# Shrinkage Estimators for Semi-parametric Proportional Hazards Mixture Cure Models

by

© **Negar Kalanpour**

A thesis submitted to the School of Graduate Studies in partial fulfillment of the requirements for the degree of Master of Science.

Department of Mathematics and Statistics

Memorial University

September 2024

St. John's, Newfoundland and Labrador, Canada

# Abstract

Survival analysis is essential for modelling time-to-event data, particularly in medical research. Mixture cure models are widely used methods to study patients' latency and incidence components. This research focuses on mixture model properties in the semi-parametric estimation of the Cox proportional hazard models in the presence of the multicollinearity problem, where the explanatory variables are linearly dependent so that the input design matrix is ill-conditioned. In the mixture of cure models, the multicollinearity issue can happen in both latency and incidence components, where the commonly used least squares (LS) method may lead to unreliable estimates for the coefficients of the underlying model. To address this issue, we propose shrinkage methods to estimate the coefficient of the underlying model. To do so, we developed new expectation-maximization (EM) algorithms to incorporate the shrinkage methods for both components.

Through various simulations, we show that the proposed shrinkage methods cope with the multicollinearity problem in latency and incidence components and lead to more reliable estimates in semi-parametric settings. Our findings indicate that Ridge and Liu-type (LT) shrinkage methods provide more reliable parameter estimates and outperform the LS estimation method in scenarios with high multicollinearity.

The developed methods are finally applied to a dataset on breast cancer, analyzing the disease prognosis and survival rates of patients with 10 or more positive lymph nodes. The results consistently show that the Ridge and LT methods offer better estimation and survival results compared to the LS method. Our numerical studies show the practical advantages of our proposed shrinkage methods in medical research.

**Keywords:** Semi-parametric models, Mixture Cure Models, Cox Proportional Hazard Model, Multicollinearity, Shrinkage Estimators, EM algorithm

I dedicate this thesis to my beloved partner and family.

# Acknowledgements

I would like to thank my supervisor, Dr. Armin Hatefi, for his continuous guidance and support throughout our research and thesis writing. His encouragement and valuable feedback made this work achievable. I would also like to express my gratitude to my supervisory committee member, Dr. Loredo-Osti, whose comments and advice improved the quality of this thesis. In addition, I would like to acknowledge Dr. Variyath, Dr. Oyet, Dr. Fan, and all the other faculty members whose courses and insights were a huge help in shaping my academic journey.

I am also grateful to Dr. Hatefi, the Department of Mathematics and Statistics, and the School of Graduate Studies at the Memorial University of Newfoundland for their financial support.

Finally, I am incredibly grateful to my family and friends, who have supported me unconditionally and helped me through every step of my academic journey.

# Table of contents

# List of tables

# List of figures

# Chapter 1

# Introduction

This thesis centers on mixture cure models, flexible statistical tools in survival analysis that can distinguish between individuals likely to be affected by an event of interest and those who are cured or unaffected by it. Our focus lies particularly in the semi-parametric proportional hazards mixture cure model, which is a mixture of two components, including logistic and Cox proportional hazards (PH) regression models.

Multicollinearity is a common issue in datasets characterized by high correlations among predictor variables. The unknown parameters in the semi-parametric mixture cure models are commonly estimated through an expectation–maximization (EM) algorithm that maximizes their likelihood function. However, these estimates may appear considerably unreliable when multicollinearity is present. In this research, we proposed shrinkage methods into the EM algorithm for mixture cure models, aiming to tackle the challenges caused by multicollinearity in parameter estimation within these models. By developing and implementing shrinkage Ridge and Liu-type (LT) methods, we aim to refine the estimation process, thereby improving the reliability and accuracy of model parameters. Through a series of simulations and analyses using real-world data, we demonstrate that the LT estimators surpass both the Ridge and the original method in accurately estimating the parameters of mixture cure models.

In order to implement the shrinkage estimators into the mixture cure models, we first need to implement them into the mixture cure models' components: Logistic and Cox PH regression models. Logistic regression is a special regression case with binary responses that distinguish between the cured and the uncured individuals. Therefore, in Section 1.1, we start by introducing linear regression to familiarize ourselves with

the simplest form of regression. After this foundational understanding, we move on to logistic and Cox PH regressions, which will also be discussed further in Chapter 2. Section 1.2 examines multicollinearity and its implications in survival analysis, proposing shrinkage estimators as a solution. To incorporate shrinkage estimators into the logistic and Cox PH models, we first implement them in the linear regression model in Section 1.3, and then incorporate them into mixture cure models and their components in Chapter 3. Finally, Section 1.4 discusses the mixture cure models.

## 1.1 Regression Models

In this section, we discuss regression models. Regression modelling is a fundamental statistical tool that gives insights into the relationships between variables, enabling the investigation of factors such as exposures or risk factors on outcomes like mortality or cancer. Regression analysis helps in understanding and predicting outcomes based on input variables. Regression methods offer a universal approach to exploring relationships between dependent (response) and independent (explanatory) variables, also known as covariates. Simple regression involves one explanatory variable, while multiple regression incorporates several.

First, we discuss linear regression, a longtime topic used to model the relationship between a dependent variable and one or more independent variables. In its simplest form, linear regression aims to find the best-fitting linear relationship between the predictors and the response variable; it is formulated as

$$y_i = \mathbf{x}_i^\top \beta + \epsilon_i, \tag{1.1}$$

where $y_i$ is the response variable of the $i$-th observation ($i = 1, ..., n$), $\beta$ is the vector of $p$ unknown coefficients, and $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top$ represents the predictor variables for the $i$-th observation. The design matrix $\mathbf{X} = (\mathbf{x}_1^\top, ..., \mathbf{x}_n^\top)^\top$ is a $(n \times p)$ matrix where $p < n$, and $\epsilon = (\epsilon_1, ..., \epsilon_n)$ is the vector of independent and identically distributed error terms that follow a standard normal distribution. The most common method for estimating linear regression coefficients is the Least Squares (LS) method [28]. This method's goal is to minimize the sum of the squared errors, namely $\epsilon$. Mathematically,

it is written as

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \beta)^2, \tag{1.2}$$

where $n$ is the number of observations. Solving (1.2) gives the LS estimates of the coefficients $\beta$, denoted as $\widehat{\beta}$

$$\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top Y,$$

where $\mathbf{X}$ is the design matrix of predictors, $Y$ is the vector of observed responses, $\mathbf{X}^\top$ denotes the transpose of $\mathbf{X}$, and $(\mathbf{X}^\top \mathbf{X})^{-1}$ denotes the inverse of the matrix $\mathbf{X}^\top \mathbf{X}$.

This unbiased linear estimation method works well when $\mathbf{X}^\top \mathbf{X}$ is not a singular matrix, meaning that its determinant is not equal to zero [24]. Least squares estimation can produce high-variance estimates of the coefficients when there is multicollinearity among the predictors. To address this issue, shrinkage methods were introduced. Shrinkage methods work by shrinking the estimated coefficients towards zero, effectively reducing their variance by adding a small value to the diagonal of the $\mathbf{X}^\top \mathbf{X}$ matrix. By using shrinkage methods, we make a balance between bias and variance in the coefficient estimates, hopefully leading to more reliable estimates, smaller variances, and smaller mean square errors [24].

In the following sections, we outline the core features of logistic regression (1.1.1) for binary outcomes and Cox proportional hazard regression (1.1.2) for time-to-event data and provide insights into model construction. In Chapter 2, we will discuss these models further.

### 1.1.1 Logistic Regression

Logistic regression is a powerful statistical tool utilized in medical research to understand the effect of predictor variables on categorical outcomes. These outcomes are binary, such as the presence or absence of a disease. When exploring multiple predictors, the logistic regression model becomes a multivariable model, allowing for the simultaneous analysis of various factors. This method is widely used in medical research because it can handle complex relationships between predictors and outcomes [4]. In medical research, logistic regression models have many applications, going even beyond epidemiology and public health. These models are useful in uncovering the complicated connections between risk factors and disease occurrence.

Other than logistic regression, statistical techniques like the Cox proportional hazards model, which will be discussed in Subsection 1.1.2, also hold significance in medical research [41, 56].

In finance, a common method for forecasting bankruptcy involves enhancing research findings to identify numerous potential predictive financial and non-financial variables. This process requires filtering through a vast array of variables, subsequently narrowing down the selection through traditional mathematical analysis to construct a bankruptcy prediction model [12]. Within this framework, logistic regression, among other traditional classification techniques, emerges as a prominent tool for predicting financial distress, particularly leveraging financial ratios [7].

Unlike linear least squares, which allows for a direct calculation of parameters through a closed-form solution, logistic regression lacks such a straightforward closed-form solution due to the non-linear nature of the logistic function. As a result, the maximum likelihood estimator of the logistic regression parameters is typically obtained using iterative methods such as the Newton-Raphson method. The formula and explanation of logistic regression, along with the parameter estimation technique, are provided in Section 2.1, allowing for a deeper understanding of logistic regression application.

### 1.1.2 Cox Proportional Hazard Regression

Survival analysis is known as an area focused on modelling and analyzing data, the primary objective of which is to determine the time until the incidence of a particular event [2]. A fundamental challenge in this field lies in scenarios where specific event outcomes become unobservable beyond a certain time threshold or where some individuals do not encounter the event of interest within the observation period. This scenario, known as censoring, causes a significant challenge, which can be effectively addressed by applying survival analysis methodologies [50].

In survival analysis, the time to the event of interest is only available for individuals where the event occurs within the observation period. However, only censored time is available for the remaining individuals due to loss of follow-up during observation or because their time to the event exceeds the observation period. Censored individuals are those where data collection ends, individuals withdraw, or tracking is lost [60].

In cancer research, particularly in our study of breast cancer data, a key focus lies in assessing the duration from diagnosis until death or recurrence, termed "survival time." Survival data rarely follow a normal distribution; instead, they tend to show skewness commonly with early events and fewer occurrences later on. These characteristics of the data necessitate the utilization of specialized techniques such as survival analysis [8].

In survival analysis, data are typically characterized and analyzed through two interconnected key concepts: survival and hazard probabilities. The survival function, denoted as $S(t)$, represents the likelihood of an individual surviving from a defined starting point, such as a cancer diagnosis, up to a specified future time t. This function is fundamental in understanding the probability of an individual remaining in a particular state over time. Complementing the survival function are two essential functions associated with the continuous random variable T, which represents the duration spent in a specific state, such as an individual's health condition. These functions are the cumulative distribution function (cdf), denoted as $F(t)$, which quantifies the probability of the event occurring before or at time t, and the probability density function (pdf), denoted as $f(t)$, which provides the likelihood of the event occurring at time t. The $F(t)$ is commonly referred to as the failure function, while the survival function, $S(t)$, is defined as the complementary function of $F(t)$, where t represents the time that has passed since entering the state at the defined starting point. Additionally, the survival function decreases steadily as $t$ increases, starting at 1 when $t = 0$. This signifies that all observed individuals survive at the beginning of the observation period, indicating that none of the events of interest have occurred [29]. The failure function is obtained by

$$F(t) = P(T \leq t), \tag{1.3}$$

that implies the survival function to be given as

$$S(t) = P(T > t) = 1 - F(t). \tag{1.4}$$

The analysis of survival data is enhanced by the hazard function. The hazard function, often denoted as h(t), explains an event's occurrence rate at time $t$, given that no event has occurred before $t$. The hazard function assumes the individual has

survived up to that moment. Understanding the concept of the hazard function is important because it shows how the risk of an event evolves over time. The hazard function is defined by the relationship between the probability density function and the survival function as given

$$h(t) = \frac{f(t)}{S(t)}. \tag{1.5}$$

The differential form of the hazard function is given by

$$h(t) = -\frac{d}{dt} \log S(t). \tag{1.6}$$

Furthermore, we can rewrite the survival function (1.4) as

$$S(t) = \exp(-H(t)), \tag{1.7}$$

where $H(t) = \int_0^t h(u)\, du$ indicates the cumulative hazard function [33, 40].

Three primary approaches exist for estimating survival functions: non-parametric, semi-parametric, and parametric methods. These methods include, for example, Kaplan-Meier (KM) and Nelson-Aalen estimator (NA), which are mainly suitable when no underlying distribution is assumed for event times. The Cox model relies on proportional hazards assumption and partial likelihood estimation. Each method offers distinct advantages based on data characteristics and research objectives [60].

In studies, individuals are often observed for different periods, resulting in time-to-event data, also known as survival or failure time data. In time-to-event data, the time until an event occurs is the focus instead of binary outcomes. The distribution of time-to-event data is typically characterized by the survival distribution function (1.4), often estimated using a method named the Kaplan-Meier method to assess the impact of one variable. However, for multiple variables, potentially continuous explanatory variables with the response variable necessitate a multiple regression model [4].

David R. Cox was the first to introduce the Cox proportional hazard regression as an approach to analyze time-to-event data [10]. This model is a primary choice and stands as one of the most commonly applied statistical models in survival analysis. Cox regression models defined in Section 2.2 have become widely popular because they do not necessitate the assumption of a specific survival distribution for the data. Instead, these models rely on a hazard function [33].

The Cox model maintains proportional hazards across predictor values, regardless of changes in the underlying hazard over time. This allows it to be defined as a linear combination of risk factors in a parametric manner. The hazard function consists of two elements: Firstly, a parametric portion influenced by risk factors, which linearly affects survival duration. Incorporating risk factors through an exponential function ensures their effects are proportional. Secondly, the baseline hazard represents inherent risk. The non-parametric baseline function indicates the hazard when no risk factor exists without presuming any specific form or relationship with time. These characteristics underscore why the Cox model is regarded as a semi-parametric model, where parameters are uniquely dedicated to describing how predictors influence the outcome. [18, 34].

Renowned for its flexibility, the Cox semi-parametric proportional hazards model is a prominent choice in both medical and business failure prediction fields. In the Cox proportional hazards model, accurate prediction models for business failure hold great value across various industries, notably in financial investment and lending sectors. Referred to by different names like bankruptcy prediction, firm failure prediction, and financial stress prediction, the field of business failure prediction treats the phenomenon as a timeline, picturing businesses through their lifetime distributions. In this approach, regression-based survival analysis models typically estimate the hazard rate, from which the survival rate is derived as needed [19].

## 1.2 Multicollinearity

Collinearity, in regression analysis, refers to the linear relationships among independent variables. When two explanatory variables closely resemble perfect linear combinations of each other, meaning that one variable ultimately determines the other, it's termed collinearity. If this relationship extends to more than two explanatory variables, it is referred to as multicollinearity. This can lead to unreliable parameter estimates and standard errors, making it harder to understand how each predictor affects the response variable [21]. Multicollinearity also leads to an increase in the standard error of some of the coefficients, resulting in certain variables becoming statistically non-significant despite their true values, potentially leading to incorrect conclusions drawn from the model. Fundamentally, having high correlations among

predictors will prevent the researcher from capturing some predictors for inclusion in the model [54].

Detecting multicollinearity relies mainly on correlation coefficient, tolerance, and the variance inflation factor (VIF). When the correlation coefficient between variables is high, it suggests the potential presence of collinearity. Tolerance represents the proportion of a predictor's variance that remains unexplained by other predictors, while VIF indicates the extent to which multicollinearity inflates the variance of coefficient estimates. Unlike tolerance, there isn't a specific threshold for VIF to ascertain multicollinearity. Typically, VIF values surpassing 10 are commonly seen as indicative of multicollinearity. When correlation is present among the predictors, the standard errors of their coefficients increase, leading to inflation in the variance of predictor coefficients. VIF serves as a tool to measure this inflation in variance [53, 54]. While diagnosing multicollinearity does not provide a direct solution to the issue, recognizing its potential impact on regression analysis findings enables a more cautious interpretation of the data [59].

Multicollinearity can significantly impact the regression parameter estimation techniques, such as the maximum likelihood (ML) estimator, leading to serious consequences. This issue is widely known across various regression models. In Section 1.3, we describe the Ridge and Liu-type (LT) shrinkage estimation methods in terms of the linear regression method for the sake of completeness following [37]. In Chapter 4, we develop shrinkage methods for logistic and Cox PH regression models and discuss how they deal with the multicollinearity problem in the underlying semi-parametric mixture cure model. These shrinkage techniques have been known to provide a reliable strategy for tackling multicollinearity issues [21].

## 1.3 Shrinkage Estimators

As noted in Section 1.2, multicollinearity increases the standard error estimate of regression coefficients, resulting in wider confidence intervals and a higher likelihood of rejecting significant test statistics. Consequently, it causes unreliable estimates of regression coefficients, displaying incorrect signs and exaggerated importance for certain predictors due to their effects becoming intertwined [20, 64]. Shrinkage estimators are the most popular methods to overcome multicollinearity. Shrinkage estimators utilize

data from the entire model to create better estimations of regression parameters. This involves adjusting the estimates derived from the full model towards those of a chosen candidate submodel. Although this process introduces some bias in the estimation, it effectively decreases the overall estimation error, thereby making up for the impact of the bias [1].

The following sections discuss Ridge and Liu-type (LT) shrinkage methods utilized in linear regression. These methods are employed to address collinearity issues effectively. We begin with linear regression to familiarize ourselves with implementing shrinkage methods. This approach is important because these methods are essential for later applying shrinkage methods to logistic regression and Cox PH regression in Chapter 3, which are generalizations of the linear regression model.

### 1.3.1 Ridge Shrinkage Estimator

Ridge regression is a popular shrinkage method used to reduce collinearity in linear regression models [24]. It addresses collinearity by adding a penalty term to the mentioned least squares (LS) method. After adding the Ridge penalty to (1.2), we'll minimize the Ridge penalized sum of the squared errors as [21]

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \beta)^2 + \lambda \beta^\top \beta / 2,$$

and the ridge regression estimator of $\beta$ will be

$$\widehat{\beta}_R = (\mathbf{X}^\top \mathbf{X} + \lambda I)^{-1} \mathbf{X}^\top Y, \tag{1.8}$$

where $\lambda > 0$. Choosing a small $\lambda$ is preferable because as it increases, the gap between $\lambda^{\frac{1}{2}}\beta$ and 0 gets larger. Consequently, incorporating the ridge penalty into the linear regression equation leads to increased bias in ridge regression. The collinearity can be measured through the condition number that is defined as

$$\kappa = \left(\frac{\lambda_{max}}{\lambda_{min}}\right)^{1/2},$$

where $\lambda_{max}$ and $\lambda_{min}$ are maximum and minimum eigenvalues of $\mathbf{X}^\top \mathbf{X}$, respectively. A large condition number tells us that $\mathbf{X}^\top \mathbf{X}$ is ill-conditioned. Therefore, in cases of

high multicollinearity, a large $\lambda$ can decrease the multicollinearity, so the shrinkage parameter chosen by the current method for ridge regression may not sufficiently solve the issue of ill-conditioning. In addition, we introduce a two-parameter estimator named Liu-type (LT) estimator to address this challenge [37].

### 1.3.2 Liu-Type Shrinkage Estimator

Liu-type (LT) shrinkage estimator offers an alternative approach to address collinearity in regression models. Liu suggested replacing the left part of the ridge penalty $0 = \lambda^{\frac{1}{2}}\beta + \epsilon$, by $\left(-d/\lambda^{1/2}\right)\widehat{\beta}$, where $\widehat{\beta}$ could be any estimate of $\beta$ [37]. Now $\lambda$ is free to be chosen a large number, and we can adjust the new parameter $d$ for a good result.

If the LT penalty $\left(-d/\lambda^{1/2}\right)\widehat{\beta} = \lambda^{\frac{1}{2}}\beta + \epsilon$ is augmented into the linear regression formula (1.1), the penalized sum of squared errors is minimized as [21]

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \beta)^2 + \left[\left(-d/\lambda^{1/2}\right)\widehat{\beta} - \lambda^{\frac{1}{2}}\beta\right]^\top \left[\left(-d/\lambda^{1/2}\right)\widehat{\beta} - \lambda^{\frac{1}{2}}\beta\right].$$

Consequently, the LT regression estimator of $\beta$ will be

$$\widehat{\beta}_{Liu} = (\mathbf{X}^\top\mathbf{X} + \lambda I)^{-1}(\mathbf{X}^\top Y - d\widehat{\beta}), \tag{1.9}$$

where $\lambda > 0$, and $-\infty < d < \infty$.

In the calculated Liu-type estimator for linear regression (1.9), we use the $\lambda$ parameter to control how well the $\mathbf{X}^\top\mathbf{X} + \lambda I$ matrix is structured. Some bias unavoidably occurs when we reduce the condition number of $\mathbf{X}^\top\mathbf{X} + \lambda I$ to the desired level. This bias can be controlled by the second parameter $d$, making our model a better fit and effectively dealing with the ill-conditioning problem [37].

## 1.4 Mixture Models

Finite Mixture models (FMM) are statistical tools used to represent complex populations made up of multiple subgroups or components. They assume that the observed data arises from a combination of different probability distributions, each representing

a different subgroup within the population. Each component in a FMM is associated with its own probability distribution, and the model parameters include the parameters of these distributions as well as the mixing proportions [38].

On the other hand, mixture cure models are a specialized form of mixture models used in survival analysis, where the population is divided into subpopulations with different probabilities of being "cured" or experiencing the event of interest [17]. These models specifically account for the possibility of individuals remaining permanently unaffected by the event of interest. Mixture cure models combine elements of mixture modelling with survival analysis, providing a powerful tool for analyzing and characterizing complex survival data, particularly in medical research [36].

Similar to regression models, multicollinearity creates challenges in mixture models, especially when the covariates present a high correlation. This issue leads to less dependable maximum likelihood estimates for all coefficients within the mixture model. To cope with the multicollinearity issue, the proposed shrinkage ridge and LT methods are employed to estimate the model coefficients more accurately in the multicollinearity [26, 37].

In this thesis, we introduce our developed ridge and LT shrinkage techniques to address multicollinearity within mixture cure models. Through comprehensive simulations and Breast Cancer data analyses, we demonstrate these methods' enhanced reliability in estimating mixture model coefficients. We concentrate exclusively on scenarios featuring multicollinearity, given the nature of LT and ridge techniques as shrinkage methods specifically designed to address the multicollinearity concerns.

The thesis is outlined as follows: Chapter 2 introduces semi-parametric mixture cure models and discusses parameter estimation. In Chapter 3, we develop shrinkage estimators for our semi-parametric mixture cure model. Chapter 4 describes the application of our developed methods through extensive simulation and breast cancer data studies. Finally, Chapter 5 provides a summary and outlines potential future research directions.

# Chapter 2

# Semi-parametric Mixture Cure Models

An important area of research in medical studies is the time until an event occurs (failure, death, etc.). Survival analysis is a statistical time-to-event data analysis method in which the outcome variable shows the time until an event occurs. The time variable shows the time that the individuals became at risk of the event of interest. It is worthwhile to mention that the event does not necessarily always occur. Even in the case of a study where an event may not happen for sure by the end of the study, we encounter some data in our studies as "censored" data.

Survival models that include a component accounting for individuals who are cured are referred to as cure models. Cure models are widely used in modelling survival data [46]. One of the most popular cure models is the mixture cure model, which defines the study population as a mixture of cured and uncured subjects. The mixture of cure models enables us to study whether the event occurs and when it will occur separately. Hence, the mixture of the cure model consists of two component models, including the proportion of cured subjects (incidence model) and the survival function of uncured subjects (latency model) to handle heterogeneity in the time-to-event population. Furthermore, we desire to allow the risk of being uncured in our model. In that case, the model will turn into a semi-parametric mixture cure model developed through an EM algorithm. The proportional hazards (PH) model is a popular model in survival analysis, and if the survival model is specified using

the Cox proportional hazard regression, the mixture cure model is called the Semi-parametric proportional hazards mixture cure (PHMC) model.

This chapter discusses the logistic regression in Section 2.1 and the Cox proportional hazard regression in Section 2.2, and the methods that estimate their coefficients. We introduce the cure models in Section 2.3 and then construct the mixture of cure models in Subsection 2.3.1. Finally, Section 2.3.2 investigates the semi-parametric mixture of cure model and its latency and incidence models.

## 2.1 Logistic Regression

The logistic regression model is one of the most popular statistics models for analyzing binary events. Logistic regression is defined as a binary classification predictive model and is used in many fields, such as medical studies and machine learning. For example, logistic regression can be used to prognose a patient's disease status based on their symptoms and the disease risk factors [58]. Based on a set of explanatory variables, logistic regression will predict the results through binary response variables that consist of only two possible outcomes such as: success/failure or true/false. The logistic regression is given by:

$$P(y_i = 1 \mid \mathbf{z}_i) = \pi(\mathbf{z}_i) = \frac{1}{1 + e^{-(b_0 + b_1 z_{i1} + b_1 z_{i2} + ... + b_q z_{iq})}}, \tag{2.1}$$

where $\mathbf{y} = (y_1, ..., y_n)$ is the vector of response variables, $\mathbf{b} = (b_0, b_1, ..., b_q)$ is the vector of the unknown coefficients, and $\mathbf{Z} = (z_1, ..., z_q)$ represents the $(n \times q)$ design matrix of $q$ explanatory variables of the logistic regression. Furthermore, while the logistic response $\mathbf{y}$ can only take 0 and 1 values, $\pi(\mathbf{z}_i)$ can take any value in the [0,1] interval. In logistic regression, the logit link (also called the log of odds) is utilized as a transformation for the dependent variable. Rather than directly modelling the binary response variable Y, logistic regression employs the logit function. The logit transformation is defined by:

$$g(\mathbf{z}_i) = \frac{\pi(\mathbf{z}_i)}{1 - \pi(\mathbf{z}_i)}. \tag{2.2}$$

The logit function is commonly used to transform the probabilities from the unit interval [0,1] to real numbers[26]. One of the popular methods to estimate the $\mathbf{b}$

parameters is the maximum likelihood estimation.

### 2.1.1 Maximum Likelihhood Estimation

Maximum likelihood (ML) is among the most common methods to estimate the coefficients of a logistic regression model. In this context, the maximum likelihood finds the coefficients' estimates, henceforth called maximum likelihood estimate (MLE), that makes the observed data most probable by maximizing its likelihood function. The likelihood function represents the probability of observing the binary set of outcomes given the coefficients and is typically expressed using the Bernoulli distribution. Assume that $Y_i$ is the $i$-th binary response variable, where $Y_i \overset{iid}{\sim} Binomial\,(1, \pi(\mathbf{z}_i))$; $i = 1, \cdots, n$. The likelihood function of $\mathbf{b}$ (assuming independent observations) is the product of the individual likelihoods given by

$$L\,(\mathbf{b}) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i}.$$

Since the logarithm of the likelihood function is a monotonically increasing function, maximizing the log-likelihood is equivalent to maximizing the likelihood, and it is a much easier procedure. Accordingly, the log-likelihood can be obtained by

$$\ell(\mathbf{b}) = \sum_{i=1}^{n} \{y_i \cdot \log(\pi_i) + (1 - y_i) \cdot \log(1 - \pi_i)\}\,.$$

If we use the exponent form of $\pi$, the log-likelihood is given by

$$\ell(\mathbf{b}) = \sum_{i=1}^{n} \left\{ y_i \mathbf{z}_i^\top \mathbf{b} - \log\left(1 + \exp(\mathbf{z}_i^\top \mathbf{b})\right) \right\}. \tag{2.3}$$

Our goal is to find the parameter $\mathbf{b}$ that maximizes the log-likelihood function. Unfortunately, there is not a closed-form solution for the MLE of the parameter $\mathbf{b}$ in logistic regression. Consequently, we use the Newton-Raphson (NR) technique to find the MLE of the coefficients of the logistic regression. The NR is an iterative optimization method which is commonly used to compute the coefficients in logistic regression. This method iteratively finds the parameter estimates by leveraging the

first and second derivatives of the log-likelihood function. The process involves updating the parameter values to maximize the log-likelihood, ultimately converging to the optimal solution. The NR is utilized for its efficiency in estimating logistic regression coefficients when a closed-form solution is unavailable. The algorithm iteratively estimates $\mathbf{b}$ as follows

$$\widehat{\mathbf{b}}^{(m+1)} = \widehat{\mathbf{b}}^{(m)} - H^{-1}\left(\widehat{\mathbf{b}}^{(m)}\right).\nabla_{\mathbf{b}}\ell\left(\widehat{\mathbf{b}}^{(m)}\right), \tag{2.4}$$

where $\widehat{\mathbf{b}}^{(m)}$ is the estimate updated from the $m$-th iteration, and $\nabla_{\mathbf{b}}\ell\left(\widehat{\mathbf{b}}^{(m)}\right)$ and $H^{-1}\left(\widehat{\mathbf{b}}^{(m)}\right)$ represent the gradient and Hessian matrix evaluated at $\widehat{\mathbf{b}}^{(m)}$, respectively. To obtain the gradient and Hessian matrix, we need to calculate the first and second derivatives of the log-likelihood function. The gradient is obtained by taking the first derivative from (2.3) with respect to $b_l$ as follows

$$\begin{aligned}\frac{\partial\ell\left(\mathbf{b}\right)}{\partial b_l} &= \sum_{i=1}^{n}\left\{y_i z_{il} - \left(\frac{e^{\mathbf{z}_i^\top\mathbf{b}}}{1+e^{\mathbf{z}_i^\top\mathbf{b}}}\right)z_{il}\right\}\\ &= \sum_{i=1}^{n}\left(y_i - p_i\right)z_{il},\end{aligned} \tag{2.5}$$

where

$$p_i = \frac{e^{\mathbf{z}_i^\top\mathbf{b}}}{1+e^{\mathbf{z}_i^\top\mathbf{b}}}. \tag{2.6}$$

We can express the gradient, which is a vector of partial derivatives as given in (2.5), in matrix form as

$$\nabla_{\mathbf{b}}\ell\left(\mathbf{b}\right) = \mathbf{Z}^\top\left(\mathbf{y}-\mathbf{p}\right). \tag{2.7}$$

We next need to obtain the Hessian matrix through the second derivative of (2.3) with respect to $b_l$ and $b_k$

$$\begin{aligned}\frac{\partial^2\ell\left(\mathbf{b}\right)}{\partial b_k\partial b_l} &= \sum_{i=1}^{n}\left\{-z_{ik}z_{il}\left(\frac{1}{1+e^{-\mathbf{z}_i^\top\mathbf{b}}}\right)^2 e^{-\mathbf{z}_i^\top\mathbf{b}}\right\}\\ &= -\sum_{i=1}^{n} z_{il}z_{ik}p_i\left(1-p_i\right)\end{aligned} \tag{2.8}$$

where $p_i$ is given by (2.6). The Hessian matrix is the inner product of the weighted

matrix and is calculated by

$$H\left(\ell\right)=-\mathbf{Z}^{\top}\mathbf{D}\mathbf{Z}, \tag{2.9}$$

where $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{ii}=p_i\left(1-p_i\right); i=1,\cdots,n$.

We can estimate the logistic regression coefficients by incorporating the gradient (2.7) and the Hessian matrix (2.9) into the NR equation (2.4) to get the following term

$$\widehat{\mathbf{b}}^{(m+1)}=\widehat{\mathbf{b}}^{(m)}+\left(\mathbf{Z}^{\top}\mathbf{D}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\top}\left(\mathbf{y}-\mathbf{p}\right). \tag{2.10}$$

### 2.1.2 IRWLS Method For Logistic Regression

One can reformulate the NR method (2.10) as an iteratively re-weighted least squares (IRWLS) algorithm [11]. This method enables us to take advantage of the weighted least squares structure in estimating the logistic regression coefficients.

We can re-write the gradient (2.7) and Hessian (2.9) in matrix form by

$$\frac{\partial\ell}{\partial\mathbf{b}}=\mathbf{Z}^{\top}\left(\mathbf{y}-g^{-1}\left(\mathbf{z};\mathbf{b}\right)\right), \tag{2.11}$$

and

$$\frac{\partial^2\ell}{\partial\mathbf{b}\partial\mathbf{b}^{\top}}=-\mathbf{Z}^{\top}\mathbf{W}\mathbf{Z}, \tag{2.12}$$

where $g^{-1}(\mathbf{z};\mathbf{b})$ is defined as $g^{-1}\left(\mathbf{z};\mathbf{b}\right)=\left[g^{-1}\left(\mathbf{z}_1;\mathbf{b}\right),...,g^{-1}\left(\mathbf{z}_n;\mathbf{b}\right)\right]^{\top}$ and $\mathbf{W}$ is a $(n\times n)$ diagonal matrix with entries

$$\mathbf{W}_{ii}=\frac{e^{-\mathbf{z_i}^{\top}\widehat{\mathbf{b}}}}{\left(1+e^{-\mathbf{z_i}^{\top}\widehat{\mathbf{b}}}\right)^2}, i=1,...,n. \tag{2.13}$$

Accordingly, the NR equation (2.4) can be updated as below

$$\begin{aligned}\widehat{\mathbf{b}}^{(m+1)}&=\widehat{\mathbf{b}}^{(m)}+\left(\mathbf{Z}^{\top}\mathbf{W}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\top}\left[\mathbf{y}-g^{-1}\left(\mathbf{z},\widehat{\mathbf{b}}^{(m)}\right)\right]\\&=\left(\mathbf{Z}^{\top}\mathbf{W}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\top}\mathbf{W}\left\{\mathbf{Z}\widehat{\mathbf{b}}^{(m)}+\mathbf{W}^{-1}\left[\mathbf{y}-g^{-1}\left(\mathbf{z},\widehat{\mathbf{b}}^{(m)}\right)\right]\right\}\\&=\left(\mathbf{Z}^{\top}\mathbf{W}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\top}\mathbf{W}\mathbf{Q}_l,\end{aligned} \tag{2.14}$$

where

$$\mathbf{Q}_l = \left\{ \mathbf{Z}\widehat{\mathbf{b}}^{(m)} + \mathbf{W}^{-1} \left[ \mathbf{y} - g^{-1}\left( \mathbf{z}, \widehat{\mathbf{b}}^{(m)} \right) \right] \right\}.$$

The above NR update represents the solution to the weighted least square problem that was originally introduced to estimate the coefficients of the logistic regression model. In each iteration, the updated response $\mathbf{Q}_l$ is regressed on the covariates, including $\mathbf{z}$.

Finally, one can use the Iterative re-weighted least squares (IRWLS) algorithm (2.14) to obtain the LS estimate, $\widehat{\mathbf{b}}_{LS}$, for the unknown coefficients of the logistic regression model.

## 2.2 Cox Proportional Hazard Regression

Cox Proportional Hazard (Cox PH) Regression is a statistical method widely used in survival analysis to analyze right censored data. It models the relationship between the time until an event, such as death or failure occurs, and some predictor variables. Cox Proportional hazard regression is widely used in many fields, such as medical research and epidemiology, to analyze survival data. This model enables us to study the effects of covariates on failure rate, and it can also work with censored data.

The Cox proportional hazard regression is written in terms of the hazard model formula as

$$\lambda(t|\mathbf{x}_i) = \lambda_0(t) \exp(\mathbf{x}_i^\top \beta), \tag{2.15}$$

where $t$ is the survival time, $\lambda_0(t)$ is the baseline hazard function, $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ is the vector of unknown coefficients corresponding to the covariates in the $(n \times p)$ design matrix $\mathbf{X} = (x_1, \ldots, x_p)$, and $e^{\mathbf{x}_i^\top \beta}$ is called the hazard ratio. The response variable of the model is the hazard function $\lambda(t)$, that is the probability that the event of interest occurred before time t, or a patient's "risk" of failure at time t [10]. The Cox model is semi-parametric when the baseline hazard function is not specified [60]. The commonly used method to estimate the $\beta$ coefficients is the maximum partial likelihood estimation, which is presented in the following.

### 2.2.1 Maximum Partial Likelihood Estimation

In the Cox PH model, we break down the likelihood function into two parts as

$$L(\beta, \lambda_0) = L_1(\beta) L_2(\beta, \lambda_0),$$

where $L_1(\beta)$ is the function of $\beta$ and $L_2(\beta, \lambda_0)$ is a function of $\beta$ and $\lambda_0$. Cox's idea was to focus on maximizing only $L_1(\beta)$ for better inference and ignore $L_2(\beta, \lambda_0)$ because it does not contain much valuable information about $\beta$ [10]. Therefore, we will refer to $L_1(\beta)$ as $L(\beta)$ henceforth while keeping in mind that $L(\beta)$ represents the partial likelihood. Thus, the partial likelihood of the Cox PH regression is

$$L(\beta) = \prod_{j=1}^{n} \left( \frac{\exp(\mathbf{x}_j^\top \beta)}{\sum_{k \in R(t_j)} \exp(\mathbf{x}_k^\top \beta)} \right)^{\delta_j},$$

where $t_j$ is the time of the $j$-th event, $R(t_j)$ is the set of individuals at risk just before time $t_j$, and $\delta_j$ is an independent censoring variable where it is 0 if the patient is censored and 1 otherwise. The latent variable $\delta_j$ in the expression makes sure that only the individuals who experience an event of interest contribute to the likelihood. The Cox partial log-likelihood is then written by

$$\ell(\beta) = \sum_{j=1}^{n} \delta_j \left[ \mathbf{x}_j^\top \beta - \log \left( \sum_{k \in R(t_j)} \exp(\mathbf{x}_k^\top \beta) \right) \right]. \tag{2.16}$$

The partial log-likelihood (2.16) can be simplified if we let $u_j = \exp(\mathbf{x}_j^\top \beta)$ and $U_j = \sum_{k \in R(t_j)} u_k$, indicating $u_j$ is the hazard ratio (i.e. probability of failure), and $U_j$ is the total hazard overall failure times. We can also identify $\pi_{ij} = Y_i(t_j) \frac{u_i}{U_j}$ as the probability of failure for the $i$-th individual. Thus, we can re-write the partial log-likelihood (2.16) by

$$\ell(\beta) = \sum_{j} \delta_j \log(u_j) - \sum_{j} \delta_j \log(U_j). \tag{2.17}$$

Similar to Section 2.1.1 for logistic regression, we apply the Newton-Raphson (NR)

technique for the Cox proportional hazard regression. The algorithm iteratively estimate $\beta$ as follows

$$\widehat{\beta}^{(m+1)} = \widehat{\beta}^{(m)} - H^{-1}\left(\widehat{\beta}^{(m)}\right).\nabla_{\beta}\ell\left(\widehat{\beta}^{(m)}\right), \tag{2.18}$$

where $\widehat{\beta}^{(m)}$ is the estimate updated from the $m$-th iteration, and $\nabla_{\beta}\ell\left(\widehat{\beta}^{(m)}\right)$ and $H^{-1}\left(\widehat{\beta}^{(m)}\right)$ represent the gradient and Hessian matrix evaluated at $\widehat{\beta}^{(m)}$, respectively. To obtain the gradient, we take the first derivative from (2.17) with respect to $\log(u_k)$ as follows

$$\frac{\partial \ell\left(\beta\right)}{\partial \log(u_k)} = \delta_k - \sum_j \pi_{kj}\delta_j,$$

where $\pi_{kj}$ represents an individual's relative risk (probability of failure) at time $j$. If the observations are ordered based on time, for an individual's likelihood calculation at a specific time point, only individuals at risk up to that time point contribute to the likelihood. Once an event occurs or an individual gets censored, they no longer contribute to the likelihood calculation for subsequent times. Thus, the result matrix of partial likelihood will be in a lower triangle structure, where all the elements above the diagonal are zero. So $\mathbf{P}$ will be a lower triangle matrix where $P_{ij} = \pi_{ij}$. Now we'll calculate the real gradient with respect to $\beta$

$$\begin{aligned}\frac{\partial \ell}{\partial \beta} &= \frac{\partial \log(u_k)}{\partial \beta} \cdot \frac{\partial \ell(\beta)}{\partial \log(u_k)} \\ &= \mathbf{X}^{\top}(\mathbf{c} - \mathbf{P}\mathbf{c}),\end{aligned} \tag{2.19}$$

where $\frac{\partial \log(u_k)}{\partial \beta} = \frac{\partial \mathbf{x}_k^{\top}\beta}{\partial \beta} = \mathbf{x}_k^{\top}$, and $\mathbf{c}$ is a vector of censoring elements ($\delta_j$'s).

In the next step, we need to calculate the Hessian matrix. To do so, based on the second derivative of (2.17), we first obtain

$$\frac{\partial^2 \ell\left(\beta\right)}{\partial \log(u_k)\partial \log(u_l)} = -\sum_j \delta_j \pi_{kj}\pi_{lj}, \tag{2.20}$$

$$\frac{\partial^2 \ell\left(\beta\right)}{\partial^2 \log(u_k)} = -\sum_j \delta_j \pi_{kj}(1 - \pi_{kj}). \tag{2.21}$$

Now we need to apply the chain rule to calculate the Hessian matrix with respect to $\beta$, so we first need to calculate the term below

$$\frac{\partial \log(u_k)\partial \log(u_l)}{\partial \beta_k \partial \beta_l^\top} = \frac{\partial(\mathbf{x}_k^\top \beta_k)\partial(\mathbf{x}_l \beta_l^\top)}{\partial \beta_k \partial \beta_l^\top} = \mathbf{x}_k^\top \mathbf{x}_l. \tag{2.22}$$

Hence, the Hessian matrix is equal to

$$H(\beta) = -\mathbf{X}^\top \mathbf{W} \mathbf{X}, \tag{2.23}$$

where $\mathbf{W}$ is calculated from (2.20) and (2.21). We can estimate the Cox PH regression coefficients by replacing by incorporating the gradient (2.19) and the Hessian matrix (2.23) into the NR equation (2.18) to get the following term

$$\widehat{\beta}^{(m+1)} = \widehat{\beta}^{(m)} + (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{c} - \mathbf{P}\mathbf{c}). \tag{2.24}$$

### 2.2.2 IRWLS Method For Cox PH Regression

Similar to subsection 2.1.2, we can re-formulated the NR algorithm (2.24) to an iteratively re-weighted least squares (IRWLS) algorithm in estimating the coefficients of the Cox PH regression. The NR equation (2.24) is re-formulated as

$$\begin{aligned}
\widehat{\beta}^{(m+1)} &= \widehat{\beta}^{(m)} + \left(\mathbf{X}^\top \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^\top (\mathbf{c} - \mathbf{P}\mathbf{c}) \\
&= \left(\mathbf{X}^\top \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W} \left\{ \mathbf{X}\widehat{\beta}^{(m)} + \mathbf{W}^{-1}(\mathbf{c} - \mathbf{P}\mathbf{c}) \right\} \\
&= \left(\mathbf{X}^\top \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{Q}_c,
\end{aligned} \tag{2.25}$$

where

$$\mathbf{Q}_c = \left\{ \mathbf{X}\widehat{\beta}^{(m)} + \mathbf{W}^{-1}(\mathbf{c} - \mathbf{P}\mathbf{c}) \right\}.$$

The above NR update represents the solution to the weighted least square problem for estimating the coefficients of the Cox PH model. In each iteration, the updated response $\mathbf{Q}_c$ is regressed on the covariates, including $\mathbf{X}$. Therefore, the LS estimator of the coefficients of the Cox PH is obtained by (2.25). Finally, one can use the Iterative re-weighted least squares (IRWLS) algorithm (2.25) to obtain the LS estimate, $\widehat{\beta}_{LS}$, for the unknown coefficients of the Cox PH regression model.

## 2.3 Cure Models

Cure models are popular tools in censored survival analysis for situations where the individuals will never experience the event of interest, no matter how long they are studied. These individuals are commonly called "cured subjects" or "long-term survivors" [45]. Cure models are widely used in clinical and medical studies. Analyzing survival data for individuals who may be cured brings on a challenge in dealing with censored data. Censoring occurs when an individual is either cured or uncured but hasn't been followed up long enough for the event of interest to happen. To analyze cured and uncured individuals simultaneously, it is recommended that a threshold time be set. Individuals are considered cured if their censored time exceeds this threshold and uncured if they have not yet reached the threshold time [10]. The most popular cure model is the mixture cure model, which consists of both the cured and uncured groups. We describe this model in the next section.

### 2.3.1 Mixture Cure Models

The mixture cure model is a type of survival analysis model that is used worldwide and has attracted a lot of attention in medical applications. It is important that some individuals may never experience the event of interest, and the mixture cure model takes this into account. As mentioned before, the mixture cure model studies the population as a mixture of two components, including cured and uncured individuals. We can detect the covariates associated with each component by studying them in two different models: incidence and latency models.

To build the mixture cure model, suppose that $U = 1$ if an individual is uncured and $U = 0$ otherwise. $T$ is defined as the failure time, $T(u)$ is the failure time of the uncured individuals, and $C$ is the censoring variable. Then, the incidence model and latency model can be defined as

$$\begin{aligned} &\text{Incidence model:} \quad P(U = 1|z) = \pi(z), \\ &\text{Latency model:} \quad P(T(u) > t|x, U = 1) = S_u(t|x), \end{aligned}$$

where $\mathbf{z} = (z_1, \ldots, z_q)$ and $\mathbf{x} = (x_1, \ldots, x_p)$ are the covariates of the incidence and latency models, respectively. Normally, there is an overlap between $\mathbf{x}$ and $\mathbf{z}$ to some

extent, and they might be nearly identical, with the only difference of an intercept term in $\mathbf{z}$. The proportional hazard mixture cure (PHMC) model assumes that the incidence model is defined by a logit link (2.1), and the latency model is specified by using the Cox proportional hazard model (2.15). The mixture cure model combines these components to provide the survival function $S(t|x,z)$ for the event $T > t$ as

$$P(T > t|x,z) = S(t|x,z) = 1 - \pi(z) + \pi(z)S_u(t|x), \tag{2.26}$$

where $S(t|x,z)$ is the survival function of all the individuals involved in the population, $\pi(z)$ is the probability of being uncured, and $S_u(t|x)$ is the survival function of uncured proportion [45, 46].

The specification of $S_u(t|x)$ can be parametric, semi-parametric, or non-parametric, which will create different mixture models [44]. The parametric approach proposes that both the incidence and latency parts are fully parametric models. A model for this approach is logistic regression, and the latency part is modelled by an exponential distribution [14]. The parametric models, however, show little flexibility because of the parametric assumptions in the incidence and latency distributions. The non-parametric approach implements both the incidence and latency components in a non-parametric approach, achieving the maximum flexibility of the models. A non-parametric method for estimating the cure rate and examining the effects of the covariates on the cure rate has been studied [43, 63]. The semi-parametric approach proposes enhanced flexibility in the aforementioned parametric method. The majority of these models concentrate on increasing the flexibility of the latency aspect while maintaining the logistic regression format for the incidence [2]. In this research, we aim to study the semi-parametric mixture cure model in Subsection 2.3.2, assuming the Cox proportional hazard regression to model the latency part of the mixture cure model.

### 2.3.2 Semi-parametric Mixture Cure Model

The specific type of semi-parametric mixture cure model may vary based on the assumptions regarding the latency part. In the latency part, if the proportional hazard (PH) assumption is used to describe the effect of $\mathbf{x}$ on the uncured individuals, and

$S_u(t|x)$ is defined by

$$S_u(t|x) = S_{u0}(t)^{\exp(x_j\beta)},$$

where $S_{u0}(t)$ is the baseline survival function that is not defined parametrically, we will be dealing with a semi-parametric proportional hazard latency model [46, 47]. The mixture cure model (2.26) is identifiable if $\exp(x_j\beta) \in \mathbb{R}$, meaning that the hazard ratio is a real number. Also, the $x$ covariates should lie within a finite number interval for identifiability [35].

To estimate the unknown parameters of the mixture cure model (2.26), we need to maximize the log-likelihood function by developing an EM algorithm. First, the incomplete log-likelihood function is obtained by

$$\ell(\theta) = \log \prod_{i=1}^{n} \left\{ [\pi(z_i) f_u(t_i|x_i)]^{\delta_i} \left[1 - \pi(z_i) + \pi(z_i) S_u(t_i|x_i)\right]^{1-\delta_i} \right\},$$

where the density function of $S_u(t|x)$ is represented by $f_u(t_i|x_i)$, $\delta_i$ is a censoring indicator with $\delta_i$=1 for uncensored observation and $\delta_i$=0 otherwise. The $\theta = (\mathbf{b}, \beta, S_{u0})$ is the vector of the unknown parameters in the mixture cure model, where $\mathbf{b}$ and $\beta$ are the unknown coefficients of the incidence and latency models, respectively [47].

To obtain the complete log-likelihood function, since the status of $\mathbf{y} = (y_1, ..., y_n)$ is unknown, we introduce the latent variable $y_i$, where $y_i$ represents the membership for individual $i$. Consequently, $y_i$ equals 1 when $\delta_i$=1, and it typically remains unknown when $\delta_i$=0. The ambiguity surrounding the status of $y_i$ arises because a subject undergoes censorship either when cured or when not cured but with the failure time T exceeding the censoring time C. Given $(y_1, ..., y_n)$, the complete likelihood function corresponding to the complete data $(t_i, \delta_i, x_i, z_i, y_i)$ for $i = 1, \ldots, n$ is given by

$$\begin{aligned} \mathcal{L}_c(\theta) &= \prod_{i=1}^{n} f_u(T_i \mid y_i = 1)^{I(y_i=1)} \cdot f_u(T_i \mid y_i = 0)^{I(y_i=0)} \\ &= \prod_{i=1}^{n} f_u(T_i \mid y_i = 1)^{y_i} \cdot f_u(T_i \mid y_i = 0)^{1-y_i}. \end{aligned} \tag{2.27}$$

We need to suppose two cases of $y_i = 0$, and $y_i = 1$ for calculating (2.27); First we consider $y_i = 0$ which means that we have a censored observations and $\delta_i = 0, T = \infty$,

thus

$$f_u(T_i \mid y_i = 0) = P(T = t_i|y_i = 0)^{\delta_i} + P(T > t_i|y_i = 0)^{1-\delta_i} = 1. \tag{2.28}$$

Then, for $y_i = 1$, we have

$$\begin{aligned} f_u(T_i \mid y_i = 1) &= P(T = t_i|y_i = 1)^{\delta_i} + P(T > t_i|y_i = 1)^{1-\delta_i} \\ &= f_u(t_i \mid x_i)^{\delta_i} \cdot S_u(t_i \mid x_i)^{1-\delta_i}. \end{aligned} \tag{2.29}$$

From (2.27)-(2.29), the complete log-likelihood function is given by

$$\begin{aligned} \ell_c(\theta) &= \log \prod_{i=1}^{n} \left[\pi(z_i)^{y_i} \cdot [1 - \pi(z_i)]^{1-y_i}\right] \\ &\quad + \log \prod_{i=1}^{n} \left[f_u(t_i \mid x_i)^{y_i\delta_i} S_u(t_i \mid x_i)^{y_i(1-\delta_i)}\right] \\ &= l_1(b \mid \mathbf{y}) + l_2(\beta, S_{u0} \mid \mathbf{y}), \end{aligned} \tag{2.30}$$

where $\theta = (\mathbf{b}, \beta, S_{u0})$ represents the collection of all the unknown parameters, $f_u$ is the probability density function, and $S_u$ is the survival function.

The expectation–maximization (EM) algorithm is an iterative approach for obtaining the maximum likelihood estimates when dealing with latent data. The EM algorithm starts with an initialization step, where initial model parameters are chosen. It then proceeds with two main steps: Expectation (E) and Maximization (M). In the E-step, the algorithm imputes missing data based on the current parameter estimates. The M-step updates the parameters by maximizing the expected log-likelihood computed using the imputed data. This process alternates between the E-step and M-step until convergence. Let $\theta^{(r)} = (\mathbf{b}^{(r)}, \beta^{(r)}, S_{u0}^{(r)})$ represent the estimate for the $(r)$-th iteration of the EM algorithm, and let $w_i^{(r+1)}$ denote the conditional expectation of $y_i$ from the $(r+1)$-th iteration. Thus, $w_i^{(r+1)}$ is computed by

$$\begin{aligned} w_i^{(r+1)} &= E(y_i \mid \mathbf{b}^{(r)}, \beta^{(r)}, S_{u0}^{(r)}) \\ &= \delta_i + (1 - \delta_i) \cdot \frac{\pi^{(r)}(z_i) S_u^{(r)}(t_i \mid x_i)}{1 - \pi^{(r)}(z_i) + \pi^{(r)}(z_i) S_u^{(r)}(t_i \mid x_i)}. \end{aligned} \tag{2.31}$$

In the M-step, the EM algorithm maximizes the conditional expectation of the

complete log-likelihood $l_1(b \mid w_i^{(r)})$ and $l_2(\beta, S_{u0} \mid w_i^{(r)})$ as defined in (2.30) in the $(r+1)$-th iteration to update $\theta^{(r+1)} = (\mathbf{b}^{(r+1)}, \beta^{(r+1)}, S_{u0}^{(r+1)})$. The EM algorithm iterates until it reaches convergence, at which point the final LS estimates of $\theta$ are obtained [47].

According to Subsection 2.1.1, $\mathbf{b}^{(r+1)}$ is updated via the logistic regression log-likelihood function (2.3) from the introduced IRWLS method in (2.14). From Subsection 2.2.1, $\beta^{(r+1)}$ is updated via Cox's partial log-likelihood function (2.16) obtained by implementing the IRWLS (2.25) iteratively in the M-step of the EM algorithm. Finally, the baseline survival function, $S_{u0}^{(r+1)}$, is updated through a Nelson–Aalen cumulative hazard estimator [47]

$$\hat{S}_{u0}^{(r+1)}(t) = \exp\left( - \sum_{j:\tau_j<t} \frac{d_j}{\sum_{i \in R_j} \exp\left( \log w_i^{(r)} + \hat{\beta'}^{(r+1)} x_i \right)} \right), \tag{2.32}$$

where $\tau_j$ is the upper time limit, $R_j$ is the risk set at time $\tau_j$, and $d_j$ is the number of observations experiencing the event of interest at time $\tau_j$ [2, 6]. The Nelson–Aalen estimator accounts for the likelihood contribution of a patient who experiences the event at $\tau_j$ and from a censored patient at time $t$, leading to a baseline distribution with probability mass focused on the uncensored observations [44].

# Chapter 3

# Shrinkage Estimators for Mixture Cure Models

In regression analysis, shrinkage methods are used to address multicollinearity, reducing variance and improving model reliability [21]. Ridge and Liu-type (LT) shrinkage estimators are proposed by researchers to offer solutions by introducing regularization penalties to combat collinearity [37, 51]. In this chapter, we investigate how ridge and LT shrinkage methods help in overcoming the collinearity problem in the logistic regression method, discussed in Subsection 3.1, and Cox proportional hazard regression method, discussed in Subsection 3.2.

## 3.1 Shrinkage Estimators for Logistic Regression

Multicollinearity among explanatory variables in logistic regression is known to increase the variances of the maximum likelihood estimator. While the iteratively reweighted least squares (IRWLS) method explained in Subsection 2.1.2 is commonly used to estimate logistic regression parameters, it faces significant issues when its covariates are linearly dependent. This problem can be addressed by the ridge estimation approach explained in Subsection 3.1.1. However, because the ridge estimator is not fully capable of addressing the ill-conditioning problem, the LT estimator is introduced later in Subsection 3.1.2 to enhance the performance of estimation [26].

### 3.1.1 Ridge Estimator for Logistic Regression

The ridge estimate is obtained by maximizing the ridge penalized log-likelihood function of logistic regression models. By augmenting the equation $0 = \lambda^{\frac{1}{2}}\mathbf{b} + \xi$, to the logistic regression equation, we can obtain the ridge penalized log-likelihood function [37, 51]

$$\ell_R(\mathbf{b}) = \ell(\mathbf{b}) - \frac{1}{2}\lambda\mathbf{b}^\top\mathbf{b}, \tag{3.1}$$

where $\ell(\mathbf{b})$ is the logistic log-likelihood function (2.3), and $\lambda$ is the tuning ridge parameter. Many researchers have introduced methods to estimate $\lambda$, such as Lawless and Wang [30] and Khalaf and Shukur [31]. Cross-validation is also a method to estimate the tuning parameter $\lambda$ [39]. However, following [26, 37], we have chosen to use the commonly used technique to estimate the ridge parameter proposed by Hoerl and Kennard [24]. This method calculates $\lambda$ as $\lambda = (1+p)/\widehat{\mathbf{b}}_{LS}^\top\widehat{\mathbf{b}}_{LS}$, where $p$ is the length of the unknown vector of coefficients, and $\widehat{\mathbf{b}}_{LS}$ is derived from (2.14) using the IRWLS method. Therefore, $\widehat{\mathbf{b}}_{LS}$ is considered optimal for constructing the tuning ridge parameter since it roughly minimizes the Weighted Sum of Squared Errors [51].

Similar to the LS estimation in Subsection 2.1.2, we apply the Newton-Raphson method to iteratively estimate the ridge logistic regression parameters as

$$\mathbf{b}^{(m+1)} = \mathbf{b}^{(m)} - H^{-1}(\ell_R\left(\mathbf{b}^{(m)}\right)).\nabla_{\mathbf{b}}\ell_R\left(\mathbf{b}^{(m)}\right). \tag{3.2}$$

For estimating the gradient $\nabla_{\mathbf{b}}\ell_R\left(\mathbf{b}^{(m)}\right)$ and the Hessian matrix $H^{-1}(\ell_R\left(\mathbf{b}^{(m)}\right))$ estimated at $\mathbf{b}^{(m)}$ in (3.2), we calculate the first and second derivatives of the ridge penalized log-likelihood function (3.1). The gradient with respect to $b_l$ is given by

$$\frac{\partial\ell_R\left(\mathbf{b}\right)}{\partial b_l} = \frac{\partial\ell\left(\mathbf{b}\right)}{\partial b_l} - \lambda\mathbf{b}. \tag{3.3}$$

According to (2.11), we can write the gradient (3.3) in a matrix form as

$$\nabla_{\mathbf{b}}\ell_R\left(\mathbf{b}\right) = \mathbf{Z}^\top\left(\mathbf{y} - g^{-1}\left(\mathbf{z},\mathbf{b}\right)\right) - \lambda\mathbf{b}. \tag{3.4}$$

Similarly, we can obtain the second derivative of ridge penalized log-likelihood function

(3.1) by

$$\frac{\partial^2 \ell_R(\mathbf{b})}{\partial b_k \partial b_l} = \frac{\partial^2 \ell(\mathbf{b})}{\partial b_k \partial b_l} - \lambda I_p.$$

Thus, according to (2.12), the Hessian matrix is obtained by

$$H\left(\ell_R(\mathbf{b})\right) = -\mathbf{Z}^\top \mathbf{W}\mathbf{Z} - \lambda I_p = -\mathbf{V}, \tag{3.5}$$

where $I_p$ is the identity matrix with size $(p \times p)$, and $\mathbf{V} = \mathbf{Z}^\top \mathbf{W}\mathbf{Z} + \lambda I_p$. Now similar to what we did to obtain (2.14), with (3.4) and (3.5) in hand, we can re-formulate the NR equation (3.2) as an IRWLS algorithm

$$\begin{aligned}
\widehat{\mathbf{b}}_R^{new} &= \widehat{\mathbf{b}}_R^{old} + \mathbf{V}^{-1}\left\{\mathbf{Z}^\top\left[\mathbf{y} - g^{-1}\left(\mathbf{z}, \widehat{\mathbf{b}}_R^{old}\right)\right] - \lambda \widehat{\mathbf{b}}_R^{old}\right\} \\
&= \mathbf{V}^{-1}\mathbf{V}\,\widehat{\mathbf{b}}_R^{old} - \mathbf{V}^{-1}\lambda \widehat{\mathbf{b}}_R^{old} + \mathbf{V}^{-1}\mathbf{Z}^\top \mathbf{W}\mathbf{W}^{-1}\left\{\left(\mathbf{y} - g^{-1}\left(\mathbf{z}, \widehat{\mathbf{b}}_R^{old}\right)\right)\right\} \\
&= \mathbf{V}^{-1}\mathbf{Z}^\top \mathbf{W}\left\{\mathbf{Z}\,\widehat{\mathbf{b}}_R^{old} + \mathbf{W}^{-1}\left[\mathbf{y} - g^{-1}\left(\mathbf{z}, \widehat{\mathbf{b}}_R^{old}\right)\right]\right\} \\
&= \left(\mathbf{Z}^\top \mathbf{W}\mathbf{Z} + \lambda \mathbf{I}_p\right)^{-1}\mathbf{Z}^\top \mathbf{W}\mathbf{Q}_{\mathrm{l}},
\end{aligned} \tag{3.6}$$

where $\mathbf{Q}_l = \left\{\mathbf{Z}\widehat{\mathbf{b}}_R^{old} + \mathbf{W}^{-1}\left[\mathbf{y} - g^{-1}\left(\mathbf{z}, \widehat{\mathbf{b}}_R^{old}\right)\right]\right\}$. This method is used to update $\widehat{\mathbf{b}}_R^{new}$ until convergence is reached, and we get the ridge logistic $\widehat{\mathbf{b}}_R$ estimate.

By taking a close look at (2.14), we can write the ridge logistic estimator (3.6) as

$$\widehat{\mathbf{b}}_R = \left(\mathbf{Z}^\top \mathbf{W}\mathbf{Z} + \lambda \mathbf{I}_p\right)^{-1}\mathbf{Z}^\top \mathbf{W}\mathbf{Z}\widehat{\mathbf{b}}_{LS}, \tag{3.7}$$

where $\widehat{\mathbf{b}}_{LS}$ is the LS estimation of $\mathbf{b}$ (2.14).

In instances of severe multicollinearity, a larger regularization parameter $\lambda$ is chosen to reduce multicollinearity. Nevertheless, the drawback is that a large $\lambda$ brings about increased bias in the ridge logistic estimator, preventing it from completely resolving the issues associated with the ill-conditioning problem. That's where the alternative LT shrinkage method is introduced [26, 24].

### 3.1.2 Liu-type Estimator for Logistic Regression

Liu-type (LT) estimator is a method offered in the case of high multicollinearity, and Liu proved that the LT estimator outperforms ridge regression in terms of mean squared error (MSE) [37]. As mentioned earlier, ridge estimation uses a small shrinkage parameter $\lambda$ for multicollinearity, but when the issue is severe, a small $\lambda$ is inadequate. On the flip side, a large $\lambda$ introduces noticeable biases to the estimates. The LT two-parameter estimator was suggested as a substitute for the ridge regression estimator. If the left part of the ridge penalty $0 = \lambda^{\frac{1}{2}}\mathbf{b} + \xi$, is replaced by $\left(-d_l/\lambda^{1/2}\right)\widehat{\mathbf{b}}$, where $\widehat{\mathbf{b}}$ could be any estimate of $\mathbf{b}$, we'll end up with the LT penalty

$$\left(d_l/\lambda^{1/2}\right)\widehat{\mathbf{b}} = \lambda^{1/2}\mathbf{b} + \xi'. \tag{3.8}$$

It is proved that the parameter $\lambda$ is employed solely to manage the conditioning of $\mathbf{Z}^\top\mathbf{W}\mathbf{Z} + \lambda I_p$. Once the condition number reaches the desired level, any bias introduced by $\lambda$ can be rectified using a second parameter, known as the bias correction parameter, denoted as $d_l$ [26, 37].

If we extend this new penalty (3.8) to the logistic regression log-likelihood equation (2.3), we'll obtain

$$\ell_{LT}(\mathbf{b}) = \ell(\mathbf{b}) - \frac{1}{2}\xi^\top\xi, \tag{3.9}$$

where $\ell(\mathbf{b})$ is the logistic log-likelihood function, and $\xi = \left(d_l/\lambda^{1/2}\right)\widehat{\mathbf{b}} - \lambda^{1/2}\mathbf{b}$ is the LT penalty. Afterwards, the Newton-Raphson (NR) technique needs to be applied to get the desired LT estimates as

$$\mathbf{b}^{(m+1)} = \mathbf{b}^{(m)} - H^{-1}(\ell_{LT}\left(\mathbf{b}^{(m)}\right)).\nabla_{\mathbf{b}}\ell_{LT}\left(\mathbf{b}^{(m)}\right). \tag{3.10}$$

We need to calculate the gradient and Hessian matrix to replace in the above equation; therefore, the first derivative of (3.9) is obtained by

$$\frac{\partial\ell_{LT}\left(\mathbf{b}\right)}{\partial b_l} = \frac{\partial\ell\left(\mathbf{b}\right)}{\partial b_l} - d_l\widehat{\mathbf{b}} - \lambda\mathbf{b}.$$

According to (2.11), this equation can be written in a matrix form by

$$\nabla_{\mathbf{b}}\ell_{LT}\left(\mathbf{b}\right) = \mathbf{Z}^\top\left[\mathbf{y} - g^{-1}\left(\mathbf{z},\mathbf{b}\right)\right] - d_l\widehat{\mathbf{b}} - \lambda\mathbf{b}. \tag{3.11}$$

In the next step, the second derivative is calculated as

$$\frac{\partial^2 \ell_{LT}(\mathbf{b})}{\partial b_k \partial b_l} = \frac{\partial \ell(\mathbf{b})}{\partial b_l} - \lambda I_p,$$

From (2.12), we can write this in matrix form as

$$H(\ell_{LT}(\mathbf{b})) = -\mathbf{Z}^\top \mathbf{W}\mathbf{Z} - \lambda I_p. \tag{3.12}$$

Afterwards, with the replacement of the calculated gradient and Hessian matrices in the NR equation (3.10), the new NR equation will be

$$\mathbf{b}^{(m+1)} = \mathbf{b}^{(m)} + (\mathbf{Z}^\top \mathbf{W}\mathbf{Z} + \lambda \mathbf{I_p})^{-1} \cdot \left[\mathbf{Z}^\top \left(\mathbf{y} - g^{-1}\left(\mathbf{z}, \mathbf{b}^{(m)}\right)\right) - \mathbf{d_l}\widehat{\mathbf{b}} - \lambda \mathbf{b}^{(m)}\right]$$

where $\widehat{\mathbf{b}}$ can represent any of the $\widehat{\mathbf{b}}_{LS}$ or $\widehat{\mathbf{b}}_R$ estimators. Let $\mathbf{V} = \mathbf{Z}^\top \mathbf{W}\mathbf{Z} + \lambda I_p$. Subsequently, by applying the IRWLS method, a new LT estimator can be calculated as

$$\begin{aligned}
\widehat{\mathbf{b}}_{\mathrm{LT}}^{\mathrm{new}} &= \widehat{\mathbf{b}}_{\mathrm{LT}}^{\mathrm{old}} + \mathbf{V}^{-1}\left\{\mathbf{Z}^\top \left[\mathbf{y} - g^{-1}\left(\mathbf{z}, \widehat{\mathbf{b}}_{\mathrm{LT}}^{\mathrm{old}}\right)\right] - \mathbf{d_l}\widehat{\mathbf{b}} - \lambda \widehat{\mathbf{b}}_{\mathrm{LT}}^{\mathrm{old}}\right\} \\
&= \mathbf{V}^{-1}\mathbf{V}\,\widehat{\mathbf{b}}_{\mathrm{LT}}^{\mathrm{old}} + \mathbf{V}^{-1}\mathbf{Z}^\top \mathbf{W}\mathbf{W}^{-1}\left\{\mathbf{y} - g^{-1}\left(\mathbf{z}, \widehat{\mathbf{b}}_{\mathrm{LT}}^{\mathrm{old}}\right)\right\} - \mathbf{V}^{-1}\mathbf{d_l}\widehat{\mathbf{b}} - \mathbf{V}^{-1}\lambda \widehat{\mathbf{b}}_{\mathrm{LT}}^{\mathrm{old}} \\
&= \mathbf{V}^{-1}\mathbf{Z}^\top \mathbf{W}\left\{\mathbf{Z}\,\widehat{\mathbf{b}}_{\mathrm{LT}}^{\mathrm{old}} + \mathbf{W}^{-1}\left[\mathbf{y} - g^{-1}\left(\mathbf{z}, \widehat{\mathbf{b}}_{\mathrm{LT}}^{\mathrm{old}}\right)\right]\right\} - \mathbf{V}^{-1}\mathbf{d_l}\widehat{\mathbf{b}} \\
&= \left(\mathbf{Z}^\top \mathbf{W}\mathbf{Z} + \lambda \mathbf{I}_p\right)^{-1}\left(\mathbf{Z}^\top \mathbf{W}\mathbf{Q}_{\mathrm{l}} - \mathbf{d_l}\widehat{\mathbf{b}}\right),
\end{aligned} \tag{3.13}$$

where $\mathbf{Q}_l = \left\{\mathbf{Z}\widehat{\mathbf{b}}_{LT}^{old} + \mathbf{W}^{-1}\left[\mathbf{y} - g^{-1}\left(\mathbf{z}, \widehat{\mathbf{b}}_{LT}^{old}\right)\right]\right\}$.

If we consider $\widehat{\mathbf{b}}$ representing the LS estimator $\widehat{\mathbf{b}}_{LS}$, the LT estimator is given by

$$\widehat{\mathbf{b}}_{LT} = \left(\mathbf{Z}^\top \mathbf{W}\mathbf{Z} + \lambda \mathbf{I}_p\right)^{-1}\left(\mathbf{Z}^\top \mathbf{W}\mathbf{Z} - d_l \mathbf{I}_p\right)\widehat{\mathbf{b}}_{\mathbf{LS}}, \tag{3.14}$$

where $0 < \lambda$, $-\infty < d_l < \infty$.

So, as mentioned, the LT estimator, with its two tuning parameters, effectively handles severe multicollinearity. It has two advantages over the ridge method: the ability to handle large $\lambda$ and a smaller MSE. The MSE of the LT estimator from

(3.14) is obtained by

$$\begin{aligned}\mathrm{MSE}\left(\widehat{\mathbf{b}}_{LT}\right) &= tr[\mathrm{Var}(\widehat{\mathbf{b}}_{LT})] + \left\|\mathrm{E}(\widehat{\mathbf{b}}_{LT} - \mathbf{b})\right\|^2 \\ &= tr\left[\left(\mathbf{Z}^T\mathbf{W}\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-1}\left(\mathbf{Z}^T\mathbf{W}\mathbf{Z} - d_l\mathbf{I}_p\right)\left(\mathbf{Z}^T\mathbf{W}\mathbf{Z}\right)^{-1}\right. \\ &\quad \left.\left(\mathbf{Z}^T\mathbf{W}\mathbf{Z} - d_l\mathbf{I}_p\right)\left(\mathbf{Z}^T\mathbf{W}\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-1}\right] \\ &\quad + \left\|\left(\mathbf{Z}^\top\mathbf{W}\mathbf{Z} + \lambda\mathbf{I}_p\right)^{-1}\left(\mathbf{Z}^\top\mathbf{W}\mathbf{Z} - d_l\mathbf{I}_p\right)\left(\mathbf{Z}^\top\mathbf{W}\mathbf{Z}\right)^{-1}\right. \\ &\quad \left.\mathbf{Z}^\top\mathbf{W}p\left(\mathbf{z}_i, \mathbf{b}\right) - \mathbf{b}\right\|^2.\end{aligned} \tag{3.15}$$

The norm $\|.\|^2$ is used here to express the importance of the bias, ensuring that both positive and negative differences between the estimated and true coefficients are considered equally, thereby providing a comprehensive measure of the overall estimation error. Determining the optimal $d_l$ value to minimize the MSE in (3.15) is done through the iterated estimation of $d_l$ [25]. Thus, using the recommended $d_l$, the LT estimator always has an MSE that is the same or smaller than the ridge estimator's MSE. At last, the presented LT estimator fully addresses the ill-conditioning problem [26, 37].

## 3.2 Shrinkage Estimators for Cox Proportional Hazard Regression

In survival analysis, the Cox proportional hazard regression model is fundamental, offering crucial insights into how covariates influence the hazard function. Commonly, the maximum partial likelihood (MPL), followed by the iteratively re-weighted least squares (IRWLS) method explained in Subsection 2.2.2, is used to estimate the Cox proportional hazard model parameters. Yet, although effective, the IRWLS estimation method encounters challenges such as unstable estimates when there is multicollinearity among covariates. Shrinkage estimators present a solution to these challenges by introducing a penalty term into the estimation process.

In this section, we thoroughly discuss the application of shrinkage estimators in the context of Cox proportional hazard regression. The ridge estimator explained in Subsection 3.2.1 is a widely used shrinkage technique that is adapted to Cox PH models to counter multicollinearity issues. Unlike LS estimators, ridge estimators offer improved

stability, making them advantageous in survival analysis. However, high collinearity remains a challenge for complete resolution. LT estimator explained in Subsection 3.2.2 is introduced as an alternative method, particularly useful in the presence of severe multicollinearity. LT estimators surpass ridge estimators by exhibiting a smaller MSE and effectively addressing the ill-conditioning problem. By examining ridge and LT estimators in subsequent sections, we aim to provide a good understanding of their roles and applications in Cox proportional hazard regression [27].

### 3.2.1 Ridge Estimator for Cox PH Model

To obtain the Cox PH ridge estimate, we ought to maximize the ridge penalized partial log-likelihood function of the Cox regression model. By augmenting the ridge penalty $0 = k^{\frac{1}{2}}\beta + \epsilon$, to the Cox partial log-likelihood equation, the ridge penalized log-likelihood function can be obtained

$$\ell_R(\beta) = \ell(\beta) - \frac{1}{2}k \cdot \beta^\top \beta, \tag{3.16}$$

where $\ell(\beta)$ is the Cox partial log-likelihood function (2.16), and $k$ is the ridge parameter. Selecting the value of $k$ is challenging because it manages the bias of the regression towards the dependent variable's mean [16]. Many methods for estimating $k$ are discussed in the literature [27, 39, 55]. We estimate the ridge parameter as $k = 1/\widehat{\beta}_{LS}^\top \widehat{\beta}_{LS}$ following Hoerl and Kennard's proposal [24], where $\widehat{\beta}_{LS}$ is obtained from (2.25) using the IRWLS method. Thus, $\widehat{\beta}_{LS}$ is considered optimal for determining the ridge tuning parameter as it can minimize the Weighted Sum of Squared Errors [51].

To calculate the ridge Cox estimators $\beta_R$, we need to apply the Newton-Raphson (NR) technique similar to the LS estimation in (2.18)

$$\beta^{(m+1)} = \beta^{(m)} - H^{-1}(\ell_R\left(\beta^{(m)}\right))\nabla_\beta \ell_R\left(\beta^{(m)}\right). \tag{3.17}$$

In the above equation, the gradient $\nabla_\beta \ell_R\left(\beta^{(m)}\right)$ and the Hessian matrix $H^{-1}(\ell_R\left(\beta^{(m)}\right))$ are estimated by calculating the first and second derivatives of the ridge penalized

log-likelihood function (3.16). The gradient with respect to $\beta_l$ is equal to

$$\frac{\partial \ell_R(\beta)}{\partial \beta_l} = \frac{\partial \ell(\beta)}{\partial \beta_l} - k\beta. \tag{3.18}$$

According to (2.19), we can find the matrix form of (3.18) by

$$\nabla_\beta \ell_R(\beta) = \mathbf{X}^\top(\mathbf{c} - \mathbf{Pc}) - k\beta. \tag{3.19}$$

The second derivative of ridge penalized partial log-likelihood function (3.16) is obtained by

$$\frac{\partial^2 \ell_R(\beta)}{\partial \beta_k \partial \beta_l} = \frac{\partial^2 \ell(\beta)}{\partial \beta_k \partial \beta_l} - kI_p.$$

So in line with (2.23), the Hessian matrix is written as

$$H\left(\ell_R(\beta)\right) = -\mathbf{X}^\top\mathbf{WX} - kI_p = -\mathbf{V}, \tag{3.20}$$

where $I_p$ is the identity matrix with size $(p \times p)$, and $\mathbf{V} = \mathbf{X}^\top\mathbf{WX} + kI_p$. Now by replacing (3.19) and (3.20) terms into the NR equation (3.17) we'll obtain the following term

$$\beta_R^{(m+1)} = \widehat{\beta}_R^{(m)} + \mathbf{V}^{-1}\left\{\mathbf{X}^\top(\mathbf{c} - \mathbf{Pc}) - k\widehat{\beta}_R^{(m)}\right\}. \tag{3.21}$$

We can reformulate (3.21) to an IRWLS algorithm by

$$\begin{aligned}
\widehat{\beta}_R^{new} &= \widehat{\beta}_R^{old} + \mathbf{V}^{-1}\left\{\mathbf{X}^\top(\mathbf{c} - \mathbf{Pc}) - k\widehat{\beta}_R^{old}\right\} \\
&= \mathbf{V}^{-1}\mathbf{V}\,\widehat{\beta}_R^{old} - k\mathbf{V}^{-1}\widehat{\beta}_R^{old} + \mathbf{V}^{-1}\mathbf{X}^\top\mathbf{WW}^{-1}\{(\mathbf{c} - \mathbf{Pc})\} \\
&= \mathbf{V}^{-1}\mathbf{X}^\top\mathbf{W}\left\{\mathbf{X}\,\widehat{\beta}_R^{old} + \mathbf{W}^{-1}(\mathbf{c} - \mathbf{Pc})\right\} \\
&= \left(\mathbf{X}^\top\mathbf{WX} + k\mathbf{I}_p\right)^{-1}\mathbf{X}^\top\mathbf{WQ}_c,
\end{aligned} \tag{3.22}$$

where $\mathbf{Q}_c = \left\{\mathbf{X}\widehat{\beta}^{old} + \mathbf{W}^{-1}(\mathbf{c} - \mathbf{Pc})\right\}$. From (2.25), we can write the ridge estimator for the Cox PH Model as

$$\widehat{\beta}_R = \left(\mathbf{X}^\top\mathbf{WX} + k\mathbf{I}_p\right)^{-1}\mathbf{X}^\top\mathbf{WX}\widehat{\beta}_{LS}, \tag{3.23}$$

where $\widehat{\beta}_{LS} = \left(\mathbf{X}^\top\mathbf{WX}\right)^{-1}\mathbf{X}^\top\mathbf{WQ}_c$.

As it can be seen, ridge regression addresses the issue of collinearity by augmenting a small constant $k$ to the diagonal of $\mathbf{X}^\top\mathbf{W}\mathbf{X}$ to improve the condition number that is defined as

$$\kappa = \left(\frac{\lambda_{max}}{\lambda_{min}}\right)^{1/2},$$

where $\lambda_{max}$ and $\lambda_{min}$ are maximum and minimum eigenvalues of $\mathbf{X}^\top\mathbf{W}\mathbf{X}$, respectively.

In general, a condition number below 10 signifies no significant collinearity issues. Condition numbers between 30 and 100 indicate moderate to strong collinearity, while a value exceeding 100 suggests severe collinearity [3]. In practical applications, the shrinkage parameter $k$ in ridge regression is typically small. Notably, the condition number of $\mathbf{V}$ is a decreasing function of $k$ because the ridge parameter $k$ affects the condition number by increasing the minimum eigenvalue of the matrix, thus reducing the condition number. Thus, to maintain a low condition number for $\mathbf{V}$, we should choose a larger $k$. When $\mathbf{X}^\top\mathbf{W}\mathbf{X}$ is highly ill-conditioned, the chosen small $k$ may not be sufficient to address the ill-conditioning problem. If the condition number of $\mathbf{V}$ indicates strong collinearity, it implies that the underlying issue persists, leaving ridge regression unstable. However, introducing a larger $k$ brings more bias to the ridge regression. Due to these challenges, the LT penalty is suggested in the next section, providing a solution to overcome the issues associated with severe collinearity in ridge Cox regression [27].

### 3.2.2 Liu-Type Estimator for Cox PH Model

The Liu-type (LT) estimator was initially introduced as a remedy for the mentioned bias generated by the shrinkage parameter in the ridge estimator. Similar to subsection 3.1.2, the LT penalty gets augmented into the Cox partial log-likelihood equation, as shown below

$$\ell_{LT}(\beta) = \ell(\beta) - \frac{1}{2}\epsilon^\top\epsilon, \tag{3.24}$$

where $\ell(\beta)$ is the Cox partial log-likelihood function (2.16), and $\epsilon = \left(d_c/k^{1/2}\right)\widehat{\beta} - k^{1/2}\beta$ is the LT penalty. Similar to the LS and ridge estimations, for obtaining the LT Cox estimators, the Newton-Raphson method is applied as

$$\beta^{(m+1)} = \beta^{(m)} - H^{-1}(\ell_{LT}\left(\beta^{(m)}\right)).\nabla_\beta\ell_{LT}\left(\beta^{(m)}\right), \tag{3.25}$$

For obtaining the gradient, we'll take the derivative of (3.24) with respect to $\beta_l$

$$\frac{\partial \ell_{LT}(\beta)}{\partial \beta_l} = \frac{\partial \ell(\beta)}{\partial \beta_l} - d_c\widehat{\beta} - k\beta, \tag{3.26}$$

Using (2.19), we can write the gradient (3.26) in a matrix form by

$$\nabla_\beta \ell_{LT}(\beta) = \mathbf{X}^\top(\mathbf{c} - \mathbf{Pc}) - d_c\widehat{\beta} - k\beta. \tag{3.27}$$

This time, we'll take the second derivative of (3.24) for obtaining the Hessian matrix

$$\frac{\partial^2 \ell_{Liu}(\beta)}{\partial \beta_k \partial \beta_l} = \frac{\partial^2 \ell(\beta)}{\partial \beta_k \partial \beta_l} - kI_p. \tag{3.28}$$

Thus, from (2.23), the Hessian matrix of LT Cox regression is given by

$$H(\ell_{LT}(\beta)) = -\mathbf{X}^\top\mathbf{W}\mathbf{X} - kI_p. \tag{3.29}$$

We have now both the gradient (3.27) and the Hessian matrix (3.29); therefore, the estimates of the LT Cox proportional hazard regression coefficients can be obtained by replacing these terms into (3.25) to get the following term

$$\beta^{(m+1)} = \beta^{(m)} + (\mathbf{X}^\top\mathbf{W}\mathbf{X} + kI_p)^{-1} \cdot [\mathbf{X}^\top(\mathbf{c} - \mathbf{Pc}) - d_c\widehat{\beta} - k\beta],$$

where $\widehat{\beta}$ can represent any of the $\widehat{\beta}_{LS}$ or $\widehat{\beta}_R$ estimators. Assuming $\mathbf{V} = \mathbf{X}^\top\mathbf{W}\mathbf{X} + kI_p$, this term can be reformulated as an IRWLS algorithm

$$\begin{aligned}
\widehat{\beta}_{LT}^{new} &= \widehat{\beta}_{LT}^{old} + \mathbf{V}^{-1}\left\{\mathbf{X}^\top(\mathbf{c} - \mathbf{Pc}) - d_c\widehat{\beta} - k\widehat{\beta}_{LT}^{old}\right\} \\
&= \mathbf{V}^{-1}\mathbf{V}\,\widehat{\beta}_{LT}^{old} + \mathbf{V}^{-1}\mathbf{X}^\top\mathbf{W}\mathbf{W}^{-1}\{(\mathbf{c} - \mathbf{Pc})\} - \mathbf{V}^{-1}d_c\widehat{\beta} - \mathbf{V}^{-1}k\widehat{\beta}_{LT}^{old} \\
&= \mathbf{V}^{-1}\mathbf{X}^\top\mathbf{W}\left\{\mathbf{X}\,\widehat{\beta}_{LT}^{old} + \mathbf{W}^{-1}(\mathbf{c} - \mathbf{Pc})\right\} - \mathbf{V}^{-1}d_c\widehat{\beta} \\
&= \left(\mathbf{X}^\top\mathbf{W}\mathbf{X} + k\mathbf{I}_p\right)^{-1}(\mathbf{X}^\top\mathbf{W}\mathbf{Q}_c - d_c\widehat{\beta}),
\end{aligned} \tag{3.30}$$

where $\mathbf{Q}_c = \left\{\mathbf{X}\widehat{\beta}^{old} + \mathbf{W}^{-1}(\mathbf{c} - \mathbf{Pc})\right\}$. As mentioned, $\widehat{\beta}$ can represent any of the $\widehat{\beta}_{LS}$ or $\widehat{\beta}_R$ estimators, so based on (2.25), and (3.23), we can obtain

$$\widehat{\beta}_{LT} = \left(\mathbf{X}^\top\mathbf{W}\mathbf{X} + kI_p\right)^{-1}(\mathbf{X}^\top\mathbf{W}\mathbf{X} - d_cI)\widehat{\beta}_{LS}, \tag{3.31}$$

$$\widehat{\beta}_{LT} = \left(\mathbf{X}^{\top}\mathbf{W}\mathbf{X} + kI_p\right)^{-1}\left(\mathbf{X}^{\top}\mathbf{W}\mathbf{X} + kI_p - d_c I\right)\widehat{\beta}_R. \tag{3.32}$$

The proposed LT estimator can handle severe multicollinearity now. One of the reasons for its superiority over the ridge method is having a smaller MSE. If we consider that $\widehat{\beta} = \widehat{\beta}_R$, $k$ can be considered as $k = (1+p)/\widehat{\beta}_{LS}^{\top}\widehat{\beta}_{LS}$, following the estimation of the bias correction parameters $d_c$ by maximizing the mean square errors of $\widehat{\beta}_{LT}$ [26]. It is simple to show that the MSE of the LT estimator for the Cox PH regression model is equal to

$$\begin{aligned}
\text{MSE}\left(\widehat{\beta}_{LT}\right) &= tr[\text{Var}(\widehat{\beta}_{LT})] + \left\|\text{E}(\widehat{\beta}_{LT} - \beta)\right\|^2 \\
&= tr\Big[\left(\mathbf{X}^T\mathbf{W}\mathbf{X} + k\mathbf{I}_p\right)^{-1}\left(\mathbf{X}^T\mathbf{W}\mathbf{X} + k\mathbf{I}_p - d_c\mathbf{I}_p\right) \\
&\quad \left(\mathbf{X}^T\mathbf{W}\mathbf{X} + k\mathbf{I}_p\right)^{-1}\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{W}\mathbf{X} + k\mathbf{I}_p\right)^{-1} \\
&\quad \left(\mathbf{X}^T\mathbf{W}\mathbf{X} + k\mathbf{I}_p - d_c\mathbf{I}_p\right)\left(\mathbf{X}^T\mathbf{W}\mathbf{X} + k\mathbf{I}_p\right)^{-1}\Big] \\
&\quad + \Big\|\left(\mathbf{X}^{\top}\mathbf{W}\mathbf{X} + k\mathbf{I}_p\right)^{-1}\left(\mathbf{X}^{\top}\mathbf{W}\mathbf{X} + k\mathbf{I}_p - d_c\mathbf{I}_p\right) \\
&\quad \left(\mathbf{X}^{\top}\mathbf{W}\mathbf{X} + k\mathbf{I}_p\right)^{-1}\mathbf{X}^{\top}\mathbf{W}\text{E}\left(\mathbf{Q}_c\right) - \beta\Big\|^2
\end{aligned} \tag{3.33}$$

where $\text{E}(\mathbf{Q}_c)$ is replaced by $\widehat{\mathbf{Q}}_c$. The calculated MSE of the LT estimator (3.33) is always the same or smaller than the ridge estimator's MSE because the bias correction parameter $d_c$ is optimized to minimize the MSE, so it naturally adjusts the bias and variance to ensure that the total MSE is always the same or smaller than that of the ridge estimator, which proves the LT estimator can address the ill-conditioning problem to a good extent [27].

## 3.3 Shrinkage Estimators for Semi-parametric Mixture Cure Models

As mentioned in chapter 2, the semi-parametric mixture cure model is defined as (2.26), where the incidence model is defined by a logit link (2.1), and the latency model is specified by using the Cox proportional hazard model (2.15). The estimates of the parameters in the semi-parametric mixture cure models are developed through an EM algorithm that maximizes their likelihood function as mentioned in Subsection 2.3.2.

However, this usual way of LS method can run into problems, like giving unreliable results when there's multicollinearity among the factors being considered. To address this issue, we propose shrinkage ridge and LT methods to estimate the coefficients of the underlying model. To do so, we developed new EM algorithms to incorporate the shrinkage methods for both components.

We have discussed the ridge and LT shrinkage estimators for logistic regression (3.1) and Cox PH regression (3.2) in previous sections. In this section, we will study the application of these shrinkage estimators in the semi-parametric mixture cure model that consists of both the logistic and Cox PH regression. The ridge estimator is first incorporated into the mixture cure model to improve our model estimates to a large extent. Then, the LT estimator is proposed to enhance the effectiveness of ridge regression even further.

### 3.3.1 Ridge Estimator for Semi-parametric Mixture Cure Models

While the EM algorithm method is commonly used for estimating parameters in mixture cure models, it becomes significantly impacted by multicollinearity when the covariates are linearly dependent. The ridge penalized EM estimation can be offered as a solution to this problem. As mentioned in Subsection 2.3.2, the complete log-likelihood semi-parametric mixture cure model (2.30) is composed of the logistic and Cox PH regression models; thus, adding the ridge penalty to each of these terms will result in the ridge penalized complete log-likelihood function:

$$\begin{aligned}\ell_R(\theta) = l_1(\mathbf{b} \mid \mathbf{y}) - \frac{1}{2}\lambda \cdot \mathbf{b}^\top \mathbf{b} \\ + l_2(\beta, S_{u0} \mid \mathbf{y}) - \frac{1}{2} k \cdot \beta^\top \beta,\end{aligned}$$

where $\ell_c(\theta) = l_1(\mathbf{b} \mid \mathbf{y}) + l_2(\beta, S_{u0} \mid \mathbf{y})$ is the complete log-likelihood semi-parametric mixture cure model (2.30), $\theta$ represents the unknown $(\mathbf{b}, \beta, S_{u0})$ parameters, and $\lambda, k$ are the ridge parameters calculated as: $\lambda = (1+p)/\widehat{\mathbf{b}}_{LS}^\top \widehat{\mathbf{b}}_{LS}$ and $k = 1/\widehat{\beta}_{LS}^\top \widehat{\beta}_{LS}$ [27]. The other new method to estimate the ridge parameters is using the cross-validation method. In that method, the performance of the tuning parameters can be evaluated by their prediction error, and the $\lambda, k$ parameters get chosen that minimize the $\theta$

estimates [39].

Expanding the EM algorithm with ridge regularization enhances the model. This variation, called the ridge EM algorithm, adds the ridge penalty to the original EM framework. Here, the iterative procedure remains unchanged, and the algorithm alternates between imputing missing data in the E-step and updating parameters in the M-step. However, in the ridge EM approach, each step is modified to include the ridge penalty term in the parameter estimation process. During the E-step, missing data is filled in using current parameter estimates, taking into account the influence of the ridge penalty. Then, in the M-step, parameters are adjusted to maximize the expected log-likelihood while considering the regularization effect of the ridge penalty. This integration of ridge regularization within the EM framework improves parameter estimation by encouraging more reliable solutions and reducing the impact of multicollinearity. This, in turn, enhances the model's ability to predict accurately and maintain stability.

Accordingly, the ridge E-step in ridge EM for the semi-parametric mixture of cure model is given by

$$\begin{aligned}\tilde{w}_i^{(r+1)} &= E(y_i \mid \mathbf{b}^{(r)}, \beta^{(r)}, S_{u0}^{(r)}) \\ &= \delta_i + (1-\delta_i) \cdot \frac{\pi^{(r)}(z_i) S_u^{(r)}(t_i \mid x_i)}{1 - \pi^{(r)}(z_i) + \pi^{(r)}(z_i) S_u^{(r)}(t_i \mid x_i)},\end{aligned} \tag{3.34}$$

where $\tilde{w}_i^{(r+1)}$ represents the updated conditional expectation of $y_i$ incorporating ridge regularization. The subsequent M-step involves maximizing the expected log-likelihood while considering the ridge penalty term to update the parameters $\mathbf{b}^{(r+1)}$ based on ridge logistic estimator (3.7), and $\beta^{(r+1)}$ based on ridge estimator for the Cox PH (3.23), and $S_{u0}^{(r+1)}$ based on the Nelson–Aalen cumulative hazard estimator (2.32). We will then end up having the updated $\widehat{\theta}^{(r+1)} = (\widehat{\mathbf{b}}^{(r+1)}, \widehat{\beta}^{(r+1)}, \widehat{S_{u0}}^{(r+1)})$. Iterations between the E-step and M-step continue until the change in parameter estimates between iterations falls below our $1 \times 10^{-7}$ threshold, indicating when convergence is achieved, providing final estimates for the $\theta$ parameters.

The final estimators obtained through the ridge EM algorithm generally outperform those from the original method without ridge regularization. This improvement results in smoother solutions. Additionally, by employing the LT shrinkage estimator instead of the ridge estimator, we can effectively address the ill-conditioning problem,

further enhancing the model's performance. Therefore, recognizing the potential for further improvement, we proceed to enhance the model in the subsequent section by incorporating the LT shrinkage estimator.

### 3.3.2 Liu-type Estimator for Semi-parametric Mixture Cure Models

The LT shrinkage estimator offers a better approach to regularization, allowing for precise control over parameter shrinkage. This capability enables the model to adapt more flexibly to the data, effectively resolving issues related to multicollinearity and improving the stability of parameter estimation.

We introduce the LT EM algorithm that incorporates the LT penalty into the core EM framework. Similar to the ridge EM algorithm, this process involves augmenting the original function with the LT penalty term. By incorporating the LT penalty to both the logistic and Cox PH regression parts of the complete log-likelihood semi-parametric mixture cure model, the LT penalized complete log-likelihood function is obtained by

$$\begin{aligned}\ell_{LT}(\theta) = l_1(\mathbf{b} \mid \mathbf{y}) - \frac{1}{2}\xi'\xi \\ + l_2(\beta, S_{u0} \mid \mathbf{y}) - \frac{1}{2}\epsilon'\epsilon,\end{aligned}$$

where $\ell_c(\theta) = l_1(\mathbf{b} \mid \mathbf{y}) + l_2(\beta, S_{u0} \mid \mathbf{y})$ is the complete log-likelihood semi-parametric mixture cure model (2.30), $\theta$ represents the unknown $(\mathbf{b}, \beta, S_{u0})$ parameters, $\xi = \left(d_l/\lambda^{1/2}\right)\widehat{\mathbf{b}} - \lambda^{1/2}\mathbf{b}$ is the LT penalty for logistic regression, $\epsilon = \left(d_c/k^{1/2}\right)\widehat{\beta} - k^{1/2}\beta$ is the LT penalty for Cox PH regression, $\lambda, k$ are the tuning ridge parameters and $d_l, d_c$ are the bias correction parameters calculated with the help of logistic MSE (3.15), and Cox PH MSE (3.33).

The LT E-step is done similarly to the ridge E-step (3.34); however, the M-step is different in how the unknown parameters are updated. The $\mathbf{b}^{(r+1)}$ parameter is updated based on LT logistic estimator (3.14), $\beta^{(r+1)}$ is updated based on LT estimator for the Cox PH (3.31), and $S_{u0}^{(r+1)}$ based on the Nelson–Aalen cumulative hazard estimator (2.32).

The LT penalty is favoured here due to its tendency to outperform the ridge

penalty. The LT penalty offers advantages such as improved model stability and better handling of multicollinearity. Thus, by incorporating the LT penalty into the EM algorithm, we anticipate achieving superior optimization results compared to the ridge counterpart.

In the upcoming chapter, we will put our developed LS, ridge, and LT methods to the test. Through simulated and real-life data examples, we aim to demonstrate the effectiveness of these methods and clarify their comparative strengths in providing accurate and reliable estimates for the parameters of the semi-parametric mixture cure model.

# Chapter 4

# Numerical Studies

In this chapter, we present the results of our numerical studies on semi-parametric mixture cure models in the presence of multicollinearity, comparing the three different estimation methods, including Least Squares (LS), Ridge, and Liu-type (LT) method discussed in Subsections 2.3.2, 3.3.1, and 3.3.2, respectively.

This chapter is organized as follows. In Section 4.1, we provide various simulation studies to evaluate the performance of the three methods under various scenarios. Subsequently, these methods are implemented on a Breast Cancer dataset in Section 4.2 to showcase their effectiveness in real-world medical settings. Analyzing real data enables us to assess the methods' performance on actual datasets and provide insights into their practical applicability.

## 4.1 Simulation Studies

This section compares the performance of the LS, Ridge, and LT methods in estimating the parameters of semi-parametric mixture cure models when faced with multicollinearity. We examine how variations in sample size and levels of multicollinearity within the mixture of cure models impact the performance of the proposed estimation techniques.

We begin by generating the design matrix with four covariates $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_4)$. In this study, we follow [26] to incorporate multicollinearity into the design matrix through two parameters $\phi$ and $\rho$ that represent the levels of correlation between the

first and last two predictors in the mixture model. Initially, we generated random numbers $\{w_{ij}, i = 1, \ldots, n; j = 1, \ldots, 5\}$ from the standard normal distribution and then simulated the correlated covariates as

$$x_{i,j_1} = (1 - \phi^2)w_{i,j_1} + \phi w_{i,5}, \quad j_1 = 1, 2,$$

$$x_{i,j_2} = (1 - \rho^2)w_{i,j_2} + \rho w_{i,5}, \quad j_2 = 3, 4,$$

where we choose $\phi = \{0.85, 0.9, 0.95\}$ and $\rho = \{0.85, 0.9, 0.95\}$ to represent three sets of different multicollinearities in the mixture of cure models. Next, we generated the response variables from the mixture of cure model in a similar way to [22], where survival data is generated based on cure and censoring rates. Accordingly, responses are provided for the mixture cure model's desired incidence and latency parts. The incidence part is derived from a logistic regression function, and the patient's cure status is determined as a Bernoulli random variable with a failure probability given by the logistic model. Survival time-to-event responses (referred to as Time) are obtained from a Weibull-Cox PH model for those not cured. Following [22], we set the Weibull shape parameter to be 1.45 and the Weibull scale parameter to be 0.25. We designated the censoring indicator (referred to as Event) to follow an exponential distribution with a rate of 0.16. We set the true values of the model parameters as $\beta_0 = (0.4, 0.2, -0.5, 0.2)$ for the Cox regression part, and $b_0 = (1, 0.4, 0.2, -0.5, 0.2)$ for the logistic regression part. The true parameter values are kept constant across all methods (LS, Ridge, and LT) to ensure that data from the same population are simulated throughout the entire simulation. We replicated the above data collection process 1000 times and then applied the LS, Ridge, and LT estimation methods as described in Chapters 2 and 3 in each replicate. We selected the sample size $n = \{20, 40, 80\}$; we opted for smaller sample sizes relative to the number of covariates, as collinearity is known to have a greater impact on parameter estimation in such cases [27].

To investigate the estimation performance of $(\widehat{\beta}, \widehat{\mathbf{b}})$, we computed the square root of the mean of squared errors ($\sqrt{\text{MSE}}$) formulas given by

$$\sqrt{\text{MSE}(\widehat{\beta})} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{\beta} - \beta_0)^\top(\widehat{\beta} - \beta_0)}\ , \tag{4.1}$$

and

$$\sqrt{\text{MSE}(\widehat{\mathbf{b}})} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathbf{b}} - \mathbf{b}_0)^\top(\widehat{\mathbf{b}} - \mathbf{b}_0)}\ . \tag{4.2}$$

Then, we computed the 2.5%, 50%, and 97.5% percentiles of the $\sqrt{\text{MSE}}$ and demonstrated the median (M), lower (L), and upper (U) bounds of the estimation intervals to show detailed results.

One of the goals of the mixture of cure models is to predict whether each individual will end up cured or uncured. Using logistic regression, we utilize the incidence part of the mixture cure model to predict whether patients will be cured or uncured. First, using training datasets of sizes $n = \{20, 40, 80\}$, we used vectors of all zeros for $b$ and $\beta$ values to initialize the LS, Ridge, and LT EM algorithms. These algorithms iteratively updated the parameters until the difference between successive estimates was less than $1 \times 10^{-7}$, thus reaching the final estimates. We then created a validation test set of 50 samples using the mixture cure model, independent of the training data. Finally, we used the estimated parameters from the training dataset to predict the binary response for the validation dataset through the incidence part. We need the prediction criteria, including Error, Sensitivity, and Specificity, to evaluate the classification performance of the methods. We computed

$$\text{Error} = \frac{FP + FN}{TP + TN + FP + FN},\ \text{Sensitivity} = \frac{TP}{TP + FN},\ \text{Specificity} = \frac{TN}{TN + FP},$$

where FP, FN, TP, and TN denote false positives, false negatives, true positives, and true negatives in the confusion matrix, respectively. To study the effect of sample size and multicollinearity, we repeated this process 1000 times for different sample sizes and levels of multicollinearity using LS, ridge, and LT methods and computed median and 95% confidence intervals for our evaluation measures.

Tables 4.1-4.9 show the simulation study results. The presented tables examine the impact of collinearity and sample size on the performance of different estimation methods. One notable finding is the noticeable effect of collinearity on smaller sample sizes, where results indicate superior performance with a sample size of 20 compared to larger sample sizes. This suggests that as the sample size decreases, the influence of collinearity becomes more notable, affecting estimation methods' accuracy. As the sample size increases, more population information is incorporated into the estimation

methods so that even the LS method can estimate the parameters adequately.

The ridge and LT methods, designed to address multicollinearity, show better performance with a lower square root of MSE and better overall results. These shrinkage methods introduce a bias into the estimation process to improve the estimation of model coefficients in the presence of multicollinearity. Multicollinearity significantly affects the LS estimates, making the coefficient estimates unreliable. In contrast to LS estimates, the ridge and LT estimates offer more accurate insights into the model's coefficients as they have significantly lower $\sqrt{\text{MSE}}$. Among the shrinkage methods, the LT method shows better reliability than the ridge method in dealing with multicollinearity. The Error, Specificity, and Sensitivity classification performances across the three methods are close; however, the ridge and LT classification rates are, on average, slightly better than the LS method. While a significant reduction in $\sqrt{\text{MSE}}$ is observed in ridge and LT estimates in almost all cases, no significant difference is observed in the classification performance between the shrinkage methods and the benchmark LS estimation method. It is worth noting that our findings align with [26, 21], who indicated that multicollinearity severely impacts the estimation accuracy of the methods, while prediction performance remains mainly unchanged.

Table 4.1: The median (M) and 95% CIs for the $\sqrt{\text{MSE}}$, Error, Sensitivity (Sen) and Specificity (Spe) of the LS, Ridge, and LT methods in estimation and prediction of the mixture of cure models when $\phi = 0.85$ and $n = 20$.

| | | | $\sqrt{\text{MSE}}$ | | | Error | | | Sen | | | Spe | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\rho$ | EM | $\boldsymbol{\Psi}$ | M | L | U | M | L | U | M | L | U | M | L | U |
| 0.85 | LS | $\beta$ | 3.20 | 0.86 | 14.05 | | | | | | | | | |
| | | $b$ | 9.74 | 1.80 | 37.35 | 0.56 | 0.36 | 0.66 | 0.42 | 0.31 | 0.64 | 0.50 | 0.36 | 0.64 |
| | Ridge | $\beta$ | 2.25 | 0.51 | 9.58 | | | | | | | | | |
| | | $b$ | 2.89 | 0.80 | 14.14 | 0.46 | 0.34 | 0.60 | 0.56 | 0.39 | 0.69 | 0.50 | 0.29 | 0.71 |
| | LT | $\beta$ | 1.38 | 0.41 | 8.90 | | | | | | | | | |
| | | $b$ | 1.40 | 0.56 | 7.31 | 0.46 | 0.34 | 0.62 | 0.53 | 0.39 | 0.69 | 0.50 | 0.29 | 0.79 |
| 0.9 | LS | $\beta$ | 3.86 | 1.01 | 16.13 | | | | | | | | | |
| | | $b$ | 10.92 | 1.80 | 39.68 | 0.56 | 0.36 | 0.66 | 0.42 | 0.33 | 0.64 | 0.50 | 0.36 | 0.64 |
| | Ridge | $\beta$ | 2.69 | 0.59 | 12.10 | | | | | | | | | |
| | | $b$ | 2.95 | 0.81 | 13.56 | 0.46 | 0.34 | 0.60 | 0.56 | 0.39 | 0.67 | 0.50 | 0.29 | 0.71 |
| | LT | $\beta$ | 1.67 | 0.49 | 12.18 | | | | | | | | | |
| | | $b$ | 1.33 | 0.60 | 6.58 | 0.48 | 0.34 | 0.62 | 0.53 | 0.36 | 0.69 | 0.50 | 0.29 | 0.79 |
| 0.95 | LS | $\beta$ | 5.41 | 1.09 | 26.10 | | | | | | | | | |
| | | $b$ | 11.36 | 1.80 | 43.34 | 0.56 | 0.38 | 0.64 | 0.42 | 0.33 | 0.64 | 0.50 | 0.36 | 0.64 |
| | Ridge | $\beta$ | 3.42 | 0.64 | 19.71 | | | | | | | | | |
| | | $b$ | 3.05 | 0.81 | 16.77 | 0.46 | 0.34 | 0.60 | 0.56 | 0.39 | 0.69 | 0.50 | 0.29 | 0.71 |
| | LT | $\beta$ | 2.02 | 0.56 | 18.06 | | | | | | | | | |
| | | $b$ | 1.35 | 0.58 | 8.02 | 0.46 | 0.34 | 0.62 | 0.53 | 0.36 | 0.69 | 0.50 | 0.29 | 0.71 |

Table 4.2: The median (M) and 95% CIs for the $\sqrt{\text{MSE}}$, Error, Sensitivity (Sen) and Specificity (Spe) of the LS, Ridge, and LT methods in estimation and prediction of the mixture of cure models when $\phi = 0.9$ and $n = 20$.

| | | | $\sqrt{\text{MSE}}$ | | | Error | | | Sen | | | Spe | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\rho$ | EM | $\boldsymbol{\Psi}$ | M | L | U | M | L | U | M | L | U | M | L | U |
| 0.85 | LS | $\beta$ | 3.75 | 1.03 | 18.44 | | | | | | | | | |
| | | $b$ | 12.04 | 1.80 | 53.73 | 0.56 | 0.38 | 0.66 | 0.42 | 0.31 | 0.64 | 0.50 | 0.36 | 0.64 |
| | Ridge | $\beta$ | 2.67 | 0.53 | 11.48 | | | | | | | | | |
| | | $b$ | 3.33 | 0.82 | 19.79 | 0.46 | 0.34 | 0.60 | 0.56 | 0.39 | 0.69 | 0.50 | 0.29 | 0.71 |
| | LT | $\beta$ | 1.61 | 0.43 | 11.81 | | | | | | | | | |
| | | $b$ | 1.42 | 0.57 | 10.45 | 0.46 | 0.34 | 0.60 | 0.53 | 0.39 | 0.69 | 0.50 | 0.29 | 0.71 |
| 0.9 | LS | $\beta$ | 4.83 | 1.17 | 17.89 | | | | | | | | | |
| | | $b$ | 12.55 | 1.80 | 52.20 | 0.56 | 0.36 | 0.66 | 0.42 | 0.33 | 0.64 | 0.50 | 0.36 | 0.64 |
| | Ridge | $\beta$ | 3.51 | 0.59 | 12.96 | | | | | | | | | |
| | | $b$ | 3.47 | 0.79 | 18.79 | 0.46 | 0.34 | 0.60 | 0.56 | 0.39 | 0.67 | 0.50 | 0.29 | 0.79 |
| | LT | $\beta$ | 2.11 | 0.49 | 12.07 | | | | | | | | | |
| | | $b$ | 1.42 | 0.62 | 8.44 | 0.46 | 0.34 | 0.60 | 0.53 | 0.39 | 0.69 | 0.50 | 0.29 | 0.79 |
| 0.95 | LS | $\beta$ | 6.54 | 1.57 | 27.49 | | | | | | | | | |
| | | $b$ | 14.16 | 1.80 | 56.94 | 0.56 | 0.38 | 0.64 | 0.42 | 0.33 | 0.64 | 0.50 | 0.36 | 0.64 |
| | Ridge | $\beta$ | 4.34 | 0.75 | 20.90 | | | | | | | | | |
| | | $b$ | 3.75 | 0.85 | 19.30 | 0.46 | 0.34 | 0.60 | 0.56 | 0.39 | 0.69 | 0.50 | 0.29 | 0.79 |
| | LT | $\beta$ | 2.66 | 0.61 | 21.55 | | | | | | | | | |
| | | $b$ | 1.44 | 0.60 | 9.69 | 0.46 | 0.34 | 0.60 | 0.56 | 0.39 | 0.69 | 0.50 | 0.29 | 0.71 |

Table 4.3: The median (M) and 95% CIs for the $\sqrt{\text{MSE}}$, Error, Sensitivity (Sen) and Specificity (Spe) of the LS, Ridge, and LT methods in estimation and prediction of the mixture of cure models when $\phi = 0.95$ and $n = 20$.

| | | | $\sqrt{\text{MSE}}$ | | | Error | | | Sen | | | Spe | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\rho$ | EM | $\boldsymbol{\Psi}$ | M | L | U | M | L | U | M | L | U | M | L | U |
| 0.85 | LS | $\beta$ | 5.56 | 1.43 | 31.29 | | | | | | | | | |
| | | $b$ | 15.15 | 1.80 | 68.03 | 0.56 | 0.38 | 0.66 | 0.42 | 0.31 | 0.64 | 0.50 | 0.36 | 0.64 |
| | Ridge | $\beta$ | 3.52 | 0.58 | 20.13 | | | | | | | | | |
| | | $b$ | 3.95 | 0.76 | 23.95 | 0.46 | 0.34 | 0.60 | 0.56 | 0.39 | 0.67 | 0.50 | 0.29 | 0.79 |
| | LT | $\beta$ | 2.29 | 0.43 | 20.04 | | | | | | | | | |
| | | $b$ | 1.55 | 0.62 | 10.87 | 0.46 | 0.34 | 0.58 | 0.56 | 0.39 | 0.67 | 0.50 | 0.29 | 0.79 |
| 0.9 | LS | $\beta$ | 6.64 | 1.61 | 26.73 | | | | | | | | | |
| | | $b$ | 16.98 | 1.80 | 87.22 | 0.56 | 0.38 | 0.66 | 0.42 | 0.33 | 0.64 | 0.50 | 0.36 | 0.64 |
| | Ridge | $\beta$ | 4.40 | 0.71 | 20.41 | | | | | | | | | |
| | | $b$ | 4.28 | 0.71 | 27.83 | 0.46 | 0.34 | 0.60 | 0.56 | 0.39 | 0.67 | 0.57 | 0.36 | 0.71 |
| | LT | $\beta$ | 2.64 | 0.53 | 19.64 | | | | | | | | | |
| | | $b$ | 1.49 | 0.62 | 13.32 | 0.46 | 0.34 | 0.60 | 0.54 | 0.39 | 0.67 | 0.50 | 0.29 | 0.79 |
| 0.95 | LS | $\beta$ | 9.27 | 1.92 | 37.54 | | | | | | | | | |
| | | $b$ | 20.82 | 1.80 | 89.28 | 0.56 | 0.36 | 0.64 | 0.44 | 0.33 | 0.64 | 0.50 | 0.36 | 0.71 |
| | Ridge | $\beta$ | 6.42 | 0.77 | 28.22 | | | | | | | | | |
| | | $b$ | 5.09 | 0.79 | 28.32 | 0.46 | 0.34 | 0.60 | 0.56 | 0.39 | 0.67 | 0.50 | 0.29 | 0.71 |
| | LT | $\beta$ | 3.71 | 0.54 | 27.59 | | | | | | | | | |
| | | $b$ | 1.60 | 0.60 | 12.96 | 0.46 | 0.36 | 0.60 | 0.56 | 0.39 | 0.67 | 0.50 | 0.29 | 0.71 |

Table 4.4: The median (M) and 95% CIs for the $\sqrt{\text{MSE}}$, Error, Sensitivity (Sen) and Specificity (Spe) of the LS, Ridge, and LT methods in estimation and prediction of the mixture of cure models when $\phi = 0.85$ and $n = 40$.

| | | | $\sqrt{\text{MSE}}$ | | | Error | | | Sen | | | Spe | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\rho$ | EM | $\boldsymbol{\Psi}$ | M | L | U | M | L | U | M | L | U | M | L | U |
| 0.85 | LS | $\beta$ | 1.65 | 0.51 | 4.09 | | | | | | | | | |
| | | $b$ | 7.86 | 1.80 | 14.87 | 0.56 | 0.38 | 0.66 | 0.42 | 0.31 | 0.64 | 0.50 | 0.36 | 0.64 |
| | Ridge | $\beta$ | 1.25 | 0.33 | 3.40 | | | | | | | | | |
| | | $b$ | 2.20 | 0.67 | 6.16 | 0.46 | 0.34 | 0.58 | 0.56 | 0.39 | 0.69 | 0.54 | 0.29 | 0.72 |
| | LT | $\beta$ | 0.81 | 0.31 | 2.90 | | | | | | | | | |
| | | $b$ | 1.25 | 0.54 | 4.11 | 0.46 | 0.34 | 0.60 | 0.53 | 0.39 | 0.67 | 0.50 | 0.29 | 0.79 |
| 0.9 | LS | $\beta$ | 1.98 | 0.55 | 4.79 | | | | | | | | | |
| | | $b$ | 8.17 | 1.80 | 17.88 | 0.56 | 0.40 | 0.64 | 0.42 | 0.33 | 0.61 | 0.50 | 0.36 | 0.57 |
| | Ridge | $\beta$ | 1.42 | 0.38 | 4.29 | | | | | | | | | |
| | | $b$ | 2.21 | 0.72 | 6.53 | 0.46 | 0.34 | 0.60 | 0.56 | 0.39 | 0.69 | 0.50 | 0.29 | 0.71 |
| | LT | $\beta$ | 0.88 | 0.36 | 3.79 | | | | | | | | | |
| | | $b$ | 1.22 | 0.53 | 4.86 | 0.46 | 0.34 | 0.60 | 0.56 | 0.39 | 0.69 | 0.50 | 0.29 | 0.79 |
| 0.95 | LS | $\beta$ | 2.90 | 0.78 | 8.18 | | | | | | | | | |
| | | $b$ | 8.75 | 1.80 | 19.38 | 0.56 | 0.42 | 0.60 | 0.42 | 0.36 | 0.61 | 0.50 | 0.43 | 0.57 |
| | Ridge | $\beta$ | 1.77 | 0.52 | 6.93 | | | | | | | | | |
| | | $b$ | 2.25 | 0.73 | 6.68 | 0.46 | 0.34 | 0.58 | 0.56 | 0.39 | 0.69 | 0.50 | 0.29 | 0.79 |
| | LT | $\beta$ | 1.13 | 0.50 | 6.52 | | | | | | | | | |
| | | $b$ | 1.22 | 0.56 | 4.12 | 0.46 | 0.34 | 0.60 | 0.53 | 0.39 | 0.67 | 0.50 | 0.29 | 0.72 |

Table 4.5: The median (M) and 95% CIs for the $\sqrt{\text{MSE}}$, Error, Sensitivity (Sen) and Specificity (Spe) of the LS, Ridge, and LT methods in estimation and prediction of the mixture of cure models when $\phi = 0.9$ and $n = 40$.

| | | | $\sqrt{\text{MSE}}$ | | | Error | | | Sen | | | Spe | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\rho$ | EM | $\boldsymbol{\Psi}$ | M | L | U | M | L | U | M | L | U | M | L | U |
| 0.85 | LS | $\beta$ | 1.97 | 0.55 | 5.00 | | | | | | | | | |
| | | $b$ | 8.55 | 1.80 | 19.84 | 0.56 | 0.38 | 0.66 | 0.42 | 0.31 | 0.64 | 0.50 | 0.36 | 0.64 |
| | Ridge | $\beta$ | 1.39 | 0.35 | 4.31 | | | | | | | | | |
| | | $b$ | 2.32 | 0.71 | 7.41 | 0.46 | 0.34 | 0.58 | 0.53 | 0.39 | 0.67 | 0.50 | 0.29 | 0.71 |
| | LT | $\beta$ | 0.82 | 0.34 | 3.63 | | | | | | | | | |
| | | $b$ | 1.23 | 0.56 | 5.13 | 0.46 | 0.34 | 0.60 | 0.53 | 0.39 | 0.67 | 0.50 | 0.29 | 0.79 |
| 0.9 | LS | $\beta$ | 2.25 | 0.67 | 5.85 | | | | | | | | | |
| | | $b$ | 9.34 | 1.80 | 20.48 | 0.56 | 0.38 | 0.62 | 0.42 | 0.36 | 0.61 | 0.50 | 0.43 | 0.64 |
| | Ridge | $\beta$ | 1.60 | 0.41 | 4.90 | | | | | | | | | |
| | | $b$ | 2.50 | 0.76 | 7.10 | 0.46 | 0.34 | 0.60 | 0.56 | 0.39 | 0.69 | 0.50 | 0.36 | 0.71 |
| | LT | $\beta$ | 0.90 | 0.39 | 4.44 | | | | | | | | | |
| | | $b$ | 1.25 | 0.55 | 4.56 | 0.46 | 0.34 | 0.60 | 0.53 | 0.39 | 0.67 | 0.50 | 0.29 | 0.71 |
| 0.95 | LS | $\beta$ | 3.35 | 0.90 | 9.23 | | | | | | | | | |
| | | $b$ | 9.98 | 1.80 | 22.28 | 0.56 | 0.40 | 0.64 | 0.42 | 0.31 | 0.61 | 0.50 | 0.36 | 0.64 |
| | Ridge | $\beta$ | 2.25 | 0.54 | 7.57 | | | | | | | | | |
| | | $b$ | 2.43 | 0.78 | 7.54 | 0.46 | 0.34 | 0.58 | 0.53 | 0.39 | 0.67 | 0.50 | 0.29 | 0.71 |
| | LT | $\beta$ | 1.28 | 0.51 | 6.96 | | | | | | | | | |
| | | $b$ | 1.25 | 0.58 | 4.27 | 0.46 | 0.34 | 0.58 | 0.53 | 0.39 | 0.69 | 0.57 | 0.36 | 0.72 |

Table 4.6: The median (M) and 95% CIs for the $\sqrt{\text{MSE}}$, Error, Sensitivity (Sen) and Specificity (Spe) of the LS, Ridge, and LT methods in estimation and prediction of the mixture of cure models when $\phi = 0.95$ and $n = 40$.

| | | | $\sqrt{\text{MSE}}$ | | | Error | | | Sen | | | Spe | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\rho$ | EM | $\boldsymbol{\Psi}$ | M | L | U | M | L | U | M | L | U | M | L | U |
| 0.85 | LS | $\beta$ | 2.85 | 0.69 | 10.12 | | | | | | | | | |
| | | $b$ | 10.32 | 1.80 | 30.54 | 0.56 | 0.40 | 0.64 | 0.42 | 0.31 | 0.61 | 0.50 | 0.36 | 0.64 |
| | Ridge | $\beta$ | 1.76 | 0.36 | 8.15 | | | | | | | | | |
| | | $b$ | 2.59 | 0.71 | 10.18 | 0.46 | 0.34 | 0.58 | 0.53 | 0.39 | 0.67 | 0.50 | 0.29 | 0.71 |
| | LT | $\beta$ | 1.10 | 0.32 | 7.76 | | | | | | | | | |
| | | $b$ | 1.29 | 0.56 | 6.18 | 0.46 | 0.34 | 0.58 | 0.53 | 0.39 | 0.69 | 0.57 | 0.36 | 0.72 |
| 0.9 | LS | $\beta$ | 3.38 | 0.88 | 8.91 | | | | | | | | | |
| | | $b$ | 11.70 | 1.80 | 29.43 | 0.56 | 0.40 | 0.64 | 0.42 | 0.33 | 0.61 | 0.50 | 0.36 | 0.57 |
| | Ridge | $\beta$ | 2.17 | 0.49 | 7.37 | | | | | | | | | |
| | | $b$ | 2.86 | 0.75 | 9.14 | 0.46 | 0.34 | 0.58 | 0.56 | 0.39 | 0.67 | 0.57 | 0.36 | 0.71 |
| | LT | $\beta$ | 1.24 | 0.35 | 6.65 | | | | | | | | | |
| | | $b$ | 1.29 | 0.57 | 4.71 | 0.46 | 0.34 | 0.60 | 0.53 | 0.39 | 0.69 | 0.50 | 0.29 | 0.72 |
| 0.95 | LS | $\beta$ | 4.67 | 1.39 | 12.35 | | | | | | | | | |
| | | $b$ | 13.72 | 1.80 | 36.58 | 0.56 | 0.40 | 0.62 | 0.42 | 0.36 | 0.61 | 0.50 | 0.43 | 0.57 |
| | Ridge | $\beta$ | 3.16 | 0.62 | 10.83 | | | | | | | | | |
| | | $b$ | 3.18 | 0.78 | 13.09 | 0.46 | 0.34 | 0.58 | 0.56 | 0.39 | 0.67 | 0.57 | 0.36 | 0.79 |
| | LT | $\beta$ | 1.63 | 0.44 | 9.86 | | | | | | | | | |
| | | $b$ | 1.32 | 0.60 | 8.28 | 0.46 | 0.34 | 0.60 | 0.53 | 0.39 | 0.69 | 0.50 | 0.29 | 0.79 |

Table 4.7: The median (M) and 95% CIs for the $\sqrt{\text{MSE}}$, Error, Sensitivity (Sen) and Specificity (Spe) of the LS, Ridge, and LT methods in estimation and prediction of the mixture of cure models when $\phi = 0.85$ and $n = 80$.

| | | | $\sqrt{\text{MSE}}$ | | | Error | | | Sen | | | Spe | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\rho$ | EM | $\boldsymbol{\Psi}$ | M | L | U | M | L | U | M | L | U | M | L | U |
| 0.85 | LS | $\beta$ | 1.00 | 0.33 | 2.20 | | | | | | | | | |
| | | $b$ | 7.15 | 1.97 | 11.00 | 0.56 | 0.44 | 0.64 | 0.42 | 0.33 | 0.61 | 0.50 | 0.43 | 0.57 |
| | Ridge | $\beta$ | 0.76 | 0.25 | 1.95 | | | | | | | | | |
| | | $b$ | 1.81 | 0.52 | 3.78 | 0.46 | 0.34 | 0.58 | 0.56 | 0.39 | 0.69 | 0.50 | 0.36 | 0.71 |
| | LT | $\beta$ | 0.66 | 0.27 | 1.72 | | | | | | | | | |
| | | $b$ | 1.21 | 0.49 | 3.14 | 0.46 | 0.34 | 0.60 | 0.53 | 0.39 | 0.69 | 0.50 | 0.29 | 0.79 |
| 0.9 | LS | $\beta$ | 1.23 | 0.39 | 2.65 | | | | | | | | | |
| | | $b$ | 7.46 | 1.88 | 11.76 | 0.56 | 0.46 | 0.62 | 0.42 | 0.36 | 0.56 | 0.50 | 0.43 | 0.57 |
| | Ridge | $\beta$ | 0.89 | 0.26 | 2.44 | | | | | | | | | |
| | | $b$ | 1.92 | 0.62 | 4.11 | 0.46 | 0.34 | 0.58 | 0.56 | 0.39 | 0.69 | 0.50 | 0.29 | 0.71 |
| | LT | $\beta$ | 0.70 | 0.27 | 2.17 | | | | | | | | | |
| | | $b$ | 1.22 | 0.51 | 3.44 | 0.46 | 0.34 | 0.60 | 0.53 | 0.39 | 0.69 | 0.50 | 0.29 | 0.79 |
| 0.95 | LS | $\beta$ | 1.73 | 0.46 | 4.51 | | | | | | | | | |
| | | $b$ | 7.78 | 1.88 | 12.23 | 0.56 | 0.42 | 0.60 | 0.42 | 0.36 | 0.58 | 0.50 | 0.43 | 0.57 |
| | Ridge | $\beta$ | 1.10 | 0.36 | 4.10 | | | | | | | | | |
| | | $b$ | 1.90 | 0.56 | 4.18 | 0.46 | 0.34 | 0.58 | 0.56 | 0.39 | 0.69 | 0.50 | 0.29 | 0.79 |
| | LT | $\beta$ | 0.79 | 0.43 | 3.88 | | | | | | | | | |
| | | $b$ | 1.24 | 0.55 | 3.11 | 0.46 | 0.34 | 0.60 | 0.53 | 0.39 | 0.69 | 0.50 | 0.29 | 0.79 |

Table 4.8: The median (M) and 95% CIs for the $\sqrt{\text{MSE}}$, Error, Sensitivity (Sen) and Specificity (Spe) of the LS, Ridge, and LT methods in estimation and prediction of the mixture of cure models when $\phi = 0.9$ and $n = 80$.

| | | | $\sqrt{\text{MSE}}$ | | | Error | | | Sen | | | Spe | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\rho$ | EM | $\boldsymbol{\Psi}$ | M | L | U | M | L | U | M | L | U | M | L | U |
| 0.85 | LS | $\beta$ | 1.22 | 0.35 | 2.82 | | | | | | | | | |
| | | $b$ | 7.83 | 1.85 | 12.31 | 0.56 | 0.42 | 0.64 | 0.42 | 0.33 | 0.61 | 0.50 | 0.43 | 0.57 |
| | Ridge | $\beta$ | 0.89 | 0.30 | 2.33 | | | | | | | | | |
| | | $b$ | 1.99 | 0.66 | 4.42 | 0.46 | 0.34 | 0.58 | 0.56 | 0.39 | 0.69 | 0.50 | 0.29 | 0.71 |
| | LT | $\beta$ | 0.67 | 0.23 | 2.02 | | | | | | | | | |
| | | $b$ | 1.25 | 0.47 | 3.15 | 0.46 | 0.34 | 0.60 | 0.53 | 0.39 | 0.69 | 0.50 | 0.29 | 0.72 |
| 0.9 | LS | $\beta$ | 1.49 | 0.43 | 3.37 | | | | | | | | | |
| | | $b$ | 8.29 | 1.88 | 13.66 | 0.56 | 0.42 | 0.62 | 0.42 | 0.36 | 0.61 | 0.50 | 0.43 | 0.57 |
| | Ridge | $\beta$ | 1.06 | 0.35 | 2.75 | | | | | | | | | |
| | | $b$ | 2.06 | 0.66 | 4.79 | 0.46 | 0.34 | 0.58 | 0.56 | 0.39 | 0.69 | 0.50 | 0.29 | 0.71 |
| | LT | $\beta$ | 0.71 | 0.30 | 2.32 | | | | | | | | | |
| | | $b$ | 1.24 | 0.51 | 3.24 | 0.46 | 0.34 | 0.60 | 0.56 | 0.39 | 0.69 | 0.50 | 0.29 | 0.79 |
| 0.95 | LS | $\beta$ | 2.01 | 0.62 | 5.07 | | | | | | | | | |
| | | $b$ | 8.67 | 1.80 | 14.66 | 0.56 | 0.42 | 0.60 | 0.42 | 0.39 | 0.58 | 0.50 | 0.43 | 0.57 |
| | Ridge | $\beta$ | 1.36 | 0.40 | 4.29 | | | | | | | | | |
| | | $b$ | 2.07 | 0.66 | 5.38 | 0.46 | 0.34 | 0.58 | 0.53 | 0.39 | 0.69 | 0.50 | 0.29 | 0.71 |
| | LT | $\beta$ | 0.86 | 0.44 | 3.75 | | | | | | | | | |
| | | $b$ | 1.20 | 0.58 | 3.65 | 0.46 | 0.34 | 0.60 | 0.53 | 0.39 | 0.69 | 0.50 | 0.29 | 0.79 |

Table 4.9: The median (M) and 95% CIs for the $\sqrt{\text{MSE}}$, Error, Sensitivity (Sen) and Specificity (Spe) of the LS, Ridge, and LT methods in estimation and prediction of the mixture of cure models when $\phi = 0.95$ and $n = 80$.

| | | | $\sqrt{\text{MSE}}$ | | | Error | | | Sen | | | Spe | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\rho$ | EM | $\boldsymbol{\Psi}$ | M | L | U | M | L | U | M | L | U | M | L | U |
| 0.85 | LS | $\beta$ | 1.69 | 0.48 | 4.47 | | | | | | | | | |
| | | $b$ | 9.04 | 1.97 | 17.92 | 0.56 | 0.42 | 0.64 | 0.42 | 0.33 | 0.58 | 0.50 | 0.43 | 0.57 |
| | Ridge | $\beta$ | 1.08 | 0.26 | 3.71 | | | | | | | | | |
| | | $b$ | 2.19 | 0.69 | 6.07 | 0.46 | 0.34 | 0.58 | 0.53 | 0.39 | 0.67 | 0.57 | 0.36 | 0.79 |
| | LT | $\beta$ | 0.72 | 0.25 | 3.56 | | | | | | | | | |
| | | $b$ | 1.25 | 0.46 | 3.80 | 0.46 | 0.34 | 0.60 | 0.53 | 0.39 | 0.67 | 0.57 | 0.29 | 0.72 |
| 0.9 | LS | $\beta$ | 2.10 | 0.53 | 5.05 | | | | | | | | | |
| | | $b$ | 9.64 | 1.80 | 20.15 | 0.56 | 0.42 | 0.62 | 0.42 | 0.33 | 0.58 | 0.50 | 0.43 | 0.57 |
| | Ridge | $\beta$ | 1.41 | 0.35 | 4.20 | | | | | | | | | |
| | | $b$ | 2.28 | 0.68 | 6.41 | 0.46 | 0.34 | 0.58 | 0.56 | 0.39 | 0.69 | 0.57 | 0.36 | 0.79 |
| | LT | $\beta$ | 0.85 | 0.28 | 3.76 | | | | | | | | | |
| | | $b$ | 1.25 | 0.51 | 3.90 | 0.46 | 0.34 | 0.58 | 0.53 | 0.39 | 0.69 | 0.57 | 0.29 | 0.79 |
| 0.95 | LS | $\beta$ | 2.90 | 0.85 | 6.31 | | | | | | | | | |
| | | $b$ | 10.97 | 1.88 | 22.54 | 0.56 | 0.42 | 0.60 | 0.42 | 0.36 | 0.61 | 0.50 | 0.43 | 0.57 |
| | Ridge | $\beta$ | 2.00 | 0.50 | 5.59 | | | | | | | | | |
| | | $b$ | 2.54 | 0.78 | 7.77 | 0.46 | 0.34 | 0.58 | 0.56 | 0.39 | 0.67 | 0.50 | 0.29 | 0.79 |
| | LT | $\beta$ | 1.00 | 0.37 | 4.84 | | | | | | | | | |
| | | $b$ | 1.27 | 0.58 | 4.14 | 0.46 | 0.34 | 0.58 | 0.53 | 0.39 | 0.69 | 0.50 | 0.29 | 0.79 |

The findings of Tables 4.1-4.9 suggest that in scenarios of high multicollinearity and smaller sample sizes, the Ridge and LT shrinkage methods offer more reliable coefficient estimates than the traditional LS method. This underscores the importance

of considering both sample size and collinearity when selecting an estimation method. Table 4.3 represents the simulation with the results when the sample size is small. In this table, when $\rho = 0.95$, the LS method shows the worst performance, with $\sqrt{\text{MSE}}$ values of 9.27 for $\beta$ and 20.82 for $\mathbf{b}$. The ridge method shows noticeable improvement compared to LS, with $\sqrt{\text{MSE}}$ values of 6.42 for $\beta$ and 5.09 for $\mathbf{b}$. However, the LT method has the best performance in handling multicollinearity, with $\sqrt{\text{MSE}}$ values of 3.71 for $\beta$ and 1.60 for $\mathbf{b}$, indicating more accurate and reliable coefficient estimates. The lower and upper bounds of $\sqrt{\text{MSE}}$ across the LS, Ridge, and LT methods show how much estimation results can vary. The LS method has the widest range between these bounds, indicating less reliable estimates. Ridge narrows this range, suggesting more reliability in estimation regardless of multicollinearity. The LT method shows the smallest range, indicating consistent and reliable estimates even under multicollinearity. These bounds reveal the reliability of estimation methods and emphasize the benefits of shrinkage methods in improving the mixture cure model's parameter estimates in the presence of multicollinearity.

To represent the results through figures, we first sorted the $\sqrt{\text{MSE}}$ of 1000 simulations and computed the median (M), lower (L), and upper (U) bounds of $\sqrt{\text{MSE}}$s by 50%, 2.5%, and 97.5% percentiles. Figures 4.1-4.6 present the median (M), lower (L), and upper (U) bounds of the 95% intervals for the $\sqrt{\text{MSE}}$ of the LS, Ridge, and LT methods in estimating the coefficients of the mixture of cure models. Three figures represent different sample sizes for each parameter ($\beta$ and $b$). These figures illustrate a continuous range of collinearity levels, from $\phi = \rho = 0.85$ to $\phi = \rho = 0.98$. Specifically, Figures 4.1-4.3 correspond to the estimation of $\beta$ for sample sizes $n = \{20, 40, 80\}$, respectively. Similarly, Figures 4.4-4.6 represent the estimation of $b$ for the same sample sizes and collinearity levels.

These plots visually provide intriguing insights into the impact of collinearity on estimation methods across different sample sizes for the mixture of cure models. Notably, as collinearity increases, a noticeable trend emerges: the ridge and LT estimators consistently outperform the LS method in terms of estimating $\beta$ and $b$ in the mixture of cure models. These findings underscore the effectiveness of ridge and LT estimators in reducing the poor effects of collinearity within our model. Their enhanced performance in minimizing the root of MSE suggests a robustness that surpasses traditional LS approaches, particularly under conditions of high collinearity. These observations underscore these alternative estimation techniques' usefulness

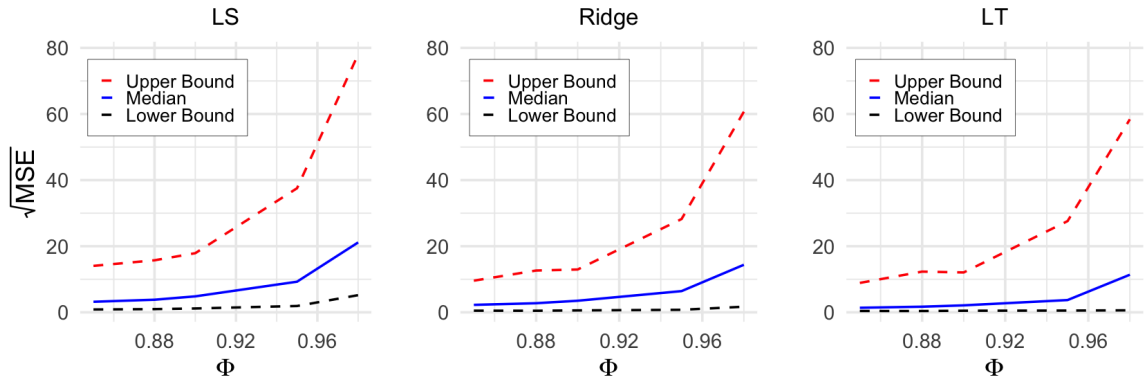


Figure 4.1: The median (blue), upper (red), and lower bound (black) of $\sqrt{\text{MSE}(\widehat{\beta})}$ across the three methods of estimation of coefficients of the mixture cure models when $n = 20$.

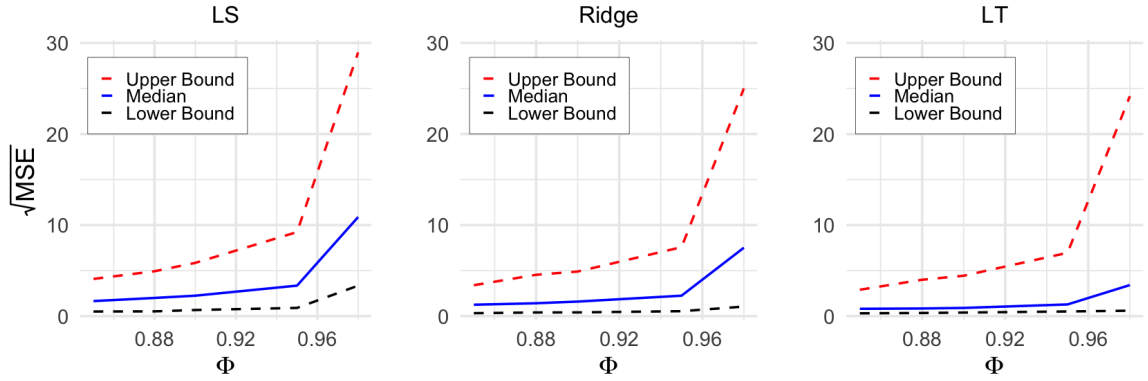


Figure 4.2: The median (blue), upper (red), and lower bound (black) of $\sqrt{\text{MSE}(\widehat{\beta})}$ across the three methods of estimation of coefficients of the mixture cure models when $n = 40$.

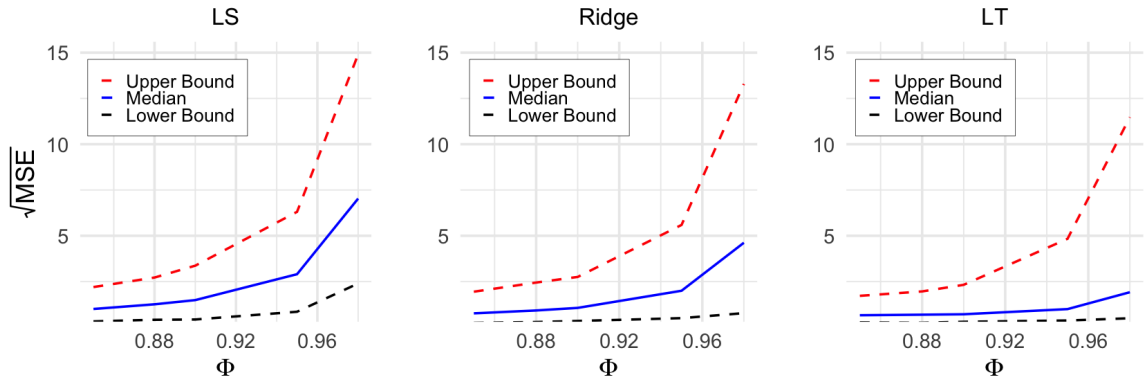


Figure 4.3: The median (blue), upper (red), and lower bound (black) of $\sqrt{\text{MSE}(\widehat{\beta})}$ across the three methods of estimation of coefficients of the mixture cure models when $n = 80$.

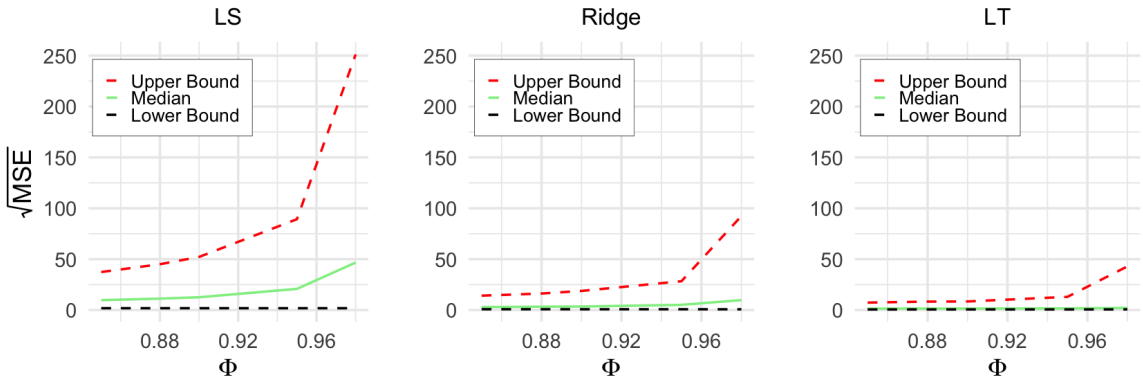


Figure 4.4: The median (green), upper (red), and lower bound (black) of $\sqrt{\text{MSE}(\widehat{\mathbf{b}})}$ across the three methods of estimation of coefficients of the mixture cure models when $n = 20$.

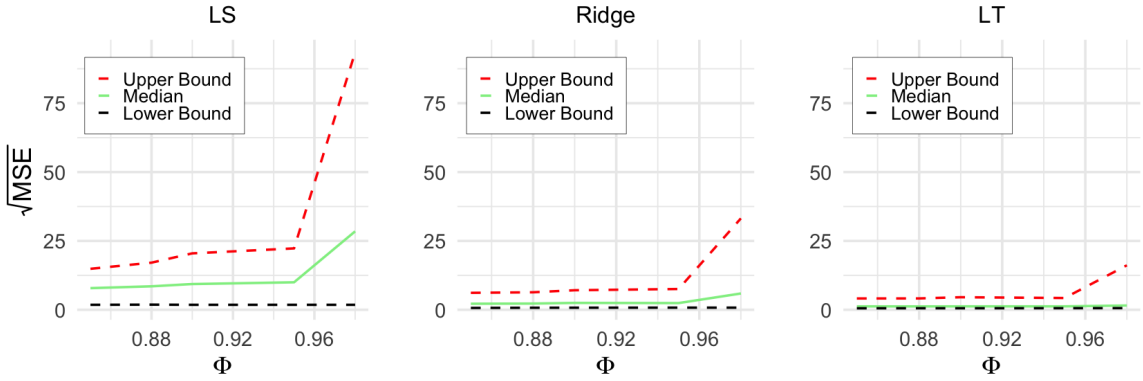


Figure 4.5: The median (green), upper (red), and lower bound (black) of $\sqrt{\text{MSE}(\widehat{\mathbf{b}})}$ across the three methods of estimation of coefficients of the mixture cure models when $n = 40$.

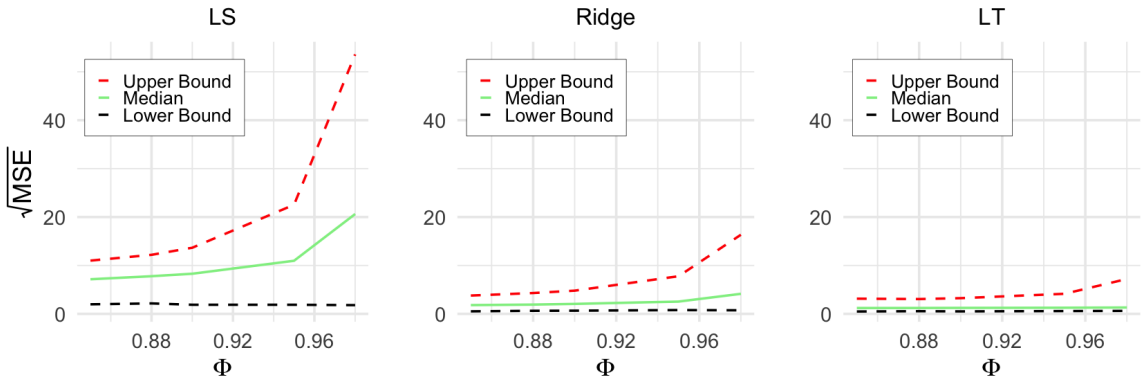


Figure 4.6: The median (green), upper (red), and lower bound (black) of $\sqrt{\text{MSE}(\widehat{\mathbf{b}})}$ across the three methods of estimation of coefficients of the mixture cure models when $n = 80$.

and potential superiority in real-world applications where multicollinearity is more frequent.

Moreover, it is worth noting that these differences are most prominent when the sample size is low. Specifically in the estimation of $\beta$, when the sample size is 20, as shown in Figure 4.1, by focusing on the y-axis, we observe notably superior performance compared to Figures 4.2 and 4.3, where larger sample sizes are considered. This difference underscores the significant impact of sample size on the performance of estimators within our study. Similarly, in the estimation of $b$ in Figure 4.4, where $n = 20$, we observe a more major variation compared to Figures 4.5 and 4.6, highlighting the sensitivity of estimator performance to sample size variations. These findings further emphasize the impact of sample size in improving estimator accuracy for both $\beta$ and $b$ estimations.

Furthermore, another objective of mixture cure models is to forecast the probability that a patient survives from the time of diagnosis until the time of interest; that is called the survival probability of individuals across the latency component of the model. Mixture cure models offer a good understanding of survival outcomes by accounting for both cured and uncured populations, thereby providing insights into the factors influencing survival beyond initial treatment. We aim to visualize these survival probabilities through survival curves plotted against time.

The estimated survival curves are shown in Figures 4.7-4.9, where differences among the methods' results are observed. Notably, both Ridge and LT methods show superior survival probabilities over the LS method, indicated by the blue (LS), red (Ridge), and green (LT) lines, with the true survival probabilities represented by the grey line. This suggests that the Ridge and LT methods provide better estimates of survival probabilities, potentially due to their ability to handle multicollinearity more effectively. The shrinkage Ridge and LT methods provide predictions closer to the true observed survival outcomes; thus, we can better predict how long a patient is likely to survive, which leads to an improved understanding of diseases and treatments. As the sample size increases, the performance of the LS method improves in predicting the survival probabilities. This indicates that the Ridge and LT methods outperform the LS method significantly, mostly when the sample sizes are small, while the LS method improves its performance when sample sizes are large. In this case, the performances of the three methods become closer to the true survival curve (See

Figure 4.9).

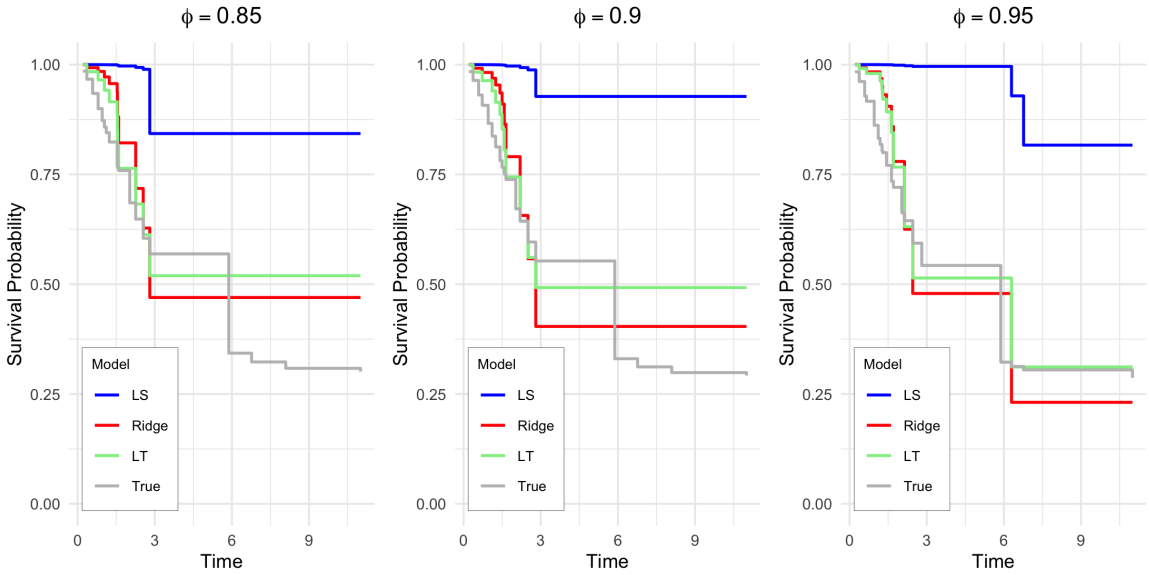


Figure 4.7: The predicted survival probabilities of the LS (blue), Ridge (red), and LT (green) methods in case of different correlation levels $\rho = \phi = (0.85, 0.9, 0.95)$ in the estimation of coefficients of the mixture of cure models when $n = 20$.

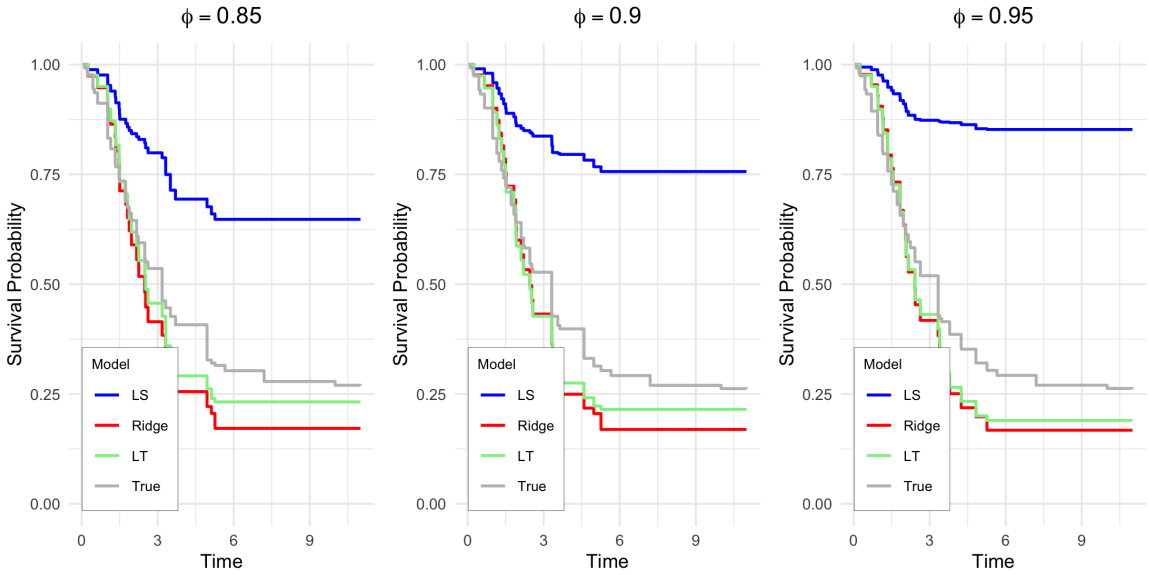


Figure 4.8: The predicted survival probabilities of the LS (blue), Ridge (red), and LT (green) methods in case of different correlation levels $\rho = \phi = (0.85, 0.9, 0.95)$ in the estimation of coefficients of the mixture of cure models when $n = 40$.

Additionally, there is a slight difference in the survival probabilities when multicollinearity increases over the sample sizes. As collinearity levels increase, the model

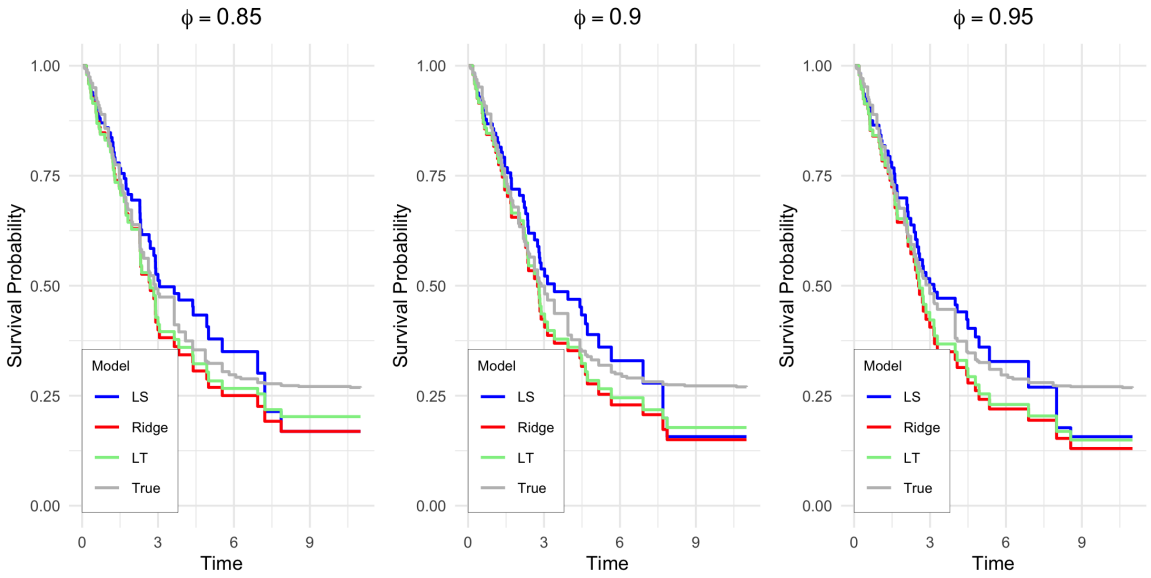


Figure 4.9: The predicted survival probabilities of the LS (blue), Ridge (red), and LT (green) methods in case of different correlation levels $\rho = \phi = (0.85, 0.9, 0.95)$ in the estimation of coefficients of the mixture of cure models when $n = 80$.

survival probabilities get slightly further from the true ones. Despite this, the overall trends in survival probabilities remain relatively consistent across the different methods, suggesting robustness in the estimation techniques employed.

## 4.2 Breast Cancer Data Analysis

Breast cancer remains one of the most common malignancies affecting women worldwide, with significant variations in prognosis and treatment response among patients. It is a globally dangerous illness that gets significant attention from medical professionals and statisticians due to its high fatality rates and unpredictable nature, influenced by various prognostic factors [40]. Breast cancer develops from an inherited genetic mutation, sparking irregular cell division within breast tissue. If left untreated, these malignant cells can spread to other areas of the body, creating a grave health risk. Symptoms may reveal themselves as breast swelling, skin irritation, or palpable lumps within the breast. Tragically, breast cancer ranks as a leading cause of death among women [42, 49]. In 2022, breast cancer affected 2.3 million women worldwide, leading to 670,000 deaths. This disease knows no geographical bounds and can affect women of any age post-puberty, with incidence rates rising with age [62].

Many predictors of breast cancer prognosis have been widely acknowledged, including various histological criteria utilized for different diagnostic and prognostic purposes [48]. These factors provide valuable insights into patient prognosis, independent of treatment modalities [23]. From July 1984 to December 1989, the German Breast Cancer Study Group (GBSG) conducted a clinical trial involving 720 patients diagnosed with primary node-positive breast cancer. Among these patients, 686 individuals with complete data on various standard prognostic factors were included in the dataset used for analysis. Recurrence-free survival (RFS) time and status (0= alive without recurrence, 1= recurrence or death) were the outcome measures. The dataset documented outcomes for 299 patients who experienced recurrence events—the follow-up period extended to 7 years. The prognostic factors assessed in the breast cancer datasets included age at primary surgery (age, years), menopausal status (meno, 0 = premenopausal, 1 = postmenopausal), tumor size (size), tumour grade (grade), number of positive lymph nodes (nodes), progesterone receptors (pgr, fmol/l), estrogen receptors (er, fmol/l), and hormonal therapy (hormon, 0 = no, 1 = yes). Grade was omitted from the analysis due to variations in measurement protocols across the datasets, and pgr did not exhibit significance at the 5% level and was also removed from consideration [49].

Drawing upon prior research highlighting the critical role of the number of positive lymph nodes in breast cancer prognosis, we have elected to narrow our focus exclusively to patients diagnosed with 10 or more positive lymph nodes [9]. Studies have demonstrated that patients with a higher burden of positive lymph nodes often face more aggressive disease courses and poorer prognoses compared to those with fewer positive nodes [32, 57]. Patients with 10 or more positive lymph nodes face a notable risk of recurrence, with event-free survival rates dropping significantly even after five years. By specifically targeting this subgroup in our analysis, we aim to provide deeper insights into the survival dynamics and treatment responses of these patients [5, 52]. Our dataset consists of 103 patients with 10 or more positive lymph nodes and 8 explanatory variables, including age, menopausal status, tumor size, number of positive lymph nodes, estrogen receptors, and hormonal therapy. As we expect, there is a high collinearity of 0.79 between the patient's age and menopausal status in our design matrix.

Semi-parametric mixture cure models are valuable for analyzing cancer survival data, offering insights into the dynamics of disease progression and treatment response.

These models are particularly adept at handling survival data where a considerable proportion of patients may experience long-term survival without recurrence or death while the remaining patients face the risk of disease recurrence or mortality. In our analysis of breast cancer survival data, we leverage the strengths of semi-parametric mixture cure models to uncover the complex interaction between patient characteristics, treatment modalities, and long-term outcomes.

Firstly, model parameters are estimated using the real breast cancer dataset. To show the estimation performance of $(\widehat{\beta}, \widehat{\mathbf{b}})$, we compute $\sqrt{\text{MSE}(\widehat{\beta})}$ and $\sqrt{\text{MSE}(\widehat{\mathbf{b}})}$ according to equations (4.1) and (4.2), respectively. Subsequently, we utilize the logistic component of the semi-parametric mixture cure model to predict the occurrence of events, whether it be recurrence or death, among breast cancer patients. We apply a 5-fold cross-validation method on our dataset to calculate the Error, Sensitivity, and Specificity evaluation metrics for the LS, ridge, and LT methods. These metrics quantify the model's performance in distinguishing cured from uncured patients. Similar to Section 4.1, these evaluation metrics are used to assess the model's performance. Additionally, comparisons with alternative methodologies presented in Table 4.10 highlight both estimation and predictive performance. While the ridge and LT models exhibit better parameter estimation performance than the LS approach, predictive performance remains consistent, mirroring our findings from the simulation section. Our finding aligns with the studies by Inan and Erdogan [26] and Ghanem et al. [20], which also reported that while multicollinearity significantly impacts the estimation accuracy of the methods, it has little effect on the prediction performance.

Table 4.10: The median (M) and 95% CIs for the $\sqrt{\text{MSE}}$; and Error, Sensitivity (Sen), and Specificity (Spe) of the LS, Ridge, and LT methods in estimation and prediction of Breast Cancer fatality using 5-fold Cross Validation.

| | | $\sqrt{\text{MSE}}$ | | | | | |
|---|---|---|---|---|---|---|---|
| EM | $\boldsymbol{\Psi}$ | M | L | U | Error | Sen | Spe |
| LS | $\beta$ | 0.34 | 0.21 | 1.14 | | | |
| | $b$ | 6.55 | 2.49 | 17.38 | 0.28 | 0.92 | 0.21 |
| Ridge | $\beta$ | 0.32 | 0.20 | 0.85 | | | |
| | $b$ | 1.79 | 1.09 | 6.74 | 0.26 | 0.95 | 0.21 |
| LT | $\beta$ | 0.18 | 0.14 | 0.25 | | | |
| | $b$ | 0.53 | 0.32 | 1.82 | 0.25 | 0.93 | 0.28 |

Moreover, similar to the survival curve Figures 4.7-4.9 in the simulation study, we

employ the Cox PH regression component of the semi-parametric mixture cure model to examine the survival probabilities of breast cancer patients over time. This allows us to visualize the survival outcomes curve in Figure 4.10. We observe that the LS, ridge, and LT techniques show almost identical survival probabilities, with LT and ridge evolving over time. This similarity occurs because we have a relatively large sample size. The simulation results show that the LS method tends to provide more reliable results with larger sample sizes. Additionally, in the breast cancer data, we have lower levels of collinearity compared to the simulation studies, which helps the survival probabilities across the different methods to be more similar.

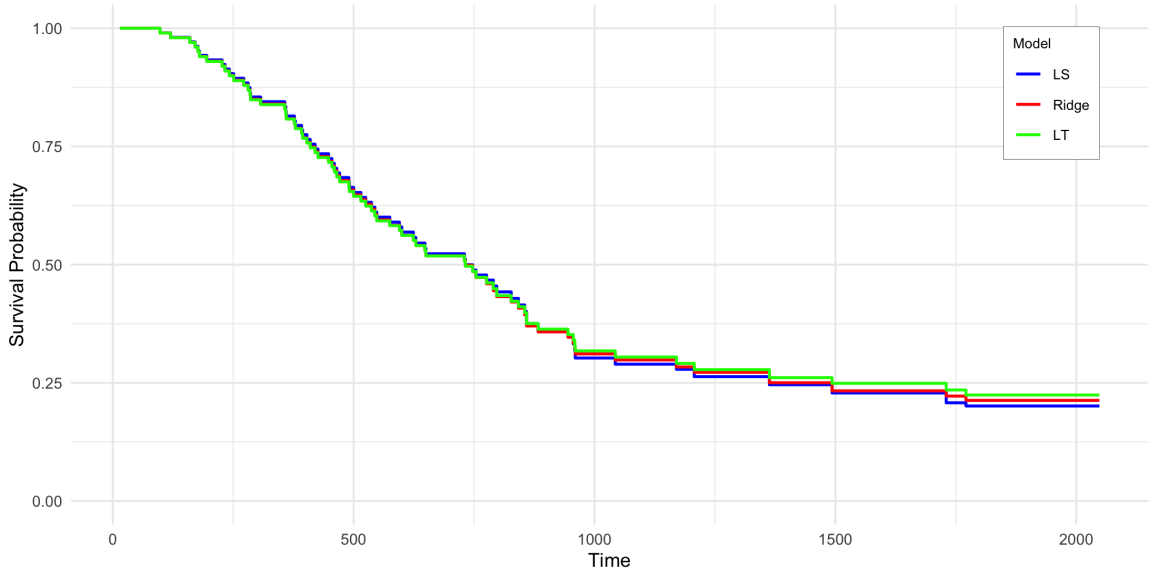


Figure 4.10: The Breast Cancer survival probabilities across the LS (blue), Ridge (red), and LT (green) methods in estimation of coefficients of the mixture of cure models.

By incorporating the incidence and latency components within the semi-parametric mixture cure models framework, we aim to provide a comprehensive understanding of breast cancer prognosis and treatment response. Through precise statistical analysis and model interpretation, we aim to reveal actionable insights that can improve clinical decision-making and patient survival.

In our analysis, we find that the LT method demonstrates greater reliability than LS and ridge methods in estimating parameters under conditions of collinearity, as evidenced by both simulation and real data studies. This suggests that, when faced with multicollinearity, the LT approach may offer more robust parameter estimates.

# Chapter 5

# Summary and Conclusion

In this thesis, we deeply studied survival analysis, focusing specifically on utilizing mixture cure models within medical research. Our primary concentration was to tackle the challenge of multicollinearity within estimating the parameters of mixture cure models. To address this issue, we developed shrinkage Ridge and LT methods for mixture cure models. To do that, we customized appropriate missing data of EM algorithms to incorporate the property of shrinkage methods in the E and M steps for both the latency and incidence components of the mixture of cure models. This chapter will summarize our research and its key findings and also draw conclusions from our study.

In Chapter 2, we provided an in-depth analysis of the incidence and latency components within the semi-parametric mixture cure models: logistic regression and Cox proportional hazard regression models. We investigated the estimation of unknown parameters in logistic and Cox regressions, explaining the process of parameter estimation in both models. Additionally, we discussed different types of cure models, with a specific focus on semi-parametric mixture cure models. We then extended our discussion to the estimation of unknown parameters within the semi-parametric mixture cure model. This involved deriving the complete log-likelihood function and implementing the EM algorithm to estimate the parameters effectively. Through this comprehensive exploration, we set the foundation for our methods contributions in Chapter 3, establishing a solid theoretical foundation for our advancements.

Mixture cure models have the potential to enhance the accuracy of survival estimates for treatments associated with statistical cure. However, when multicollinearity

is present, the reliability of the common LS approach for parameter estimation in mixture cure models is in doubt. Therefore, in Chapter 3, we introduced Ridge and LT shrinkage estimation methods into the mixture cure models to address this issue and improve our model estimates. We first implemented the shrinkage models into logistic and Cox regression separately, which laid the foundation for incorporating shrinkage estimators into the semi-parametric mixture cure model. We explored the effectiveness of ridge and LT shrinkage methods in estimating the unknown parameters of mixture cure models, which are best suited for scenarios involving multicollinearity. By undertaking these steps, we prepared the shrinkage estimators for semi-parametric proportional hazards mixture cure models to be implemented on input data to derive meaningful outcomes.

Our numerical studies in Chapter 4 revealed that when multicollinearity is present, ridge and LT approaches show superior performance in handling multicollinearity, surpassing the LS method. Through simulation studies, we observed that as collinearity rises, particularly in small sample sizes, the traditional LS estimation method's accuracy worsens, whereas ridge and LT shrinkage methods yield more reliable estimates of model coefficients. As a result of their inherent bias, shrinkage estimators are intended to address the issues arising from poorly conditioned design matrices, even though this may introduce bias into the estimation process. Therefore, they are recommended primarily for situations with high levels of multicollinearity.

Finally, we implemented our proposed methods to a dataset on breast cancer, analyzing disease prognosis and survival rates of patients with 10 or more positive lymph nodes. Our analysis consistently showed that ridge and LT techniques offer better survival probabilities compared to the LS method. This is consistent with the findings from Inan and Erdogan [26] that multicollinearity mainly affects the estimation results while we can still present the prediction power with the commonly used LS method. The prediction metrics remained consistent across all methods. This real-data exploration not only validated the accuracy of our simulation studies but also provided valuable insights into disease prognosis.

## 5.1 Future Research

In future research, we aim to extend the methodologies developed for the Cox proportional hazards (PH) model to the accelerated failure time (AFT) model [61]. The PH model is used for analyzing covariate effects on survival time with constant hazards, which is often violated in practice, leading to misinterpretation. Unlike the PH model, the AFT model does not assume constant hazards and is an alternative when this assumption is violated [15]. The research will involve adapting the same techniques to the AFT framework, including the LS, Ridge, and LT shrinkage methods. The goal is to address similar challenges within the AFT model's context, such as multicollinearity and parameter estimation, thereby enhancing its robustness and reliability in survival analysis.

Additionally, we plan to incorporate variable selection with the least absolute shrinkage and selection operator (LASSO) into the mixture cure models and combine it with the LT shrinkage method [13]. This approach will address multicollinearity in high-dimensional datasets, improving the prediction accuracy and stability of parameter estimates. By integrating LASSO's ability to perform variable selection with LT shrinkage's effectiveness in handling multicollinearity, we aim to create a more robust and flexible tool for analyzing complex survival data.

# Bibliography


[1] S. E. Ahmed and S. E. Raheem. Shrinkage and absolute penalty estimation in linear regression models. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(6):541–553, 2012.

[2] M. Amico and I. Van Keilegom. Cure models in survival analysis. *Annual Review of Statistics and Its Application*, 5:311–342, 2018.

[3] D. A. Belsley, E. Kuh, and R. E. Welsch. Regression diagnostics new york. *NY Wiley*, 1980.

[4] R. Bender. Introduction to the use of regression models in epidemiology. *Cancer epidemiology*, pages 179–195, 2009.

[5] A. U. Buzdar, S.-W. Kau, G. N. Hortobagyi, F. C. Ames, F. A. Holmes, G. Fraschini, V. Hug, R. L. Theriault, M. D. McNeese, and S. E. Singletary. Clinical course of patients with breast cancer with ten or more positive nodes who were treated with doxorubicin-containing adjuvant therapy. *Cancer*, 69(2):448–452, 1992.

[6] 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*, 108(3):1255–1260, 2012.

[7] M.-Y. Chen. Predicting corporate financial distress based on integration of decision tree classification and logistic regression. *Expert systems with applications*, 38(9):11261–11272, 2011.

[8] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman. Survival analysis part i: basic concepts and first analyses. *British journal of cancer*, 89(2):232–238, 2003.

[9] M. A. Cobleigh, B. Tabesh, P. Bitterman, J. Baker, M. Cronin, M.-L. Liu, R. Borchik, J.-M. Mosquera, M. G. Walker, and S. Shak. Tumor gene expression and prognosis in breast cancer patients with 10 or more positive lymph nodes. *Clinical Cancer Research*, 11(24):8623–8631, 2005.

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

[11] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 63(1):1–38, 2010.

[12] A. I. Dimitras, S. H. Zanakis, and C. Zopounidis. A survey of business failures with an emphasis on prediction methods and industrial applications. *European journal of operational research*, 90(3):487–513, 1996.

[13] F. Emmert-Streib and M. Dehmer. High-dimensional lasso-based computational regression models: regularization, shrinkage, and selection. *Machine Learning and Knowledge Extraction*, 1(1):359–383, 2019.

[14] V. T. Farewell. A model for a binary variable with time-censored observations. *Biometrika*, 64(1):43–46, 1977.

[15] A. Faruk. The comparison of proportional hazards and accelerated failure time models in analyzing the first birth interval survival data. In *Journal of Physics: Conference Series*, volume 974, page 012008. IOP Publishing, 2018.

[16] T. S. Fayose and K. Ayinde. Different forms biasing parameter for generalized ridge regression estimator. *International Journal of Computer Applications*, 181(37):2–29, 2019.

[17] F. Felizzi, N. Paracha, J. Pöhlmann, and J. Ray. Mixture cure models in oncology: a tutorial and practical guidance. *PharmacoEconomics-open*, 5:143–155, 2021.

[18] B. George, S. Seals, and I. Aban. Survival analysis and regression models. *Journal of nuclear cardiology*, 21(4):686–694, 2014.

[19] A. Gepp and K. Kumar. The role of survival analysis in financial distress prediction. *International research journal of finance and economics*, 16(16):13–34, 2008.

[20] E. Ghanem, A. Hatefi, and H. Usefi. Liu-type shrinkage estimators for mixture of logistic regressions: An osteoporosis study. *arXiv preprint arXiv:2209.01731*, 2022.

[21] E. Ghanem, A. Hatefi, and H. Usefi. Unsupervised liu-type shrinkage estimators for mixture of regression models. *Statistical Methods in Medical Research*, 2024.

[22] O. Gressani, C. Faes, and N. Hens. Laplacian-p-splines for bayesian inference in the mixture cure model. *Statistics in Medicine*, 41(14):2602–2626, 2022.

[23] I. C. Henderson and A. J. Patek. The relationship between prognostic and predictive factors in the management of breast cancer. *Prognostic variables in node-negative and node-positive breast cancer*, pages 351–378, 1998.

[24] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[25] A. E. Hoerl and R. W. Kennard. Ridge regression iterative estimation of the biasing parameter. *Communications in Statistics-Theory and Methods*, 5(1):77–88, 1976.

[26] D. Inan and B. E. Erdogan. Liu-type logistic estimator. *Communications in Statistics-Simulation and Computation*, 42(7):1578–1586, 2013.

[27] D. İnan and M. Tez. A new estimator for cox proportional hazard regression model in presence of collinearity. *Communications in Statistics-Theory and Methods*, 41(13-14):2437–2444, 2012.

[28] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor. Linear regression. In *An introduction to statistical learning: With applications in python*, pages 69–134. Springer, 2023.

[29] S. P. Jenkins. Survival analysis. *Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK*, 42:54–56, 2005.

[30] L. JF and W. P. A simulation study of ridge and other regression estimators. *Communications in Statistics-theory and Methods*, 5(4):307–323, 1976.

[31] G. Khalaf and G. Shukur. Choosing ridge parameter for regression problems. 2005.

[32] E. Koca, T. Y. Kuzan, O. Dizdar, T. Babacan, I. Sahin, E. Ararat, and K. Altundag. Outcomes of locally advanced breast cancer patients with 10 positive axillary lymph nodes. *Medical oncology*, 30:1–7, 2013.

[33] E. T. Lee and J. Wang. *Statistical methods for survival data analysis*, volume 476. John Wiley & Sons, 2003.

[34] J. G. Lee, S. Moon, and K. Salamatian. Modeling and predicting the popularity of online contents with cox proportional hazard regression model. *Neurocomputing*, 76(1):134–145, 2012.

[35] C.-S. Li, J. M. Taylor, and J. P. Sy. Identifiability of cure models. *Statistics & Probability Letters*, 54(4):389–395, 2001.

[36] B. G. Lindsay and M. L. Lesperance. A review of semiparametric mixture models. *Journal of statistical planning and inference*, 47(1-2):29–39, 1995.

[37] K. Liu. Using liu-type estimator to combat collinearity. *Communications in Statistics-Theory and Methods*, 32(5):1009–1020, 2003.

[38] G. J. McLachlan, S. X. Lee, and S. I. Rathnayake. Finite mixture models. *Annual review of statistics and its application*, 6:355–378, 2019.

[39] R. J. Meijer and J. J. Goeman. Efficient approximate k-fold and leave-one-out cross-validation for ridge regression. *Biometrical Journal*, 55(2):141–155, 2013.

[40] M. Nadjib Bustan, Arman, M. K. Aidid, F. A. Gobel, and Syamsidar. Cox proportional hazard survival analysis to inpatient breast cancer cases. In *Journal of Physics: Conference Series*, volume 1028, page 012230. IOP Publishing, 2018.

[41] T. G. Nick and K. M. Campbell. Logistic regression. *Topics in biostatistics*, pages 273–301, 2007.

[42] R. K. Peddarapu, V. Nunemuntala, S. Koora, S. Kuntala, J. Neeradi, and N. S. Annam. Breast cancer survival prediction and analysis. In *2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, pages 843–850. IEEE, 2023.

[43] M. Pedrosa-Laza, A. López-Cheda, and R. Cao. Cure models to estimate time until hospitalization due to covid-19: A case study in galicia (nw spain). *Applied Intelligence*, 52(1):794–807, 2022.

[44] Y. Peng and K. B. Dear. A nonparametric mixture model for cure rate estimation. *Biometrics*, 56(1):237–243, 2000.

[45] Y. Peng and J. M. Taylor. Cure models. *Handbook of survival analysis*, 34:113–134, 2014.

[46] Y. Peng and J. M. Taylor. Residual-based model diagnosis methods for mixture cure models. *Biometrics*, 73(2):495–505, 2017.

[47] Y. Peng and B. Yu. *Cure models: methods, applications, and implementation*. CRC Press, 2021.

[48] J. Poorolajal, N. Nafissi, M. E. Akbari, H. Mahjub, and N. Esmailnasab. Breast cancer survival analysis based on immunohistochemistry subtypes (er/pr/her2): a retrospective cohort study. *Archives of Iranian medicine*, 19(10):0–0, 2016.

[49] P. Royston and D. G. Altman. External validation of a cox prognostic model: principles and methods. *BMC medical research methodology*, 13:1–15, 2013.

[50] S. Salerno and Y. Li. High-dimensional survival analysis: Methods and applications. *Annual review of statistics and its application*, 10:25–49, 2023.

[51] R. Schaefer, L. Roi, and R. Wolfe. A ridge logistic estimator. *Communications in Statistics-Theory and Methods*, 13(1):99–113, 1984.

[52] C. Schmoor, W. Sauerbrei, G. Bastert, H. Bojar, M. Schumacher, G. B. C. S. Group, et al. Long-term prognosis of breast cancer patients with 10 or more positive lymph nodes treated with cmf. *European Journal of Cancer*, 37(9):1123–1131, 2001.

[53] N. Senaviratna and T. A Cooray. Diagnosing multicollinearity of logistic regression model. *Asian Journal of Probability and Statistics*, 5(2):1–9, 2019.

[54] N. Shrestha. Detecting multicollinearity in regression analysis. *American Journal of Applied Mathematics and Statistics*, 8(2):39–42, 2020.

[55] K. R. Smith, M. L. Slattery, and T. K. French. Collinear nutrients and the risk of colon cancer. *Journal of clinical epidemiology*, 44(7):715–723, 1991.

[56] S. Sperandei. Understanding logistic regression analysis. *Biochemia medica*, 24(1):12–18, 2014.

[57] F. Tonellotto, A. Bergmann, K. de Souza Abrahão, S. S. de Aguiar, M. A. Bello, and L. C. S. Thuler. Impact of number of positive lymph nodes and lymph node ratio on survival of women with node-positive breast cancer. *European journal of breast health*, 15(2):76, 2019.

[58] M. van Smeden, K. G. Moons, J. A. de Groot, G. S. Collins, D. G. Altman, M. J. Eijkemans, and J. B. Reitsma. Sample size for binary logistic prediction models: beyond events per variable criteria. *Statistical methods in medical research*, 28(8):2455–2474, 2019.

[59] K. P. Vatcheva, M. Lee, J. B. McCormick, and M. H. Rahbar. Multicollinearity in regression analyses conducted in epidemiologic studies. *Epidemiology (Sunnyvale, Calif.)*, 6(2), 2016.

[60] P. Wang, Y. Li, and C. K. Reddy. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6):1–36, 2019.

[61] L.-J. Wei. The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in medicine*, 11(14-15):1871–1879, 1992.

[62] World Health Organization. Breast cancer. https://www.who.int/news-room/fact-sheets/detail/breast-cancer, 2024. [Online; accessed 17-April-2024].

[63] J. Xu and Y. Peng. Nonparametric cure rate estimation with covariates. *Canadian Journal of Statistics*, 42(1):1–17, 2014.

[64] W. Yoo, R. Mayberry, S. Bae, K. Singh, Q. P. He, and J. W. Lillard Jr. A study of effects of multicollinearity in the multivariable analysis. *International journal of applied science and technology*, 4(5):9, 2014.