsymbol{z})) = \frac{1}{M}\sum_{k=1}^{M}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\widehat{\pi^{(k)}}(\boldsymbol{z}_i) - \pi^{(k)}(\boldsymbol{z}_i)\right\}\right];$$

$$\text{Bias}(\hat{S}p) = \frac{1}{M}\sum_{k=1}^{M}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\widehat{S_p^{(k)}}(L_i, R_i; \boldsymbol{x}_i, \boldsymbol{z}_i) - S_p^{(k)}(L_i, R_i; \boldsymbol{x}_i, \boldsymbol{z}_i)\right\}\right];$$

$$\text{MSE}(\hat{\pi}(\boldsymbol{z})) = \frac{1}{M}\sum_{k=1}^{M}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\widehat{\pi^{(k)}}(\boldsymbol{z}_i) - \pi^{(k)}(\boldsymbol{z}_i)\right\}^2\right];$$

$$\text{MSE}(\hat{S}_p) = \frac{1}{M}\sum_{k=1}^{M}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\widehat{S_p^{(k)}}(L_i, R_i; \boldsymbol{x}_i, \boldsymbol{z}_i) - S_p^{(k)}(L_i, R_i; \boldsymbol{x}_i, \boldsymbol{z}_i)\right\}^2\right],$$

where $\pi^{(k)}(\boldsymbol{z}_i)$ and $S_p^{(k)}(L_i, R_i; \boldsymbol{x}_i, \boldsymbol{z}_i)$ are the true uncured probability and population survival probability, respectively, corresponding to the i-th subject and the k-th Monte Carlo run. Similarly, $\widehat{\pi^{(k)}}(\boldsymbol{z}_i)$ and $\widehat{S_p^{(k)}}(L_i, R_i; \boldsymbol{x}_i, \boldsymbol{z}_i)$ are the estimated uncured probability and population survival probability, respectively, corresponding to the i-th subject and the k-th Monte Carlo run.

Table 3.1: Comparison of Bias and MSE of the uncured probability and overall survival probability

		Uncured Probability				Overall Survival Probability			
n	Scenario	Bias		MSE		Bias		MSE	
		SVM	LOGISTIC	SVM	LOGISTIC	SVM	LOGISTIC	SVM	LOGISTIC
400	1	-0.126	-0.002	0.083	0.002	-0.000	0.002	0.015	0.002
	2	-0.063	0.132	0.042	0.209	0.028	-0.050	0.014	0.063
	3	-0.020	0.089	0.019	0.080	0.010	-0.029	0.008	0.013
300	1	-0.126	-0.001	0.088	0.002	-0.001	0.001	0.018	0.002
	2	-0.063	0.130	0.046	0.210	0.028	-0.049	0.016	0.063
	3	-0.023	0.087	0.022	0.080	0.013	-0.027	0.010	0.014

From Table 3.1 and looking at the results related to the uncured probability, it is clear that the bias and MSE of the logistic based EM algorithm is smaller than the proposed SVM based EM algorithm when logistic regression is the correct model (Scenario 1). However, when the true model for the uncured probability is not the logistic regression, i.e., when the models considered in Scenarios 2 and 3 are the true models, the proposed SVM based EM algorithm produces smaller bias and MSE. In Figure 3.3, we present the bias of the estimates of the uncured probabilities when plotted against each covariate. On the other hand, in Figure 3.4, we present the bias of the estimates of the uncured probabilities when plotted against both covariates.

Looking at the results corresponding to the overall survival probability, it turns out that when the logistic regression model (Scenario 1) is the true model for the uncured probability, both SVM based EM algorithm and logistic based EM algorithm produce similar biases, but the logistic based EM algorithm produces smaller MSE. On the other hand, when the true model for the uncured probability is non-logistic (Scenarios 2 and 3), the SVM based EM algorithm results in smaller bias and MSE of the overall survival probability when compared to the logistic based EM algorithm. These findings clearly indicate that the SVM based EM algorithm is able to capture more complex and non-linear classification boundaries, where the standard logistic based EM algorithm produces relatively larger bias and MSE. In Figure 3.5, we present the bias of the estimates of the overall survival probabilities when plotted against each covariate, whereas, in Figure 3.6, we present the bias of the estimates of the survival probabilities when plotted against both covariates.

In Table 3.2, we present the estimation results corresponding to the latency parameters. In particular, we compare bias, standard error (SE) and MSE of the estimates of the latency parameters corresponding to the proposed SVM based mixture cure rate model and the traditional logistic regression based mixture cure rate model. From Table 3.2, we can see that the bias, SE and MSE corresponding to the logistic regression based EM algo-

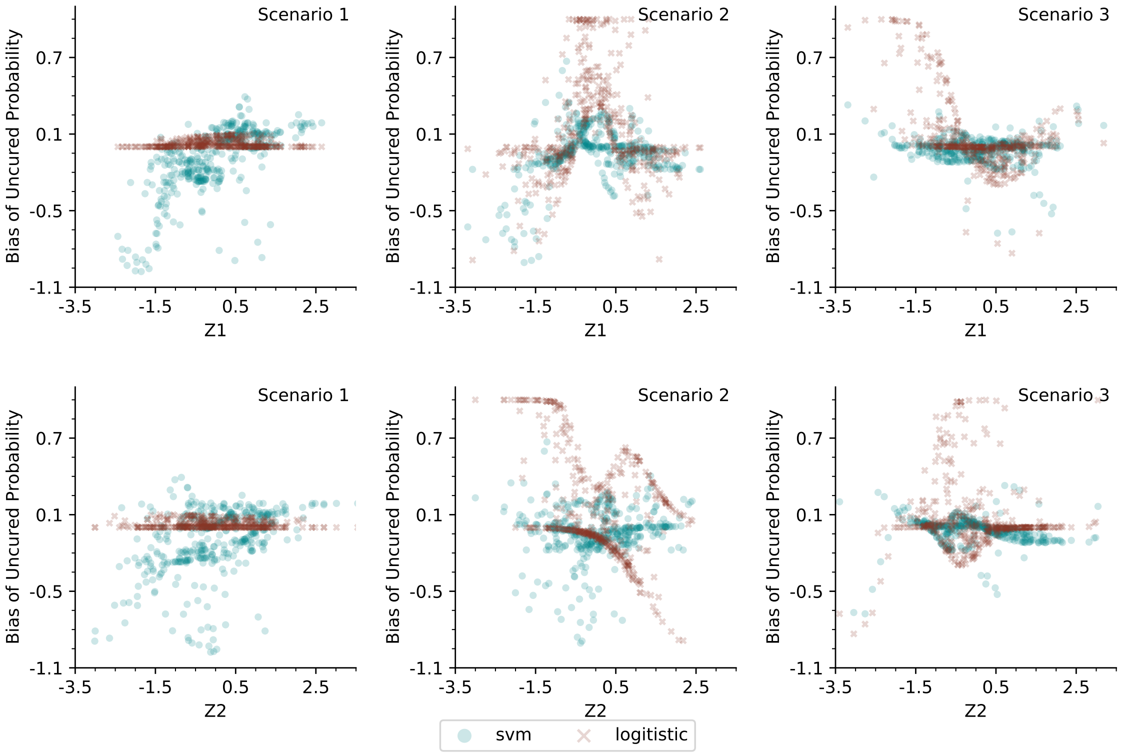


Figure 3.3: Bias of the uncured probabilities with respect to each covariate for the three considered scenarios

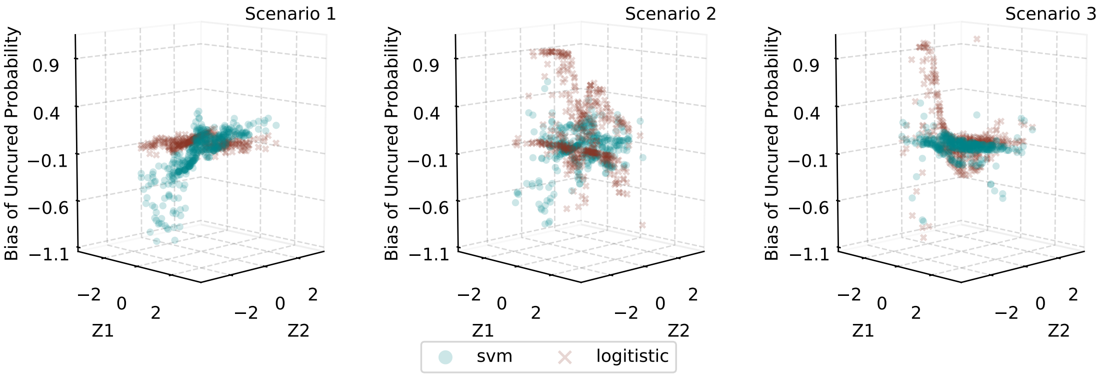


Figure 3.4: Bias of the uncured probabilities with respect to both covariates for the three considered scenarios

rithm is smaller when the logistic regression is the true model for the uncured probabilities (i.e., Scenario 1 is true). In this case, bias corresponding to the SVM based EM algorithm is relatively higher. However, when the true model for the uncured probabilities is non-

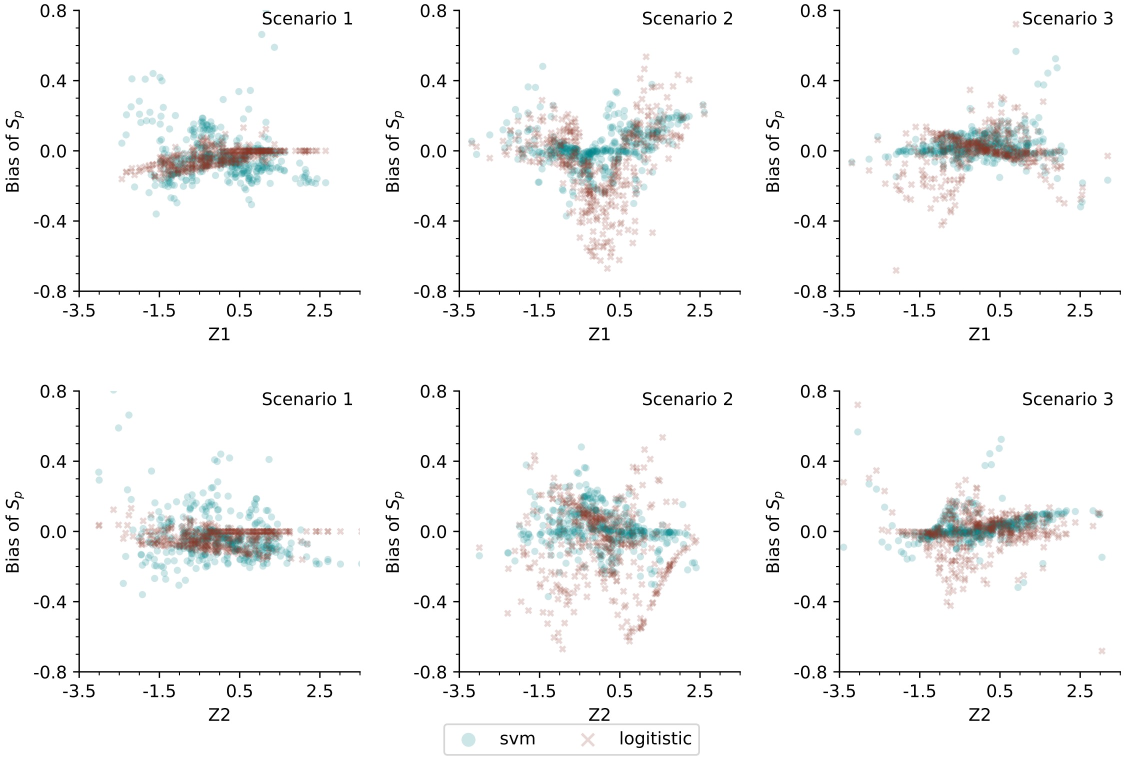


Figure 3.5: Bias of the overall survival probabilities with respect to each covariate for the three considered scenarios

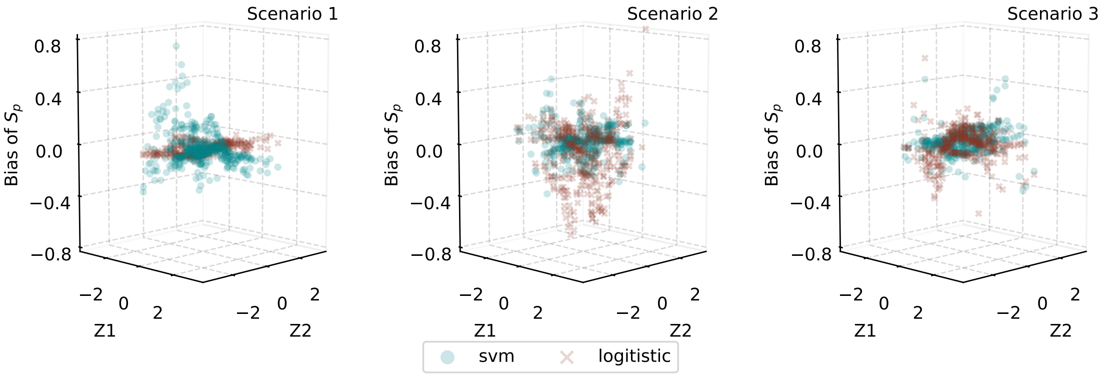


Figure 3.6: Bias of the overall survival probabilities with respect to both covariates for the three considered scenarios

logistic (i.e., Scenarios 2 and 3 are the true models), the SVM based EM algorithm results in smaller bias, SE and MSE. With an increase in the sample size, the bias, SE and MSE tend to decrease further, which is what we would expect.

Summarizing the findings from both Table 3.1 and Table 3.2, we can conclude that the proposed SVM based EM algorithm performs better than the standard logistic regression based EM algorithm, both in terms of the incidence part and the latency part of the mixture cure rate model, when the true classification boundary is non-liner and complex. This clearly demonstrates the ability of the proposed SVM based model to handle complex non-linear classification boundaries.

Although, in practice, the cured status is unobserved for a real data, we do know which observations can be considered as cured when we simulate data. Using such information on the cured status for a simulated data, we can easily compare the proposed SVM based mixture model with the logistic regression based mixture model using the receiver operating characteristic (ROC) curves and the area under the curves (AUCs) for different scenarios we have considered. In Figure 3.7, we present the ROC curves under different scenarios for a particular simulated data of size 400. The corresponding AUC values are presented in Table 3.3. It is once again clear that under Scenarios 2 and 3 (i.e., when the classification boundaries are non-linear), the performance (or the accuracy) of the SVM based model is better than the logistic regression based model. Note, in particular, that the performance of the SVM based model is significantly better under Scenario 2. However, under scenario 1 (i.e., when the classification boundary is linear), the logistic regression based model performs slightly better than the SVM based model.

3.4 Illustrative example: smoking cessation data analysis

We further demonstrate our proposed methodology using a dataset on smoking cessation study [39, 40]. Out of those who relapsed, most did so in the first year of their smoking cessation trial (see Figure 3.8). Figure 3.9 presents the Kaplan-Meier curve. Clearly, we

Table 3.2: Estimation results corresponding to the latency parameters

n	Scenario	Latency Parameter	Bias		SE		MSE	
			SVM	LOGISTIC	SVM	LOGISTIC	SVM	LOGISTIC
400	1	$\alpha = 0.5$	0.103	0.008	0.052	0.050	0.014	0.003
		$\gamma_1 = 1.0$	-0.498	0.010	0.139	0.123	0.270	0.018
		$\gamma_2 = 0.5$	-0.269	0.004	0.109	0.105	0.086	0.012
	2	$\alpha = 0.5$	0.074	-0.117	0.056	0.038	0.008	0.016
		$\gamma_1 = 1.0$	-0.099	-0.111	0.102	0.109	0.018	0.026
		$\gamma_2 = 0.5$	-0.012	0.740	0.167	0.132	0.022	0.574
	3	$\alpha = 0.5$	0.047	-0.010	0.049	0.045	0.005	0.002
		$\gamma_1 = 1.0$	-0.037	0.257	0.141	0.120	0.018	0.082
		$\gamma_2 = 0.5$	0.085	0.079	0.121	0.106	0.017	0.018
300	1	$\alpha = 0.5$	0.107	0.007	0.062	0.060	0.015	0.004
		$\gamma_1 = 1.0$	-0.526	0.006	0.164	0.143	0.303	0.021
		$\gamma_2 = 0.5$	-0.281	-0.004	0.131	0.125	0.096	0.014
	2	$\alpha = 0.5$	0.067	-0.116	0.067	0.047	0.009	0.017
		$\gamma_1 = 1.0$	-0.093	-0.102	0.123	0.129	0.022	0.029
		$\gamma_2 = 0.5$	0.009	0.722	0.198	0.164	0.033	0.598
	3	$\alpha = 0.5$	0.056	-0.004	0.060	0.053	0.007	0.003
		$\gamma_1 = 1.0$	-0.036	0.252	0.162	0.141	0.021	0.085
		$\gamma_2 = 0.5$	0.092	0.073	0.142	0.125	0.021	0.022

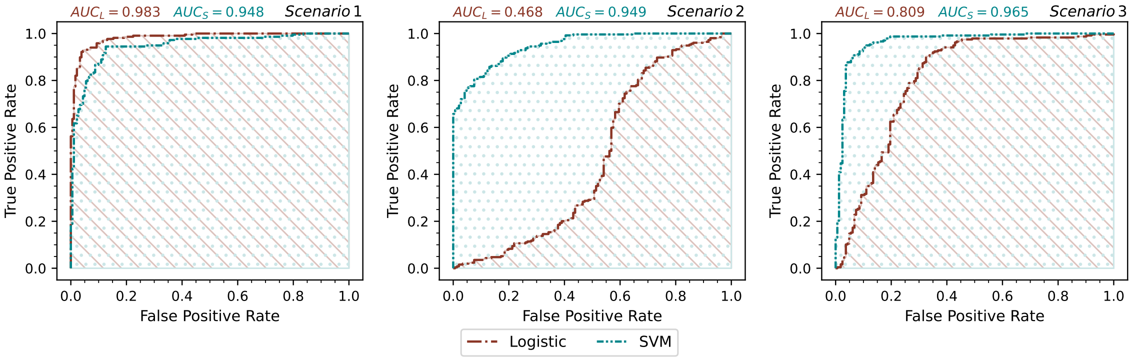


Figure 3.7: ROC curves based on simulated data for different scenarios

Table 3.3: AUC values under different scenarios

Scenario	LOGISTIC	SVM
1	0.983	0.948
2	0.468	0.949
3	0.809	0.965

can see that the curve levels off to a significant non-zero proportion. This indicates that there could be a greater likelihood of the presence of cured fraction in the data.

In Table 3.4, we present few important descriptive statistics related to the study. The proportion of relapses in the SI group for males and females are, respectively, 0.329 and 0.219. The proportion of relapses in the UC group for males and females are, respectively, 0.357 and 0.375. A two-sample z-test reveals statistically non-significant (p-value = 0.1959) difference in proportion of relapses between the SI and UC groups. A Pearson's χ^2 test for independence results in no significant association between gender and relapse rate (p-value = 0.3426). Further, no significant differences in duration of smoking (p-value = 0.088) and average number of cigarettes smoked per day (p-value = 0.369) before the study period are found between the relapsed and non-relapsed categories by non-parametric Mann-Whitney U-test. The distributions of DUR and AVGCIG, categorized by whether relapsed or not, are given in Figure 3.10.

Table 3.4: Distribution of proportion of relapse, average duration and average number of cigarettes smoked per year by gender and treatment group

Treatment Group	Measure	Gender	
		Female	Male
SI	n (%)	73 (32.735)	96 (43.049)
	$\hat{p}_r$ (95% CI)	0.329 (0.221, 0.437)	0.219 (0.136, 0.301)
	Avg Dur (SD)	29.506 (6.390)	25.246 (9.667)
	Avg Cig (SD)	30.343 (7.115)	29.375 (12.552)
UC	n (%)	14 (6.278)	40 (17.937)
	$\hat{p}_r$ (95% CI)	0.357 (0.106, 0.608)	0.375 (0.224, 0.525)
	Avg Dur (SD)	28.214 (8.833)	22.714 (9.160)
	Avg Cig (SD)	30.750 (7.502)	26.875 (9.915)

SI: smoking intervention, UC: usual care, n: sample size, %: percentage of the total, $\hat{p}_r$: proportion of relapse, CI: confidence interval, Avg Dur: average of DUR, Avg Cig: average of AVGCIG, SD: standard deviation

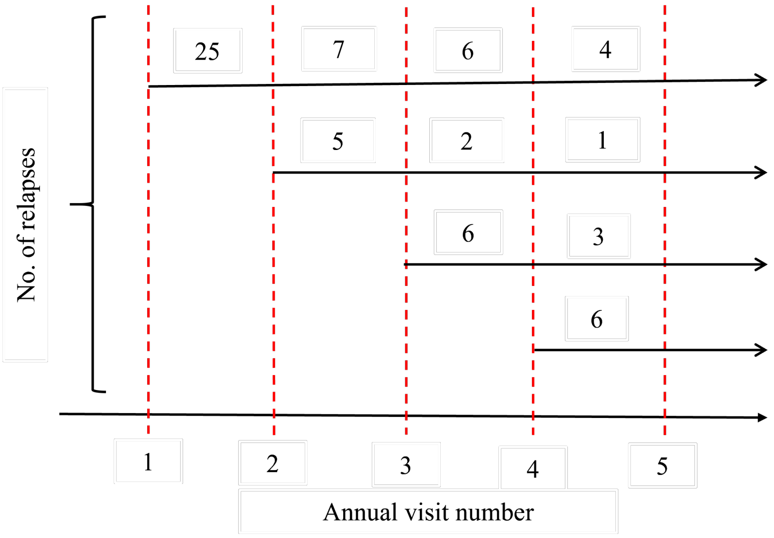


Figure 3.8: Number of relapses in between every consecutive annual visits from study entry

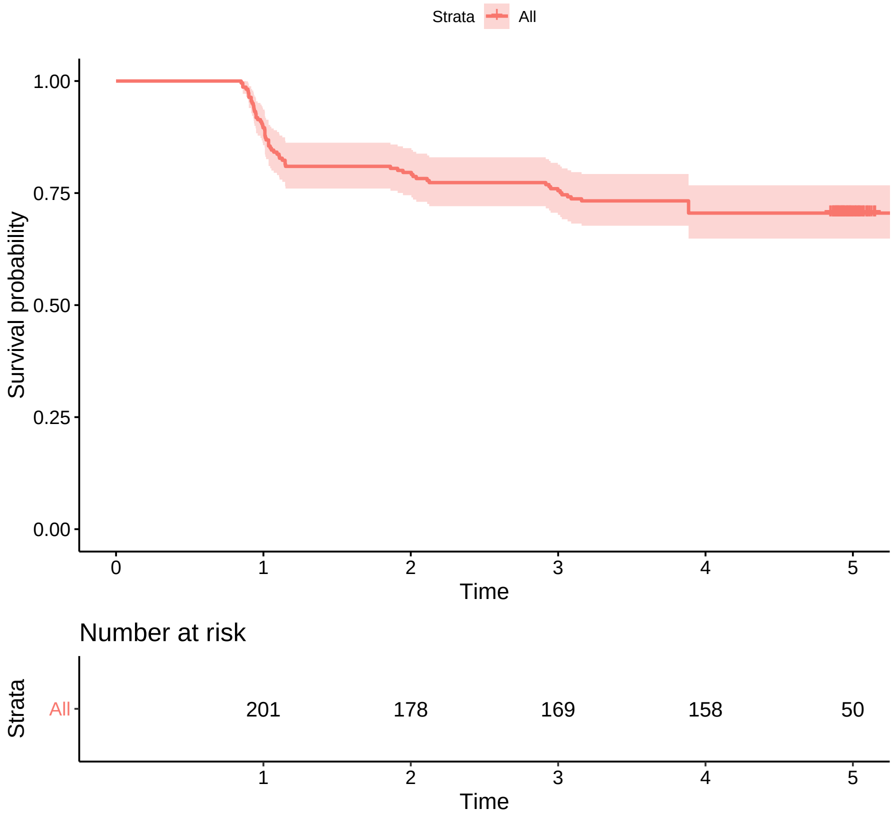


Figure 3.9: Kaplan Meier curve for the smoking cessation data

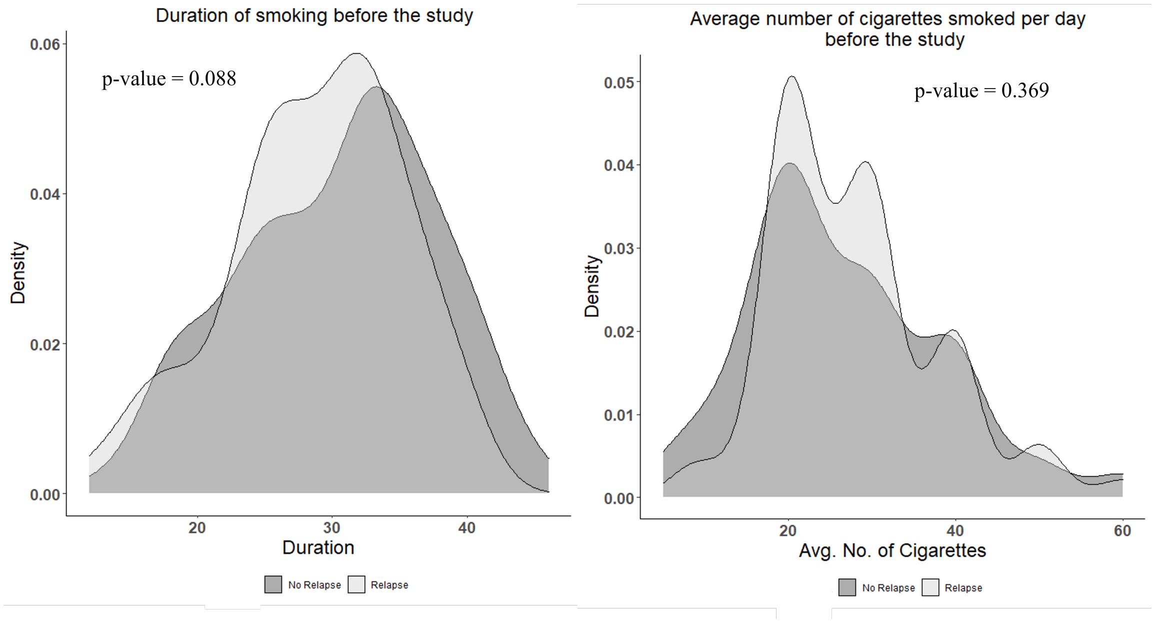


Figure 3.10: Distributions of DUR and AVGCIG categorized by the relapsed category
DUR: duration of smoking, AVGCIG: average number of cigarettes smoked per day

In our application, we consider duration of smoking and average cigarettes smoked as two covariates of interest. We fit the proposed SVM based mixture cure rate model and, for comparison, we also fit the logistic regression based mixture cure rate model. First, we draw inference on the incidence part of the model. In Figure 3.11, we plot the estimates of the uncured probabilities against the two covariates for both models. Clearly, under the proposed SVM based model, the change in the estimates of the uncured probabilities is non-monotonic with respect to duration of smoking and average cigarettes smoked. The estimates of the uncured probabilities vary from 21% – 38%. Under the logistic regression based model, owing to its rigid model assumption, the estimates of the uncured probabilities decrease with an increase in both duration of smoking and average cigarettes smoked. In this case, the estimated uncured probabilities vary from 16% – 51%. However, for fixed average cigarettes smoked, the uncured probability is a decreasing function of duration of smoking ($\hat{\beta}_1 = -0.319$). This may sound counter intuitive, but our finding is in line with the finding reported in [55]. On the other hand, for fixed duration of smoking, the uncured probability increases with an increase in average cigarettes smoked ($\hat{\beta}_2 = 0.181$).

Now, we turn our attention to the latency part of the model. In Table 3.5, we present the estimates of the latency parameters and their standard errors for both SVM based and logistic regression based models. The effects of duration of smoking and average cigarettes smoked on the latency part is the same for both models. Clearly, duration of smoking and average cigarettes smoked turns out to be significant as far as the time to relapse of uncured patients is concerned. Since the estimate of γ_1 is positive, the hazard of smoking relapse increases with longer duration of smoking. On the other hand, since the estimate of γ_2 is negative, it implies that those who smoked less cigarettes tend to relapse faster. In Figure 3.12, we plot the predicted survival probabilities of uncured subjects for fixed duration of smoking and different values of average cigarettes smoked. In Figure 3.13, we plot the predicted survival probabilities of uncured subjects for fixed average cigarettes smoked and different values of duration of smoking.

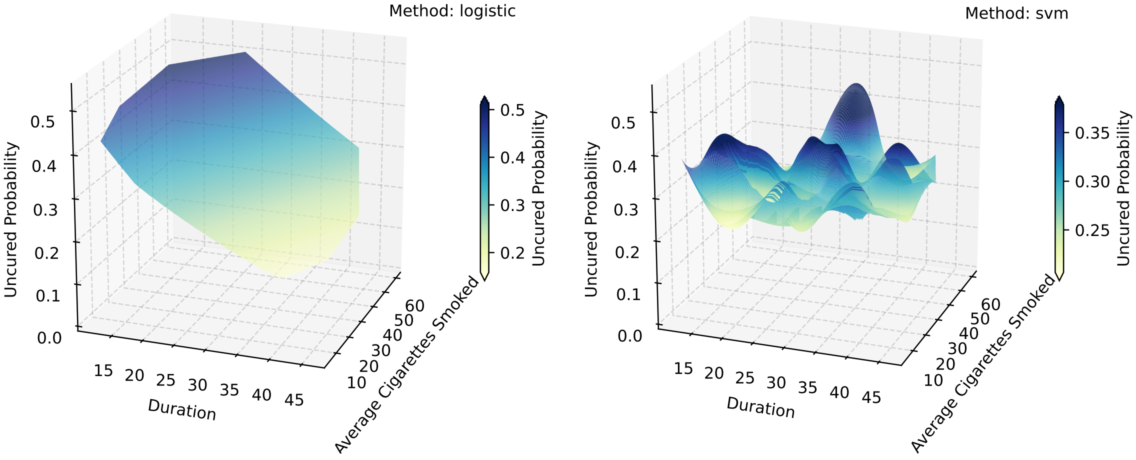


Figure 3.11: Estimates of uncured probabilities as a function of duration of smoking and average cigarettes smoked

Table 3.5: Estimation results corresponding to the latency parameters for the smoking cessation data

Parameter	Estimates		SE		p-value	
	SVM	LOGISTIC	SVM	LOGISTIC	SVM	LOGISTIC
α	0.895	0.875	0.072	0.064	–	–
γ_1 (DUR)	0.229	0.277	0.124	0.138	0.064	0.044
γ_2 (AVGCIG)	-0.214	-0.255	0.106	0.130	0.044	0.050

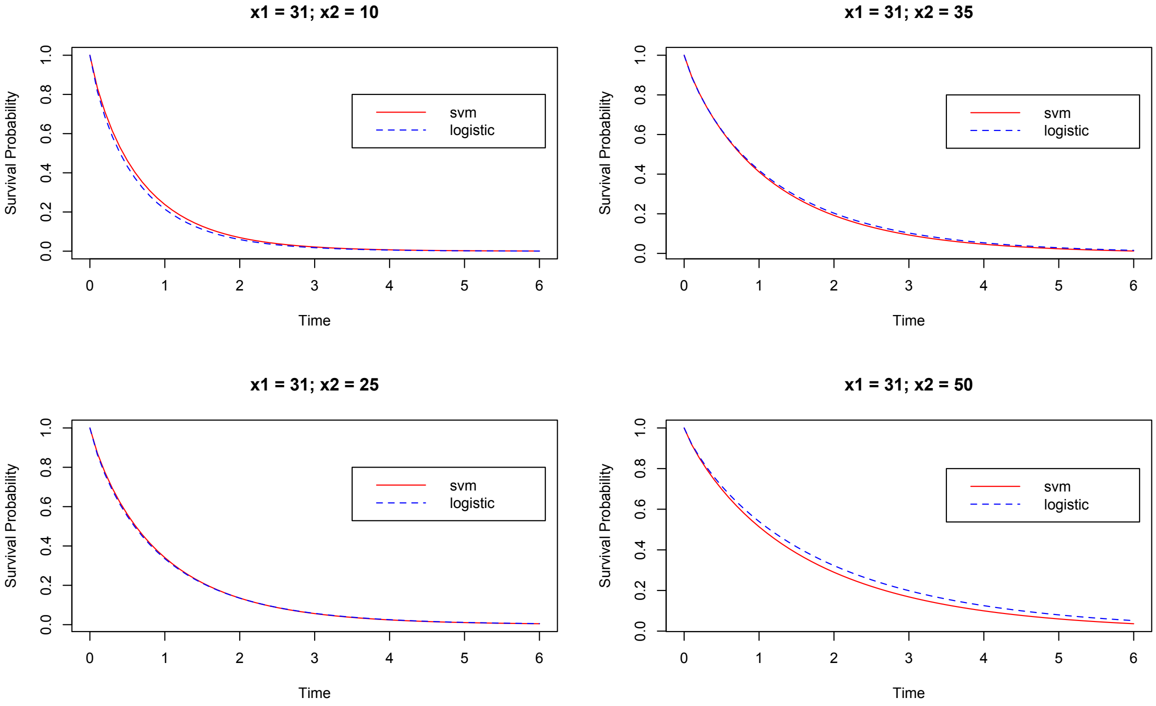


Figure 3.12: Predicted survival probability of the susceptible for fixed duration as smoker (x_1) and different values of average cigarettes smoked per day (x_2)

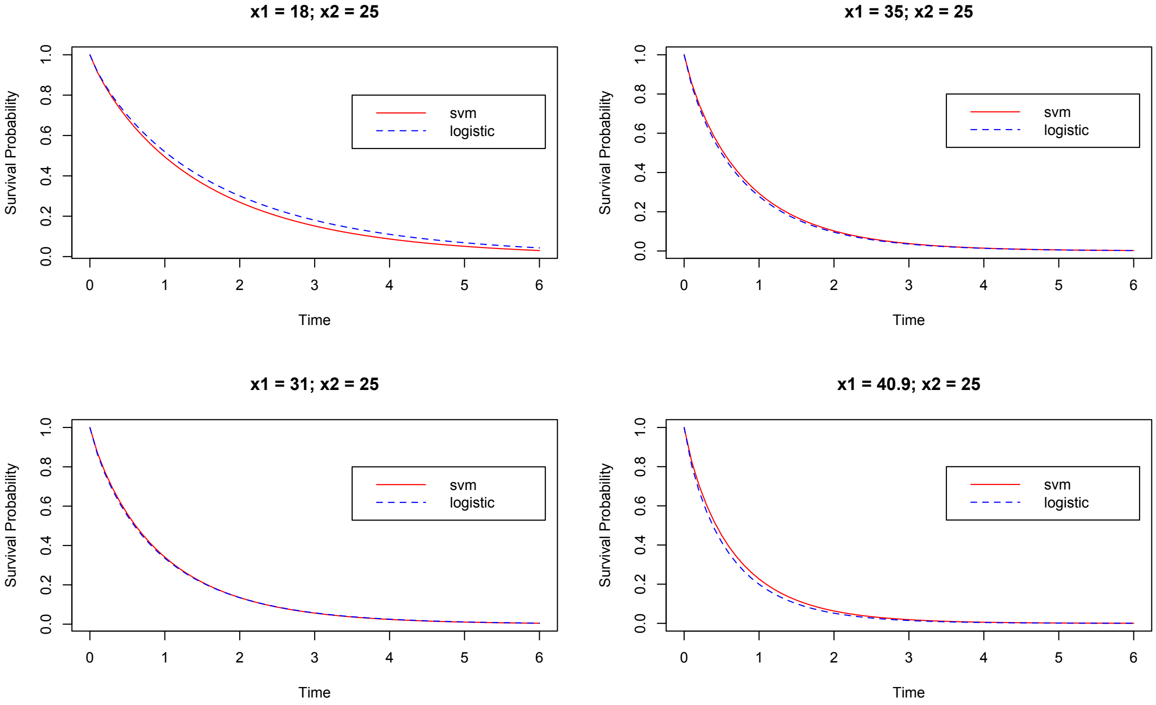


Figure 3.13: Predicted survival probability of the susceptible for fixed average cigarettes smoked per day (x_2) and different values of duration as smoker (x_1)

CHAPTER 4

Concluding remarks

4.1 Summary of research

The mixture cure rate model and the promotion time cure rate model are the two most commonly used cure rate models in the literature. Through a Box-Cox transformation on the population survival function, the Box-Cox transformation cure rate model provides an elegant way to unify the two commonly used cure rate models. In chapter 2, we proposed the use of generalized gamma distribution as the distribution of the lifetime for a particular transformation cure rate model, known as the Box-Cox transformation cure rate model, which contains the mixture and promotion time cure rate models as special cases. The generalized gamma distribution contains the commonly used Weibull, lognormal, and gamma distributions as its particular cases, and hence has substantial flexibility to capture the characteristics in a distribution that may be easily missed when using these particular cases. This allows us to carry out formal tests of hypotheses to check for the suitability of the mixture and the promotion time cure rate models. On the other hand, the generalized gamma distribution introduces flexibility that allows us to test for the suitability of the commonly used lifetime distributions. The proposed GGBCT model provides two-way flexibility in selecting a correct cure rate model (within the family of Box-Cox transformation cure rate models) together with a correct lifetime distribution (within the wider class of generalized gamma distribution) that jointly provides the best fit to a given data. Through model discrimination studies, we have shown that the likelihood ratio test can discriminate between models both within the Box-Cox family and within the generalized gamma family. The sensitivity analysis results clearly suggest that model mis-specification can lead to highly biased and highly inefficient inference on the cure rates. Thus, for a given time-to-event data, it is important to correctly specify the cure rate model as well as the lifetime distribution, assuming a completely parametric framework. Through the real breast cancer data, we are able to illustrate the importance of the Box-Cox family and the wider class of generalized gamma distribution. The Box-Cox family allows us to reject the specification of the mixture cure rate model ($\phi = 1$) for the breast cancer data. Furthermore, the generalized gamma family allows us to reject all the commonly used lifetime distributions, i.e., gamma, lognormal and Weibull distributions. It turns out that the promotion time cure rate model ($\phi = 0$) together with the generalized gamma lifetime ($q = 0.084$) provides the best fit to the breast cancer data. When compared to the piecewise exponential approach of Yin and Ibrahim [45], we found that our proposed parametric approach results in a better model fit.

Support vector machine has received a great amount of interest in the past two decades. It has been shown that SVM performs well in a wide array of problems including face detection, text categorization and pedestrian detection. However, the use of SVM in the context of cure rate models is new and not well explored. In chapter 3, we proposed a new cure rate model that uses the SVM to model the incidence part and a proportional hazards structure to model the latency part, given that the form of available data is interval censored. The new cure rate model inherits the properties of the SVM and can capture more complex classification boundaries. For the estimation purpose, we proposed an EM algorithm, where sequential minimal optimization together with Platt scaling method were employed to estimate the uncured probabilities. In this regard, due to the non availability of the training data, in the sense that the cured statuses are unknown, we made use of the multiple imputation based approach to generate missing cured statuses. Due to the complexity of the proposed model and the estimation method, we calculated the standard errors of the estimated parameters using non-parametric bootstrapping. Through simulation study, we have shown that when the true classification boundary is non-linear the proposed SVM based model performs much better than the standard logistic regression based model. This is true with respect to both incidence and latency parts of the model.

4.2 Future work

In this section, we discuss some future research problems that are along the lines of the research work carried out in this thesis.

4.2.1 Developing new optimization algorithms

In chapter 2, we used a readily available optimization method to calculate the maximum likelihood estimates of the GGBCT model parameters. Such an in-built optimizer may not perform well when the number of parameters to be estimated is high or when there are parameter(s) that make the likelihood surface flat. As such, it is of interest to explore other optimization algorithms that can handle the aforementioned issues [20, 60, 61].

4.2.2 Integrating other machine learning techniques with cure rate model

In this thesis, we considered a SVM-based approach in chapter 3 to capture non-linearity in the data. In this regard, it will be of great interest to study other machine learning algorithms such as the neural network, k-nearest neighbours and decision trees, among others. Furthermore, it will be of interest to employ the aforementioned machine learning algorithms to study more flexible and biologically motivated cure rate models such as those that look at the elimination or destruction of competing risks after a passage of time [62, 43, 63].

4.2.3 Semiparametric machine learning technique-based cure model

In chapter 3, we assumed a particular parametric form for the baseline hazard function when modeling the latency part of the mixture cure model. In such a case, as we have seen, the lifetime distribution of the susceptible subjects reduced to a parametric Weibull distribution. In this regard, we may think of modeling the baseline hazard using a semi-parametric approach such as the piecewise linear or piecewise exponential function.

CHAPTER 5

Appendix

5.1 Appendix for Introduction

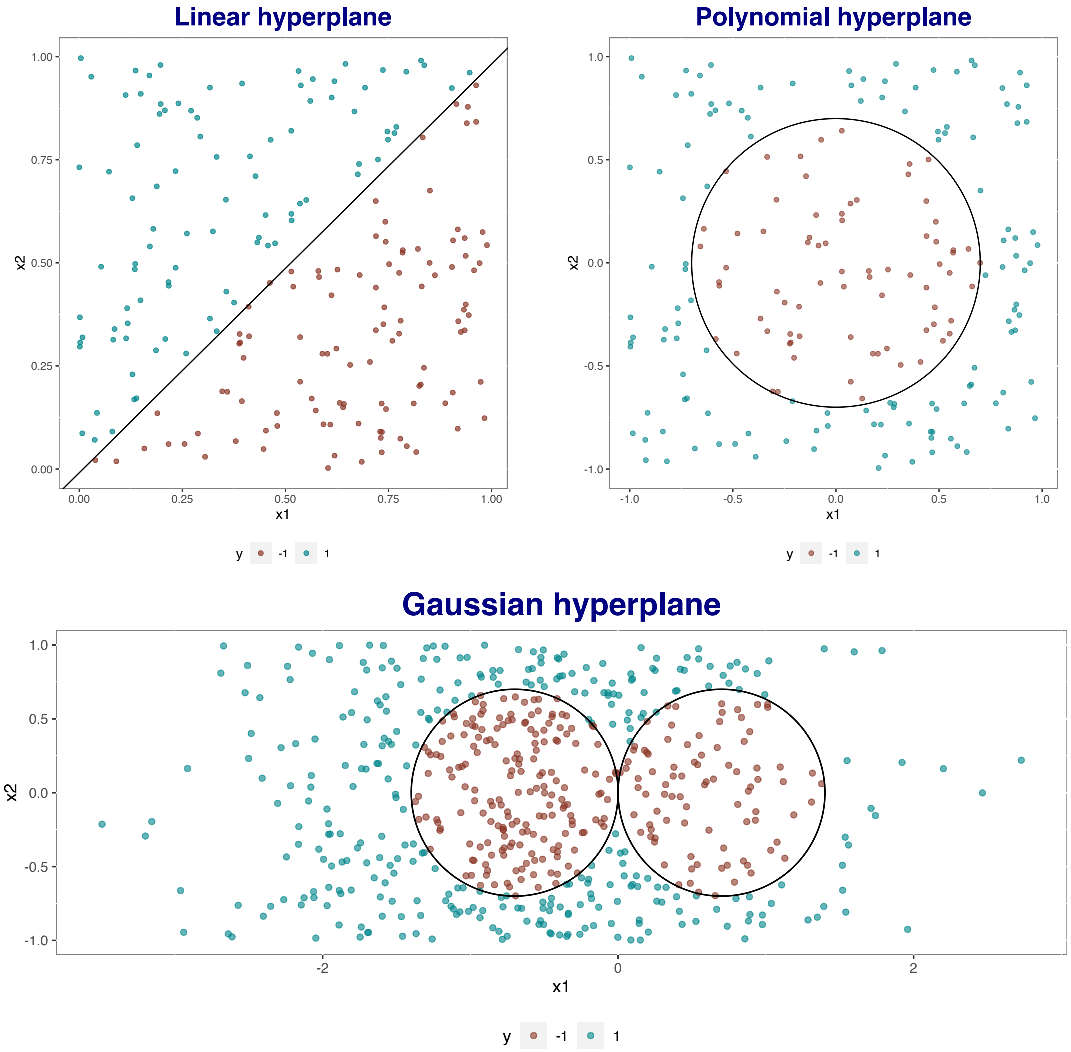


Figure 5.1: Three different hyperplanes for SVM

5.2 Appendix for Chapter 1

5.2.1 Generalized Gamma data generation verification

Let $Y \sim f(y;\boldsymbol{\gamma})$, which is defined as in eqn.(2.14). Let $S(y;\boldsymbol{\gamma})$ be the survival function of Y, which is given as in eqn. (2.15). Then, we have

$$V = Y^{\frac{q}{\sigma}} \sim \Gamma\left(\frac{1}{q^2}, \frac{q^2}{\lambda^{q/\sigma}}\right).$$

Therefore, $F(y) = 1 - S(y;\boldsymbol{\gamma}) = 1 - \dfrac{\Gamma\left(q^{-2}, q^{-2}(\lambda y)^{q/\sigma}\right)}{\Gamma\left(q^{-2}\right)} = A$. Note that

$$\Gamma\left(q^{-2}\right) = \Gamma\left(q^{-2}, q^{-2}(\lambda y)^{q/\sigma}\right) + \gamma\left(q^{-2}, q^{-2}(\lambda y)^{q/\sigma}\right).$$

Hence,

$$F(y) = \frac{\gamma\left(q^{-2}, q^{-2}\lambda^{q/\sigma}y^{q/\sigma}\right)}{\Gamma\left(q^{-2}\right)}.$$

Let $y^{q/\sigma} = v$, then we have

$$y = v^{\sigma/q},$$

and

$$A = F(y) = \frac{\gamma\left(q^{-2}, q^{-2}\lambda^{q/\sigma}v\right)}{\Gamma\left(q^{-2}\right)} = F_{gamma}\left(v; q^{-2}, \frac{q^2}{\lambda^{q/\sigma}}\right).$$

Then,

$$y^{q/\sigma} = v = \texttt{qgamma}\left(A, \texttt{shape} = q^{-2}, \texttt{rate} = q^{-2}\lambda^{q/\sigma}\right),$$

which implies

$$v = y^{q/\sigma} \sim \Gamma\left(\frac{1}{q^2}\frac{q^2}{\lambda^{q/\sigma}}\right).$$

Therefore,

$$y = \left(\texttt{qgamma}\left(A, \texttt{shape} = q^{-2}, \texttt{rate} = q^{-2}\lambda^{q/\sigma}\right)\right)^{\frac{\sigma}{q}}.$$

Bibliography

[1] J. P. Sy and J. M. Taylor, “Estimation in a cox proportional hazards cure model,” *Biometrics*, vol. 56, no. 1, pp. 227–236, 2000.

[2] Y. Peng and B. Yu, *Cure Models: Methods, Applications, and Implementation*. CRC Press, 2021.

[3] J. W. Boag, “Maximum likelihood estimates of the proportion of patients cured by cancer therapy,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 11, no. 1, pp. 15–53, 1949.

[4] V. T. Farewell, “The use of mixture models for the analysis of survival data with long-term survivors,” *Biometrics*, pp. 1041–1046, 1982.

[5] A. Y. Kuk and C.-H. Chen, “A mixture model combining logistic regression with proportional hazards regression,” *Biometrika*, vol. 79, no. 3, pp. 531–541, 1992.

[6] A. I. Goldman, “Survivorship analysis when cure is a possibility: a Monte Carlo study,” *Statistics in Medicine*, vol. 3, pp. 153–163, 1984.

[7] J. M. Taylor, “Semi-parametric estimation in failure time mixture models,” *Biometrics*, pp. 899–907, 1995.

[8] Y. Peng and K. B. Dear, “A nonparametric mixture model for cure rate estimation,” *Biometrics*, vol. 56, no. 1, pp. 237–243, 2000.

[9] A. Y. Yakovlev, A. B. Cantor, and J. J. Shuster, “Parametric versus non-parametric methods for estimating cure rates based on censored survival data,” *Statistics in Medicine*, vol. 13, no. 9, pp. 983–986, 1994.

[10] J. G. Ibrahim, M.-H. Chen, and D. Sinha, "Bayesian semiparametric models for survival data with a cure fraction," *Biometrics*, vol. 57, no. 2, pp. 383–388, 2001.

[11] M.-H. Chen, J. G. Ibrahim, and D. Sinha, "A new bayesian model for survival data with a surviving fraction," *Journal of the American Statistical Association*, vol. 94, no. 447, pp. 909–919, 1999.

[12] J. Berkson and R. P. Gage, "Survival curve for cancer patients following treatment," *Journal of the American Statistical Association*, vol. 47, no. 259, pp. 501–515, 1952.

[13] R. A. Maller and S. Zhou, "Estimating the proportion of immunes in a censored sample," *Biometrika*, vol. 79, no. 4, pp. 731–739, 1992.

[14] L. Zhao, D. Feng, E. L. Bellile, and J. M. Taylor, "Bayesian random threshold estimation in a cox proportional hazards cure model," *Statistics in Medicine*, vol. 33, no. 4, pp. 650–661, 2014.

[15] S. Pal and N. Balakrishnan, "Expectation maximization algorithm for box–cox transformation cure rate model and assessment of model misspecification under weibull lifetimes," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 3, pp. 926–934, 2017.

[16] R. A. Maller and X. Zhou, *Survival Analysis with Long-term Survivors*. Wiley New York, 1996.

[17] J. Rodrigues, M. de Castro, V. G. Cancho, and N. Balakrishnan, "Com–poisson cure rate survival models and an application to a cutaneous melanoma data," *Journal of Statistical Planning and Inference*, vol. 139, no. 10, pp. 3605–3611, 2009.

[18] N. Balakrishnan and S. Pal, “Expectation maximization-based likelihood inference for flexible cure rate models with weibull lifetimes,” *Statistical Methods in Medical Research*, vol. 25, no. 4, pp. 1535–1563, 2016.

[19] N. Balakrishnan, M. Koutras, F. Milienos, and S. Pal, “Piecewise linear approximations for cure rate models and associated inferential issues,” *Methodology and Computing in Applied Probability*, vol. 18, no. 4, pp. 937–966, 2016.

[20] S. Pal and S. Roy, “A new non-linear conjugate gradient algorithm for destructive cure rate model and a simulation study: illustration with negative binomial competing risks,” *Communications in Statistics-Simulation and Computation*, pp. 1–15, 2020.

[21] S. Pal and S. Roy, “On the estimation of destructive cure rate model: A new study with exponentially weighted poisson competing risks,” *Statistica Neerlandica*, 2021.

[22] G. Yin and J. G. Ibrahim, “Cure rate models: a unified approach,” *Canadian Journal of Statistics*, vol. 33, no. 4, pp. 559–570, 2005.

[23] G. E. Box and D. R. Cox, “An analysis of transformations (with discussion),” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 26, no. 2, pp. 211–243, 1964.

[24] J. Xu and Y. Peng, “Nonparametric cure rate estimation with covariates,” *Canadian Journal of Statistics*, vol. 42, no. 1, pp. 1–17, 2014.

[25] G. Diao and G. Yin, “A general transformation class of semiparametric cure rate frailty models,” *Annals of the Institute of Statistical Mathematics*, vol. 64, no. 5, pp. 959–989, 2012.

[26] M. Koutras and F. Milienos, “A flexible family of transformation cure rate models,” *Statistics in Medicine*, vol. 36, no. 16, pp. 2559–2575, 2017.

[27] D. Zeng, G. Yin, and J. G. Ibrahim, “Semiparametric transformation models for survival data with a cure fraction,” *Journal of the American Statistical Association*, vol. 101, no. 474, pp. 670–684, 2006.

[28] B. Yu, R. C. Tiwari, K. A. Cronin, and E. J. Feuer, “Cure fraction estimation from the mixture cure models for grouped survival data,” *Statistics in Medicine*, vol. 23, no. 11, pp. 1733–1747, 2004.

[29] N. Balakrishnan and Y. Peng, “Generalized gamma frailty model,” *Statistics in Medicine*, vol. 25, no. 16, pp. 2797–2816, 2006.

[30] N. Balakrishnan and S. Pal, “An em algorithm for the estimation of parameters of a flexible cure rate model with generalized gamma lifetime and model discrimination using likelihood-and information-based methods,” *Computational Statistics*, vol. 30, no. 1, pp. 151–189, 2015.

[31] S. Pal, H. Yu, Z. D. Loucks, and I. M. Harris, “Illustration of the flexibility of generalized gamma distribution in modeling right censored survival data: Analysis of two cancer datasets,” *Annals of Data Science*, vol. 7, no. 1, pp. 77–90, 2020.

[32] Y. Peng, “Fitting semiparametric cure models,” *Computational Statistics & Data Analysis*, vol. 41, no. 3-4, pp. 481–490, 2003.

[33] C. Cai, Y. Zou, Y. Peng, and J. Zhang, “smcure: An R-package for estimating semiparametric mixture cure models,” *Computer Methods and Programs in Biomedicine*, vol. 108, no. 3, pp. 1255–1260, 2012.

[34] E. N. Tong, C. Mues, and L. C. Thomas, “Mixture cure models in credit scoring: If and when borrowers default,” *European Journal of Operational Research*, vol. 218, no. 1, pp. 132–139, 2012.

[35] A. López-Cheda, R. Cao, M. A. Jácome, and I. Van Keilegom, “Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models,” *Computational Statistics & Data Analysis*, vol. 105, pp. 144–165, 2017.

[36] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[37] P. Li, Y. Peng, P. Jiang, and Q. Dong, “A support vector machine based semiparametric mixture cure model,” *Computational Statistics*, vol. 35, no. 3, pp. 931–945, 2020.

[38] W. Sauerbrei and P. Royston, “Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials,” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 162, no. 1, pp. 71–94, 1999.

[39] R. P. Murray, N. R. Anthonisen, J. E. Connett, R. A. Wise, P. G. Lindgren, P. G. Greene, M. A. Nides, L. Group *et al.*, “Effects of multiple attempts to quit smoking and relapses to smoking on pulmonary function,” *Journal of Clinical Epidemiology*, vol. 51, no. 12, pp. 1317–1326, 1998.

[40] P. Wiangnak and S. Pal, “Gamma lifetimes and associated inference for interval-censored cure rate model with COM–Poisson competing cause,” *Communications in Statistics-Theory and Methods*, vol. 47, no. 6, pp. 1491–1509, 2018.

[41] S. Banerjee and B. P. Carlin, “Parametric spatial cure rate models for interval-censored time-to-relapse data,” *Biometrics*, vol. 60, no. 1, pp. 268–275, 2004.

[42] Y.-J. Kim and M. Jhun, “Cure rate model with interval censored data,” *Statistics in Medicine*, vol. 27, no. 1, pp. 3–14, 2008.

[43] S. Pal, J. Majakwara, and N. Balakrishnan, "An EM algorithm for the destructive COM-Poisson regression cure rate model," *Metrika*, vol. 81, no. 2, pp. 143–171, 2018.

[44] I. L. MacDonald, "Does newton–raphson really fail?" *Statistical Methods in Medical Research*, vol. 23, no. 3, pp. 308–311, 2014.

[45] G. Yin and J. G. Ibrahim, "A general class of bayesian survival models with zero and nonzero cure fractions," *Biometrics*, vol. 61, no. 2, pp. 403–412, 2005.

[46] J. Sun, *The statistical analysis of interval-censored failure time data*. Springer, 2007.

[47] J. C. Lindsey and L. M. Ryan, "Methods for interval-censored data," *Statistics in Medicine*, vol. 17, no. 2, pp. 219–238, 1998.

[48] S. Ma, "Cure model with current status data," *Statistica Sinica*, pp. 233–249, 2009.

[49] S. Ma, "Mixed case interval censored data with a cured subgroup," *Statistica Sinica*, pp. 1165–1181, 2010.

[50] L. Xiang, X. Ma, and K. K. Yau, "Mixture cure model with random effects for clustered interval-censored survival data," *Statistics in Medicine*, vol. 30, no. 9, pp. 995–1006, 2011.

[51] B. A. Aljawadi, M. R. A. Bakar, and N. A. Ibrahim, "Nonparametric versus parametric estimation of the cure fraction using interval censored data," *Communications in Statistics-Theory and Methods*, vol. 41, no. 23, pp. 4251–4275, 2012.

[52] A. Tsodikov, J. Ibrahim, and A. Yakovlev, "Estimating cure rates from survival data: an alternative to two-component mixture models," *Journal of the American Statistical Association*, vol. 98, no. 464, pp. 1063–1078, 2003.

[53] D. G. Kleinbaum and M. Klein, *Survival Analysis*. Springer, 2010.

[54] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. John Wiley & Sons, 2007, vol. 382.

[55] S. Pal and N. Balakrishnan, “Likelihood inference for the destructive exponentially weighted poisson cure rate model with weibull lifetime and an application to melanoma data,” *Computational Statistics*, vol. 32, no. 2, pp. 429–449, 2017.

[56] C. C. Chang and C. J. Lin, “LIBSVM: a library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[57] J. Platt, “Sequential minimal optimization: A fast algorithm for training support vector machines,” 1998.

[58] J. Platt *et al.*, “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” *Advances in Large Margin Classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[59] Y. Wu and G. Yin, “Cure rate quantile regression for censored data with a survival fraction,” *Journal of the American Statistical Association*, vol. 108, no. 504, pp. 1517–1531, 2013.

[60] Y.-h. Dai and Y. Yuan, “An efficient hybrid conjugate gradient method for unconstrained optimization,” *Annals of Operations Research*, vol. 103, no. 1, pp. 33–47, 2001.

[61] R. Fletcher and C. M. Reeves, “Function minimization by conjugate gradients,” *The Computer Journal*, vol. 7, no. 2, pp. 149–154, 1964.

[62] S. Pal and N. Balakrishnan, "Destructive negative binomial cure rate model and EM-based likelihood inference under Weibull lifetime," *Statistics & Probability Letters*, vol. 116, pp. 9–20, 2016.

[63] S. Pal and N. Balakrishnan, "Likelihood inference based on EM algorithm for the destructive length-biased Poisson cure rate model with Weibull lifetime," *Communications in Statistics - Simulation and Computation*, vol. 47, no. 3, pp. 644–660, 2018.

BIOGRAPHICAL STATEMENT

Pei Wang is from Yancheng, Jiangsu, China. She received her B.S. degree in Finance from China Agriculture University in 2010, her M.S. and Ph.D. degrees in General Statistics from the University of Texas at Arlington in 2020 and 2021, respectively. During her three and a half years at UT Arlington, she received several schoolarships and fellowships from the College of Science.