Xi'an Jiaotong-Liverpool University

Final Year Project

Estimation methods for the semiparametric accelerated failure time mixture cure model

半参数加速失效时间混合固化模型的估计方法

Wei Zhou ID 1825310

DEPARTMENT OF SCIENCE

Supervised by

Mu He

May 4, 2022

Abstract

Survival analysis is an important part of modern statistics. It refers to a series of statistical methods used to explore the occurrence time of events of interest. The common ones are survival time analysis of cancer patients and failure time analysis in engineering and so on. This paper aims to introduce the basic theory of survival analysis and conventional application models, including proportional risk model and accelerated failure time model. However, in real life, the original model has considerable limitations. Due to the lack of a variety of data, it is difficult for us to apply it directly. Therefore, based on this problem, this paper first reviews the basic research methods and research steps. Then, the semi parametric accelerated failure time model is studied by using the Expectation-Maximum method through the maximum likelihood method and Newton method. Then, a new research method is proposed by using Gehan-weight function and convex function, and the theoretical research from semi parametric AFT model to AFT mixture cure model is completed

Key words: Linear regression, Maximum likelihood method, EM method , AFT model

摘要

生存分析是现代统计学的重要组成部分，指的是一系列用来探究所感兴趣的事件的发生的时间的统计方法。常见的有癌症患者生存时间分析和工程中的失败时间分析等等。本文旨在通过对多篇研究论文的贡献和挑战的研究，初步介绍了生存分析的基本理论以及常规应用模型，包括比例风险模型和加速失效时间模型。然而，在现实生活中，原始的模型具有相当大的局限性，由于多种数据的缺失，我们很难直接应用。因此，基于此问题，本文先回顾了生物统计的基本研究方法和研究步骤。然后通过最大似然法和牛顿方法来实现期望最大化算法，着重研究了半参数加速失效时间模型。之后，通过应用格汉权函数和凸函数提出了一种新的研究方法，完成了从半参数加速失效时间模型到半参数加速失效时间混合固化模型的拓展的理论研究

关键词：线性回归，最大似然法，EM算法，AFT模型

Contents

1 Introduction

We are interested to study how risk factors associated with disease exist or not exist in logistic regression. Sometimes, we look for the treatment of risk factors or interested in how to influence the time or other incidents of disease. Meanwhile, we might have to have study dropout, so we are not sure whether they will suffer from diseases of the subjects. In these cases, the logistic regression is inappropriate.

To begin with, survival analysis is applied to the analysis of the incident before the time the data of interest. Tolley, Barnes & Freeman (2016) claimed that survival analysis is analyzing the data according to time of one of the main statistical methods. Such data analysis is very important to many aspects, including the estimate loss of life years, evaluate drug safety, measurement of medical treatment and the feasibility of the device, an actuarial loss, product reliability, etc. This experience science branch needs to collect and analyze data, until a failure or death.

It is usually determined that loss or damage is what the survival analysis needs to solve. "How long can the patient live? "Or "How much has the victim's life shortened?" are common questions. However, an accurate answer is impossible for any question. We can only get an objective answer as a statistical probability. It requires an average or "expected value" with a relevant uncertainty level. Usually, we illustrate this level of uncertainty by using confidence intervals.

How to obtain more information from data is the goal that statistics has been pursuing tirelessly. However, many times, it is a bit regrettable to delete some data, and it is inaccurate to leave it. In fact, this kind of data that does not contain the exact value of the data and only contains its upper or lower bound is called Censored Data. The main goal of survival analysis is to deal with the situation where the corresponding variable is censored data.

Censored data is very common in daily life, for example: a member of the test withdraws from the experiment during the experiment; some samples are incorrectly measured due to the inaccuracy of the instrument; the questionnaire is distorted

due to social reasons, etc. These data can all be viewed as censored cases. Survival analysis is the most powerful way to deal with this situation.

In the data of survival analysis studies, the response variable is censored, which is common and unavoidable. Therefore, the traditional mean, standard deviation, and related t tests cannot be used. Survival analysis focuses on two questions. One is to identify differences between groups. The main methods used are Kaplan-Meier estimation, Log-Rank test, and regression. The second is to make predictions. The main method used is regression, including parametric models and semi-parametric models.

2 Literature Review

2.1 Basic concepts

First of all, what exactly is survival analysis? Goel, Khanna & Kishore (2010) proposed that survival analysis is a collection of statistical procedures for data analysis. Among them, the outcome variable is the time before the event and the outcome variable of interest is the time before the event. In survival analysis, we take the time variable as survival time, and usually define the event as failure (Deborah et al. 2021). Most of the time here refers to days months, years or from the start of the follow-up of the individual to the occurrence of the event. Alternatively, time can also refer to the age of the individual at the time of the event. The event here may be death, onset of illness, remission, recovery or any specified experience of interest that may occur to the individual.

However, in many cases, we cannot know exactly when the event occurred, so we need to introduce a new concept, namely censoring data. Censorship means that we have some information about the survival time of an individual, but we don't know exactly the survival time (Goel, Khanna & Kishore 2010). The censored may occur due to the following three reasons: no events experienced before the end of the study; lost to follow-up during the study; withdrawal from the study due to death (if the death is not an event of interest) or other reasons. In addition,

the review can be clearly divided into three types. Right-censoring was defined as a situation where the observed survival time was less than or equal to the true survival time. The opposite is left-censored, where the observed survival time is greater than or equal to the true survival time. There is also a case of interval censoring, which means that the known time interval contains the true survival time.

Next, we need to identify terms and symbols to facilitate future model understanding. First, T is a random variable of survival time, and t is any particular interest value of the random variable capital T. Secondly, d is a (0,1) random variable used to indicate failure or review. When d=1, the event occurred during the study period, which means failure; if the survival time is censored at the end of the study period, d=0. At the same time, here we need to introduce the definitions of two basic functions, the survival function $S(t)$ and the hazard function $h(t)$. The survivor function refers to the probability that a individual will survive beyond a specified time t. The hazard function describes the instantaneous potential of an event per unit time, assuming the individual has survived until time t..

Also, there are several relationships in the above terminology and notation. On the one hand, survivor function gives the probability that the random variable T exceeds the specified time t, which can be expressed as $S(t) = P(T > t)$.

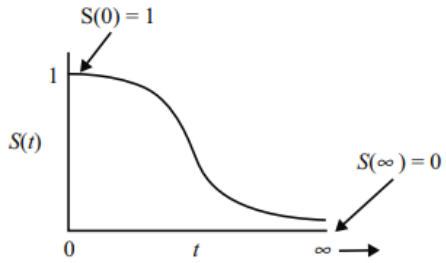


Figure 1: Theoretical $S(t)$

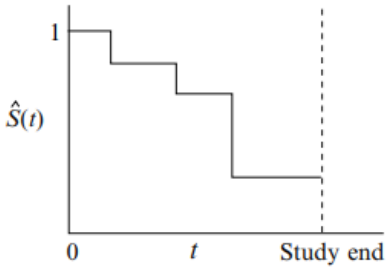


Figure 2: $\hat{S}(t) in practice$

On the other hand, as for $S(t)$ and $h(t)$, knowing one can determine the other. Their conversion formulas are: $S(t) = e^{-\int_0^t h(u)\,\mathrm{d}u}$ and $h(t) = -\frac{\mathrm{d}S(t)/\mathrm{d}t}{S(t)}$

2.2 Cox proportional hazards (PH) model

Goel et al. (2010) claimed that Kaplan-Meier estimation is the easiest way to calculate survival rates over a period of time. Various situations can be assumed to create a survival curve that involves calculating the probability of events occurring at a certain point in time and combining these consecutives. Multiply the probability by any previously calculated probability to obtain the final estimate.

However, the Kaplan-Meier survival estimation method is univariate analysis, which means model only describes the relationship between a single variable and survival, without considering or ignoring the impact of other variables. But in real life, we need to consider multiple variables. For example, when comparing the survival rates of two groups of patients with different genders, the patients in one group are older. Therefore, the survival rate may be affected by gender or age. This is why we need to introduce the Cox proportional hazards regression model next.

The Cox proportional hazard model supposes that the potential hazard rate is a function of the independent variables (and covariates). It is essentially a commonly used statistical regression model for investigating the association between a patient's survival time and one or more predictors(Cox 1972). The model is expressed as follows:

$$h(t, \mathbf{X}) = h_0(t) \times exp \sum_{i=1}^{p} X_i \beta_i; \qquad X = (X_1, X_2, ...X_P) \tag{1}$$

Among them, $h_0(t)$ is called the hazard function, that is, the probability of the event of interest occurring at time t, assuming that the subject survives at time t and beyond. The term $h_0(t)$ is called the baseline risk; when all independent variable valus are zero, it is the risk of each individual. The terms $X_1, X_2, X_3, ...X_k$ are covariates, and $\beta_1, \beta_2, ...\beta_k$are the corresponding regression coefficients Deborah, Derek and Bruce, 2021).

The model gives an individual risk expression at time t, which contains a set

of given specifications of explanatory variables denoted by **X**. That is, the **X** indicates that a set of predictor variables (sometimes called a vector) is being modeled to predict the individual's risk.

Additionally, the baseline risk $h_0(t)$ is an important feature of the Cox model. It is an unspecified function, and this feature makes the Cox model a semi-parametric model. A model whose function form is fully specified is called a parametric model, except for the values of unknown parameters.

For instance, the Weibull hazard model is a parametric model with the following form, where the unknown parameters are λ, p, and β_i.

$$Weibull: \qquad h(t, X) = \lambda p t^{p-1} \tag{2}$$

where $\lambda = exp\sum_{i=1}^{p} X_i\beta_i$ *and* $h_0(t) = pt^{p-1}$

However, how can we apply Cox PH model into Survival Curves?
When using the Cox model to fit survival data, we can get a survival curve, which is called an adjusted survival curve. It can be adjusted for explanatory variables used as predictors. The hazard function formula of the model can be converted to the corresponding survival function formula as shown below.

$$Cox\ model\ survival\ function: \qquad S(t, X) = S_0(t)exp\sum_{i=1}^{p} X_i\beta_i \tag{3}$$

$$Estimated\ survival\ function: \qquad \hat{S} = \hat{S}_0(t)exp\sum_{i=1}^{p} X_i\hat{\beta}_i \tag{4}$$

The survival function formula is the basis for determining the adjusted survival curve. $\hat{S}_0(t)$ *and* $\hat{\beta}_i$ ca provided by the computer program. The X_i are specified by the investigator.

2.3 Accelerated failure time (AFT) model

For the AFT and PH models, the interpretation of the parameters is different. The AFT assumption is suitable for the comparison of survival time, and the PH value hypothesis is suitable for the comparison of hazards(Fu, Yang, Zhou & Wang 2021). To put it succinctly, AFT is the multiplicative effect of survival time, and PH is the multiplicative effect of hazard. Now we discuss the AFT assumption.
We already know that the basic assumption of the AFT model is that the influence of the covariate is multiplied by the survival time (proportional). For a random event time T, an AFT model proposes the following relationship between covariates and $Y = logT$:

$$Y_i = X_i\beta + W_i \tag{5}$$

Where $W_i \sim f$ are the error, or residual, terms. The above framework describes a class of general models: the distribution we specify for W allows us to obtain different models, but all models actually have the same structure. For instance, a basic possibility is to assume $W_i \sim N(0, \sigma_i)$. If it is assumed that Y obeys a normal distribution, from the formula $Y = logT$, it can be inferred that T obeys a logarithmic normal distribution. Therefore, Therefore, we can fit the model and obtain the confidence interval by using the ordinary least squares regression method. For any AFT, we have $T = e^{\eta}T_0$, $where\ T_0 = e^W\ and\ \eta_i = X_i\beta$. In other words, the objects of the two models are different. As for the proportional hazard (PH) model, the covariate acts on the risk by multiplication, while in the AFT model, the covariate acts on the time by multiplication. Also,

$$Survival\ function: \qquad S_i(t) = S_0[exp(-\beta_i t)] \tag{6}$$

$$Hazard\ function: \quad \lambda_i(t) = \lambda_0[exp(-\eta_i t)]exp(-\eta_i t) \tag{7}$$

It's worth mentioning that Weibull returns to satisfy both AFT and PH. Since it is a linear distribution, the vertical movement of the line will correspond to the horizontal movement(Cox 1972). According to the extreme value distribution, $\lambda(y) = e^y$, which is linear on this scale. In addition, the Weibull distribution represents its series of positional scales. Therefore, the Weibull distribution is the

only distribution that satisfies both PH and AFT assumptions. Finally, let us briefly consider Maximum likelihood estimation, the likelihood is:

$$L(\beta, \sigma|y, d) = \prod_i \sigma^{-1} f(w_i)^{d_i} S(w_i)^{1-d_i} = \prod_i \sigma^{-1} \lambda(w_i)^{d_i} S(w_i) \tag{8}$$

where f represents the density, λ and S represent hazard and survival functions respectively for the error distribution. $w_i = (y_i - X_i\beta)/\sigma$.

3 Methodology Analysis

The classic models for analyzing failure time data, including the Cox proportional hazard model and accelerated failure time model, we have already introduced. Now we want to apply it to medical research. However, In the real problem we find that because of the limitation of the AFT estimation method, it is difficult for us to solve the problem directly. Therefore, our purpose is to develop a new estimation method for the AFT model. First, we introduce the EM method which be used to optimize and solve the problem.

3.1 EM method

3.1.1 Introduction of EM method

The EM method, also known as the expectation maximization method, is the most common latent variable estimation method and is used in machine learning widely (Kolaczyk 2009). For instance, it is often used to learn Gaussian mixture models, hidden Markov algorithms, etc. In this paper, the principle of EM method is summarized in detail.

The EM algorithm is an iterative optimization strategy. For each iteration in its calculation method, it can be divided into two steps, the expected step size (E-step) and the maximum step size (M-step). It was originally designed to solve the problem of how to estimate parameters in the case of missing data . Its algorithmic basis and convergence effectiveness have been discussed by mathematicians, and

the basic idea is iteration. First, the values of the model parameters are estimated from the known observed data. The values for the missing data portion are then estimated based on the parameter values just estimated. Add the estimated missing data to the value of the previously estimated missing data. The new parameter values are re-estimated using the observed data, and the iteration is repeated until it finally converges and the iteration ends.

3.1.2 Preliminary knowledge

3.1.2.1 Maximum likelihood estimation

"Likelihood" and "probability" are similar in meaning, both refer to the possibility of a certain event, but in statistics, "likelihood" and "probability" (probability) have a clear distinction(Efron & Stein 2007): Probability, which is used to predict the result of the next observation given some parameters; Likelihood is used to estimate parameters about the properties of things when the results obtained from some observations are known. In this sense, the likelihood function can be understood as the inverse of the conditional probability. Given a certain parameter B, the probability that event A will occur is written as:$P(A|B) = \frac{P(A,B)}{P(B)}$ Using Bayes' theorem,$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$ Therefore, we can construct a method for expressing the likelihood in reverse: given that an event A has occurred, using the likelihood function $L(B|A)$, we estimate the likelihood of the parameter B. Formally, the likelihood function is also a conditional probability function, but the variable we care about is changed: $b \mapsto P(A|B = b)$. Note that the likelihood function is not required to be normalized here: $\sum_{b \in B} P(A|B = b) = 1$. A likelihood function multiplied by a positive constant is still a likelihood function. For all $\alpha > 0$, there can be a likelihood function:$L(b|A) = \alpha P(A|B = b)$

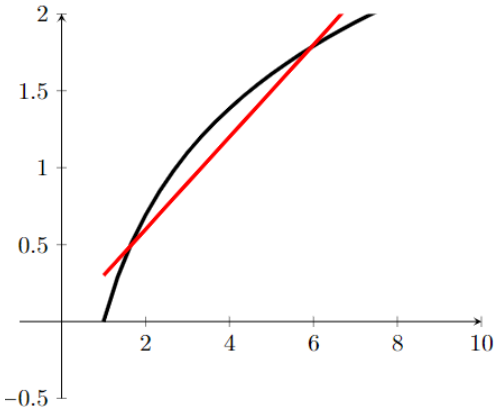
So, why do we want to estimate the likelihood of a parameter? Back to the definition, the purpose is to estimate the parameter theta, not to find the similarity What is the value of probability. The same result can correspond to many different parameters theta, and the probability of each parameter is different, that is, each parameter θ has a corresponding likelihood $L(\theta)$. Maximum likelihood es-

timation, a method used to estimate the parameters of a probabilistic model, is the original and most natural application of the likelihood function. The method of maximum likelihood estimation is to first select a likelihood function (usually a probability density function or a probability mass function), and then find the maximum point after sorting.

3.1.2.2 Jensen's inequality

First, let's understand what a convex function is. Fygenson & Ritov (1994) mentioned that a convex function is a real-valued function f defined on a convex subset C (interval) of a vector space, if any two points x_1, x_2, $0 \leq t \leq 1$ on its domain C have

$$tf(x_1) + (1-t)f(x_2) \geq f(tx_1 + (1-t)x_2) \tag{9}$$



That is to say, the secant of any two points of a convex function is above the function graph, which is also the two-point form of Jensen's inequality. The Left is an example.(f(x)=lnx and f(x)= 0.3x)

Figure 3: Convex Function

If for any set of points x_i, if $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$, using mathematical induction, it can be proved that the convex function f (x) satisfies:

$$f(\sum_i \lambda_i x_i) \leq \sum_i \lambda_i f(x_i) \tag{10}$$

Equation (10) is called Jensen's inequality, which is a generalized form of Equation (9). If f is a convex function, where E(x) represents the mathematical expectation of x. x is a random variable, then we can get $E[f(x)] \geq f(E[x])$. This formula takes the equal sign if and only if x is a constant.

3.1.3 EM algorithm

Our goal is to maximize a likelihood function as following:

$$\theta_{MLE} = argmax_{\theta} P(X|\theta) \tag{11}$$

Here θ is the model parameter, and X is the observable data, but $P(X|\theta)$ may be more complicated, we can split $P(X|\theta)$ into two parts:

$$P(X|\theta) = \sum_{z} P(Z|\theta)P(X|\theta, Z) \tag{12}$$

Here Z is an latent variable defined by us, which cannot be directly observed. The above equation is equivalent to a full-probability expansion of $P(X|\theta)$. Our loss function is generally The logarithm of the likelihood function, as follows:

$$L(\theta) = logP(X|\theta) = log(\sum_{Z} P(Z|\theta)P(X|\theta, Z)) \tag{13}$$

Next, it will be very difficult to solve by directly taking the partial derivative of the parameter and making it 0, which is the difficulty of solving the maximum likelihood estimation with latent variables.
The idea of solving EM is to find a sequence of parameters that can gradually improve the likelihood estimate, namely:

$$\theta^1 > \theta^2 > \cdots > \theta^i \rightarrow L(\theta^1) < L(\theta^2) < \cdots < L(\theta^i)$$

If the current round is the i-th round and the current parameter is θ^i, then the next step is to find a θ^{i+1} on this basis such that $L(\theta^{i+1}) > L(\theta^i)$ According to the conditional probability formula, we can transform $P(X|\theta)$ as follows: $P(X|\theta) = \frac{P(X,Z|\theta)}{P(Z|X,\theta)}$ so:

$$L(\theta) = logP(X|\theta) = logP(X, Z|\theta) - logP(Z|X, \theta) \tag{14}$$

So, how to use the information of the previous step θ^{i+1}? Since we learned θ^i in the previous step, then we can find a distribution of Z on this basis, that is, $P(Z|X, \theta^i)$, so we can find The expectation of $L(\theta)$ on the distribution $P(Z|X, \theta^i)$:

$\sum_Z logP(X|\theta)P(Z|X,\theta^i) = \sum_Z logP(X,Z|\theta)P(Z|X,\theta^i) - \sum_Z logP(Z|X,\theta)P(Z|X,\theta^i)$

$\rightarrow logP(X|\theta) = \sum_Z logP(X,Z|\theta)P(Z|X,\theta^i) - \sum_Z logP(Z|X,\theta)P(Z|X,\theta^i)$

$\rightarrow L(\theta) = \sum_Z logP(X,Z|\theta)P(Z|X,\theta^i) - \sum_Z logP(Z|X,\theta)P(Z|X,\theta^i)$

$\rightarrow L(\theta) = Q(\theta,\theta^i) - H(\theta,\theta^i)$

The last step is:

$$Q(\theta,\theta^i) = \sum_Z logP(X|Z,\theta)P(Z|X,\theta^i)$$

$$H(\theta,\theta^i) = \sum_Z logP(Z|X,\theta)P(Z|X,\theta^i)$$

Another point to note is that θ is an unknown quantity, and θ^i is a known quantity. At this point, we can already see the direction of the next optimization, namely:

$$\theta^{i+1} = argmax_\theta Q(\theta,\theta^i) - H(\theta,\theta^i)$$

However, only the maximum value of the Q function is solved in the actual solution:

$$\theta^i = argmax_\theta Q(\theta,\theta^i)$$

This is because there must be $H(\theta^{i+1},\theta^i) \leq H(\theta^i,\theta^i)$ for the H function, so $L(\theta^{i+1}) \geq L(\theta^i)$, proof is below

3.1.4 EM Convergence Proof

$$\begin{aligned} H(\theta^{i+1},\theta^i) - H(\theta^i,\theta^i) &= \sum_Z (log\frac{P(Z|X,\theta^{i+1}}{P(Z|X,\theta^i)})P(Z|X,\theta^i) \\ &\leq log(\sum_Z \frac{P(Z|X,\theta^{i+1})}{P(Z|X,\theta^i)}P(Z|X,\theta^i)) \\ &= log(\sum_Z P(Z|X,\theta^i)) = log1 = 0 \end{aligned}$$

Jensen's inequality is used, which can be obtained from the definition of the convex function. log(x) is a convex function, and the above x_i is regarded as $P(Z_i|X,\theta^{i+1})P(Z_i|X,\theta^i)$, λ_i can be proved as $P(Z_i|X,\theta^i)$.

3.1.5 ELBO and KL

The purpose of the expectation-maximization algorithm is to solve the parameter estimation of a mixed model with latent variables (maximum likelihood estimation. MLE estimates the parameter of $P(x|\theta)$ as: $\theta_{MLE} = argmax_\theta P(X|\theta)$. The algorithmic solution to this problem is an iterative approach:

$$\theta^{t+1} = argmax_\theta \int_Z logP(X|\theta)P(Z|X,\theta^t)$$

This formula consists of two steps of iteration:

E-step: Calculate the expectation of $\log P(X|\theta)$ under the probability distribution $P(Z|X,\theta^t)$

M-step: Calculate the parameters that maximize this expectation to get the input for the next EM step Prove: $logP(X|\theta^t) \leq logP(X|\theta^{t+1})$
Proof:
$logP(X|\theta) = logP(Z,X|\theta) - logP(Z|X,\theta)$, integrate the left and right sides:

Left:

$$\int_z P(Z|X,\theta^t) logP(X|\theta) dz = logP(x|\theta)$$

Right:

$$\int_z P(Z|X,\theta^t) logP(X,Z|\theta) dz - \int_z P(Z|X,\theta^t) logP(Z|X,\theta) dz = Q(\theta,\theta^t) - H(\theta,\theta^t)$$

Therefore:

$$logP(x|\theta) = Q(\theta,\theta^t) - H(\theta,\theta^t)$$

Since

$$Q(\theta,\theta^t) = \int_z P(Z|X,\theta^t) logp(x,z|\theta) dz$$

and

$$\theta^{t+1} = \underset{\theta}{argmax} \int_z log[P(X,Z|\theta)] P(Z|X,\theta^t) dz$$

so

$$Q(\theta^{t+1},\theta^t) \geq Q(\theta^t,\theta^t)$$

To prove $logP(X|\theta^t) \leq logP(X|\theta^{t+1})$, we need to prove:

$$H(\theta^t,\theta^t) \geq H(\theta^{t+1},\theta^t)$$

$$\begin{aligned} H(\theta^{t+1},\theta^t) - H(\theta^t,\theta^t) &= \int_z P(Z|X,\theta^t) \log P(Z|X,\theta^{t+1}) dz - \int_z P(Z|X,\theta^t) \log P(Z|X,\theta^t) dz \\ &= \int_z P(Z|X,\theta^t) \log \frac{P(Z|X,\theta^{t+1})}{P(Z|X,\theta^t)} \\ &= -KL(P(Z|X,\theta^t), P(Z|X,\theta^{t+1})) \leq 0 \end{aligned}$$

Above all:

$$logP(X|\theta^t) \leq logP(X|\theta^{t+1})$$

From the proof above, we would see that the likelihood function increases at each step. Then, let's see how the formula in the EM iteration process comes from:

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

Find the expectation on both sides $Eq(z)$:

Left:

$$\int_z q(z) \log p(x|\theta) dz = \log p(x|\theta)$$

Right:

$$\int_z q(z) \log \frac{p(z,x|\theta)}{q(z)} dz - \int_z q(z) \log \frac{p(z|x,\theta)}{q(z)} dz = ELBO + KL(q(z), p(z|x,\theta))$$

In the above formula, Evidence Lower Bound(ELBO), is a lower bound , so $\log p(x|\theta) \geq ELBO$, the equal sign is taken when the KL divergence is 0, that is: $q(z) = p(z|x,\theta)$, the EM algorithm The purpose is to maximize ELBO.

According to the above proof process, after each step of EM, the largest ELBO is obtained, and the parameter that maximizes ELBO is substituted into the next step:

$$\hat{\theta} = argmax_\theta ELBO = argmax_\theta \int_z q(z) \log \frac{p(x,z|\theta)}{q(z)} dz$$

Since $q(z) = p(z|x,\theta^t)$, the maximum value of this step can take the equal sign,

$$\begin{aligned}
\hat{\theta} &= argmax_\theta ELBO = argmax_\theta \int_z q(z) log \frac{P(X,Z|\theta)}{q(z)} dz \\
&= argmax_\theta \int_z P(Z|X,\theta^t)(log \frac{P(X,Z|\theta)}{P(Z|X,\theta^t)}) dz \\
&= argmax_\theta \int_z P(Z|X,\theta^t) log P(X,Z|\theta)
\end{aligned}$$

This formula is the above Equation during EM iteration. Starting from Jensen's inequality, this formula can also be derived, which we proved in 3.1.2

3.1.6 Generalized EM

The EM model solves the problem of parameter estimation of the probabilistic generation model. It learns θ by introducing the latent variable z, and the specific

model has different assumptions about z. For the learning task $P(X|\theta)$, it is the learning task $\frac{P(X,Z|\theta)}{P(Z|X,\theta)}$. In this formula, we assume that in step E, $q(z) = p(z|x,\theta)$, but if this $P(Z|X,\theta)$ cannot be solved, it must be Use methods such as sampling (MCMC) or variational inference to approximate this posterior. We observe the expression of KL divergence, in order to maximize ELBO, we need to minimize KL divergence at a fixed θ, so:

$$\hat{q}(z) = \underset{q}{argmin}\, KL(p,q) = \underset{q}{argmax}\, ELBO$$

This is the basic idea of generalized EM:

E-step:

$$\hat{q}^{t+1}(z) = \underset{q}{argmax} \int_z q^t(z) \log \frac{p(x,z|\theta)}{q^t(z)} dz, fixed\ \theta$$

M-step:

$$\hat{\theta} = \underset{\theta}{argmax} \int_z q^{t+1}(z) \log \frac{p(x,z|\theta)}{q^{t+1}(z)} dz, fixed\ \hat{q}$$

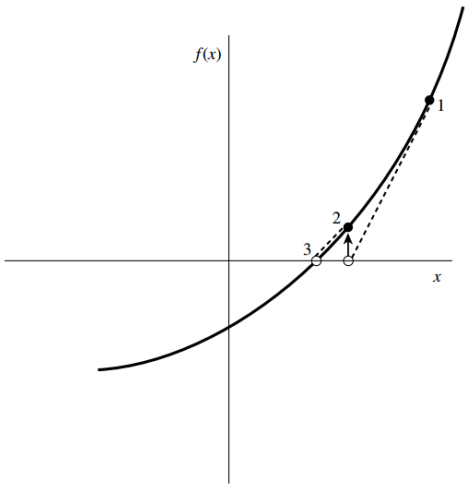
For the above integral:

$$ELBO = \int_z q(z) \log \frac{p(x,z|\theta)}{q(z)} dz = \mathbb{E}_{q(z)}[p(x,z|\theta)] + Entropy(q(z))$$

Therefore, we see that generalized EM is equivalent to adding the term entropy to the original formula.

3.2 Newton–Raphson algorithm

The Newton-Raphson algorithm, referred to as Newton's method, is one of the most famous one-dimensional root-finding methods(Deisenroth et al. 2020). Its characteristic is that the function value and its first derivative value need to be calculated at the same time(Hussein 2011). From a geometrical interpretation, Newton's method is to extend the tangent at the current point to make it intersect with the horizontal axis, and then use the value at the intersection as the next estimate point.



Newton's method is to let x descend in the direction of the gradient of f(x), similar to the gradient descent method in the optimization method. The Left is an example.

Newton's method can be obtained from the Taylor expansion of the function. The Taylor expansion of f(x) can be expressed as:

$$f(x+\delta) = f(x) + f'(x)\delta + f''(x)^2\delta^2 + O(\delta^3)$$

For a sufficiently small δ, only the first-order terms on the right-hand side of the above equation can be kept:

$$\delta = \frac{-f(x)}{f'(x)}$$

Then

$$x_i + 1 = x_i\delta = xi - \frac{-f(x_i)}{f'(x_i)}$$

Compared with the bisection method and the truncated chord method, the advantage of Newton's method is that the convergence speed could reach the second order, and there is no iteration near the root, and the effective number of the result can be almost doubled. Of course, Newton's method may also fail, such as converging to a local extreme value whose tangent direction is horizontal to the horizontal axis, so that the next iteration value cannot be calculated. In addition, the implementation of Newton's method requires the user to provide a function to calculate the function value $f(x)$ and its first derivative value $f'(x)$, so it is more suitable for the situation where the derivative of the function can be obtained analytically. Then the convergence speed and accuracy of Newton's method will be affected.

3.3 Semi-parametric Model

3.3.1 Parametric model

Before introducing the semi-parametric moedel, let's take a look at the parametric Model. ℓ stands for log likelihood, L stands for likelihood. take the Cox model as an example, to estimate $h_0(t)$ together with $X_i\beta_i$, using hazard function $f(t_i|x)$ and survival function $S(t_i;x)$ to estimate, according to Kleinbaum & Klein (1996) the corresponding likelihood function that needs to be optimized is:

$$\Theta_{MLE} = argmax_{\Theta} L(\Theta) = \prod f(t_i|x)^{\delta_i} S(t_i|x)^{1-\delta_i}$$

This likelihood function is to model the likelihood estimation of all the information. When we use it, we actually use $t_i, i = 1 \cdots, n$, the information at all time points. It includes the order relationship of death time and The time point when the specific s died, so we estimated both the time function $h_(t)$ and β

3.3.2 Introduction of semi-parametric Model

The PH Model and the AFT Model are exponential models of Hazard and Time, respectively(Chen & Ren 2020). However, parametric models mostly rely on strong assumptions. For example, assumptions about the baseline Hazard formula in the PH Model and assumptions about the residual formula in the AFT Model. When

the assumptions are not satisfied, the model generally fits poorly. Therefore, we introduce a semi-parametric model: Semi-parametric AFT Model, which is also known as Rank Regression

First, what is a semi-parametric model? Since the parametric model has high requirements for model assumptions, it is not robust enough. Therefore, we hope that the model assumptions can be appropriately reduced. Meanwhile, if we want to be able to preserve part of the interpret-ability of the model, we need to preserve the parameters appropriately. This is why it is called a semi-parametric model.

For instance, for a normal AFT model, $Y_i = X_i\beta + W_i$, $W_i \sim f$. If we want to apply the AFT model, we need to first assume that the residuals W follow a certain distribution, and then solve it using iterations. However, if we keep the parameters β, but relax our assumptions about the distribution of the residuals W (assuming that the residuals follow some distribution that is unknown). Then, this is a semi-parametric model. Set another example, as for the normal pH model, $h_i(t) = h_0(t)exp(X_i\beta)$, if we want to apply the PH model, we need to assume $h_0(t)$. However, if our assumptions are not true, the model performance will generally be unsatisfactory. Therefore, we keep the parameter β, but relax the assumptions about the base hazard function h_0. The above is also a semi-parametric model.

In conclusion, in survival analysis, we always keep the linear parametric part for the parametric part, i.e. the way the covariates affect the distribution. Relax the assumptions about the survival distribution, that is, the non-parametric part, relax the specific assumptions about the survival distribution, and keep only the basic form.

3.3.3 Application of semi-parametric Model

The full parameter model relies too much on the assumptions about the specific functional form of h(t) and $e^{X\beta}$, we now choose to estimate only h(t) or only the parameters of $e^{X\beta}$, that is, a part is regarded as MLE The non-parameter part that cannot be processed, one is the parameter part. Such a compromise method

is called the semi-parameter method. Generally, we can't guess the baseline risk rate

However, we prefer to study the effect of covariates on the risk rate, so we use the semi-parametric method to estimate only β, Again, we take Cox model as an example, and the constructed likelihood function that needs to be optimized is:

$$\hat{\beta}_{MLE} = argmax_{\beta} \prod_{i}^{n} \frac{h_i(t_i, x_i)}{\sum_{j=1}^{n} h_j(t_j, x_j)} = argmax_{\beta} \prod_{i}^{n} \frac{e^{\sum_i^n X_i \beta}}{\sum_{j=1}^{n} e^{\sum_j^n X_j \beta}} \tag{15}$$

This likelihood function is called partial likelihood. It does not use all the death time information, but only uses the order of the death time of different people, that is, only the order relationship of $t_1 \leq \cdots t_n$ is used. Among them, $\frac{h_i(t_i,x_i)}{\sum_{j=1}^{n} h_j(t_j,x_j)}$ represents the probability that we have n-i-11 people left, and now it is the i-th person's turn to die. Since we only use the ordinal relationship to construct the likelihood, we cannot estimate the time function baseline risk function $h_0(t)$ without time point data, but it is enough to estimate β, so it seems that we are missing more data information, but compared to the hard-to-guess the huge negative impact of the shape of h_0 on the estimate, it may be better to lose point information to estimate β The partial likelihood function is not a real likelihood, because the distributions corresponding to all events are not continuously multiplied, but only for the In this case, we only care about the multiplicative probability corresponding to the order relationship, but Cox proved that this likelihood function and its estimator satisfy most of the properties satisfied by the ordinary likelihood function L

3.4 Rank Statistics

The event-rank statistic is a commonly used concept in mathematical statistics (Efron & Stein 2007). Its advantage lies in the weak assumption about the distribution of the random variable, which is very suitable for our current requirements.

First, we define: $y^{(i)}$ is the ith largest corresponding variable after sorting; $x^{(i)}$ is the corresponding explanatory variable. Then Rank Statistics can be expressed as $\sum_i (i - \bar{x})(x_(i) - \bar{x})$

where $(i - \bar{i})$ represents statistics for y. Since we have no distribution assumptions about y, its actual value is not important, but its relative position i is what we import. The $(x_(i) - \bar{x})$ represents the statistics of x. The whole represents whether the position of y is regular with respect to the value of x. If x and y are in a perfectly positive relationship, this statistic should achieve a large value. Therefore, the rank statistic shows the correlation between x and y. When the rank statistic is 0, it can be considered that there is no correlation between x and y.

Then, given a set of parameters $\{\beta^j\}_{j=1}^K$, we can test the rank statistic of $y_i - X_i\beta^j$ versus x_i . If the rank statistic is close to 0, it means that we have extracted enough information x_i. At the same time, since the rank statistic has asymptotic normality, it can be tested to determine whether the model is significant compared to $\beta = 0$

We consider the case that x is one-dimensional, and the multi-dimensional case can be directly obtained in a similar way. First, we give the first null hypothesis and alternative hypothesis. This test is mainly aimed at whether the model is significant:

$$H_0 : \beta = 0 \ or \ H_1 : \beta \neq 0$$

Then, $U = \sum_j (x_{(j)} - \bar{x}_{(j)})$is the corresponding rank statistic. Among them, $x_{(j)}$ represents the individuals who died at the time point $t_{(j)}$, and $\bar{x}_{(j)}$ represents the individuals who were still alive at the time point $t_{(j)}$.

Why is this a rank statistic? First of all, the information of y is implicit in the time point $t_{(j)}$. Since the individual $x_{(j)}$ at each time point $t_{(j)}$ is the first place in the order, it does not need to be multiplied by $(i-\bar{(}i))$. If the null hypothesis holds i.e. $\beta = 0$, then there should be no information about y in x, and the deaths of individuals should be random and in no predetermined order. However, in reality we do observe a set of orders of individual deaths, so this statistic tests whether the order of individual deaths is indeed independent of the explanatory variable x.

Therefore, if the null hypothesis holds, we can calculate the variance $V = \sum_j (x_{(j)} - \bar{x}_{(j)})^2$ of the rank statistic U, and then use $\frac{U^2}{V} \sim \chi^2$ to test it.

Further, we can give a second hypothesis test:

$$H_0 : \beta = \beta_0 \ or \ H_1 : \beta \neq \beta_0$$

Similar to the above method, $U = \sum_j (x_{(j)} - \bar{x}_{(j)})$ is also used as the rank statistic. It's just that our sorting is no longer by time of death t_j, but by $W_i = Y_i - X_i\beta_0 = log(t_i) - X_i\beta$. The meaning of this test is to check whether the residuals also contain information about the explanatory variable $x_{(j)}$. Further, we can use the same method as above to check. Of course, we have reason to believe that the smaller the chi-square value corresponds the less the information of the $X_i\beta$ in the residuals.

However, the above method can only choose from a set of parameters $\{\beta\}$, there is no way to directly obtain the optimal $\hat{\beta}$. It cannot be solved using iteration, nor using Wald to make confidence intervals. Therefore, in practical applications, rank-based estimation is very limited.

4 Estimation

According to Zhang & Peng (2009), we call the subjects who have never experienced the event as cured subjects (not susceptible). The remaining subjects that uncured are susceptible. T represents the failure time of interest, and the survival function of T is $S_(t|x,z)$. x and z are the observed values of the two covariate vectors. Therefore, the mixture cure model can be written as:

$$S(t|x,z) = \pi(z)S(t|x) + 1 - \pi(z) \tag{16}$$

Where $\pi(z)$ denotes the probability that the patient is not cured, which depends on z. the survival function of the failure time distribution of uncured patients$S(t|x)$ is , which depends on x. "Incidence" refer to the $\pi(z)$ model, and "delay" refer to the $S(t|x)$ model.
Applied into Semi-parametric AFT mixture cure model we can get the incidence component is:

$$\pi(z) = \frac{e^{(bz)}}{1 + e^{(bz)}}$$

where b represents unknown parameters (a row vector). $1-\pi(z)$ is the proportion of cured patients. From (5) we can get the latency component function:

$$log(T) = \beta x + \varepsilon$$

S is the corresponding survival function. Given that the patient is not cured, the conditional survival function of T is $S(log(t) - \beta x)$. If we let $(t_i, \delta_i, z_i, x_i)$denote the observed data of the i-th individual $i = 1, 2, \ldots n$. And here t_i is the observed survival time of the $i-th$ patient. δ_i is the censored index, into detail, $\delta_i = 1$ means uncensored time, $\delta_i = 0$ means censored time. z_i and x_i are possible covariates. Using $\mathbf{O} = (t_i, \delta_i, z_i, x_i)i = 1, 2, \ldots, n$ to represent the observed data, since the censoring is independent, applied into (8), from $i-th$ patient, the contribution to the likelihood is:

$$\pi(z_i)f(log(t_i) - \beta X_i)/t_i \qquad when \quad \delta_i = 1 \tag{17}$$

$$1 - \pi(z_i) + \pi(z_i)S(log(t_i) - \beta X_i) \quad when \quad \delta_i = 0 \tag{18}$$

In order to put it together, we write the observed likelihood function as following:

$$L(b,\beta,S(\cdot);O) \propto \prod_{i}^{n}[\pi(z_i)f(log(t_i)-\beta X_i)]^{\delta_i}[1-\pi(z_i)+\pi(z_i)S(log(t_i)-\beta X_i)]^{1-\delta_i}$$

Generally, we can directly maximize $L(b,\beta,S(\cdot);O)$. However, due to the lack of $f(\cdot)$ and $S(\cdot)$, it does not fit the semiparametric AFT mixture cure model any more.

Therefore, we need to find another method to solve the problem, which is exact EM method. To estimate unknown parameters(b,β) and the survival function $S(\cdot)$ in the AFT model, we define y_i be a latent random variable:

$$\begin{aligned} y_i &= 1 \qquad if\ the\ ith\ individual\ is\ not\ cured \\ y_i &= 0 \qquad if\ the\ ith\ individual\ is\ cured \end{aligned}$$

Compared with the definition of δ_i before(equation(17)(18)), we can find that

$$\begin{aligned} y_i &= 1 \qquad\qquad if \quad \delta_i = 1 \\ y_i\ &is\ unknown \qquad if \quad \delta_i = 0 \end{aligned}$$

Then, we try to write the complete likelihood function:

$$L_c(b,\beta,S(\cdot);O,y) = L_{c1}(b;O.y)\cdot L_{c2}(\beta,S(\cdot);O,y)$$

where

$$L_{c1}(b;O,y) = \prod_{i=1}^{n}[\pi(z_i)]^{y_i}[1-\pi(z_i)]^{1-y_i}$$

$$L_{c2}(\beta,S(\cdot);O,y) = \prod_{i=1}^{n}[h(log(t_i)-\beta x_i)]^{y_i\delta_i}[S(log(t_i)-\beta x_i)]^{y_i}$$

To make estimation easier, we change them into complete log likelihood

$$l_c(b,\beta,S(\cdot);O,y) = l_{c1}(b;O.y) + l_{c2}(\beta,S(\cdot);O,y)$$

where

$$l_{c1}(b; O, y) = \sum_{i=1}^{n} y_i log[\pi(z_i)] + (1 - y_i) log[1 - \pi(z_i)] \tag{19}$$

$$l_{c2}(\beta, S(\cdot); O, y) = \sum_{i=1}^{n} y_i \delta_i log[h(log(t_i) - \beta x_i)] + y_i log[S(log(t_i) - \beta x_i)] \tag{20}$$

Here $h(\cdot)$ is the hazard function of ε and $h(\cdot) = f(\cdot)/S(\cdot)$. We can find that after changing likelihood function into log, the latent random variable y_i become linear, which benefit to our following estimation.

The basic idea of the EM method we have mentioned in 3.1. In this problem, given the current estimates $\Theta^{(m)} = \{b^{(m)}, \beta^{(m)}, S^{(m)}(t)\}$ and observed data O.

The E-step computes the conditional expectation of the full log-likelihood of the latent variable y_i, $i = 1, 2, 3 \cdots n$. $E(y_i|\Theta^{(m)}, O)$ is the conditional probability that the ith individual is still not cured in the mth iteration of the algorithm

$$E(y_i|\Theta^{(m)}, O) = \delta_i + (1 - \delta_i) \frac{\pi(z_i) S(log(t_i) - \beta x_i)}{1 - \pi(z_i) + \pi(z_i) S(log(t_i) - \beta x_i)} |_{(\Theta_{(m)}, O)} \tag{21}$$

$$E(l_c|\Theta_{(m)}, O) = l_{c1}(b) + l_{c2}(\beta, S(\cdot)) \tag{22}$$

we denote $E(y_i|\Theta^{(m)}$ as $w_i^{(m)}$. According to equation(19)(20), we would get the following:

$$l_{c1}(b) = \sum_{i=1}^{n} w_i^{(m)} log[\pi(z_i)] + (1 - w_i^{(m)}) log[1 - \pi(z_i)] \tag{23}$$

$$l_{c2}(\beta, S(\cdot)) = \sum_{i=1}^{n} w_i^{(m)} \delta_i log[h(log(t_i) - \beta x_i)] + w_i^{(m)} log[S(log(t_i) - \beta x_i)] \tag{24}$$

The M-step is to maximize (23)(24), we see that b, β and $S(\cdot)$ are unknown. For b, we can use the Newton–Raphson algorithm that mentioned before to maximize

(23). However, since the survival function $S(\cdot)$ is not a fixed function, it is difficult for us to achieve the maximization task of β and $S(\cdot)$.

Therefore, we came up with an alternative method to convert the semi-parametric AFT mixture curing model to the log-likelihood function of the standard semi-parametric AFT model, excluding constant $w_i^{(m)} \equiv 1$. Also, $\delta_i log w_i^{(m)} \equiv 0$ and $\delta_i w_i^{(m)} \equiv \delta_i$, we now would rewrite equation(26) as following:

$$l_{c2}(\beta, S(\cdot)) = \sum_{i=1}^{n} \delta_i log[w_i^{(m)} h(log(t_i) - \beta x_i)] + w_i^{(m)} log[S(log(t_i) - \beta x_i)] \quad (25)$$

Then take the anti-log of (25), we can get:

$$\begin{aligned} L_{c2}(\beta, S(\cdot)) &= \prod_{i}^{n} [w_i^{(m)} h(log(t_i) - \beta x_i)]^{\delta_i} [S(h(log(t_i) - \beta x_i))]^{w_i^{(m)}} \\ &= \prod_{i}^{n} [w_i^{(m)} h(log(t_i) - \beta x_i)]^{\delta_i} [S(h(log(t_i) - \beta x_i)^{w_i^{(m)}}] \end{aligned}$$

According to:

Hazard function: $w_i^{(m)} h(log(t_i) - \beta x_i)$

Survival function: $S(log(t_i) - \beta x_i)^{w_i^{(m)}}$

We can take (25) as log likelihood function of the AFT model. $log(T_i) = \beta x_i + \varepsilon^*$ where $w_i^{(m)} h(\varepsilon^*)$ is the hazard function of ε^*. We are now able to estimate β in M-step based on the method of the semi-parametric AFT model.

Next, we continue to apply the partial likelihood principle rank estimation method to consider the usual PH model with regression coefficients ξ

$$\hat{h}(\varepsilon_i^*) = w_i^{(m)} h(\varepsilon_i^*) e^{\xi x} \quad (26)$$

We can see that when the coefficient $\xi = 0$, this formula perfectly satisfies the risk

function in the AFT model.

Therefore, our general idea is as following:
First, assume that we don't know ξ. Then, use the partial likelihood method to derive a system of equations for $\xi and \beta$. After that, we re-substitute $\xi = 0$ into the equation system. Thus, now we can get the estimated equation system only about β.

As for ξ, the derivative of the log-partial likelihood function for equation(26) is:

$$G(\xi) = \sum_{i}^{n} \delta_i (x_i - \frac{\sum_j x_j w_j^{(m)} e^{\xi x_j} I(\varepsilon_i^* \leq \varepsilon_j^*)}{\sum_j w_j^{(m)} e^{\xi x_j} I(\varepsilon_i^* \leq \varepsilon_j^*)}) \tag{27}$$

where I$(\cdot)$ is the indicate function. When ξ is 0, just like we mentioned before, $G(0) = 0$ can be used as a linear rank-like estimating equation, which means β is linear to the above equation now(Efron & Stein 2007).

Now, we want to extend this formula. If we are under the condition that it can be shown that I(0) is an average 0 martingale . Now, this formula can be added as a function containing weights $k(\cdot)$. That is, we extend G(0) to $G(\beta; k(\cdot))$:

$$G(\beta; k(\cdot)) = \sum_{i}^{n} \delta_i k(\varepsilon_i^*)(x_i - \frac{\sum_j x_j w_j^{(m)} e^{\xi x_j} I(\varepsilon_i^* \leq \varepsilon_j^*)}{\sum_j w_j^{(m)} e^{\xi x_j} I(\varepsilon_i^* \leq \varepsilon_j^*)}) \tag{28}$$

To push the estimation, we need to introduce a new method, using Gehan weight function, $k(u) = \sum_j I(u \leq \varepsilon_j^*)/n$, which lead equation(28) to be monotone (Fygenson & Ritov 1994). By imitating the function and adding $w^{(m)}$, we define a new function as following: $k(u) = \sum_j I(u \leq \varepsilon_j^*) w_j^{(m)}/n$
Bringing the weight formula back to equation (28), we can get a new monotonic

estimating function:

$$\begin{aligned} G(\beta; k(\cdot)) &= \sum_i^n \delta_i k(u)(x_i - \frac{\sum_j x_j w_j^{(m)} e^{\xi x_j} I(\varepsilon_i^* \le \varepsilon_j^*)}{\sum_j w_j^{(m)} e^{\xi x_j} I(\varepsilon_i^* \le \varepsilon_j^*)}) \\ &= \frac{1}{n} \sum_i^n \delta_i \sum_j I(\varepsilon_i^* \le \varepsilon_j^*) w_j^{(m)} (x_i - \frac{\sum_j x_j w_j^{(m)} e^{\xi x_j} I(\varepsilon_i^* \le \varepsilon_j^*)}{\sum_j w_j^{(m)} e^{\xi x_j} I(\varepsilon_i^* \le \varepsilon_j^*)}) \\ &= \frac{1}{n} \sum_i^n \sum_j^n \delta_i w_j^{(m)} (x_i - x_j) I(\varepsilon_i^* \le \varepsilon_j^*) \end{aligned}$$

From above we know that this estimation function is monotonic. Therefore, from the properties of monotone function solutions, if there is a solution for $G(b; k(\cdot)) = 0$, it is unique and consistent. At the same time, $G(b; k(\cdot))$ has asymptotic normality because it can be written as a U-statistic with a symmetric kernel(Kunstner et al. 2011)

Meanwhile, using the Gehan-type weight function, the estimating function $G(\beta; k(\cdot)) = \frac{1}{n} \sum_i^n \sum_j^n \delta_i w_j^{(m)} (x_i - x_j) I(\varepsilon_i^* \le \varepsilon_j^*)$ can be taken as the gradient of a convex function.

$$g(\beta) = \frac{1}{n} \sum_i^n \sum_j^n \delta_i w_j^{(m)} |\varepsilon_i^8 - \varepsilon_j^*| I(\varepsilon_i^* \le \varepsilon_j^*) \tag{29}$$

According to Minty (1964), if we want to minimizing this convex function, what we need is finding the root of G(b; k(·)) = 0. By using the linear programming method, we can find out it easily. Also, Zhang & Peng (2009) claimed that to other choices of the weight function, the estimation method can be extended too.

Now β has been non-parametric estimated, therefore, accoding to the residuals function $\varepsilon = log(t_i) - \beta x_i$, $S(\cdot)$ can be an estimated based on the complete log-likelihood function(24) (Kleinbaum & Klein 1996).

We now need to go into detail. Zhang & Peng (2009) give us an example: Set the distinct uncensored failure residuals be $r_1 < r_2 < \cdots < r_k$ and the set of failures be K_{r_j}. $R(r_j)$ denote the risk set at r_j

M-step: $\hat{S}^{(m+1)}(\varepsilon) = exp(-\sum_{j:r_j<\varepsilon}(\frac{K_{r_j}}{\sum_{i\in R(r_j)} w_i^{(m)}}))$

where $\hat{S}^{(m+1)}(\varepsilon)$ is an estimate of S$(\cdot)$ in the current.

E-step: If $\varepsilon > r_k$, then let $S^{(m+1)}(\varepsilon) = 0$. $w_i^{(m)}$ in (6) is updated, that is, the baseline distribution in semi-parametric AFT mixture curing model is updated.

The above reflects the change process from the semi-parametric AFT mixture curing model to the original semi-parametric AFT model. If there are no cured patients ($y_i \equiv 1; w_i \equiv 1,$) anymore, the E step in the EM algorithm stops, at which point the program is reduced to the method of the semi-parametric AFT model.

5 Conclusion and Discussion

Through the research on the contributions and challenges of many research papers, this paper preliminarily introduces the basic theory of survival analysis and conventional application models, including proportional risk model and accelerated failure time model. However, in real life, the original model has considerable limitations. Due to the lack of a variety of data, it is difficult for us to apply it directly. Therefore, based on this problem, this paper first reviews the basic research methods and research steps. Then, the semi-parametric accelerated failure time model is studied by using the EM method through the maximum likelihood method and Newton method. Then, a new research method is proposed by using Gehan-weight function and convex function, which completes the theoretical research from semi parametric AFT model to AFT mixture cure model.However, since we have not applied specific data for the practical application of the model, the convincing is not strong enough, and this paper is limited to theoretical.

6 References

Chen, W. & Ren, F. L. (2020), 'Polynomial-based smoothing estimation for a semi-parametric accelerated failure time partial linear model', *Open Access Library Journal* **7**, 1–15.
URL: *10.4236/oalib.1106824.*

Cox, D. R. (1972), 'Regression models and life-tables', *Journal of the Royal Statistical Society* **34**, 187–220.
URL: *http://www.jstor.org/stable/2985181*

Deborah, V. D., Derek, R. B. & P, B. L. (2021), 'Burt and eklund's dentistry, dental practice, and the community", *ScienceDirect* pp. 131–153.
URL: *https://www.sciencedirect.com/science/article/pii/B9780323554848000137*

Deisenroth, M. P., Faisal, A. A. & Ong, C. S. (2020), *Mathematics for Machine Learning*, Cambridge University Press.
URL: *https://mml-book.github.io/book/mml-book.pdf*

Efron, B. & Stein, C. (2007), 'The jackknife estimate of variance', *The Annals of Statistics* **9(3)**, 586–596.
URL: *https://doi.org/10.1214/aos/1176345462*

Fu, L. Y., Yang, Z. R., Zhou, Y. & Wang, Y. G. (2021), 'An efficient gehan-type estimation for the accelerated failure time model with clustered and censored data', *Lifetime Data Analysis* **27**, 679–709.
URL: *https://doi.org/10.1007/s10985-021-09526-4*

Fygenson, M. & Ritov, Y. (1994), 'Monotone Estimating Equations for Censored Data', *The Annals of Statistics* **22**, 732–746.
URL: *https://doi.org/10.1214/aos/1176325493*

Goel, M. K., Khanna, P. & Kishore, J. (2010), 'Understanding survival analysis: Kaplan-meier estimate', *International Journal of Ayurveda Research* pp. 274–278.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3059453/*

Hussein, E. M. (2011), *Computed Radiation Imaging*, Elsevier, London.
URL: *https://www.sciencedirect.com/science/article/pii/B9780123877772000100*

Kleinbaum, D. G. & Klein, M. (1996), *Survival Analysis*, Springer.
URL: *http://www.springer.com/series/2848*

Kolaczyk, E. D. (2009), *Statistical Analysis of Network Data: Methods and Models*, Springer New York, New York.
URL: *https://doi.org/10.1007/978-0-387-88146-1_1*

Kunstner, F., Kumar, R. & Schmidt, M. (2011), 'Homeomorphic-invariance of EM: non-asymptotic convergence in KL divergence for exponential families via mirror descent', *CoRR* .
URL: *https://arxiv.org/abs/2011.01170*

Minty, G. (1964), 'On the monotonicity of the gradient of a convex function.', *Pacific Journal of Mathematics* .

Tolley, H. D., Barnes, J. M. & Freeman, M. D. (2016), 'Forensic epidemiology', *ScienceDirect* pp. 261–284.
URL: *https://www.sciencedirect.com/science/article/pii/B9780124045842000100*

Zhang, J. & Peng, Y. (2009), 'Accelerated hazards mixture cure model', *Lifetime data analysis* **15**, 455–467.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2903637/*