Subgroup Analysis in Nonparametric Mixture Model for Cure Rate Analysis

基于子群分析的非参混合模型在治愈率分析上的应用

Wenbo Zhang
ID: 1508618
Supervisor: Xiaojun Zhu

December 2018

Abstract

Nonparametric Mixture models have been employed to estimate cure rate with failure time data. However, one critical issue is that the cure fraction and parameters of survival function are same for every patient in these model. In this paper, we propose a developed nonparametric model partially based on subgroup analysis. This model assumes that there are two hidden subgroups for patients, treatment favorable and unfavorable, and different groups have different parameters to estimate cure rate or other survival properties. The EM algorithm is utilized to estimate parameters in this model. We also apply Cox's proportional hazard regression and investigated the influence of covariates on survive. Model is built by utilizing breast cancer data and compared with one nonparametric model from literature proposed by previous researchers to show its advantages.

Keywords: Subgroup Analysis; Survival Data; Nonparametric Model; Proportional Hazards Assumption; Cure Rate Model; EM Algorithm; Logistic Regression

Abstract

非参混合模型已经被广泛的用于估计失败时间数据的治愈率。但是，一个很关键的问题：在以前的模型中，对任何不同患者，他的治愈率和生存函数的参数都是相同。在这篇文章中，我们提出了一个改良过后的基于子群分析的非参模型。这个模型假设在患者中存在两个隐藏的子群，治疗友好和不友好组。在不同的组中都有不同的参数用于估计治愈率和其他的生存性质。我们用了EM算法去估计这些参数。同时，还采用了Cox的比例危险回归并且分析了变量对生存的影响。最后，我们用一组乳癌患者的数据去构建模型并且把它和之前的一个模型进行比较分析。

关键字:生存分析; 生存数据; 非参模型; 比例风险假设; 治愈率模型; EM 算法; 逻辑回归

Contents

Chapter 1

Introduction

1.1 Overview

In some clinic studies, in order to examine the effects of the treatment, doctors tend to record patients' condition at a certain time, death or survive. Based on these data, biostatisticians have utilized statistical assumption and methods to construct survival models. They focus on the estimation of the cure rate or the death time of patients. In 2000, Peng and Dear [1] proposed a nonparametric mixture model for cure rate analysis, which attracted researches' attention and provided an applicable estimation approach. In their model, patients could be cured, which was different from traditional models. Cure rate analysis is meaningful in real life in terms of advanced medical condition.

However, this model has one main limitations. In Peng's paper, one assumption behind model is that all the patients could obtain relatively similar feedbacks from treatment. In real life, this assumption is not greatly prices because the treatment effects vary from person to person, which depends on physiological conditions or other effects. Some people may get positive feedbacks but the others don't or even obtain negative effects after they receive the same treatment. For people who could not benefit from the operation, it is better for them to receive expectant treatment. Therefore, it is proper to consider that whether people should be divided into two groups with respect to different treatment effects.

There is one critical question: what's the optimal treatment plan for patients? In other words, is operation better than expectant treatment or vice versa? The answer is highly important because doctors can apply targeted approaches to individual patients rather than neglect their distinct property, which is helpful to increase cure rates and survival time for patients. This could be an innovative and significant development in clinical trials.

1.2 Objectives

This project mainly focuses on built the model which includes personal and group information based on data from breast cancer patients. Then the model can be utilized to estimate cure rate, death time and survival period. It can be a useful predictive tool to help doctor to evaluate outcome for patients and chose the best treatment method. In addition, this project studies the influence of age and gender to the model. For example, what's the significance of these factors to decide people's groups? How does these factor effect the survival condition of patients? These questions can also be answered through this research project.

1.3 Literature Review

The statistical researches embody the concept of proportion of cured patients can be found in many studies related to survival analysis. ([2] [3] [4] [5] [6].)

In previous research, statisticians focus on parametric mixture model. In parametric models, there are assumptions about the failure time distribution of uncured patients. The density function of death and survival functions can be derived directly from the given distribution. (Studies of parametric models can be found in [7] [8] [9] [10] [11].)

However, there is a problem with these parametrical models. The problem is that the verification of assumptions behind these models is difficult to conduct. Yamaguchi [12] applied generalized Gamma distribution to reduce the parameter constraints of his model. Peng, Dear and Denham [13] proposed generalized F distribution in their model.

In recent years, employing nonparametric mixture models to estimate cure rate estimation has been studied. Kuk and Chen [14] proposed a proportional hazards(PH) assumption to the failure time distribution of uncured patients and constructed a nonparametric model, which is similar to Cox's PH model. Nevertheless, their results are obtained through Monte Carlo approximation of the likelihood function, which is not convenient to use to a large extent. Then Taylor [15] employed Kaplan-Meirer survival estimator to estimate the failure time distribution of uncured patients and EM algorithm to estimate hidden parameters in the model. Although their method can be used to estimate necessary distributions and parameters, there are no covariates in the failure time distribution of uncured patients.

Peng and Dear [1] investigated the nonparametric mixture model further and proposed a new nonparametric estimation method. They built a PH mixture model based on EM algorithm. Although this idea is similar to that of Kuk and Chen [14], their estimation approaches are different. They also compared their model with existing nonparametric and parametric models with same data in terms of common statistical metrics, which reveals the goodness of fit of their model.

However, as mentioned before that, in Peng and Dear's model, they assume that all the patients can receive the similar feedbacks of treatment based on the factors with the same weight. This assumption is not applicable in real world. In order to solve this issue, subgroup analysis is started to be considered in clinical trails.

Peter [16] discussed the feasibility, importance, causes and interpretation of subgroup analysis in randomised controlled trials. He mentioned that Subgroup analyses is highly necessary when there are potentially huge differences between groups in terms of risks with or without treatment. Peng, Talor and Yu [17] proposed a marginal regression model which cluster patients into subgroups based on the institutions. Altstein and Li [18] studied a semiparametric accelerated failure time mixture model to estimate treatment effect on a latent subgroup in randomized clinical trials. Shen and He [19] built a structured logistic-normal mixture model to identify a subgroup that has an enhanced treatment effect. Based on their work, Wu, Zhang and Yu [20] employed a Logistic-Cox mixture model to analysis the existence of subgroup. Kim et al. [21] considered to utilize semiparametric Bayesian latent model and secondary data to test its utility for an exploratory subgroup effect analysis.

In this project, based on Pend and Dear's model which employs proportional hazards assumption, we utilize subgroup analysis in clinical trial and present a developed nonparametric mixture model.

Chapter 2

Methodology

2.1 Design

2.1.1 Survival Analysis

Survival analysis is a branch of statistics, which analyzes the expected time duration until one or more events occur, such as failure in manufacturing engineering and death of patients. Survival analysis also contains the modelling of time to event data; in this condition, failure or death is considered as an ”event”. Moreover, at most one single event occurs for each subject, after which the mechanism or patient is broken or dead. Particularly, survival analysis is widely used in bio-statistics.

Survival Function

The survival function is defined as:

$$S(t) = Pr(T > t)$$

where t is certain time, T is a random variable which represents the death time, $Pr(T > t)$ stands for the probability that the time of death T is later than t. It usually assumed that $S(0) = 1$. It is also natural that the survival function should be monotonically decreasing function: $S(u) \leq S(v)$ if $u \geq v$. Furthermore, this function tends to be 0 as t goes to infinite, which is realistic: people have less chance to survival when they get older. However, when $t \to \infty$, it assumes that finally patients will die under the influence of illness, which is not acceptable when people can be cured without recurrence. Therefore, cure rate is needed.

Lifetime Distribution Function and Event Density

The lifetime distribution function, denoted by F, is defined as :

$$F(t) = Pr(T \leq t) = 1 - S(t)$$

where $S(t)$ is survival function. When F is differentiable, its derivative is the probability density function of the lifetime distribution, which is denoted by f:

$$f(t) = F'(t) = \frac{d}{dt}F(t)$$

The function f also represents death rate or failure events per unit time.

Hazard Function

The hazard function, denoted by λ, is defined as the event rate at time t conditional on survival until time t. Suppose that an item has survived for time t and we want the probability that it will die in a following time dt:

$$\lambda(t) = \lim_{dt \to 0} \frac{Pr(t < T < t + dt)}{dtS(t)} = \frac{f(t)}{S(t)}$$

where $f(t)$ is lifetime distribution function and $S(t)$ is survival function.

Cumulative Hazard Function

The cumulative hazard function is defined as:

$$\Lambda(t) = \int_0^t \lambda(u)du = -\log[S(t)]$$

where $S(t)$ is survival function.

2.1.2 Proportional Hazards Mixture Model

Let T be non-negative random variable representing failure time, $f(t|\mathbf{x})$ be the probability density function of T, $S(t|\mathbf{x})$ be the survival function, where $\mathbf{x}$ is the vector of covariates.In the mixture model, there is an assumption that the survival rate of patients will not goes to 0 as t goes to infinity, which is different from other models. Following definitions of $f(t|\mathbf{x})$ and $S(t|\mathbf{x})$, we can figure out that how this assumption can be achieved:

$$f(t|\mathbf{x}) = \pi(\mathbf{x})f_u(t|\mathbf{x})$$

$$S(t|\mathbf{x}) = \pi(\mathbf{x})S_u(t|\mathbf{x}) + 1 - \pi(\mathbf{x})$$

where $\pi(\mathbf{x})$ is the proportion of uncured patients, $f_u(t|\mathbf{x})$ and $S_u(t|\mathbf{x})$ are the density of failure time distribution and survival functions of uncured patients. The Proportional Hazards(PH) assumption introduce covariates to the model and then separate the effects of regressors and baseline hazard function. The failure time distribution of uncured patients in this model can be obtained by employing PH assumption with covariates $\mathbf{x_i}$ and the hazard function is modeled by:

$$h_u(t_i) = h_{u0}(t_i)\exp(\boldsymbol{\beta'}\mathbf{x_i})$$

where $h_{u0}(t)$ is the baseline hazard function, which can be arbitrary function but unrelated to x_i of $\mathbf{x_i}$.

This assumption also implies that:

$$S_u(t_i|\mathbf{x_i}) = S_{u0}(t_i)^{\exp(\boldsymbol{\beta'}\mathbf{x_i})}$$

where $S_{u0}(t) = \exp\{-\int_0^t h_{u0}(w)dw\}$.

Cure Rate Model

In cure rate model, it is assumed that some subjects are reasonably believed to be cured during the process. This model is often used in time-to-event data. There are two types of cure rate models for estimating cure rate. The first one is the Mixture Cure Model (MCM). In this model, it is assumed that the whole population is consist of susceptible subjects and cured subjects. The second cure model type was non-mixture model that sets a asymptote for cumulative hazard function and cure rate. In fact, these two models are related to each other and when the cure rate specially specified, the NMCM can be changed into the MCM. In this project, we use the first type cure rate model.

Censoring Data

Censoring data different from death events. It is kind of data which are missed due to out of touch or other uncontrollable reasons in the observing process, which is kind of missing value problem. In terms of this condition, it is not realistic to consider these missing patients to be dead because there exist cured patients. Following this cure rate, missing patients have some probability of being cured. They may also be uncured and follow a survival function and corresponding distribution.

2.1.3 EM Algorithm

Likelihood and Partial Likelihood

Likelihood is a function of parameters which describes the probability to obtain the observed data. A partial likelihood is an changed form of the full likelihood such that only a part of the parameters (the parameters of interest) are in the function.

In statistics, an expectation-maximization (EM) algorithm is an iterative method to find maximum likelihood of parameters in statistical models which depends on unobserved variables [22]. This algorithm is consist of two steps, E-step and M-step. In E-step, a function is created to calculate the expectation of the log-likelihood of a model by using the initial or current estimate for the expected parameters. Then in M-step, the updated parameters can be obtained by maximizing the likelihood functions obtained in E-step. These parameters will be used in the next E-step as the initial parameters. This process will stop until the error or relative error is decreased to an acceptable value.

Description

Given the statistical model, a set of observed data $\mathbf{X}$, a set of unobserved latent values $\mathbf{Z}$, a maximum likelihood function $L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z}) = P(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$, a set of unknown parameters $\boldsymbol{\theta}$, the maximum likelihood estimate(MLE) is obtained by maximizing marginal likelihood of data:

$$L(\boldsymbol{\theta}; \mathbf{X}) = P(\mathbf{X}|\boldsymbol{\theta}) = \int P(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) dZ$$

The goal of EM algorithm is to find the MLE of the marginal likelihood by iteratively applying these two steps:

E-step: Define $M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ as the expected value of the loglikelihood function of $\boldsymbol{\theta}$ in the t-th step, in terms of the current conditional distribution of $\mathbf{Z}$ given $\mathbf{X}$ and the current estimate of $\boldsymbol{\theta}^{(t)}$:

$$M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E_{\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}(t)} \log L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z})$$

M-step: Find the parameters that maximize this function:

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$$

2.2 Model Description

EM Algorithm for developed mixture model

Suppose the data are in the form $(t_i, \delta_i, \mathbf{x_i})$, where t_i represents the observed survival time for the ith patients, δ_i indicates whether the record is censored (1 is censoring, 0 is death) and x_i is the vector which consists of the ith patient's property, like age and gender.

Here the EM algorithm is employed to estimate hidden variables. Let c_i be an indicator of cure rate, meaning that $c_i = 1$ if the ith patient is cured and $c_i = 0$ otherwise. As for g_i, the value is 1 for the ith patient in the treatment favorable group, 0 for unfavorable group. There is a table for notations:

Group name	Gruop number	Cure Rate	Probability density function	Survival function	Grouping Probability
Treatment Unfavorable	0	1-π_1	$f_1(t_i)$	$S_1(t_i)$	α
Treatment Favorable	1	1-π_2	$f_2(t_i)$	$S_2(t_i)$	1-α

Table 2.1: The Notation Used

By combing mixture model and logistic regression, the cure rates can be obtained:

$$\boldsymbol{\beta_1'}\mathbf{x_i} = \log \frac{\pi_{i,1}}{1-\pi_{i,1}} \qquad \boldsymbol{\beta_1'}\mathbf{x_i} + \lambda_1 = \log \frac{\pi_{i,2}}{1-\pi_{i,2}}$$

where $\mathbf{x}$ is a vector of covariates, representing one patient's information. In treatment favorable group, since it is assumed that patients in this group can get better feedbacks from operation, then λ_1 is regarded as a positive effect brought by treatment and its value is always positive.

As for α_i, the probability of ith patient belonging to treatment unfavorable group, it can be derived from another logistic regression model:

$$\boldsymbol{\theta'}\mathbf{x_i} = \log \frac{1-\alpha_i}{\alpha_i}$$

Therefore complete log likelihood function can be expressed by using above notations:

$$l = \log \prod_{i=1}^{n} \{[\pi_{i,2} f_{i,2}(t_i)(1-\alpha_i)]^{(1-c_i)(1-\delta_i)}[(1-\pi_{i,2})(1-\alpha_i)]^{c_i\delta_i}[\pi_{i,2}S_{i,2}(t_i)(1-\alpha_i)]^{(1-c_i)\delta_i}\}^{g_i}$$
$$\times\{[\pi_{i,2} f_{i,1}(t_i)\alpha_i]^{(1-c_i)(1-\delta_i)}[(1-\pi_{i,1})\alpha_i]^{c_i\delta_i}[\pi_{i,1}S_{i,1}(t_i)\alpha_i]^{(1-c_i)\delta_i}\}^{1-g_i}$$

Then it is the sum of the following functions:

$$
\begin{aligned}
l_1 &= \log \prod_{i=1}^{n} \pi_{i,2}^{m_{i,1}} (1-\pi_{i,2})^{\delta_i m_{i,2}} \pi_{i,1}^{m_{i,3}} (1-\pi_{i,1})^{\delta_i m_{i,4}} \\
l_2 &= \log \prod_{i=1}^{n} S_{i,2}(t_i)^{m_{i,1}} h_{i,2}(t_i)^{d_i(1-\delta_i)} S_{i,1}(t_i)^{m_{i,3}} h_{i,1}(t_i)^{(1-d_i)(1-\delta_i)} \\
l_3 &= \log \prod_{i=1}^{n} (1-\alpha_i)^{d_i} \alpha_i^{1-d_i}
\end{aligned}
$$

where

$$
\begin{aligned}
m_{i,1} =& E[g_i(1-c_i)] \\
=& \frac{\delta_i S_{i,2}(t_i)\pi_{i,2}(1-\alpha_i)}{(S_{i,2}(t_i)\pi_{i,2}+1-\pi_{i,2})(1-\alpha_i)+(S_{i,1}(t_i)\pi_{i,1}+1-\pi_{i,1})\alpha_i} \\
&+ \frac{f_{i,2}(t_i)(1-\alpha_i)(1-\delta_i)}{f_{i,2}(t_i)(1-\alpha_i)+f_{i,1}(t_i)\alpha_i} \\
m_{i,2} =& E[g_i c_i] = \frac{\delta_i(1-\pi_{i,2})(1-\alpha_i)}{(S_{i,2}(t_i)\pi_{i,2}+1-\pi_{i,2})(1-\alpha_i)+(S_{i,1}(t_i)\pi_{i,1}+1-\pi_{i,1})\alpha_i} \\
m_{i,3} =& E[(1-g_i)(1-c_i)] \\
=& \frac{\delta_i(1-\pi_{i,1})\alpha_i}{(S_{i,2}(t_i)\pi_{i,2}+1-\pi_{i,2})(1-\alpha_i)+(S_{i,1}(t_i)\pi_{i,1}+1-\pi_{i,1})\alpha_i} \\
&+ \frac{(1-\delta_i)f_{i,1}(t_i)\alpha_i}{f_{i,1}(t_i)\alpha_i+f_{i,2}(t_i)(1-\alpha_i)} \\
m_{i,4} =& E[(1-g_i)c_i] = \frac{\delta(1-\pi_{i,1})\alpha_i}{(S_{i,2}(t_i)\pi_{i,2}+1-\pi_{i,2})(1-\alpha_i)+(S_{i,1}(t_i)\pi_{i,1}+1-\pi_{i,1})\alpha_i} \\
d_i =& E[g_i] = E[g_i(1-c_i)] + E[g_i c_i] = m_{i,1} + m_{i,2}
\end{aligned}
$$

Therefore, when we utilize EM algorithm for this problem, the E-step is to compute above expectations of hidden variables, because it is impossible to obtain the value of the random variables and we replace them with their expectations.

In the M-step, l_1, l_2 and l_3 are maximized respectively in terms of fixed hidden variables.

Estimation of $\boldsymbol{\beta}_2$ in the M-step

Let $t_1 < .... < t_k$ represents distinct uncensored failure times. D_j denotes the set of d_j tied uncensored failures at t_j. Let E_j denotes the censored data in the time interval $[t_j, t_{j+1})$, j from 0 to k, where $t_0 = 0$ and $t_{k+1} = \infty$. R_j is the risk set at time t_j, which consists of individuals alive and uncensored just prior to t_j. Assuming that $d_i = 1$ for all i and following

Kalbfleisch and Prentice [23], partial likelihood is used here and l_2 can be transformed to:

$$
\begin{aligned}
&\log \int_0^\infty \int_{t_1}^\infty \cdots \int_{t_{k-1}}^\infty \prod_{i=1}^{k} \{h_0(t_i) \exp(\boldsymbol{\beta_2'}\mathbf{x_i} + d_i\lambda_2) \\
&\quad \exp\Bigg[-\Bigg(\sum_{j \in E_i} m_{j,1} \exp(\boldsymbol{\beta_2'}\mathbf{x_j} + \lambda_2) + m_{j,3} \exp(\boldsymbol{\beta_2}\mathbf{x_j})\Bigg] \int_0^{t_j} h_0(w)dw\Bigg)\} dt_k \cdots dt_1 \\
&\quad = \log \prod_{i=1}^{k} \frac{\exp(\boldsymbol{\beta_2'}\mathbf{x_i} + d_i\lambda_2)}{\sum_{j \in R_i} m_{j,1} \exp(\boldsymbol{\beta_2'}\mathbf{x_j} + \lambda_2) + m_{j,3} \exp(\boldsymbol{\beta_2'}\mathbf{x_j})}
\end{aligned}
$$

If there are ties in some time intervals, it is convenient to utilize method proposed by Breslow [24] and the changed form is:

$$
\log \prod_{i=1}^{k} \frac{\exp(\boldsymbol{\beta_2'}\mathbf{s_i} + d_i\lambda_2)}{\{\sum_{j \in R_i} m_{j,1} \exp(\boldsymbol{\beta_2'}\mathbf{x_j} + \lambda_2) + m_{j,3} \exp(\boldsymbol{\beta_2'}\mathbf{x_j})\}^{d_j}}
$$

where $\mathbf{s_i} = \sum_{j \in D_i} \mathbf{x_j}$ is the sum of covariates vectors for individuals in D_i. In this case, maximizing this function is equivalent to maximizing l_2, but here the estimation of baseline hazard function is eliminated, which can help Newton method obtain more accurate results.

Estimation of the Survival function in the M-step

We can employ a method used similarly in Cox's PH model to estimate the baseline survival function $S_{u0}(t)$. Function l_2 can be written as:

$$
\begin{aligned}
l_2 = &\log \prod_{i=1}^{n} S_{i,2}(t_i)^{m_{i,1}} h_{i,2}(t_i)^{d_i(1-\delta_i)} S_{i,1}(t_i)^{m_{i,3}} h_{i,1}(t_i)^{(1-d_i)(1-\delta_i)} \\
= &\log \prod_{i=1}^{k} \prod_{j \in D_i} S_{j,2}(t_i)^{m_{j,1}} h_{j,2}(t_i)^{d_j} S_{j,1}(t_i)^{m_{j,3}} h_{j,1}(t_i)^{1-d_j} \prod_{j \in E_i} S_{j,2}(t_i)^{m_{j,1}} S_{j,1}(t_i)^{m_{j,3}} \\
= &\log \prod_{i=1}^{k} \prod_{j \in D_i} S_{u0}(t_i)^{\exp(\boldsymbol{\beta_2'}\mathbf{x_j}+\lambda_2)d_j} S_{u0}(t_i)^{\exp(\boldsymbol{\beta_2'}\mathbf{x_j})(1-d_j)} h_0(t_i)^{d_j} \exp((\boldsymbol{\beta_2'}\mathbf{x_j} + \lambda_2)d_j) \\
&\quad h_0(t_i)^{1-d_j} \exp((\boldsymbol{\beta_2'}\mathbf{x_j})(1-d_j)) \prod_{j \in E_i} S_{u0}(t_i)^{\exp(\boldsymbol{\beta_2'}\mathbf{x_j}+\lambda_2)m_{j,1}} S_{u0}(t_i)^{\exp(\boldsymbol{\beta_2'}\mathbf{x_j})(m_{j,3})} \\
= &\log \prod_{i=1}^{k} \prod_{j \in D_i} S_{u0}(t_i)^{\exp(\boldsymbol{\beta_2'}\mathbf{x_j})(\exp(\lambda_2)d_j+1-d_j)} h_0 \exp(\boldsymbol{\beta_2'}\mathbf{x_j} + \lambda_2 d_j) \\
&\quad \prod_{j \in E_i} S_{u0}(t_i)^{\exp(\boldsymbol{\beta_2'}\mathbf{x_j})(\exp(\lambda_2)m_{j,1}+m_{j,3})}
\end{aligned}
$$

Using μ to replace $S_{u0}(t)$ and $h_0(t)$,

$$= \log \prod_{i=1}^{k} \prod_{j \in D_i} 1 - \mu_i^{(\exp(\lambda_2) d_j + 1 - d_j) \exp(\boldsymbol{\beta_2} \mathbf{x_j})} \prod_{j \in R_i - D_i} \mu_i^{(\exp(\lambda_2) m_{j,1} + m_{j,3}) \exp(\boldsymbol{\beta_2} \mathbf{x_j})}$$

Therefore, the estimation of $S_{u0}(t)$ can be obtained:

$$\hat{S}_{u0}(t_i) = \exp \left\{ - \sum_{i:t_i<t} \frac{a_i}{\sum_{j \in R_i} (\exp(\lambda_2) m_{j,1} + m_{j,3}) \exp(\boldsymbol{\beta_2} \mathbf{x_j})} \right\}$$

Where a_i are the ties of death at time t_i.

2.3 Biases, Variances and Confidence Intervals

Bias is the difference between the estimated value and true underlying quantitative parameter, measuring the error of fitting. As for variance, it is used to measure how far a set of (random) numbers are spread out from their average value. Confidence Interval is a type of interval estimate and this can be employed to estimate rough range of parameters.

Jackknife, a resampling method, is useful to estimate mean and variance of parameters. The jackknife estimator is obtained by systematically deleting each observation from a dataset. Then the estimates of parameters from each step and the average of these values can also be calculated. Given sample with n observed data, the jackknife estimate is found by combing the estimates of each sub-sample with size ($n - 1$).

2.4 Model Selection

When we obtain several candidate statistical models. It is necessary to select one if we want to know which model is better. Under this condition, A suitable metric to measure the goodness of models is important. Likelihood can be regarded as one metric. However, it has some limitations in many cases. Therefore statisticians often utilize the adaption of likelihood: AIC and BIC.

Akaike information criterion (AIC)

AIC is presented in information theory. When a process that generates data is represented by a statistical model, the representation will never be exactly precise. It means that when the model is employed to represent the process, some information will be lost. AIC estimates the relative information lost by a given model: the less information a model loses, the higher the quality of that model. It is definition is defined as:

$$AIC = 2k - 2\log \hat{L}$$

where k is the number of parameters in the model and $\hat{L}$ represents the maximum value of likelihood function for the model.

AIC measures the goodness of fit of models and it includes a penalty term for the number of parameters because increasing number of parameters can cause overfitting.

log-likelihood ratio test (LRT)

LRT is a statistical test used to select models when one is a special case of another. The test is based on the likelihood ratio, which expresses that how many times more likely the events occur under one model than the other. The likelihood ratio is defined as :

$$R(\mathbf{x}) = \frac{L(\boldsymbol{\theta_0}|\mathbf{x})}{L(\boldsymbol{\theta_1}|\mathbf{x})}$$

where $\boldsymbol{\theta_0}$ is a set of restricted parameters, $\boldsymbol{\theta_1}$ is a set of unrestricted parameters. $\boldsymbol{\theta_1}$ has more parameters than $\boldsymbol{\theta_0}$. When the logarithm form of the likelihood ratio is utilized, the test is called LRT. If this ratio value is larger than a critical value, the model with $\boldsymbol{\theta_0}$ has better goodness of fit. Otherwise, the model with $\boldsymbol{\theta_1}$ is better.

Chapter 3

Illustrating Example

In order to make further inferences and comparisons, we employ our method to the breast cancer data. These data can be found in R package ”Rsmcure”. The raw data consists of 248 patients with their age and gender and other information. As for illustration, we pick the control group from the data set, consisting of 140 patients.

In the beginning of using EM Algorithm, we set the initial values by reasonable assumption. After finishing modeling, we found that uncure rates for almost all patients in treatment unfavorable group were close to 100%. Therefore, we set cure rate in this group a fixed value 0, which could reduce the estimation of one parameter in terms of cure rate. Therefore in this updated model, the expression of uncure rate of ith patient can be:

$$\pi_i = \begin{cases} 1 & g_i = 0 \\ \dfrac{1}{1 + \exp(\beta_{10} + \beta_{11} \cdot age_i + \beta_{12} \cdot sex_i)} & g_i = 1 \end{cases}$$

Where β_{10}, β_{11} and β_{12} are the components of the vector $\boldsymbol{\beta_1}$, age_i and sex_i are the age and gender for ith patient respectively.

Furthermore, from our results, it seemed that age had no significant influence on division of groups. In order to make our model more concise and efficient, we delete this factor and only consider the effects of age in estimation of grouping probability. The probability of division group for ith patient is shown below:

$$\alpha_i = \begin{cases} \dfrac{1}{1 + \exp(\theta_0 + \theta_2 \cdot sex)} & g_i = 0 \\ \dfrac{\exp(\theta_0 + \theta_1 \cdot sex)}{1 + \exp(\theta_0 + \theta_1 \cdot sex)} & g_i = 1 \end{cases}$$

Therefore, the parameter vector $\boldsymbol{\theta} = (\theta_0, \theta_1)$.

In conclusion, the final model is shown as:

$$P(g_i = 1) = \frac{\exp(\theta_0 + \theta_1 \cdot sex)}{1 + \exp(\theta_0 + \theta_1 \cdot sex)}$$

$$S_{i,1}(t) = [S_0(t)]^{\beta_{21} age_i + \beta_{22} sex_i}$$

$$S_{i,2}(t) = [S_0(t)]^{\beta_{21} age_i + \beta_{22} sex_i + \lambda}$$

$$P(c_i = 1 | g_i = 0) = 0$$

$$P(c_i = 0 | g_i = 1) = \frac{1}{1 + \exp(\beta_{10} + \beta_{11} \cdot age_i + \beta_{12} \cdot sex_i)}$$

Then We apply Peng and Dear's model on this dataset to compare with our approach. Here we utilize AIC and LRT. For LRT, following Wilks' theorem, we chose statistics of chi-squared distribution with 3 degrees of freedom and P-value 0.05. The results are shown in Table 3.1. In terms of both metrics, our model achieves better performance than previous one, showing the effectiveness of our method.

Table 3.1: Comparing Two Models with Two Metrics

	Peng's model	Proposed
AIC	1090.103	1083.046
LogLR Test	Ratio	13.057
	Critical Value	7.810
	P-value	0.05

Then we employ jacknife to estimate the inference of parameters in our model, shown in Table 3.2, except intercepts because these parameters provide little information for us and we don't want to consider them to a large extent. It can be seen that all confidence intervals do not cover 0, so we can confidently deduce that all parameters have significant meaning. Furthermore, θ_1 shows that sex is an important factor when dividing the group because males tends to be in the treatment favorable group. From $\hat{\beta}_{11}$, $\hat{\beta}_{12}$ reveal that in treatment favorable group, it is more possible for older people or females to be cured.

In Figure 3.1, there is a set of Kaplan-Meier survival curves of the last failed person in our dataset. Treatment favorable, unfavorable curves show the survival trend when the patient in either group. The unconditional curve estimates the unconditional survival rate and No subgroup curve is from Peng's model and the two curves are very similar. It can be seen that there is a huge discrepancy between treatment favorable and unfavorable and it can tell us the division of subgroup can make better treatment effects in this case.

Table 3.2: Estimates of Parameters and Their Biases and Variances

Parameter	Estimate	Variance	Confidence Interval
θ_0	-0.5249	0.0015	(-0.5585,-0.5061)
θ_1	1.1142	0.1371	(1.0431,1.1638)
$\hat{\beta}_{10}$	1.7834	0.0260	(1.7513,1.8455)
$\hat{\beta}_{11}$	0.0482	0.0016	(0.0441,0.0514)
$\hat{\beta}_{12}$	-3.1186	0.0689	(-3.1902,-2.9842)
$\hat{\beta}_{21}$	-0.0268	0.0006	(-0.0283,-0.0259)
$\hat{\beta}_{22}$	2.3799	0.0323	(2.3585，2.4349)
$\hat{\lambda}$	-3.6566	0.0467	(-3.7654,-3.6190)

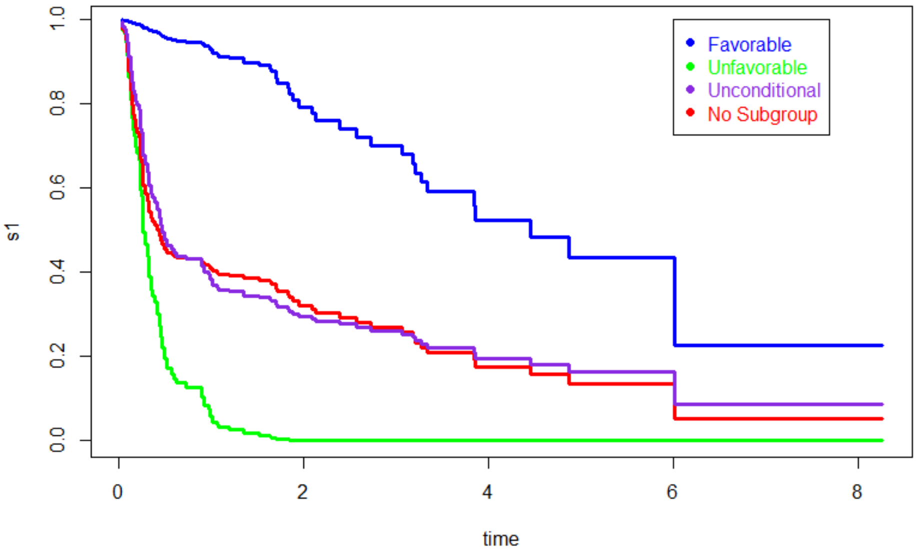

Figure 3.1: Survival Curve for One Patient

Chapter 4

Discussion and Feature Work

In this project, we propose a nonparametric mixture model with two subgroups to analyze cure rate. However, a limitation of our model is that we only consider two conditions in terms of treatment effects. There are more status for patients in real life. Further study is needed to determine the number of subgroups. Furthermore, in our research, the way of setting initial value is fixed and it is depending on the data. We can consider to use other methods, like grid search, to select initial value.

Bibliography

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

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

[3] J. L. Haybittle, "A two-parameter model for the survival curve of treated cancer patients," *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 16–26, 1965.

[4] J. Zhang and Y. Peng, "Accelerated hazards mixture cure model," *Lifetime data analysis*, vol. 15, no. 4, p. 455, 2009.

[5] N. Balakrishnan and S. Pal, "Expectation maximization-based likelihood inference for flexible cure rate models with weibull lifetimes," *Statistical methods in medical research*, vol. 25, no. 4, pp. 1535–1563, 2016.

[6] E. Z. Martinez, J. A. Achcar, A. A. Jácome, and J. S. Santos, "Mixture and non-mixture cure fraction models based on the generalized modified weibull distribution with an application to gastric cancer data," *Computer methods and programs in biomedicine*, vol. 112, no. 3, pp. 343–355, 2013.

[7] V. T. Farewell, "The use of mixture models for the analysis of survival data with long-term survivors." *Biometrics*, vol. 38, no. 4, p. 1041, 1982.

[8] V. T. Farewell, "Mixture models in survival analysis: Are they worth the risk?" *Canadian Journal of Statistics-revue Canadienne De Statistique*, vol. 14, no. 3, pp. 257–262, 1986.

[9] A. B. Cantor and J. J. Shuster, "Parametric versus nonparametric methods for estimating cure rates based on censored survival data," *Statistics in Medicine*, vol. 11, no. 7, pp. 931–937, 1992.

[10] M. E. Ghitany, R. A. Maller, and S. Zhou, “Exponential mixture models with long-term survivors and covariates,” *Journal of Multivariate Analysis*, vol. 49, no. 2, pp. 218–241, 1994.

[11] J. Denham, E. Denham, K. Dear, and G. Hudson, “The follicular non-hodgkin’s lymphomas——i. the possibility of cure,” *European Journal of Cancer*, vol. 32, no. 3, pp. 470–479, 1996.

[12] K. Yamaguchi, “Accelerated failure-time regression models with a regression model of surviving fraction: An application to the analysis of permanent employment in japan,” *Journal of the American Statistical Association*, vol. 87, no. 418, pp. 284–292, 1992.

[13] Y. Peng, K. B. G. Dear, and J. W. Denham, “A generalized f mixture model for cure rate estimation,” *Statistics in Medicine*, vol. 17, no. 8, pp. 813–830, 1998.

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

[15] J. M. G. Taylor, “Semi-parametric estimation in failure time mixture models,” *Biometrics*, vol. 51, no. 3, pp. 899–907, 1995.

[16] P. M. Rothwell, “Subgroup analysis in randomised controlled trials: importance, indications, and interpretation,” *The Lancet*, vol. 365, no. 9454, pp. 176 – 186, 2005.

[17] Y. Peng, J. M. Taylor, and B. Yu, “A marginal regression model for multivariate failure time data with a surviving fraction,” *Lifetime data analysis*, vol. 13, no. 3, pp. 351–369, 2007.

[18] L. Altstein and G. Li, “Latent subgroup analysis of a randomized clinical trial through a semiparametric accelerated failure time mixture model,” *Biometrics*, vol. 69, no. 1, pp. 52–61, 2013.

[19] J. Shen and X. He, “Inference for subgroup analysis with a structured logistic-normal mixture model,” *Journal of the American Statistical Association*, vol. 110, no. 509, pp. 303–312, 2015.

[20] R. Wu, M. Zheng, and W. Yu, “Subgroup analysis with time-to-event data under a logistic-cox mixture model,” *Scandinavian Journal of Statistics*, vol. 43, no. 3, pp. 863–878, 2016.

[21] H. J. Kim, B. Lu, E. J. Nehus, and M. O. Kim, "Estimating heterogeneous treatment effects for latent subgroups in observational studies," *Statistics in medicine*, 2018.

[22] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society series b-methodological*, vol. 39, pp. 1–38, 1976.

[23] J. D. Kalbfleisch and R. L. Prentice, "Marginal likelihoods based on cox's regression and life model," *Biometrika*, vol. 60, no. 2, pp. 267–278, 1973.

[24] N. E. Breslow, "Covariance analysis of censored survival data." *Biometrics*, vol. 30, no. 1, pp. 89–99, 1974.