Xi’an Jiaotong-liverpool University

Department of Science

MTH301

Final Year Project

REIMPLEMENT MIXTURE CURE MODELS in ONCOLOGY with R

用 R 重新实现肿瘤学中的混合治疗模型

Student name: Zhaobing.Fu

Student ID: 1824353

Supervisor: Mu He

Date:4/Mar/2022

Contents

List of Figures

Abstract

The mixed cure model divides the patients into cured group and uncured group by cure rate, then calculates the survived individuals in the uncured group by the survival probability, reflected in plateaus in overall Kaplan-Merier curves. In this paper, the data ”brim3_simulated.csv” is collected by the data collation method compiled by Felizzi et al in 2021. Then combing the KM survival curve obtained by using survfit function and the mixed cure model estimated by the smcure package in R to analysis the data.

摘要：混合治愈模型是将经过治疗的患者用治愈概率分成已治愈组和未治愈组，再用生存概率计算未治愈组中生存下的个体，在 KM 模型中显示为平台期。本文应用了 Felizzi 等人在 2021 整理的整理数据的方法，收集了数据 “brim3_simulated.csv”。然后运用 survfit 方程得到 KM 生存曲线，并联合 R 的估计混合治愈模型 smcure 包来分析数据。

Keywords: survival analysis, EM algorithm, likelihood function, Proportional hazards model, mixture cure model, Oncology, R package

Introduction

Survival analysis aims to study the time to occurrence of the topic and influential factors. The events usually refer to failure, recurrence or death. Survival function and hazards function are basic concepts in survival analysis. The two functions and the corresponding probability density function have relationships which means if one of them is available, you can obtain the other two. Survival analysis usually constructs the models describing when the event occurs. In other words, survival time. One of the most popular models is **Cox proportional hazard model(PH)** which is semi-parametric that estimates the effect of interested covariates under PH assumption without using specified baseline function. However, according to Liu and Liao (2020) [12], to analyze the data in clinical study and medical research, parametric models are chosen to estimate risk and the number of survivors in treatment groups. While not all the parametric models are PH models, many of them are **Acceleration Failure time models(AFT)**. Both of them are well-known in survival analysis models.

Additionally, Liu and Liao state that if the experimenters prefer a smother estimation and tend to avoid the result like a step functions, they may choose parametric models instead of non-parametric ones. Although parametric models have specific baseline hazard functions and more stricter assumption, they are not capable to handle complicated survival functions in clinical and medical applications. What Liu and Liao are trying to say in 2020 [12], non-parametric modes are more flexible to solve his problem.

One usual assumption in survival analysis is that all individuals would experience the interested event if the follow-up time is long enough. However, as the advancement of medical technology, more diseases are curable so we need to estimate the survival time. In order to solve the problem, Boag(1949) [2] firstly introduced the original definition of the cure rate model. The model has been improved by Berkson and Gag in 1952 [1]. They divide the studied population into two groups, susceptible individuals who may experience the event and nonsusceptible individuals who are cured and never experience the event in long-term follow-up. The model aims to study the cure rate and survival function which are named by incident and latency. In 1982, Farewell [6] proposed a mixture cured model consisting of the binary distribution for latency and Weibull distribution for the time to the target event. Kuk and Chen (1992) [11] constructed a mixture model which is the combination of the logistic model and proportional hazard model which is the generalized form of Farewell's model.

Literature Review

3.1 Introduction to Survival Analysis

Survival analysis is a method that considers both outcome and survival time and it can fully use by the censored data. It can finally estimate the distribution of time until the specific event happens and analyze different factors that may affect the results. There is three main types of regression model to estimate the time, nonparametric model, parametric model, and semiparametric model. Streib and Dehmer(2019) [5] state that two methods make a significant contribution in this field, Kaplan Meier estimator and Cox Proportional Hazards Model. Kaplan

Meier estimator is a nonparametric model and Cox PH is a typical semiparametric model. Both models will introduced in the this section.

3.1.1 Functions

Survival Function The probability that the event has not happens until time t

$$S(t) = 1 - F(t) = P(T \geq t)$$

Hazard Function Instantaneous failure rate: in the short interval $[t, t + \Delta t]$, the event happens given that the the event has not occur before time t.

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

Deviation Process:

$$\lambda(t) = \frac{1}{P(T \geq t)} \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t)}{\Delta t} = \frac{f(t)}{S(t)}$$

Probability density function the failure happens at time t

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t)}{\Delta t}$$

Cumulative Hazard Function

$$\Lambda(t) = \int_0^t \lambda(u) du$$

3.1.2 Transformation among the functions

$$S(t) = exp\left[-\int_0^t \lambda(u) du\right]$$

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

$$f(t) = \lambda(t) S(t)$$

$$S(t) = exp[-\int_0^t \lambda(t) dx] = exp^{-\Lambda(t)}$$

$$f(t) = \frac{dS(t)}{dt} \Leftrightarrow S(t) = \int_t^\infty f(u) du = exp\left[-\int_0^t \lambda(u) du\right] \Leftrightarrow \lambda(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} log(S(t))$$

3.1.3 Data Types for Survival Analysis

Complete data In the research process, complete data X is the survival time of a research object or the specific time of endpoint event that can be observed and recorded. Survival time refers to the period experienced from the specified observation starting point (initial event) to the occurrence of a specific endpoint event.

Censored and Truncated In contrast to the complete data, at the end of the study, if at the endpoint the event cannot be recorded clearly due to the occurrence of other events or survive in the study, we call this type of data censored data(C_r) or **incomplete data**. In survival analysis, censored and truncated are two common sorts of data categories.
There are three main reasons for censoring:

1. When the study ends, no event happened since then.
2. Because of the loss of connection or other reasons, the subjects dropped out of the study and could not continue the follow-up observation
3. The subjects died from other events, such as traffic accidents, or other diseases.

Censoring Assumption Classifying by Kleinbaum and Klein(2012) [10], there are three assumptions about censoring: independent, random, and non-informative. The independent assumption is the most useful one which affects the feasibility to compare the survival experience of two or more groups. For example, the treatment group and placebo groups. Random censoring is a stronger and more restrictive assumption than independence. On the other hand, the non-informative assumption means the distribution of survival times (F) provides no information about the distribution of censorship times (C), and vice versa.

Right censoring we can be sure that failure will occur some time after the censoring time(C_r) and it is perhaps the most commonly sort of censoring. In this report, all the censor property only refers to right-censor. It means the event that has not happen in the experience but can not be guarantee that they won't happen in the future. In other words, $T = min(X, C_r)$

Left censoring Suppose that the research object enters the study for observation at a certain time, but before this time point(C_r), the time point of interest in the study has already happened, but the specific time cannot be specified. $T = max(X, C_l)$

Interval censoring Similar to other censored type, interval censoring refer to the loss of subjects in a interval and failure occurs in this interval (L_i, R_i), so we cannot know exactly when failure occurs.

Truncation Truncated is another critical data property. It refers to the omission of samples that meet the conditions. This section will introduce manifold categories of censoring and truncation except for right censoring. Set a observation window (Y_L, Y_R). If all the observations on the (Y_L, ∞), it will be called left-truncation. Otherwise, it is right-truncation. The conditional probability of left-truncation would be $P(X|X > Y_L)$. For left-censoring, we know that the failure happens before C_l. However, when it comes to left-truncation, we never collect data before Y_L.

Likelihood construction for censored and truncated data Assuming that the lifetime and censored time are independent of each other. Exact event time observations tell us the time when an event occurs. Such data is most useful for estimating the overall event distribution. However, other truncated data can also provide partial information, and it is wasteful to throw it away directly. The following table is constructed by Klein and Moeschberger (1997) [9] showing the censoring schemes and the Likelihood Function. Note that C_r and C_l are right and left censoring times for censored individuals and Y_l and Y_r are the left and right boundaries of follow-up time for truncated data. The likelihood function of censored data.

Censoring Scheme	Likelihood Contribution
Exact lifetime	$f(t)$
Right-censoring	$S(C_r)$
Left-censoring	$1 - S(C_l)$
Interval-censoring	$S(C_r) - S(C_l)$
Right-truncation	$f(t)/[1 - S(Y_r)]$
Left-truncation	$f(t)/[S(Y_l)]$
Interval-truncation	$f(t)/[S(Y_l) - S(Y_r)]$

Table 1: Likelihood Function of incomplete data

$$\mathcal{L} \propto \prod_{i\in F}^{n} f(x_i) \prod_{i\in R}^{n} S(C_r) \prod_{i\in L}^{n} 1 - S(C_l) \prod_{i\in I}^{n} [S(C_r) - S(C_l)] \tag{3.1}$$

The difference between truncation and censoring is that we can give some censored individuals partial information about the time of their event, while truncation is a property that restricts our observations to a part of the target population and subjects whose activity time meets certain criteria. An example criterion is like: living longer than a certain age.

3.2 Survival Model

Proportional Hazard Model Regression is the most basic problem in statistics, and almost all statistical problems can be regarded as special kinds of regression. For example, generalized linear models are dealing with regression; Survival analysis deals with a special kind of Y with the censorship. Time series can also be viewed as a special kind of time-based regression. In survival analysis, the simplest and most basic regression model is the **Proportional Hazard Model** (PH Model). In order to introduce Proportional Hazard Model, we need to start from the most common linear regression: $Y_i = x_i'\beta + \epsilon$. Instead of modeling for y, we need to build for hazard function $\lambda(t)$ since $\lambda(t) \geq 0$, so we use the exponential form in the modeling process:

$$\lambda(t, X) = \lambda_0(t)\ exp\left[\sum_{i=1}^{p} \beta_i X_i\right] = \lambda_0(t) exp(X'\beta) \tag{3.2}$$

Accelerated Failure Time Model In this paragraph, I will introduce the knowledge structured by AFT Saikia and Barman [14]. Contracted to the PH Model, which is based on Hazard, the AFT Model is based on survival time T. Assuming that $Y_i = log(T_i)$, the regression model of failure time T will be $Y_i = x_i'\beta + W_i$ (W_i's are i.i.d residuals). By simple transformations, we obtain $T_i = T_0\ exp(x_i'\beta)$ and $T_0 = exp(W)$. When the j-th variable changes Δ_j, the survival time will change $exp(\Delta_j\beta_j)$. For example, AFT models compare survival functions between treatment($S_0(t)$) and placebo group($S(t)$). The AFT assumption expressing the survival function and hazard function of both groups is

$$S(t, X) = P(T_i \geq t) = P(T_0 \geq exp(-x_i'\beta)t) = S_0(exp(-X'\beta)t) \tag{3.3}$$

$$\lambda(t, X) = exp(-X'\beta)\lambda_0(exp(-X'\beta)t) \tag{3.4}$$

$$f(t, X) = f_0(exp(-X'\beta)t)exp(-X'\beta) \tag{3.5}$$

Let acceleration factor $\gamma = exp(-X'\beta)$, comparing survival time for patients in the treatment group and placebo group. If $exp(X'\beta) < 1$, the multiplicative effect of covariates in X is decelerated. On the contrary, if $exp(X'\beta) > 1$, the multiplicative effect of covariates in X is accelerated. Among $Y_i = x_i'\beta + W_i$, the baseline survival time t of the population obeys a probability distribution, and W is another random variable obeying a probability distribution. The common distribution is as follows:

t	W
Weibull	Extreme Value
log-normal	Normal
log-logistic	Logistic

The relationship between PH and AFT The most difference between the two model is that AFT is a method modeling based on time while PH based on hazards.

Comparison between PH and AFT Model Assumption Kleinbaum and Klein [10] state that PH model is applied when comparing hazards and it can describes the multiplicative effect with hazard. When the model is parametric, they may not fit PH assumption any more. Many parametric models are AFT model, which can be used comparing survival times and PH model describes the **multiplicative effect with survival time**.
The hazard function for PH(3.2) and AFT model(3.4):
PH Model:

$$\lambda(t, X) = \lambda_0(t)\ exp(X'\beta)$$

AFT Model:

$$\lambda(t, X) = exp(-X'\beta)\ \lambda_0(exp(-X'\beta)t)$$

Intuitively, the larger $exp(X'\beta)$ is, the longer the survival time will be. Since S(t) is a monotone decreasing function, the survival curve will be higher. Since $T_i = exp(X'\beta)T_0$,

Logarithmically:
PH Model:

$$log\lambda(t, X) = log\lambda_0(exp(log\ t)) + X'\beta$$

AFT Model:

$$log\lambda(t, X) = log\lambda_0(exp(log\ t - X'\beta)) - X'\beta$$

If the baseline model is $log\lambda_{base} = log\lambda_0(exp(log\ t)) - X'\beta$. PH Model can be regarded as the upward and downward translation of the benchmark Model, while AFT Model can be regarded as a left and right translation. Only when the reference model is linear, the model after left-right translation and up-down translation can coincide. From the comparison between Weibull distribution and exponential distribution in the previous section, we know that Weibull distribution is a linear model, so Weibull regression is the only one that satisfies both AFT and PH assumptions.

Semi-parametric ⇒ parametric model Peng and Yu(2021)[15] introduce the latency part needs to use the assumption of PH and AFT. The parametric function needs to specify the baseline partition. Operating survival PH function as an example:

$$S(t|x) = S_0(t)^{exp(-X'\beta)}$$

Let $S_0(t) = e^{-\lambda t}$ (exponential distribution with rate λ) ⇒ $S(t) = e^{-\lambda exp(-X'\beta)t}$ (exponential distribution with rate $\lambda \times exp(-X'\beta)$)
Let $S_0(t) = e^{-\lambda t^p}$ (Weibull distribution with shape parameter p and scale parameter λ) ⇒ $S(t) = e^{-\lambda exp(-X'\beta)t^p}$ (Weibull distribution with shape parameter p and scale parameter $\lambda \times exp(-X'\beta)$)

3.2.1 Parametric Survival Models

The section elaborates on the parametric models of time to occurrence of the event which is the specific distribution with unknown parameters. In contrast to semi-parametric models, the parametric models have specified $\lambda_0(t;S)$ for hazard function as well as $S_0(t;S)$ for survival function. The following sections will provide the different parametric models. The section will list parametric models substitute the sepcific $\lambda_0(t;S)$ into PH or AFT models.

Exponential Model Let $\lambda_0(t, S) = 1/S$ and obtain the simplest parametric survival model, the **exponential regression model** and it indicates the density function of survival data is also an exponential distribution. The exponential distribution was the simplest distribution utilized to model lifetime data. It is a model for the life of products with a constant failure rate with memoryless property. Consequently, it is inappropriate that employ the model in the actual survival applications. It has only parameter S given k =1 in the Weibull distribution.

$$f(t|S) = \frac{1}{S}exp\left(-\frac{t}{S}\right), t \geq 0 \tag{3.6}$$

$$S_0(t;S) = exp\left(-\frac{t}{S}\right) \tag{3.7}$$

$$\lambda_0(t;S) = 1/S \tag{3.8}$$

Scale Parameter - S The scale parameter adjusts the scale of the density function along the time axis. Therefore, the transition of this parameter has the same influence as the change in time scale.

Exponential PH Model The exponential PH model will be $\lambda(t,X) = \lambda_0(t;S)exp(X'\beta)$. Since $exp(X'\beta)$ exist the intercept term β_0 which have the same effect as $\lambda_0(t;S)$. Therefore, we need to reduce the model into $\lambda(t,X) = exp(X'\beta)$.

Exponential AFT Model Similar to exponential PH model, set $\lambda_0(t;S) = 1/S$. A

$$\lambda(t,X) = exp(-X'\alpha)/S$$

After simplified the intercept term, we will get the exponential PH model in case of hazard function:

- $\lambda(t,X) = exp(-X'\alpha)$
- S(t) $= exp(-\frac{1}{S}t)$
- $t = [-ln(S(t))] \times S = [-ln(S(t))] \times exp(-X'\alpha)$ seting $S = exp(-X'\alpha)$.

Comparison between exponential PH model and AFT model :
Exponential AFT Model ⇔ Exponential PH Model

- Using maximum likelihood estimation to obtain the value of parameter(R: summary()).The distribution is assumed normally distributed.
- There is a relation between two models: $\beta_j = -\alpha_j$

HR > 1	exposure harmful survival	$\gamma > 1$	exposure benefits survival
HR = 1	no effect	$\gamma = 1$	no effect
HR < 1	exposure benefits to survival	$\gamma > 1$	exposure harmful survival

Table 2: general factor comparison between PH and AFT model

Weibull Model Let $\lambda_0(t,S) = kt^{k-1}/S^k$, then it will obtain the Weibull regression model. Weibull regression is the only regression that simultaneously satisfies the assumptions of both PH and AFT models. In order to obtain the meaning of the model, the derivation process is followed.

Derivation process for CDF of Weibull distribution : $P(T < t) = F(t)$ the probability that the event will happen before t. Therefore, the survival function $S(t) = 1 - F(t) = P(T \geq t)$. Assuming that if one people does not experience until t $\rightarrow \infty$ so do all n people. Total number of n people are probability independent.

$$1 - P_k = (1 - P)^k$$

Let $F(t) = 1 - e^{-a(t)}$ with the purpose of get the simplest form for n person's failure rate

$$1 - F(t) = 1 - P_k = (1 - P)^k = e^{-ka(t)}$$

$$a(t) = \left(\frac{t - D}{S}\right)$$

$$F(t) = 1 - exp\left[-\left(\frac{t-D}{S}\right)^k\right]$$

If $D = 0$ that will lead to two parameter Weibull distribution:

$$f(t|S,k) = \frac{k}{S}\left(\frac{t}{S}\right)^{k-1} exp\left[-\left(\frac{t}{S}\right)^k\right], t \geq 0 \tag{3.9}$$

$$S_0(t|S,k) = exp\left[-\left(\frac{t}{S}\right)^k\right] \tag{3.10}$$

$$\lambda_0(t|S,k) = \frac{kt^{k-1}}{S^k} \tag{3.11}$$

Scale Parameter - S The shape parameter controls the overall shape of the density function that ranging between 0.5 and 8.0. The estimated standard errors and confidence limits displayed by the program are only valid when B > 2.0.

Shape Parameter - k The shape parameter controls the tread of hazard function. When k>1, the hazard function will be increase and convex. When k<1, the hazard function will be decrease and concave. If k = 0, the hazard function will be constant and the Weibull model will reduce to **exponential model**.

Threshold Parameter - D This parameter sets the minimum time for the model.

Weibull PH Model Here we emphasize that the $X'\beta$ contains no intercept term, otherwise, the problem would still be unrecognized. Similar to the exponential distribution, we can reduce it to the following form

- $\lambda(t) = [kt^{k-1}/S^k]exp(X'\beta)$

k > 1	time ↑ Hazard ↑
k = 1	constant Hazard
k < 1	time ↑ Hazard ↓

Table 3: monotonic of Weibull Hazard

Weibull AFT Model: By equation 4.18, here is the expression of t:

$$\begin{aligned} t &= [-lnS(t)]^{1/k}) \times S \\ &= [-lnS(t)]^{1/k}) \times exp(\alpha_0 + \alpha_1 X_1) \end{aligned} \tag{3.12}$$

- $S = exp(\alpha_0 + \alpha_1 X_1)$
- The accelerator factor $\gamma = exp(\alpha_1)$ is direct effect of an exposure which depends on survival time.
- Weibull Linearity transformation: $S(t) = ln[-ln(S(t))] = k[ln(t) - ln(S)]$.The parameter k is the slope term of while ln(S) is the intercept term.

Comparison between PH model and AFT model:

- It is a unique property that Weibull AFT Model ⇔ Weibull PH Model
- There is a relation between two models: $\beta_j = -\alpha_j p$

3.2.2 Nonparametric Survival Models

Kaplan-Meier Estimator In 1958, Kaplan and Meier have introduced a nonparametric statistics to estimate the survival function which have censored data. r_j are element in the risk set $R(t_j)$, which is the collection of survival and uncensored individuals. And each individuals in the risk set survive longer than t_j. The variable d_j are number of failure subject in the time interval $[t_{j-1}, t_j)$.

$$\hat{S}(t_{(j)}) = \hat{S}(t_{(j-1)}) \times \hat{P}(T > t_{(j)} | T \geq t_{(j)}) = \prod_{i=i}^{j-1} \hat{P}(T > t_{(i)} | T \geq t_{(i)})$$

$$\hat{S}(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{r_j}\right)$$

When experience hasn't start, $\hat{S}(0) = 1$. The long plateau stands for the cured patients after long follow-up time.

Nelson-Aalen Estimator Nelson-Aalen Estimator is a non-parametric estimator to estimate the cumulative hazard rate function from censored survival data. it can be used to check if the parametric models graphically appropriate.

$$\hat{\Lambda}(t) = \sum_{t_j \leq t} \left(1 - \frac{d_j}{r_j}\right)$$

Then the survival sunction is: $S(t) = e^{\Lambda(t)}$

3.2.3 Semi-parametric Model

Parametric models have higher requirements on model assumptions, in other words, they are not robust enough. So we want to lower the model assumptions appropriately. At the same time, if we want to preserve some of the explanatory of the model, we also need to preserve parameters appropriately. This is called a semi-parametric model. The most common and important model in survival analysis: Cox regression model. Cox proportional hazard model has both the advantage of parametric and nonparametric model.

The basic assumption for Cox Regression

$$\lambda(t, X) = \lambda_0(t)\ exp\left[\sum_{i=1}^{p} \beta_i X_i\right]$$

In Cox PH model, we classify the influence of explanatory variables into the parameter part, and the specify $\lambda_0(t)$ form of the model into the non-parameter part. Xi = $(X_{i1},\ X_{i2}\ ,\cdots,\ X_{ip})$ are explanatory variables. $log\lambda(t, X)$ changes linearly with $beta$'s.

Assumption

- The hazard ratio is assumed to remain the same for the entire follow-up and independent to time t. e^{β} is the hazard ratio which means the percentage changes for per unit change in X.
$$\frac{\lambda(t|X=x+1)}{\lambda(t|X=x)}=e^{\beta}$$
- The survival time of every patients are independent.
- the censoring is uninformative about the target outcome. Let us write it in a formula: $T_i \perp C_i|X_i$

Advantages of PH Model

- It is a semiparametric and the baseline function is arbitrary unspecified.
- If the model met the Cox PH assumption, then we will get robust result.

Hazard Ratio Here is the expansion form of HR:

$$HR=\frac{\lambda(t,X^*)}{\lambda(t,X)}=exp\left[\sum_{i=0}^{p}\beta_i(X_i^*-X_i)\right]$$

β is referred to the hazard ratio (HR). (X_i^*: group with larger hazard; X_i: group with larger hazard). β is the log hazard ratio (HR).Therefore, for each unit of increase in X:

$$\log\left[\frac{\lambda(t,X^*)}{\lambda(t,X)}\right]=\sum_{i=0}^{p}\beta_i(X_i^*-X_i)$$

Transformation model: The fomula $\lambda(t,X)=\lambda_0(t)\ e^{X'\beta}$ inplies the survival function:

$$S(t,X)=S_0(t)^{exp(X'\beta)}$$

Then transform the survival function into a linear form:

$$\log[\log\{S(t,X)\}]=\log[\log\{S_0(t)\}]+X'\beta$$

$$g\{S(t,X)\}=g\{S_0(t)\}+X'\beta$$

The function $g():(0,1)\to(-\infty,\infty)$ is a smooth monotone function.

- There are two unknown parameters: $\{\beta,\lambda_0(t)\}$
- The likelihood function is:
$$\mathcal{L}(\beta,\lambda_0(t))=\prod_{i=1}^{n}\lambda(X_i)^{\delta_i}S(X_i)=\prod_{i=1}^{n}\lambda_0(t_i)^{\delta_i}e^{\delta_i X'\beta}e^{-\int_0^{t_i}\lambda(u)du}$$

Note that the three likelihood form for the Cox PH model are shown in the section 4.1.

3.3 Logistic Model

Logistic model is a relationship between predictor values and categorical response variable. If the response is binary(if Y = 1 or Y = 0), we will use the binary logistic regression models.

$$P(Y=1)=\pi(z)=\frac{exp(Z'b)}{1+exp(Z'b)}, Z'=[z_1,z_2,...,z_n]'$$

Note that $\pi(z)$ is stand for the "success probability", which means the probability of an observation of the specific category. For n sample set, the likelihood function of the likelihood function of the binary logistic regression is shown following:

$$\begin{aligned}\mathcal{L}(b;y,Z)&=\prod_{i=1}^{n}\pi(z_i)^{y_i}[1-\pi(z_i)]^{1-y_i}\\&=\prod_{i=1}^{n}\left(\frac{exp(z_i'b)}{1+exp(z_i'b)}\right)^{y_i}\left(\frac{1}{1+exp(z_i'b)}\right)^{1-y_i}\end{aligned}\tag{3.13}$$

The log-likelihood is:

$$\begin{aligned}l(b)&=\sum_{i=1}^{n}(y_i log\ \pi(z_i)+(1-y_i)log[1-\pi(z_i)])\\&=\sum_{i=1}^{n}(y_iz_ib-log(1+exp(z_ib))\end{aligned}\tag{3.14}$$

Metholody

Let the joint distribution of $X_1,X_2,...,X_n$ is $f(X_1,X_2,...,X_n|\theta)$ and give each of the random variable a observed value, $X_1=x_1,X_2=x_2,...,X_n=x_n$. θ is the collection of unknown parameter that effect the contribution of different factors $X_1,X_2,...,X_n$. The latent variable Z_i's are the same. X and Z are the data set of X_i and Z_i.

4.1 Maximum Likelihood Estimation

The most common way to estimate the optimal parameter space θ is finding the MLE by differentiating $ln\mathcal{L}(\theta)$ and let equal to 0. Sometimes, some complicated likelihood function so I need to introduce two iterative procedure to obtain a stable parameter value, Newton-Raphson algorithm and EM algorithm.

Likelihood Likelihood function is very important in statistical inference and it has similar meaning as probability. Probability means forecasting the possibility of the results with known parameters. Likelihood, using the existing outcome $x_1,x_2,...,x_n$, estimates the probability of a specific parameter values. In other words, likelihood function is the probability density function of parameters. Therefore, likelihood function can be seen as the opposite of conditional probability, which can be written as $P(A|B)=\frac{P(A,B)}{P(B)}$. According to Bayes' theorem,

$$P(B|A)=\frac{P(A|B)P(B)}{P(A)}$$

The opposite step to construct a likelihood is:

We need to apply $\mathcal{L}(B|A)$ to estimate the parameter B(β) in condition that the event $\mathbf{A}(x_1, ..., x_n)$ happened. However, different from the aim conditional probability, we usually focus on the parameter b. For all $\alpha > 0$ we will have likelihood function $L(b|A) = \alpha P(A|B = b)$. Note that we are not assumed that the likelihood function to satisfy $\sum_{b \in B} P(A|B = b) = 1$. (Question: α = P(B)/P(A))

4.1.1 Likelihood Function

Let the probability distribution function for x_i is have a parameter set θ is $f(X|\theta)$. The joint probability distribution for parameter θ.

$$f(X_1, X_2, ..., X_n|\theta)$$

To obtain the likelihood function , replace each random variable X_i with a specific value x_i .

$$\mathcal{L}(\theta|x) = \mathcal{L}(\theta|x_1, x_2, ..., x_n) = f(x_1, x_2, ..., x_n|\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

One commonly used process of MLE:

$$\hat{\theta}_{MLE} = \max_{\theta} l(\theta|x) \tag{4.1}$$

$$l(\theta|x) = \left(\prod_{i=1}^{n} f(x_i; \theta)\right) = \sum_{i=1}^{n} lnf_i(x_i; \theta) \tag{4.2}$$

$$\frac{\partial ln\mathcal{L}}{\partial \theta_i} = 0 \ (i = 1, \ldots, p) \tag{4.3}$$

Example Using the likelihood function construction methods, we aim to obtain the likelihood function of the Cox PH model: On account of the all the data in this report is right-censored, the data set X_i will be divided into two categories:
When X_i is censored: According to table 1:

$$\mathcal{L}_j(\beta, \lambda_0(t)) = S(X_i)$$

When X_i is a observed:

$$\mathcal{L}_j(\beta, \lambda_0(t)) = f(X_i) = S(X_i)\lambda(X_i)$$

Therefore, the full likelihood is following:

$$\mathcal{L}(\beta, \lambda_0(t)) = \prod_{i=1}^{n} \lambda(X_i)^{\delta_i} S(X_i) = \prod_{i=1}^{n} \lambda_0(t_i)^{\delta_i} e^{\delta_i X'\beta} e^{-\int_0^{t_i} \lambda(u)du}$$

4.1.2 Partial Likelihood Function

If we are interest in the parameter set $\theta = [\alpha, \beta]$, we may obtain the likelihood function as:

$$\mathcal{L}(\alpha, \beta|x) = \mathcal{L}_1(\alpha|x) + \mathcal{L}_2(\beta|x) \tag{4.4}$$

When there are no relationship in the data set, the partial likelihood is effective, which means no two subjects have the same time. Otherwise, if there are relevance in the data set, the true partial log-likelihood function involves permutations and can be time-consuming to compute. In this case, the Breslow approximations to the partial log-likelihood can be used.

Example For logistic regression, the maximum of the likelihood function is able to estimate the Cox PH model parameter. For the Cox model, $\mathcal{L}(b)$ is a partial likelihood function we only consider the failure time of the susceptible individuals. b is the set of all the parameters of the covariates. Then the maximization process is:

- Partial likelihood are conditional probabilities of the product of observed failure times, seeing the observed failure, given the risk set at that time and a failure will occur. In other words, these are the conditional probabilities of the observed individual, chosen from the risk set. In summary, for failure time for j-th individual X_j, the contribution to the likelihood is

$$\begin{aligned}\mathcal{L}_j(b) &= P(individual\ j\ fails|\ one\ failure\ from\ R(X_j)) \\ &= \frac{P(individual\ j\ fails|\ at\ risk\ at\ X_j)}{\sum_{k\in R(X_j)} P(individual\ k\ fails|\ at\ risk\ at\ X_j)} \\ &= \frac{\lambda(X_j|Z_j)}{\sum_{k\in R(X_j)} \lambda(X_j|Z_k)}\end{aligned} \tag{4.5}$$

- Note that the procedure reasoning procedure applies the definition of hazard function $\lambda(t)$ and probability density function f(t): $k \in R(t_j)$ stands for the k-th survival individual at time point t_j-.In other word, $R(t) = \{i : X_i \geq t\}$. Therefore, $P(R(X_j)) = P(X_j \geq t) = S(t)$.

$$P(individual\ j\ fails|\ at\ risk\ at\ X_j) = \frac{the\ failure\ happens\ at\ time\ t}{S(t)} = \frac{f(t)}{S(t)} = \lambda(t)$$

In that case, the probability of death occurring at that time point is:

$$\frac{exp(X_j'b)}{\sum_{k\in R(t_j)} exp(X_k b)}$$

- Since the form of Cox PH model is not complicated , we can use derivatives to get the maximum value of the likelihood function.

1. form $\mathcal{L}$
2. maximize ln$\mathcal{L}$ (Solve iteratively $\frac{\partial ln\mathcal{L}}{\partial b_i} = 0(i = 1, \ldots, p)$)

Likelihood function:

$$\mathcal{L}(b|x_1, x_2, ..., x_n) = \prod_{j=1}^{n} \frac{exp(Z_j'\beta)}{\sum_{k\in R(t_j)} exp(Z_k\beta)}$$

.

4.1.3 Profile likelihood function

Similarly to the parameter set $\theta = [\alpha, \beta]$. However, the parameter α can be substitute by $\beta = g(\alpha)$. Then the likelihood function would be:

$$\mathcal{L}(g(\beta), \beta|x) \tag{4.6}$$

For example, if we want to get the likelihood function for normal distribution $X \sim N(\mu, \sigma^2)$. Then we can get the likelihood function: $\mathcal{L}(\mu|x) = \mathcal{L}(\mu, \widehat{\sigma_\mu}|x) = \mathcal{L}(\mu, \sum_i^n (x_i - \mu)^2/n|x)$.

4.2 Newton-Raphson Algorithm

The Newton-Raphson algorithm is a iterative method solving the equation with its derivatives which can be applied in calculating extremely complicated nonlinear functions obtaining from observed data. This section will start with maximizing the likelihood function which ia easiler to understand and the process can be visualized.

4.2.1 The Newton Raphson Algorithm to Maximum 1 Variable Likelihood Function

Taylor Series Approximation The first step of developing the Newton-Raphson algorithm is to approximate the likelihood function with a quadratic form which is easily to maximized. Therefore, the Taylor Series of the function f is

$$f(x+h) = f(x) + f'(x) + \frac{1}{2}f''(x)h^2 + \cdots$$

The first order of Taylor approximation of f is

$$f(x+h) \approx f(x) + f'(x)h$$

Similarly, there is the order of Taylor approximation has smaller h and more accurate approximation.

$$f(x+h) \approx f(x) + f'(x)h + \frac{1}{2}f''(x)h^2$$

To simplified the calculation in the next step, we need to rewrite the Taylor approximation equations.

$$f(x+h) \approx a + bh$$

$$f(x+h) \approx a + bh + \frac{1}{2}ch^2 \tag{4.7}$$

where $a = f(x)$, $b = f'(x)$ and $c = f''(x)$. The equation (4.4) is the second order polynomial of h.

Maximized the Second Order Polynomial After reduce the function in to a polynomial function, we can simulate the algorithm and generalized it into higher dimension. Recall the fomula 4.4

$$f(x+h) \approx a + bh + \frac{1}{2}ch^2 \tag{4.8}$$

$$f'(x+h) \approx b + ch \tag{4.9}$$

$$f''(x+h) \approx c \tag{4.10}$$

Let $b + c\hat{h} = 0$, we can get $\hat{h} = -\frac{b}{c}$ is the extreme point of f. In condition that $f''(x+h) \approx c < 0$, which means $f(-\frac{b}{c})$ would maximum. (Note that $b = f'(x)$ and $c = f''(x)$.) To sum up, the x value that maximizes the second order Taylor approximation of f:

$$\begin{aligned} x + \hat{h} &= x - \frac{b}{c} \\ &= x - \frac{1}{f''(x)}f'(x) \end{aligned}$$

General Form We can compute the MLE iteratively by

$$\widehat{\theta}^{k} = \widehat{\theta}^{k-1} + \frac{U(\widehat{\theta}^{k-1})}{I(\widehat{\theta}^{k-1})}$$

$U(\widehat{\theta}^{k-1})$ is the derivative of $\mathcal{L}(\theta)$ with respect to θ and $I(\widehat{\theta}^{k-1}) = -U'(\widehat{\theta}^{k-1})$.

4.3 EM Algorithm

When we try to get the θ which is the set all parameters, we need to use maximum likelihood estimation. However, we can get ideal solution in the model which have latent variables. Hence EM Algorithm has become a popular to estimate the parameter. Let L(θ) = log P(x|θ), the maximum likelihood function would be

$$\theta_{MLE} = log\ f(x|\theta)$$

X = $\{x_1, x_2, ..., x_n\}$ is the observation values and Z is latent variable. The folmula of EM algorithm is

E-step: get the value of $P(z|x, \theta^{(t)}$ and take it into

$$E_{z|x,\theta^{(t)}}[log\ P(x, z|\theta)]$$

M-step: ;Maximize the the formula above and get the

$$\theta^{(t+1)} = arg \max_{\theta} \int_{Z} E_{z|x,\theta^{(t)}}[log\ P(x, z|\theta)] = arg \max_{\theta} \int_{Z} P(z|x, \theta^{(t)})log\ P(x, z|\theta)dz$$

Equation Deducing Process

$$log\ P(x|\theta) = log\ P(x, z|\theta) - log\ P(z|x, \theta)$$

$$log\ P(x|\theta) = log\ \frac{P(x, z|\theta)}{q(z)} - log\ \frac{P(z|x, \theta)}{q(z)}$$

Get the expectation of both sides with q(z):

$$Left : \int q(z)\ log\ P(x|\theta)\ dz = log\ P(x|\theta) \int q(z)\ dz = log\ P(x|\theta)$$

$$Right : \int q(z)\ log\ \frac{P(x, z|\theta)}{q(z)} - \int q(z)\ log\ \frac{P(z|x, \theta)}{q(z)} = ELBO\ + KL(q(z)||P(z|x, \theta))$$

Therefore we can get

$$log\ P(x|\theta) = ELBO\ + KL(q(z)||P(z|x, \theta))$$

ELBO stands for evidence lower bound because $P(x|\theta) \geq ELBO$.

$$\hat{\theta} = arg \max_{\theta}\ ELBO = arg \max_{\theta} \int q(z)\ log\ \frac{P(x, z|\theta)}{q(z)}\ dz$$

In order to maximize ELBO, we need let ELBO equals to its upper bound value $P(x|\theta)$. Since $KL[\ q(z)\ ||\ P(z|x, \theta)\ ] \geq 0$ and $KL(\ q(z)\ ||\ P(z|x, \theta)\ ) = 0$ if $q(z) = P(\ z\ |x,\ \theta^{(t)})$.

$$\hat{\theta}^{(t+1)} = arg \max_{\theta} \int P(\ z\ |x,\ \theta^{(t)})\ log\ \frac{P(x, z|\theta)}{P(\ z\ |x,\ \theta^{(t)})}\ dz$$

Since $P(\ z\ |x,\ \theta^{(t)})$ is not related with parameter θ, we can simplified it

$$\hat{\theta}^{(t+1)} = arg \max_{\theta} \int P(\ z\ |x,\ \theta^{(t)})\ log\ P(x, z|\theta)\ dz$$

4.4 The Proportional Hazards Cure Model

4.4.1 Mixture Cure Model

The section will focus on the cure rate, survival distribution and covariates' effects. If the disease is not cured, that mean the event will still happens eventually, and we can call this **incidence**. Given that there is a probability that the event will occur, we need to find the survival time using the **latency** submodel.

Survival Data Set For the mixture cure model, there are two random variables follows, t and δ. t_i represents the follow-up time of the $i-th$ observations, it is the minimum value between failure time (F_i) and censoring time (C_i). In summary , $t_i = min(F_i, C_i)$. The censorship status δ_i , have alternative values. If $t_i = T_i$, $\delta_i = 1$ or $\delta_i = 0$, if $t_i = C_i$. Both $t_i = min(F_i, C_i)$ and δ_i are Bernoulli random variables indicating that $\delta = 0$ if the object is censored and $\delta = 1$ otherwise.

The possible covariates x_i and z_i influence the latency and incidence regression corespondingly. The cure rate depends on z_i and the survival probability of uncured patients depends on x_i. Note that $X' = [x_0, x_1, ..., x_n]'$ and $Z' = [z_0, z_1, ..., z_n]'$
Overall, we can usually represent the observation for i-th individual into $\Theta = \{t_i, \delta_i, x_i, z_i\}_{i=1}^{n}$.
Let another Bernoulli random variable Y be the indicator to show if a subject is cured or not. If Y = 1, the subject is susceptible which means it hasn't been cured and will experience the event in the long term. On the other hand if Y = 0, that means it has been cured and never experience the event in other word, nonsusceptible.
So a good question is, what is the difference between the two indicators δ and Y. All the individuals with uncensored data $\delta = 1$ must be in the uncured group$\delta = 1$. That because only susceptible subjects have probabilities to experience the event. If the patients are not cured, there will be sectional individuals $[\delta_i = 1] \subset [Y_i = 1]$ in the experiment period. The detailed explanation are shown followed:

$$\begin{cases} Y = 1, \quad uncured\ (susceptible) \begin{cases} \delta_i = 1, the\ event\ happens \\ \delta_i = 0, censored\ data \end{cases} \\ Y = 0, \qquad\qquad cured\ (insusceptible)\ all\ censored \end{cases} \tag{4.11}$$

Observed and Censored groups The sample can be divided in two groups: By the table 1, the likelihood construction of the failure time and right-censored time are equal to f(t) and S(t) respectively. Multiplying the cure rate, we can get:
Observed Group ($\delta_i = 1$)

$$\begin{aligned} P(Y = 1|z) \times f(t) \\ \pi(z) \times f(t) \end{aligned} \tag{4.12}$$

Censored Group($\delta_i = 0$)

$$\begin{aligned} S(t|x,z) &= P(Y = 0|z) + P(Y = 1|z)S(t|x, Y = 1) \\ S(t|x,z) &= (1 - \pi(z)) + \pi(z)S(t|x, Y = 1) \end{aligned} \tag{4.13}$$

Therefore, the likelihood function of i-th subject combining (4.12) and (4.13) the equation.

Cure Rate $\pi(z)$ Since the dependent variable Y is binary because $P(Y=0)+P(Y=1)=1$. The independent variable x_i and z_i has minor co-linearity and independent with each other. x_i and z_i are possible covariates in latency and incidence. Therefore the model fit the logistic regression assumption and the incident $\pi(z)$ can be defined below with unknown parameter b.

$$P(Y=1)=\pi(z)=\frac{exp(Z'b)}{1+exp(Z'b)}, Z'=[z_1,z_2,...,z_n]'$$

There are some other function that can be used in next section.

$$\begin{aligned}
f(t|x,Y=1) &= S(t|x,Y=1)\times\lambda(t_i|Y=1) && (4.14)\\
\lambda(t_i|Y=1) &= \lambda_0(t_i|Y=1)exp(X'\beta) && (4.15)\\
\Lambda(t_i|Y=1) &= \Lambda_0(t_i|Y=1)exp(X'\beta) && (4.16)\\
S(t|x,Y=1) &= exp\left[-\int_0^t \lambda(t_i|Y=1)dx\right] = exp[-\Lambda(t_i|Y=1)] && (4.17)
\end{aligned}$$

Therefore we can get the probability density function and the survival function,

$$\begin{aligned}
S(t|x,Y=1) &= exp[-\Lambda_0(t_i|Y=1)exp(X'\beta)] && (4.18)\\
f(t|x,Y=1) &= exp[-\Lambda_0(t_i|Y=1)exp(X'\beta)]\times\lambda_0(t_i|Y=1)exp(X'\beta) && (4.19)
\end{aligned}$$

Model Estimation

5.1 Likelihood Construction

The likelihood function of survival models need to consider if the observation is censored or failed. If the event occurs($\delta=1$), we need to use probability density function to estimate the time to event of interest. Or if the individual is censored($\delta=0$), survival function can be used to evaluate the probability.

Likelihood Function The likelihood function $\mathcal{L}(\theta)$ is consist of two components. Let α be a vector of unknown parameter in $\Lambda_0(t_i|Y=1)$(or $\lambda_0(t_i|Y=1)$, $S_0(t_i|Y=1)$) and set the joint set of likelihood parameter space $\theta=(b,\beta,\alpha)$. According to the likelihood construction in the equation (4.12) and (4.13), here is the likelihood function format following.

$$\mathcal{L}(\theta)\propto\prod_{i\in F}^{n} f_\theta(x_i)\propto\prod_{i\in C}^{n} S_\theta(x_i)$$

where F (for $\delta=1$) and C (for $\delta=0$) are the data sets of observed lifetime and censored time respectively. According to the formula (4.1) and (4.2):

$$\mathcal{L}(\theta)=\prod_{i=1}^{n}[\pi(z_i)f(t_i|x,Y=1)]^{\delta_i}\times[(1-\pi(z_i))+\pi(z_i)S(t_i|x_i,Y=1)]^{1-\delta_i} \tag{5.1}$$

Substitute equation (4.18) and (4.19) into the formula, then we can get the full likelihood function introduce by Sy and Taylor in 2000[8].

$$\begin{aligned}\mathcal{L}(\theta)=\prod_{i=1}^{n}\Big[\pi(z_i)e^{-\Lambda_0(t_i|Y=1)e^{X'\beta}}]\times\lambda_0(t_i|Y=1)e^{X'\beta}\Big]^{\delta_i}\times\\
\Big[(1-\pi(z_i))+\pi(z)e^{-\Lambda_0(t_i|Y=1)e^{X'\beta}}\Big]^{1-\delta_i}\end{aligned} \tag{5.2}$$

We need to get optimal $\hat{b}$ and $\hat{\beta}$ by maximizing the likelihood function. The next section was concerned with two methods that commonly used to estimate the models.

5.2 Direct Maximization Method

Then we can get the log-likelihood function of mixture cure model.

$$l(\theta) = log\mathcal{L}(\theta) = log\prod_{i=1}^{n}[\pi(z_i)f(t_i|z,Y=1)]^{\delta_i} \times [(1-\pi(z_i)) + \pi(z_i)S(t_i|x_i,Y=1)]^{1-\delta_i}$$

The log-likelihood function need to be maximized. One direct method is Newton-Raphson. Let

$$U(\theta) = \frac{\partial l(\theta)}{\partial \theta} = \begin{pmatrix} \frac{\partial l(\theta)}{\partial b} \\ \frac{\partial l(\theta)}{\partial \alpha} \\ \frac{\partial l(\theta)}{\partial \beta} \end{pmatrix}, I(\theta) = \frac{\partial^2 l(\theta)}{\partial \theta \partial \theta'} = \begin{pmatrix} \frac{\partial^2 l(\theta)}{\partial b \partial b'} & \frac{\partial^2 l(\theta)}{\partial b \partial \alpha'} & \frac{\partial^2 l(\theta)}{\partial b \partial \beta'} \\ \frac{\partial^2 l(\theta)}{\partial \alpha \partial b'} & \frac{\partial^2 l(\theta)}{\partial \alpha \partial \alpha'} & \frac{\partial^2 l(\theta)}{\partial \alpha \partial \beta'} \\ \frac{\partial^2 l(\theta)}{\partial \beta \partial b'} & \frac{\partial^2 l(\theta)}{\partial \beta \partial \alpha'} & \frac{\partial^2 l(\theta)}{\partial \beta \partial \beta'} \end{pmatrix}$$

Then taking iterated steps unsing the general form:

$$\widehat{\theta}^k = \widehat{\theta}^{k-1} + \frac{U(\widehat{\theta}^{k-1})}{I(\widehat{\theta}^{k-1})}$$

5.3 Applying the EM Algorithm

The EM algorithm is another method to maximize the likelihood function to obtain the corresponding parameters. Y is a latent variable in the model and we need to show the effect two indicators δ_i and y_i. There are three main categories. Since the set $\{\delta_i = 1\}$ is the subset of $\{Y_i = 1\}$, y need to be considered first.

$$\begin{cases} Y = 1, \quad uncured\ (susceptible) \begin{cases} \delta_i = 1, the\ event\ happens \\ \delta_i = 0, censored\ data \end{cases} \\ Y = 0, \quad\quad cured\ (insusceptible)\ all\ censored \end{cases} \tag{5.3}$$

$$\begin{aligned} P(z|Y=1) &= \pi(z_i)^{y_i} \\ P(z|Y=0) &= (1-\pi(z_i))^{1-y_i} \\ S(x|Y=1) &= \prod_{i=1}^{n}\left[\lambda_0(t_i|Y=1)e^{X'\beta}\right]^{\delta_i y_i} \times e^{-\Lambda_0(t_i|Y=1)e^{X'\beta}} \end{aligned} \tag{5.4}$$

Therefore, substitute the (5.4) into (5.2)
Then we will get the complete-data likelihood.

$$\begin{aligned} \mathcal{L}(\theta;y) &= \prod_{i=1}^{n}\pi(z_i)^{y_i}[1-\pi(z_i)]^{1-y_i}\prod_{i=1}^{n}\left[\lambda_0(t_i|Y=1)e^{X'\beta}\right]^{\delta_i y_i} \times e^{-\Lambda_0(t_i|Y=1)e^{X'\beta}} \\ &= \prod_{i=1}^{n}\pi(z_i)^{y_i}[1-\pi(z_i)]^{1-y_i}\lambda(t_i|Y=1)^{\delta_i y_i} \times S(t_i|Y=1,x_i)^{y_i} \\ &= \mathcal{L}_1(b;y) \times \mathcal{L}_2(\beta,\alpha;y) \end{aligned} \tag{5.5}$$

In order to simplified the calculation, we need to take logarithm of the both sides of the function which will be similar to the format of partial likelihood function.

$$l\ (\theta;y) = l_1\ (b;y) + l_2\ (\beta,\alpha;y) \tag{5.6}$$

Incidence log-likelihood

$$l_1(b;y) = log\mathcal{L}_1(b;y) = \sum_{i=i}^{n} y_i log[\pi(z_i)] + (1-y_i)log[1-\pi(z_i)] \tag{5.7}$$

Latency log-likelihood

$$l_2(\beta,\alpha;y) = log\mathcal{L}_2(\beta,\alpha;y) = \sum_{i=i}^{n} y_i\delta_i log[\lambda(t_i|Y=1,x_i)] + y_i log[S(t_i|Y=1,x_i)] \tag{5.8}$$

According to Peng and Yu, $w_i^{(m)}$ be the probability of a survival rate among the susceptible subjects. Given condition that the parameter value set $\theta^{(m)} = (b^{(m)}, \beta^{(m)}, \alpha^{(m)})$, we need to take conditional expectation of y_i using the linear function of y_i in equation (5.7) and (5.8). [15]

$$w_i(t) = E(y_i|\Theta, \theta^{(m)}) = \frac{\pi(z_i)S(t_i|x)}{(1-\pi(z_i)) + \pi(z)S(t_i|x)}|_{\theta=\theta^{(m)}}$$

5.3.1 E-Step

The E-step takes the conditional expectation of $\mathcal{L}(b,\beta,\alpha;y)$ with respect to the latent variable $y_i's$. The E-step in the k-th iteration and get posterior expectation of y_i as following:

$$w_i^{(m)} = E(y_i|\theta^{(m)}) = \delta_i + (1-\delta_i)\ w_{0i}(t)\ |_{b=b^{(m-1)},\beta=\beta^{(m-1)},\alpha=\alpha^{(m-1)}}$$

$w_i^{(m)} = E(y_i|\theta^{(m)})$ stand for the conditional probability The equation above show that $w_i^{(m)} = 1$ if $\delta_i = 1$ and $w_i^{(m)}$ is the probability value of uncured patient if $\delta_i = 0$.
Since we need to taking expectation of partial likelihood function. According to the equation (5.6).

$$log\mathcal{L}(\theta;y,w^{(m)}) = log\mathcal{L}_1(b;y,w^{(m)}) + log\mathcal{L}_2(\beta,\alpha;y,w^{(m)}) \tag{5.9}$$

which $w^{(m)} = \{w_i^{(m)}|i = 1,2,...\}$ represent the fraction that belongs to the susceptible group. Since $\delta_i log w^{(m)} = 0$ and $\delta_i w^{(m)} = \delta_i$, the expectation of (5.8) and (5.9) follows:

$$E[\log\mathcal{L}_1(b;y)] = \sum_{i=i}^{n} w_i^{(m)} log[\pi(z_i)] + (1-w_i^{(m)})log[1-\pi(z_i)] \tag{5.10}$$

$$E[log\mathcal{L}_2(\beta,\alpha;y)] = \sum_{i=i}^{n} \delta_i log[w_i^{(m)}\lambda(t_i|Y=1,x_i)] + w_i^{(m)} log[S(t_i|Y=1,x_i)] \tag{5.11}$$

5.3.2 M-Step

The M-Step maximize the equation in (5.10) and (5.11) respecting the parameter b, β and α with condition of $w^{(m)}$. Cai, Zhou and Peng etc. applied "glm" functions in R to estimate the parameters in equation (5.10) and different link function can be used in the mixture cure model.[3] In order to deal with the nuisance function $\lambda_0(t_i|Y=1)$ and $S_0(t_i|Y=1)$, we perform an additional maximization step in the M step using profile likelihood techniques. There are two methods to handling the Cox PH model : the **Breslow-type estimator** for $\lambda_0(t_i|Y=1)$ and the **product-limit estimator** for $S_0(t_i|Y=1)$.

PHMC Model Peng and Dear (2000) and Sy and Taylor(2000) utilize this method to obtain optimal β which involves the profile likelihood method originated from the standard PH model without specific baseline hazard function $\lambda_0(t, X)$[13, 8]. According to the format of equation (5.11), it can be rewritte with:

$$\prod_{i=1}^{n}[\lambda_0(t_i)exp((\beta x_i) + log(w_i^{(m)}))]^{\delta_i} S_0(t_i)^{exp[(\beta x_i)+log(w_i^{(m)})]} \tag{5.12}$$

The "coxph" function in R is practical to estimate the parameters in equation (5.11) and it is similar to the standard PH model with the additional offset variable $\log(w_i^{(m)})$. In order to take E-step progressively, the value of $\widehat{S}_0(t|Y = 1)$ and $\widehat{\lambda}_0(t|Y = 1)$ need to be updated for each step number m. The next section of the report was concerned with two estimation methods, Breslow-type estimator and product-limit estimator, presented by Sy and Taylor in 2000. [8] After geting the value of $\widehat{\lambda}_0(t|Y = 1)$, we need to erase the parameter α since it is the vector describing the unknown $\Lambda_0(t|Y = 1)$.

Breslow-type estimator In order to continue of the EM-Algorithm, we need to update the survival function in the discrete uncensored failure timeline $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$. $d_{t(j)}$ denotes the number of event happened at t_i and $R(t_{(j)})$ denote the risk set at time $t(j)$ The profile likelihood estimator use the Nelson-Aalen estimator with sight modification:

$$\widehat{S}_0(t|Y = 1) \quad = \quad exp\left(-\sum_{j:t_{(j)}\leq t\leq y} \frac{d_{t(j)}}{\sum_{i\in R_{(j)}} w_l^{(m)} e^{X'\beta}}\right) \tag{5.13}$$

$$\widehat{\Lambda}_0(t|Y = 1) \quad = \quad \sum_{j:t_{(j)}\leq t\leq y}\left(\frac{d_{t(j)}}{\sum_{i\in R_{(j)}} w_l^{(m)} e^{X'\beta}}\right) \tag{5.14}$$

If t→ ∞, the value of $\widehat{S}_0(t|Y = 1)$ would approach to 0. If we set $\widehat{S}_0(t|Y = 1) = 0$ for $t > t(k)$, we will get $\widehat{S}(t|Y = 1) = \widehat{S}_0(t|Y = 1)^{exp(\widehat{X}'\beta)}$ Substituting equation (5.14) into the $\mathcal{L}_2(\beta, \alpha; y)$, then we can get the paritial likelihood function of β:

$$\mathcal{L}_3(\beta; w^{(m)}) = \prod_{j:t_{(j)}\leq t\leq y}\left(-\frac{e^{X'\beta}}{\sum_{i\in R_{(j)}} w_l^{(m)} e^{X'\beta}}\right)^{\delta_i} \tag{5.15}$$

Except for the weight $w_i^{(m)}$, the format is similar to the partial likelihood function $\frac{exp(X_j'b)}{\sum_{k\in R(t_j)} exp(X_k b)}$. Giving the known $w_i^{(m)}$, maximizing $\mathcal{L}_3$ concerning β.

Application

R code The R package smcure can fit semiparametric PH mixture cure model or AFT mixture cure model by the EM algorithm.Therefore, the target form, PHMC model with logit like function can be estimating by:

smcure (formula, cureform, offset = NULL, data, na.action = na.omit, model = "ph", link = "logit", Var = TRUE, emmax = 50, eps = 1e-07, nboot = 100)

6.1 Mixture Cure Models in Oncology

Felizzi and colleagues have designed a tutorial for practical usage of the mixture cure model providing step-by-step instructions for the entire implementation workflow. [7] It includes collecting and combining data from different sources, regressing the model using maximum likelihood estimation methods, and interpreting the model results. This section introduced the procedure of implementation and interpretation of mixture cure models in R which will be widely used as new cancer treatments enter the market, this model will be widely used in health economics analysis. The following steps will provide innumerable .csv files and we need to choose countries that we are interested.

Step 1: Choosing countries of interest The first step is listing all the target country names into "country_list" file and adding the corresponding code names. Then, we can use the function "hazard_function" by applying the "funs_hazard.R" and "funs_long_term_survival.R". The algorithm traverses the age and sex distribution of the selected country/region and builds a general population background mortality curve, which is a weighted average of the curves constructed for age groups, weighted according to their proportions. The next step is obtaining the background survival country-specific data.

Step 2: Acquiring background Mortality data The background survival data is the survival time for the cancer-free individuals. The overall survival function ($S_o(t)$)and mortality rate($\lambda_o(t)$):

$$S_o(t) = S_b(t) \times ((1 - \pi(z)) + \pi(z)S_u(t))$$

$$\lambda_o(t) = \lambda_b(t) \times \frac{\pi(z) \times f_u(t)}{1 - \pi(z) + \pi(z)S_u(t)}$$

- $S_u(t)$ and $\lambda_u(t)$ denote the mortality and survival function uncured groups. While $f_u(t)$ represents the probability density function of $S_u(t)$.
- $S_b(t)$ and $\lambda_b(t)$ is the estimate background hazard and survival function in cured patients.

Preparing for the analysis procedure in the following parts, I need to access the mortality data. Exerting the "funs_load_mort_table.R" for downloading data from Human Mortality Database (HMD)and "mortal_table_wrap.R" to prepare for analysis.

Step 3: Clinical Trial Data In the tutorial, Felizzi et al. identified six characteristics to acquire patient demographics and survival data from clinical trials of interest. For the purpose of build a mixed cure model, fairly number of patients is required to examine such as baseline age, sex and country of each patient, build a mixed cure model indicators, time of observation prior to build a mixed cure model, and year of trial registration. This segment aim to simulated the data set BRAF Inhibitor in Melanoma 3 (BRIM-3). BRIM-3 is a phase 3 of a randomized controlled trial (PCT) experience that compares therapeutical efficacy result of melanoma using dacarbazine and vemurafenib. With the intention of protect the privacy for the patients, Felizzi et al. design a method to decide the ages by adding a random Gaussian noise which the mean is equal to 0 and variance is equal to 3 years while adding another one on mortality which

has a mean of 0 and a variance of 0.01 years. After the steps above, we can obtain the file "brim3_simulated.csv" .
The next step exhibits the sadistically cured fraction in metastatic melanoma and it is displayed in the plateaus in overall Kaplan-Merier (KM) curves.
Note that in the first 3 steps, Felizzi and colleagues(2021) have revealed a rigorous procedure to apply MC Model contributes to the invention of novel cancer therapies[7]. However, the coming steps choose to utilize PHMC model instead of comparing different parametric models. Consequently, the following steps are introduced by Peng etc.(2012)[4].

Step 4: Estimating and Analysis Kaplan-Meier curves The data set includes 200 males, 137 females and in total of 337 patients from different countries in the research. The diagram of the age distribution are shown below:

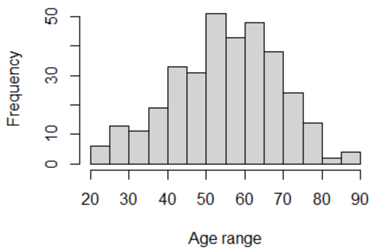


Figure 6.1: The diagram of patients' age

Kaplan Meier survival curve is the most intuitive approach to analysis the trend, median and differences between the curves. Subsequently, we will draw 3 KM cures graph related to gender, age group and countries.

KM1:Estimating Kaplan-Meier curves respecting to gender Starting with the KM curves 6.2 corresponding to gender, it has been observed that the three curves are closed to each other. Referring to the hazard table, there are strong evidence that gender only effect the mortality slightly. It should be notice that in the legend of the KM curves, SEX.F means the patient is female while SEX.M stands for the patient is male.

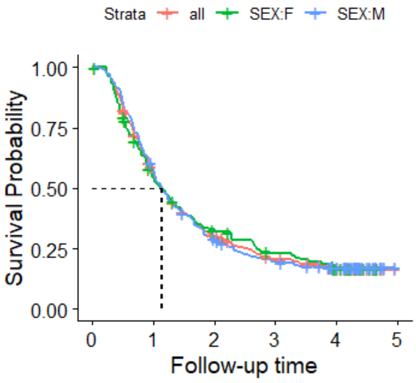
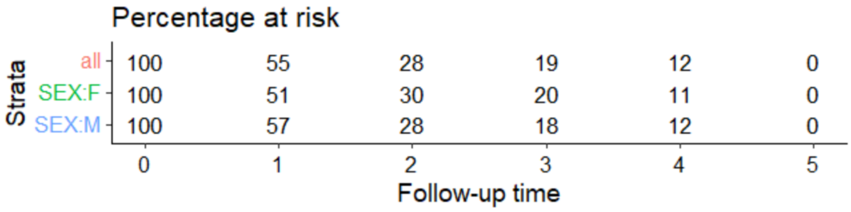


Figure 6.2: KM Curve 1(gender)



Percentage at risk

Strata	0	1	2	3	4	5
all	100	55	28	19	12	0
SEX:F	100	51	30	20	11	0
SEX:M	100	57	28	18	12	0

Follow-up time

KM2:Estimating Kaplan-Meier curves respecting to country If we want to examine relations between the survival probabilities and countries, choosing 3 countries since patients are from 19 countries. In order to help to select the object, the diagram of 19 countries are necessary:

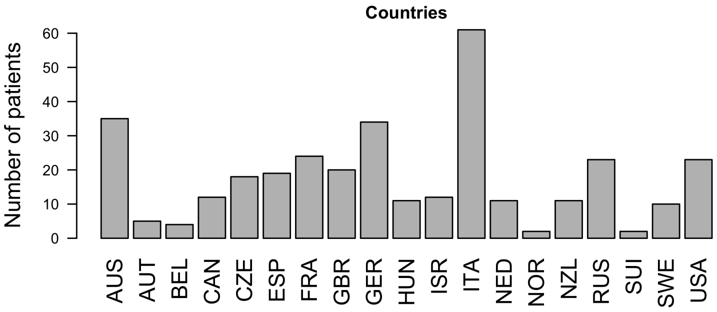


Figure 6.3: The diagram of patients' country

From the diagram above, we choose three countries for the KM estimation such as Italy, USA and Germany for the reason that the Italy group has the most subjects and United States has the most progressive medical technology. In the figure 6.4, the legend COUNTRY.ITA, COUNTRY.USA and COUNTRY.GER are refer to Italy, USA and Germany respectively. In addition, the legend "all" implies the average of the three curves. The KM curve for the variable country is displayed:

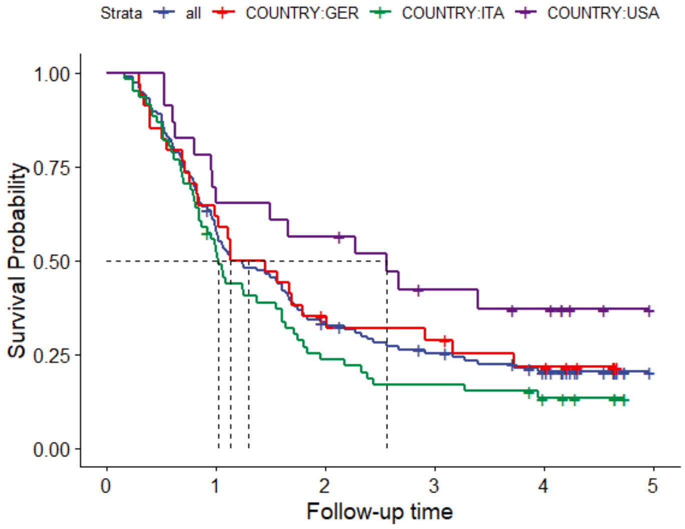


Figure 6.4: KM Curve 2(country)

The diagram show strong evidence that patients in different countries have different survival rate of melanoma. Hence the country is related to the final survival rate. However, the variable country is hard to numeric, the variable is hard to contain in the PHMC model.

KM3:Estimating Kaplan-Meier curves respecting to age group The diagram 6.1 is close to a unimodel. Intended for getting the KM curves relating to age, we need to regroup the patients according to the standard introduced by WHO.

20-44	45-59	60-74	$75 - \infty$
Young Adult	Middle-Aged Adult	Young Elderly	the Elderly

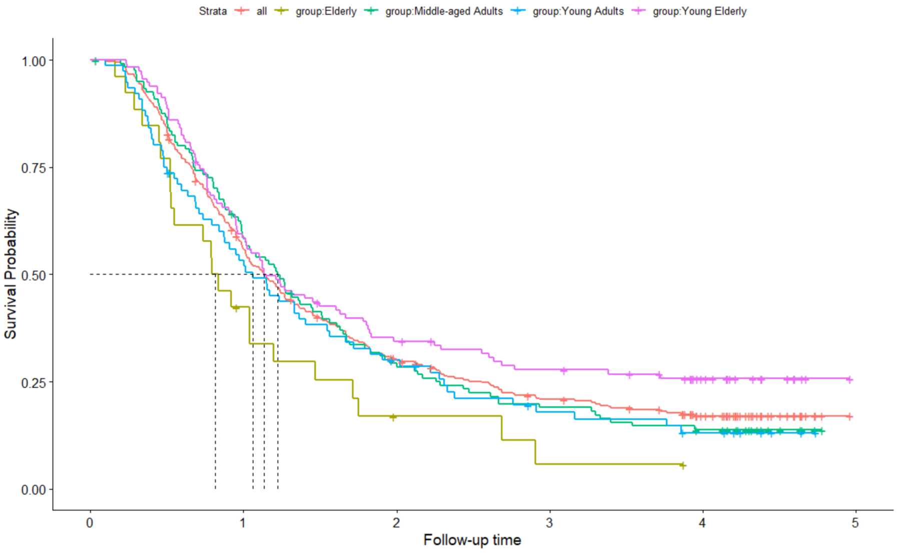


Figure 6.5: KM Curve 3(age group)

From the curve, there are substantial distinction between the KM line for each group. People older than 75 has the lowest survival rate and the other three are really closed to each other. Contrary to common sense, the 60-74 years old group have the largest survival rate in the plateau.

Step 5: Implementing the PHMC model Fitting the Brim-3 data into the semiparametric PHMC model. The model insists of two fractions: ”Cure probability model” and ”Failure time distribution model”. Thinking of the process as an equation, the cure probability can be approximated by $1-\pi(z)$ while the survival function will be S(t,X)(4.13). Consequently, the R package smcure will be used for dataset ”brim3” and get the estimation and interpretation.
Note that the default value nboot is equal to 100, if we increase the the number of bootstrap samplings, the standard error term for the estimation will decrease. **Call:**
smcure(formula = Surv(as.numeric(TIME), CNSR) $\sim$ AGE + SEX, cureform = $\sim$ AGE + SEX, data = brim3, model = ”ph”, link = ”logit”)
Cure probability model:

	Estimate	Std.Error	Z value	Pr(>|Z|)
(Intercept)	2.31789555	0.66905151	3.46445007	0.0005313168
AGE	-0.01294646	0.01128204	-1.14752832	0.2511633336
SEX	-0.01821865	0.30998916	-0.05877191	0.9531337830

Failure time distribution model:

	Estimate	Std.Error	Z value	Pr(>|Z|)
AGE	0.005484636	0.004565488	1.2013254	0.2296250
SEX	-0.094985191	0.133634612	-0.7107829	0.4772188

Step 6: Model Assessment In fact, the PHMC analysis procedure don't compare the effective of dacarbazine and vemurafenib because of the file "brim3_simulated.csv" provided Felizzi et al.(2021)[7] is not contain the corresponding variable. The drawback leads to the error. The p-value is small also because of the variable "COUNTRY" are highly correlated with the survival probability.
The cure rate can be calculated by $1-\pi(z) = 1-\frac{e^{(2.31789555+-0.01294646+-0.01821865)}}{1+e^{(2.31789555+-0.01294646+-0.01821865)}}$. From the two distribution model , if the patient is older, they will have lower probability to be cured and have larger mortality.

Conclusion

In oncology, the survival analysis methods are appropriate to develop the novel cancer therapies including survival curve with the remaining censored subjects after the experience ends (Felizzi et al., 2021)[7]. On that occasion, a portion of patients was considered statistically cured and finally die for other reasons which includes in the nonsusceptible group. On the opposite, the patients will belong to the susceptible group so cancer affects the time negatively. Both partition consists of the **Mixture Cure Model(MC)** which is more accurate than the standard survival estimate methods. To introduce the method, we first introduce survival analysis bases and models, including parametric, semi-parametric, and non-parametric models. Besides, maximum the complete likelihood functions of Mixture Cure Model need to use EM algorithm and then introduced PHMC and AFTMC for the M-step. Felizzi et al.(2021) outline a step-by-step MC model applying process to spread the model to further cutting-edge cancer therapies development. Finally, by analyzing BRIM-3 trial data with the process and eventually come to a preliminary conclusion that country and age effect the outcome obviously.

References

[1] Joseph Berkson and Robert P. Gage. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47(259):501–515, 1952.

[2] John W. Boag. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(1):15–53, 1949.

[3] Chao Cai, Yubo Zou, Yingwei Peng, and Jiajia Zhang. smcure: An r-package for estimating semiparametric mixture cure models. *Computer Methods and Programs in Biomedicine*, 108(3):1255–1260, 2012.

[4] Yingwei Peng Jiajia Zhang Chao Cai, Yubo Zou. smcure: An r-package for estimating semiparametric mixture cure models. *Computer Methods and Programs in Biomedicine*, 108(3):1255–1260, 2012.

[5] Frank Emmert-Streib and Matthias Dehmer. Introduction to survival analysis in practice. *Machine Learning and Knowledge Extraction*, 1(3):1013–1038, 2019.

[6] V. T. Farewell. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38(4):1041–1046, 1982.

[7] Federico Felizzi, Noman Paracha, Johannes Pöhlmann, and Joshua Ray. Mixture cure models in oncology: A tutorial and practical guidance. *PharmacoEconomics - Open*, 5:1–13, 02 2021.

[8] Sy Judy P. and Taylor Jeremy M. G. Estimation in a cox proportional hazards cure model. *Biometrics*, 56(1):227 – 236, 2000.

[9] John P. Klein and Melvin L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. 1997.

[10] David G. Kleinbaum and Mitchel Klein. *Survival analysis : a self-learning text*. Statistics for biology and health. Springer, 2012.

[11] ANTHONY Y. C. KUK and CHEN-HSIN CHEN. A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 79(3):531–541, 09 1992.

[12] Guanghan Frank Liu and Jason J. Z. Liao. Analysis of time-to-event data using a flexible mixture model under a constraint of proportional hazards. *Journal of Biopharmaceutical Statistics*, 30(5):783–796, 2020. PMID: 32589509.

[13] Yingwei Peng, Keith B. G. Dear, and J. W. Denham. A generalized f mixture model for cure rate estimation. *Statistics in Medicine*, 17(8):813–830, 1998.

[14] Rinku Saikia and Manash Pratim Barman. A review on accelerated failure time models. *International Journal of Statistics and Systems*, 12(2):311–322, 2017.

[15] Binbing Yu Yingwei Peng. *Cure Models Methods, Applications, and Implementation*. 2021.