ANALYSING INCIDENT AND PREVALENT COHORT SURVIVAL DATA

HONORS RESEARCH PROJECT

SUMMER 2025

An Introduction to Survival Analysis

Author: Jiajun ZHANG
Supervisor: Professor Masoud ASGHARIAN

Contents

Chapter 1

Preface and Motivation

Fall 2024, an ordinary semester for many McGill students, but extraordinary for me. It was my third year at McGill University, but my very first semester as a student in Honors Applied Mathematics program. I transferred from the Honors Statistics and Computer Science program to the Honors Applied Mathematics program after my second year, it was when I realized I might not be a good match for computer science program, since I had very little programming knowledges before attending my first computer science class. Also several computer science classes made my GPA low, and I didn't enjoy them.

However, I was always passionate about mathematics, even at a very young age. I had Gaokao (Chinese College Entrance Examination) back in 2022, where my passion in mathematics reached its top, I viewed solving an extremely hard program as ultimate achievement, and always got immersed in math problems that I constantly forgot the time to eat and to sleep. Confidence enough to say, math gave me the confidence and courage to score an excellent score on my Gaokao grades.

My appreciation in math not only stays on the fact that I simply enjoy it and think it is fun, but I also enjoy the process of learning, and spread my knowledges to other people. I signed up for several tutoring programs and served as a tutor, mentor to helping students with difficulties in math. I remembered that when I was tried of fixing bugs in my program, I would always open a math book or simply do some exercises from my math class or going to office hours to help others to keep me energetic and alive.

For me, I enjoy standing in front of the blackboard and lecturing, I enjoy the process of speaking out something that I know but maybe other people don't. I stood up and created my own video channel on Bilibili, a video platform just like YouTube and share everything I know about the subject itself. Despite having few views and little profit, my passion and appreciation persuaded me to move on.

After a throughtful consideration I made the decision to drop my program in Computer Science and I joined a brand-new program in Honors Applied Mathematics as a third year student, and I walked into Professor Masoud Asgharian's probability class as my first class that year.

I still remembered during my first class, he asked us " Why do we want to study probability?", someone answered "Gambling" and we all laughed, and it was also then I developed a strong interest in probability and was eager to know how things would work, so I reached out to Professor Masoud Asgharian for potential research topcs.

Professor Masoud Asgharian is extremely kind and knowledgeable, we discussed several topics, but what truly fasincated me was survival analysis. We agreed on this topic in early 2025.

This report was written in 2025 as part of my undergraduate Honors Research Project at McGill University under Professor Masoud Asgharian's supervision starting from Febuary 2025 to August 2025. This report serves as an introduction to survival analysis and summarizes some modern approaches to survival data. Below is a list of topics I included in this article:

- **Chapter 2**: Censoring and Truncation, introduces key survival features that distinguish survival analysis from other statistical methods.
- **Chapter 3**: Parametric inference via likelihood methods, including derivations of likelihood function and maximum likelihhod estimator under different types of censoring and truncation.
- **Chapter 4**: Non-parametric methods, featuring Kaplan-Meier, Nelson-Aalen, and Aalen-Johansen estimators, variance formula and multi-states models.
- **Chapter 5**: A theoretical framework includes asymptotic behaviors of non-parametric estimators using counting processes and martingale theory, based on Fleming and Harrington, 2005.
- **Chapter 6**: Non-parametric hypothesis testing methods for comparing survival and hazard functions.
- **Chapter 7**: Cox's semi-parametric proportional hazards model.

I sincerely wish the reader finds this article both informative and inspiring.

Best,

Jiajun Zhang

Chapter 2

Censoring and Truncation

2.1 Introduction

In survival analysis, we are interested in time to event data, i.e we are interested in the time until the occurrence of some events which are of our interests. For example, we would like to study the effectiveness of using the drug 6-mercaptopurine for children with leukemia in a given time period (Freireich et al. ,1963), where our interests would be whether the children who received the treatment would have leukemia remission or not. However, due to cost it is sometimes not possible to study the entire sample space. For example, the time until the occurrence may vary in a very large sample and we usually may encounter extreme values, thus sometimes it is not possible to obtain the data for the entire sample as well. So our set up would be studying a given observable sample with appropriate observation time, for example 3 years. Then any individual from the observable sample whose occurrence of interest does not appear in this given period are said to be censored. That is, we only know the occurrence does not happen in our study, but it could occur anytime after the study, or we might just simply lose track of the data. On the other hand, there would also be individuals that are not in our sample. Those individuals could be systematically excluded from our study or we have no information on them. Those individuals are said to be truncated.

2.2 Definitions of Censoring and Truncation and Analysis of Examples

Briefly speaking, time to event data often present a characteristic feature, known as censoring and truncation. Censoring happens when the time to event is incompletely determined for some subjects, i.e. for some individuals we may only know the occurrence of our interest happened in a certain period, whereas for other individuals we will know the exact time of their occurrence. Truncation occurs when only a group of chosen individuals are observed by the investigator, those individuals whose are not of the interest are systematically excluded.

There are three main types of censoring, namely *left censoring, right censoring and interval censoring*. For a right censored observation, we only know the time of occurrence of our interest is larger than some value. For example, during $1982 - 1992$, 863 patients had their kidney transplant performed at the Ohio state university transplant center, and researchers investigated the survival time of those patients after the transplant, with the maximum follow-up time to be 9.47 years. Thus any individual who survived beyond this follow-up time are right censored. Also during the study, some patients were moved from

Columbus to other places where the researchers lost to follow up, those individuals are also right censored (Klein and Moeschberger, Chapter 1.7, 2003). There are two main types of right censoring, what we have discussed is called type (i) censoring, where a specific censoring time is known. Type (ii) censoring will be discussed in the next chapter, where we will set up a failure index. A left censored observation is one that is known only to be less than some value. For example, Turnbull and Weiss (1978) studied the first usage of marijuana among California high school boys, some answered with "I used it before, but I can't remember when." Those individuals are left censored (Klein and Moeschberger, Chapter 1.17, 2003). An interval censored observation is the one that is known to occur in a specified given interval. For example, the National Longitudinal Survey of Youth investigated a random sample of females aged 14 to 21 yearly from 1979 to 1988. Females in the survey were asked about any pregnancies that have occurred since they were last interviewed. Let's say an individual reported she didn't experience any pregnancies in the year 1984, but was reported pregnancy in the following year's survey. Then this individual is interval censored (Klein and Moeschberger, Chapter 1.14, 2003).

Truncation occurs when subjects have been at risk (event occurs) before entering the study or after, this means we do not have access to any information on them and they are said to be truncated. This also includes the case where we systematically exclude some part of our samples, or the part of population that we could not be observed. A study of ages of death was carried out among a group of individuals in a retirement center from January 1964 to July 1975. All individuals who did not fall in this study were truncated, since in order to enter the study they must survive to a sufficient age to enter the retirement community. Those who died prior to the study were not observed and were excluded from the study (Klein and Moeschberger, Chapter 1.16, 2003). In another study, Lagakos et al. reported data on the infection and induction times of AIDS among a group of individual. The data consists of the time (measured from April 1, 1978), when individuals were infected by AIDS from a contaminated blood transfusion, and the waiting time to the development of AIDS by June 30, 1986. In their sampling scheme only those whose waiting time from transfusion to AIDS was less than the time from transfusion to June 30, 1986 were in the study. Individuals transfused prior to June 30, 1986 and who developed AIDS after June 30, 1986 were not observed, and they did not lie in the study and were left truncated (Klein and Moeschberger, Chapter 1.19, 2003).

The key difference between censoring and truncation is that, in the event of censoring, we are guaranteed to know that the event time for some individual lies in a certain time interval, those samples are still in our study but due to cost and some other reasons we are not able to track the exact event time. However in the event of truncation, truncated samples never entered our study, or they have been excluded from our study due to some reason, and we usually have no information on them.

It is also important to note the difference between left censoring and left truncation. It may seem that every sample we have observed is already truncated, since we do not have the whole access to the entire sample. But that's not the case, and it really depends on the interests of our study. Like the example where Freireich et al. studied the treatment of drugs on children with leukemia. Then of course many samples were automatically excluded, like those adults with leukemia. But notice this is not considered a truncation since our interest only focus on the group of children. On the other hand, if the study is on the treatment among all individuals, then by only focusing on the group of children is considered as a design flaw, and then it is considered as a truncation.

2.3 Basic Quantities of Survival Data

In our observable sample, we will denote the life time of ith individual as X_i, and the entire population to be the random sample $\{X_1, \cdots, X_n\}$, which means they are independent and identically distributed

random variables (iid), with common probability density function (pdf) or probability mass function (pmf) $f(x)$ and cumulative distribution function (cdf) $F(x)$. Also by definition we are sure to know that X is non-negative. Given above, we define the survival function of any of those individual (denote by X), by

Definition 1. *Let X be a non-negative random variable which denotes the time to event of interests, let $F(x)$ be the cumulative distribution function (cdf) of X, then the survival function is defined as*

$$S(x) = \mathbb{P}(X > x) = 1 - F(x). \tag{2.3.1}$$

The survival function gives the probability that a subject will survive beyond the indicated time $t = x$. By the uniqueness of a distribution function, the survival function also uniquely defines a distribution. Another characteristic is called the hazard function, denoted by $h(x)$ or $\lambda(x)$, which is the instantaneous rate of experiencing the event of interest at time right after $t = x$ given that at the indicated time $t = x$ the event does not occur.

Definition 2. *Let X be a non-negative random variable that denotes the time to event of interests, then the hazard function of X, denoted by $h(x)$ or $\lambda(x)$ is defined as*

$$h(x) = \lim_{dx \to 0} \frac{\mathbb{P}(x \leq X \leq x + dx \big| X \geq x)}{dx}. \tag{2.3.2}$$

From definition 2, we have

$$\begin{aligned}
h(x) &= \lim_{dx \to 0} \frac{\mathbb{P}(x \leq X \leq x + dx \big| X \geq x)}{dx} && (2.3.3)\\
&= \lim_{dx \to 0} \frac{\mathbb{P}((x \leq X \leq x + dx) \bigcap (X \geq x))}{P(X \geq x) dx} && (2.3.4)\\
&= \lim_{dx \to 0} \frac{\mathbb{P}(x \leq X \leq x + dx)}{[1 - F(x)] dx} && (2.3.5)\\
&= \lim_{dx \to 0} \frac{F(x + dx) - F(x)}{[1 - F(x)] dx} && (2.3.6)\\
&= \frac{F'(x)}{1 - F(x)} = \frac{f(x)}{S(x)}. && (2.3.7)
\end{aligned}$$

Thus, the hazard function $h(x)$ can be viewed as the ratio of the pdf or pmf of X and the survival function $S(x)$. In addition, if we apply the Chain rule, we have

$$h(x) = -\frac{d}{dx} \log(S(x)). \tag{2.3.8}$$

Integrating both sides and interchanging the sign, we have

$$-\int h(x) dx = \log S(x) \tag{2.3.9}$$

then we apply a one-to-one transformation, we get

$$S(x) = \exp\left\{ -\int h(x) dx \right\}. \tag{2.3.10}$$

The integral term is denoted as the cummulative hazard function.

Definition 3. *The cumulative hazard function $H(x)$ (or $\Lambda(x)$) is the total risk up to a specified time $x = t$, which is defined by*

$$H(x) = \int_0^x h(t)dt. \tag{2.3.11}$$

Then by the fundamental theorem of calculus, we have

$$H(x) = \int_0^x -\frac{d}{dx}\log(S(t))dt = -\log(S(x)). \tag{2.3.12}$$

This results in the same formula as what we got before, and this relation would be very helpful later.

Definition 4. *The mean residual life function $m(x)$ is the expected value of the remaining lifetime after a specified time $x = t$, which is defined by*

$$m(x) = E(X - x \big| X > x) = \frac{\int_x^{+\infty} S(t)dt}{S(x)}. \tag{2.3.13}$$

which exists for all t if and only if $m(0) < +\infty$.

Above are the fundamental statistical features of survival data. In the next chapter, we will use those features to construct the likelihood inference based on a parametric family of distributions.

Chapter 3

Likelihood Inference for Survival Data Based on Parametric Families

3.1 Introduction

In this chapter, we will construct likelihood inferences for survival data based on parametric families, so our key assumption in this chapter is that in our sample $X_1, \cdots, X_n \overset{i.i.d}{\sim} f$ we know the exact function f up to some unknown parameters. Without any censoring or truncation, this would be a very straightforward parametric estimation problem, where maximum likelihood estimation (MLE) method is widely used, and we may also use the asymptotic normality of MLE to construct the confidence interval as well as likelihood ratio (LR) tests to gain further knowledge of the sample. With censoring and truncation introduced, we shall take them into consideration and the goal of this chapter is to construct the corresponding likelihood inference when censoring and truncation are present.

3.2 Likelihood Function

As we had demonstrated before, survival data is typically censored and as a result, the estimation of serval key features might be inaccurate. The likelihood, on the other hand, is a very versatile tool for quantifying whether a parameter value is consistent with the data, and this versatility makes it particularly well suited to survival analysis (Patrick Breheny, 2019). Instead of investigating on X, we are interested in the parameter θ in that family and would like to determine which θ fits the best.

Definition 5. *Let X be a random variable with distribution $f(x,\theta)$ where θ is a unknown parameter, then the likelihood function for θ is defined by*

$$\mathscr{L}(\theta) = \mathbb{P}(X = x, \theta). \tag{3.2.1}$$

The likelihood function measures the support provided by the data for each possible value of the parameter. We see that if X is a discrete random variable with corresponding probability mass function, the likelihood function is simply the probability mass function with θ being the variable. When X is a continuous random variable, we don't have the conclusion that $\mathbb{P}(X = x) = f(x)$, however we can show that if we choose small ε,

$$\mathbb{P}(x - \varepsilon \leq X \leq x + \varepsilon) = \int_{x-\varepsilon}^{x+\varepsilon} f(s)ds \approx 2\varepsilon f(x) \tag{3.2.2}$$

and instead of measuring the support on a specific point x, we measure the support over a small neighborhood of x, and since ε is a constant with respect to θ, and we may then use $\mathscr{L}(\theta) = f(x,\theta)$ as well for continuous random variables with a bit loss of rigor. Note that the likelihood function is not a distribution, since the value θ will depend on the sample. So we introduce relative likelihood function where we have a standardized measure:

Definition 6. *Let X be a continuous random variable with distribution $f(x,\theta)$, and suppose the maximal likelihood estimate of θ (MLE) is given by $\widehat{\theta}$ (i.e, $\arg\max \mathscr{L}(\theta) = \mathscr{L}(\widehat{\theta})$), then the relative likelihood function is defined as*

$$\mathscr{R}(\theta) = \frac{\mathscr{L}(\theta)}{\mathscr{L}(\widehat{\theta})} \tag{3.2.3}$$

It explains again why we can replace the probability by density for continuous random variables. Since we know rigorously speaking $\mathscr{L}(\theta)$ is proportional to $f(x,\theta)$ but then taking the ratio will cancel out the constant. It is also easy to see that $\mathscr{R}(\theta) \leq 1$. Also it is extremely useful when comparing the ratio of the likelihood function evaluated at different values of θ. If we find out that $\mathscr{L}(\theta_1) > \mathscr{L}(\theta_2)$, then the sample we actually observed is more likely to have occurred if $\theta = \theta_1$ than $\theta = \theta_2$, which can also be interpreted as saying θ_1 is a more plausible value for the true value of θ than is θ_2.

Suppose that we have an observable model $X_1, \cdots, X_n \overset{i.i.d}{\sim} f(x,\theta)$ where θ is an unknown parameter, then without any censoring or truncation, the likelihood for θ is simply

$$\mathscr{L}(\theta) == \prod_{i=1}^{n} \mathbb{P}(X_i = x_i, \theta) = \prod_{i=1}^{n} f(x_i, \theta) \tag{3.2.4}$$

and the MLE of θ is $\widehat{\theta} = \arg\max_{\theta\in\Theta}\{\mathscr{L}(\theta)\}$.

3.3 Likelihood Construction for Samples with Censored Data

3.3.1 Right Censored Samples with Fixed Censoring Time

For censored data, we know that the occurrence of events for some individuals lie in a certain interval but we do not know the exact time. We will first consider the general case for right censoring. Suppose we have a random sample of size n, where X_i denotes the event time for individual i, which also has a distribution f_{X_i}. Also C_i is the censoring time for individual i. For simplicity, We first fix each C_i to be a known constant, then we know that we would be able to observe X_i if and only if $X_i \leq C_i$. We define

$$T_i = \min\{X_i, C_i\} \text{ and } \delta_i = \begin{cases} 1 & X_i \leq C_i \\ 0 & \text{otherwise} \end{cases}, \tag{3.3.1}$$

where $t_i = \min\{x_i, C_i\}$ is the actual observable time we have, so in this case we have

$$t_i = \begin{cases} x_i & \delta_i = 1 \\ C_i & \delta_i = 0 \end{cases} \tag{3.3.2}$$

then we are interested in the probability

$$\mathbb{P}(T_i \ ; \ \delta_i = \{0,1\}). \tag{3.3.3}$$

In this case, we will be working with conditional probability. Note that this is different from the sense of truncation, we are conditioning on the value of δ_i. If $\delta_i = 1$, meaning the data is uncensored, so we are able to observe the exact occurrence of the event $X_i = x_i$, in this case we have

$$\mathbb{P}(T_i \ ; \ \delta_i = 1) = \mathbb{P}(T_i = X_i | \delta_i = 1)\mathbb{P}(\delta_i = 1) \tag{3.3.4}$$

Also this case $\delta_i = 1$ implies $X_i \leq C_i$, so we are interested in the conditional probability $P(T_i = X_i | X_i \leq C_i)$, which is given by

$$\mathbb{P}(T_i = X_i | X_i \leq C_i) = \frac{\mathbb{P}(T_i = X_i, X_i \leq C_i)}{\mathbb{P}(X_i \leq C_i)} \tag{3.3.5}$$

$$= \frac{\mathbb{P}(X_i = x_i)}{\mathbb{P}(X_i \leq C_i)} \tag{3.3.6}$$

$$= \frac{f_{X_i}(x_i)}{1 - S_{X_i}(C_i)} \tag{3.3.7}$$

where $S_{X_i}(C_i)$ is the survival function of X_i evaluated at C_i and $f_{X_i}(x_i)$ is the probability distribution function (pdf) of X_i evaluated at x_i. So we have

$$\mathbb{P}(T_i, \delta_i = 1) = f_{X_i}(x_i). \tag{3.3.8}$$

If $\delta_i = 0$, meaning the data is censored, and we only know the occurrence of the event lie in the interval $(C_i, +\infty)$, and in this case

$$\mathbb{P}(T_i, \delta_i = 0) = \mathbb{P}(X_i | \delta_i = 0)\mathbb{P}(\delta_i = 0) \tag{3.3.9}$$

Also we have $X_i \geq C_i$, and thus the observable event time is the censoring time C_i, thus

$$\mathbb{P}(X_i | \delta_i = 0) = \frac{\mathbb{P}(T_i = C_i, X_i \geq C_i)}{\mathbb{P}(X_i \geq C_i)} \equiv 1 \tag{3.3.10}$$

and we now have

$$\mathbb{P}(T_i, \delta_i = 0) = 1 \cdot \mathbb{P}(X_i \geq C_i) = S_{X_i}(C_i). \tag{3.3.11}$$

where $S_{X_i}(C_i)$ is the survival function of X_i. Then for the entire sample $X_1, \cdots, X_n$, we may derive the likelihood function (with parameter θ) as

$$\mathscr{L}(\theta) = \prod_{t_i \in \mathscr{C}} \mathbb{P}(T_i, \delta_i = 0) \cdot \prod_{t_i \in \mathscr{U}} \mathbb{P}(T_i, \delta_i = 1) = \prod_{t_i \in \mathscr{C}} S_{X_i}(C_i, \theta) \cdot \prod_{t_i \in \mathscr{U}} f_{X_i}(t_i, \theta) \tag{3.3.12}$$

where $\mathscr{C}$ is the set of all censored observations, and $\mathscr{U}$ is the set of all uncensored observations. If we use the δ_i notation as we introduced before, then the likelihood function in the sense of right censoring can be written as

$$\mathscr{L}(\theta) = \prod_{i=1}^{n} \mathbb{P}(T_i, \delta_i) = \prod_{i=1}^{n} [f_{X_i}(t_i, \theta)]^{\delta_i} \cdot [S_{X_i}(t_i, \theta)]^{1-\delta_i}. \tag{3.3.13}$$

We know that $f(x) = h(x)S(x)$ where $h(x)$ is the hazard function and $S(x)$ is the survival function. So we may rewrite the likelihood function as

$$\mathscr{L}(\theta) = \prod_{i=1}^{n} [h_{X_i}(t_i, \theta) S_{X_i}(t_i, \theta)]^{\delta_i} \cdot [S_{X_i}(t_i, \theta)]^{1-\delta_i} \tag{3.3.14}$$

which gives us

$$\mathscr{L}(\theta) = \prod_{i=1}^{n} [h_{X_i}(t_i, \theta)]^{\delta_i} \cdot S_{X_i}(t_i, \theta). \tag{3.3.15}$$

After we have obtained the likelihood function, it is then straightforward to compute the MLE. Suppose we have the model $X_1, \cdots, X_n \overset{i.i.d}{\sim} Exponential(\theta) = \theta e^{-\theta x}$, then by direct computation we have $S(x) = e^{-\theta x}$, and hence the log-likelihood function is given by

$$\log \mathscr{L}(\theta) = \sum_{t_i \in \mathscr{U}} \Big(\log \theta - \theta x_i \Big) - \sum_{t_i \in \mathscr{C}} \theta x_i \tag{3.3.16}$$

and the partial derivative yields

$$\frac{\partial \log \mathscr{L}(\theta)}{\partial \theta} = \sum_{t_i \in \mathscr{U}} \Big(\frac{1}{\theta} - x_i \Big) - \sum_{t_I \in \mathscr{C}} x_i \tag{3.3.17}$$

so it is easy to obtain the MLE of θ given by

$$\widehat{\theta}_n = \frac{|\mathscr{U}|}{\sum\limits_i X_i} \tag{3.3.18}$$

where $|\mathscr{U}|$ represents the number of uncensored observations.

3.3.2 Right Censored Samples with Undetermined Censoring Time

Now we would like to study the case when C_i is no longer fixed, and is often referred as the random censoringship model. In this case we will assume mutually independence of random samples $X_1, \cdots, X_n$ and $C_1, \cdots, C_n$, where each sample has its own probability density function (pdf) or probability mass function (pmf) $f_{X_i}(x_i), f_{C_i}(c_i)$ and the survival function $S_{X_i}(x_i), S_{C_i}(c_i)$. With the δ_i notation being the same, we first consider the case when $\delta_i = 0$ (censored observation), in this case we have

$$\mathbb{P}(T_i = t_i, \delta_i = 0) = \mathbb{P}(C_i = t_i, X_i > C_i) \tag{3.3.19}$$

From here, we may rewrite the above probability in terms of the joint density function of X_i, C_i, thus we have

$$\mathbb{P}(C_i = t_i, X_i > C_i) = \frac{d}{dt_i} \int_0^{t_i} \int_{c_i}^{+\infty} f_{X_i, C_i}(x_i, c_i) dx_i dc_i \tag{3.3.20}$$

Since X_i, C_i are independent, so we have

$$\begin{aligned} \mathbb{P}(C_i = t_i, X_i \geq C_i) &= \frac{d}{dt_i} \int_0^{t_i} f_{C_i}(c_i) dc_i \cdot \int_{c_i}^{+\infty} f_{X_i}(x_i) dx_i & (3.3.21) \\ &= f_{C_i}(t_i) \cdot S_{X_i}(t_i) & (3.3.22) \end{aligned}$$

where

$$t_i = \begin{cases} x_i & X_i \leq C_i \ \ (\delta_i = 1) \\ c_i & X_i > C_i \ \ (\delta_i = 0) \end{cases}. \tag{3.3.23}$$

If $\delta_i = 1$, which means now we have

$$\mathbb{P}(T_i = t_i, \delta_i = 1) = \mathbb{P}(X_i = t_i, X_i \leq C_i) \tag{3.3.24}$$

Thus we have

$$\mathbb{P}(X_i = t_i, X_i \leq C_i) = \frac{d}{dt_i}\int_0^{t_i}\int_{x_i}^{\infty} f_{X_i}(x_i, c_i)dc_i dx_i \tag{3.3.25}$$

Again by independence, we have

$$\mathbb{P}(X_i = t_i, X_i \leq C_i) = \frac{d}{dt_i}\int_0^{t_i} f_{X_i}(x_i)dx_i \cdot \int_{x_i}^{+\infty} f_{C_i}(c_i)dc_i \tag{3.3.26}$$

$$= f_{X_i}(t_i) \cdot S_{C_i}(t_i) \tag{3.3.27}$$

In this case, if we assume θ to be the unknown parameter, the likelihood function is given by

$$\mathscr{L}(\theta) = \prod_{i=1}^{n}[f_{X_i}(t_i) \cdot S_{C_i}(t_i)]^{\delta_i} \cdot [f_{C_i}(t_i) \cdot S_{X_i}(t_i)]^{1-\delta_i} \tag{3.3.28}$$

where we can further derive the expression above as

$$\mathscr{L}(\theta) = \left(\prod_{i=1}^{n}[S_{C_i}(t_i)]^{\delta_i} \cdot [f_{C_i}(t_i)]^{1-\delta_i}\right) \cdot \left(\prod_{i=1}^{n}[f_{X_i}(t_i)]^{\delta_i} \cdot [S_{X_i}(t_i)]^{1-\delta_i}\right) \tag{3.3.29}$$

and we may derive the MLE when f_{X_i}, f_{C_i} are known using standard techniques. If the distribution of C_i does not depend on the parameter θ, then the first term in above will be a constant with respect to the likelihood function $\mathscr{L}(\theta)$, and hence we obtain a general form of the likelihood function given by

$$\mathscr{L}(\theta) = K \cdot \prod_{i=1}^{n}[f_{X_i}(t_i)]^{\delta_i} \cdot [S_{X_i}(t_i)]^{1-\delta_i} \tag{3.3.30}$$

for some constant K which does not depend on θ. Previously where we discussed the likelihood function for fixed censoring time, we have $K \equiv 1$.

3.3.3 Construction of Type (II) Right Censoring

Previously, we mainly focused on type (i) right censoring, where we set a censoring time C_i for each individual X_i and we collect the data given by $T_i = \min\{X_i, C_i\}$. For type (ii) right censoring, we do not have a specific censoring time, instead we set up a failure index for our random sample with size n. That is, the study will continue until we observe the event time for the first r individuals where $r < n$. This type of censoring is commonly used in testing the expected lifetime of electronic devices. In this case, we can derive an order statistics with failure index r from the random sample:

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(r)} \leq X_{(r+1)} \leq \cdots \leq X_{(n)} \tag{3.3.31}$$

So we know that the set given by $\mathscr{U} := \{X_{(1)}, \cdots, X_{(r)}\}$ contains all the uncensored observations, and $\mathscr{C} := \{X_{(r+1)}, \cdots, X_{(n)}\}$ contains all the censored observations. we would first like to study the joint distribution of order statistics:

Lemma 1. *Let $X_i, \cdots, X_n \overset{i.i.d}{\sim} f$ with also a distribution function F and assume X_i is continuous. Then given $r < n$, the joint distribution of the first r order statistics of $X_{(1)}, \cdots, X_{(r)}$ is given by*

$$\mathfrak{f}_r(x_{(1)}, \cdots, x_{(r)}) = \frac{n!}{(n-r)!}\left(\prod_{i=1}^{r} f(x_{(i)}) \cdot \big(1 - F(x_{(r)})\big)^{n-r}\right) \tag{3.3.32}$$

Proof. It is not hard to derive the joint distribution of the order statistics $X_{(1)},\cdots,X_{(n)}$, where we have

$$\mathfrak{f}_n(x_{(1)},\cdots,x_{(n)}) = n!\prod_{i=1}^{n} f(x_{(i)}) \tag{3.3.33}$$

Thus it suffices to find the marginal distribution of $X_{(1)},\cdots,X_{(r)}$ from $\mathfrak{f}_n$, where by definition we have

$$\mathfrak{f}_r = n!\prod_{i=1}^{r} f(x_{(i)})\cdot\int_{x_{(r)}}^{\infty}\cdots\int_{x_{(n-1)}}^{\infty} f(x_{(n)})\cdots f(x_{(r+1)})dx_{(n)}\cdots dx_{(r+1)}. \tag{3.3.34}$$

Denote the integral part of above to be $\mathscr{I}$, we first integrate $\mathscr{I}$ with respect to $x_{(n)}$, we get

$$\mathscr{I} = \int_{x_{(r)}}^{\infty}\cdots\int_{x_{(n-2)}}^{\infty}(1-F(x_{(n-1)})f(x_{(n-1)})\cdots f(x_{(r+1)})dx_{(n-1)}\cdots d_{x_{(r+1)}} \tag{3.3.35}$$

Now integrate with respect to $x_{(n-1)}$, we get

$$\mathscr{I} = \int_{x_(r)}^{\infty}\cdots\int_{x_{(n-3)}}^{\infty}\frac{(1-F(x_{(n-2)}))^2}{2}f(x_{(n-2)})\cdots f(x_{(r+1)})dx_{(n-2)}\cdots d_{x_{(r+1)}} \tag{3.3.36}$$

If we again integrate with respect to $x_{(n-2)}$, we get

$$\mathscr{I} = \int_{x_{(r)}}^{\infty}\cdots\int_{x_{(n-4)}}^{\infty}\frac{(1-F(x_{(n-3)}))^3}{6}f(x_{(n-3)})\cdots f(x_{(r+1)})dx_{(n-3)}\cdots d_{x_{(r+1)}} \tag{3.3.37}$$

Thus in general we would have

$$\mathscr{I} = \frac{(1-F(x_{(r)}))^{n-r}}{(n-r)!} \tag{3.3.38}$$

hence we would get

$$\mathfrak{f}_r = \frac{n!}{(n-r)!}\prod_{i=1}^{r} f(x_{(i)})(1-F(x_{(r)})^{n-r}. \tag{3.3.39}$$

□

Notice that in the sense of survival analysis, the above expression becomes

$$\mathfrak{f}_r(x_{(1)},\cdots,x_{(r)}) = \frac{n!}{(n-r)!}\left(\prod_{i=1}^{r} f(x_{(i)})\cdot\left(S(x_{(r)}\right)^{n-r}\right) \tag{3.3.40}$$

where S denotes the survival function. In this case the likelihood function can be written as (assume θ to be our parameter)

$$\mathscr{L}(\theta) = \mathfrak{f}_r(x_{(1)},\cdots,x_{(n)})\cdot(S(x_{(r)}))^{n-r}. \tag{3.3.41}$$

Same as generalized type (i) censoring, we can also derive a generalized type (ii) censoring where we define different failure index, i.e we pre-decide some fixed integers n_i, r_i. In the study of n samples, firstly we observe the first r_1 occurrences, then we would have $n-r_1$ samples left. From those remaining samples we would remove n_1 samples, making our new samples of size $n-r_1-n_1$. Then we observe the occurrence of next r_2 samples, and n_2 of those remaining samples are removed, and we continue the study until our pre-decided series of repetitions is completed, i.e, we terminate the study when we

have observed r_k. In this case the censoring time $T_{r_1},\cdots,T_{r_k}$ are random and no longer fixed. Now for simplicity, we would only construct the likelihood function with 2 repetitions and fixed integers r_1,n_1,r_2. We let the first r_i occurrences to be the order statistics

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(r_1)} \tag{3.3.42}$$

When we are left with $n-r_1-n_1$ samples, and denote the next r_2 occurrences to be the order statistics

$$X^*_{(1)} \leq X^*_{(2)} \leq \cdots \leq X^*_{(r_2)} \tag{3.3.43}$$

After that, the study is terminated at $X_{(r_2)}$ with the remaining $n-r_1-n_1-r_2$ samples being censored. Note that those n_1 samples which were removed by us were not truncated: they are still in our interests and we know their occurrence lie inside $(X_{(r_1)},+\infty)$. The likelihood function can be written as

$$\mathscr{L}(\theta) = \mathbb{P}(X_{(1)},\cdots,X_{(r_1)})\cdot\mathbb{P}(X^*_{(1)},\cdots,X^*_{(r_2)}|X_{(1)},\cdots,X_{(r_1)}), \tag{3.3.44}$$

where by Lemma 1, the first term on the right hand side of (2.42) is given by

$$\mathbb{P}(X_{(1)},\cdots,X_{(r_1)}) = \frac{n!}{(n-r_1)!}\prod_{i=1}^{r_1} f(x_{(i)})\cdot(S(x_{(r_1)}))^{n-r_1} \tag{3.3.45}$$

Lemma 2. *Denote $P(X^*_{(1)},\cdots,X^*_{(r_2)}|X_{(1)},\cdots,X_{(r_1)}) = \mathscr{P}$, we have*

$$\mathscr{P} = \frac{(n-r_1-n_1)!}{(n-r_1-n_1-r_2)!}\prod_{i=1}^{r_2} f^*(x^*_{(i)})\cdot[S^*(x^*_{(r_2)})]^{n-r_1-n_1-r_2}. \tag{3.3.46}$$

where

$$f^*(x) = \frac{f(x)}{S(x_{(r_1)})}, S^*(x) = \frac{S(x)}{S(x_{(r_1)})}, x \geq x_{(r_1)} \tag{3.3.47}$$

are called the truncated probability density function f^* and the truncated survival function S^*.

So now by Lemma 1,2 the likelihood function is given by

$$\mathscr{L}(\theta) = K\cdot\prod_{i=1}^{r_1} f(x_{(i)})\cdot(S(x_{(r_1)}))^{n-r_1}\cdot\prod_{j=1}^{r_2}\frac{f(x^*_{(j)})}{S(x_{(r_1)})}\cdot\left(\frac{S(x^*_{(j)})}{S(x_{(r_1)})}\right)^{n-r_1-n_1-r_2} \tag{3.3.48}$$

where

$$K \equiv \frac{n!(n-r_1-n_1)!}{(n-r_1)!(n-r_1-n_1-r_2)!}. \tag{3.3.49}$$

3.3.4 Construction of Double Censoring and Truncation

Double censoring is a special case of interval censoring we discussed before, where in the case we only know the event time t satisfies $t<T$ or $T>t$ for some determined T. Recall the study where Turnbull and Weiss studied the usage of marijuana among California high school boys, and the question might be "Have you used marijuana?". if they can only answer with "yes" or "no", then the entire sample is

double censored. Here if we still have a common density function $f(x,\theta)$, distribution function $F(x,\theta)$ and survival function $S(x,\theta)$ for the random sample $X_1,\cdots,X_n$, then we have

$$\mathscr{L}(\theta) = \prod_{i=1}^{n} P(X_i > T)^{\delta_i^*} P(X_i < T)^{1-\delta_i^*} \tag{3.3.50}$$

$$= \prod_{i=1}^{n} [S(T)]^{\delta_i^*} \cdot [F(T)]^{1-\delta_i^*} \tag{3.3.51}$$

where

$$\delta_i^* = \begin{cases} 1 & X_i > T \\ 0 & X_i < T \end{cases} \tag{3.3.52}$$

Now in the sense of truncation, things becomes a bit different. Since we have systematically excluded some observations, what we observe now becomes the conditional probability and conditional distribution. We usually work with left truncation, meaning that we will only be able to observe the sample which satisfies $X_i > T$ for the pre-determined truncation time T. In this case the distribution time for the event time X_i is the conditional distribution of X_i given $X_i > T$, i.e

$$f(X_i = x_i | X_i > T_i) = \frac{f(x_i)}{S(T)} \tag{3.3.53}$$

Thus the likelihood is given by

$$\mathscr{L}(\theta) = \prod_{i=1}^{n} \frac{f(x_i)}{S(T)}. \tag{3.3.54}$$

3.4 Likelihood Inference on Single Parameters

After we constructed the likelihood function $\mathscr{L}(\theta)$, it is sometimes easier to work with the log-likelihood defined by $\ell(\theta) = \log(\mathscr{L}(\theta))$, and we define the *score* by

Definition 7. *Consider the random sample $X_1, \cdots, X_n \sim f(x, \theta)$, we define the score of X_i, denote by $U_i(\theta)$, as the derivative of the log-likelihood with respect to variable θ:*

$$U_i(\theta) = \frac{\partial}{\partial \theta} \ell_i(\theta) = \frac{\partial}{\partial \theta} \log\left(f(x_i, \theta)\right). \tag{3.4.1}$$

The score of the entire sample is just $U(\theta) = \sum_{i=1}^{n} U_i(\theta)$.

We would obtain some nice properties of $U(\theta)$ once we have the following regularity conditions:

(i) The family $\left\{ f(\boldsymbol{x}, \theta) : \theta \in \Theta \subset \mathbb{R} \right\}$ has a common support $\mathscr{S}$ that does not depend on θ, and θ is one-dimensional.

(ii) $\dfrac{\partial}{\partial \theta} \log f(\boldsymbol{x}, \theta)$ always exists.

(iii) For any statistic $h(\boldsymbol{X})$ such that $\mathbb{E}\{|h(\boldsymbol{X})|\} < \infty$,

$$\frac{\partial}{\partial \theta} \int_{\mathscr{S}} h(\boldsymbol{x}) f(\boldsymbol{x}, \theta) d\boldsymbol{x} = \int_{\mathscr{S}} h(\boldsymbol{x}) \frac{\partial}{\partial \theta} f(\boldsymbol{x}, \theta) d\boldsymbol{x}. \tag{3.4.2}$$

Assume the parametric family satisfy the above conditions, then we introduce two identities, known as the *Bartlett's Identities*.

Theorem 1. *Under regularity conditions, let $U(\theta)$ to be the score of the random sample $X_1, \cdots, X_n \sim f(x, \theta)$, then $\mathbb{E}[U(\theta)] = 0$ (First Bartlett's Identity), and $\mathbb{V}(U(\theta)) = -\mathbb{E}\left\{ \dfrac{\partial}{\partial \theta} U(\theta) \right\}$ (Second Bartlett's Identity).*

Proof. To show that $\mathbb{E}[U(\theta)] = 0$, we note that

$$\mathbb{E}[U(\theta)] = \mathbb{E}\left\{ \frac{\partial}{\partial \theta} \log f(\boldsymbol{X}, \theta) \right\} = \int_{\mathscr{S}} f(\boldsymbol{x}, \theta) \cdot \frac{1}{f(\boldsymbol{x}, \theta)} \cdot \frac{\partial}{\partial \theta} f(\boldsymbol{x}, \theta) d\boldsymbol{x}, \tag{3.4.3}$$

then using regularity conditions one can see that

$$\mathbb{E}[U(\theta)] = \frac{\partial}{\partial \theta} \int_{\mathscr{S}} f(\boldsymbol{x}, \theta) d\boldsymbol{x} = 0 \tag{3.4.4}$$

Then, to show $\mathbb{V}(U(\theta)) = -\mathbb{E}[U'(\theta)]$, note that

$$U'(\theta) = \frac{\partial^2}{\partial \theta^2} \log f(\boldsymbol{x}, \theta) = \frac{f''(\boldsymbol{x}, \theta)}{f(\boldsymbol{x}, \theta)} - \left(\frac{f'(\boldsymbol{x}, \theta)}{f(\boldsymbol{x}, \theta)} \right)^2 \tag{3.4.5}$$

where $f'(\boldsymbol{x}, \theta), f''(\boldsymbol{x}, \theta)$ are the first, second derivatives with respect to θ, and note that by regularity conditions, we have

$$\int_{\mathscr{S}} \frac{\partial^2}{\partial \theta^2} f(\boldsymbol{x}, \theta) \cdot \frac{1}{f(\boldsymbol{x}, \theta)} d\boldsymbol{x} = \frac{d}{d\theta} \mathbb{E}\left\{ \frac{d}{d\theta} \log f(\boldsymbol{X}, \theta) \right\} = 0 \tag{3.4.6}$$

Thus by taking the expectation of $U'(\theta)$, we have

$$\mathbb{E}[U'(\theta)] = -\mathbb{E}\left\{\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X},\theta)\right)^2\right\}. \tag{3.4.7}$$

Since $\mathbb{V}(U(\theta)) = \mathbb{E}[U^2(\theta)] - (\mathbb{E}[U(\theta)])^2$, the desired result follows trivially.

□

Definition 8. *The Fisher information of a random variable $X \sim f(x,\theta)$, denoted as $\mathscr{I}_1(\theta)$, is defined as*

$$\mathscr{I}_1(\theta) = -\mathbb{E}\left\{\frac{\partial^2}{\partial\theta^2}\log f(x,\theta)\right\}. \tag{3.4.8}$$

For a random sample $X_1,\cdots,X_n$, the total Fisher information is $\mathscr{I}(\theta) = \sum_{i=1}^{n}\mathscr{I}_i(\theta) = n\mathscr{I}_i(\theta)$.

Sometimes, we may use the observed Fisher information (or empirical Fisher information) $\widehat{\mathscr{I}(\theta)}$ to replace the actual Fisher information.

Definition 9. *The empirical Fisher information of one variable in a random sample $X_1,\cdots,X_n \sim f(x,\theta)$ is defined as*

$$\widehat{\mathscr{I}_1(\theta)} = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log f(X_i,\theta). \tag{3.4.9}$$

Theorem 2. *The empirical Fisher information converges to the actual one almost surely: $\widehat{\mathscr{I}_1(\theta)} \xrightarrow{a.s} \mathscr{I}_1(\theta)$.*

Proof. This result follows from *Glivenko-Cantelli* lemma:

$$\sup_{x\in\mathbb{R}}\left|\widehat{\mathscr{I}_1(\theta)} - \mathscr{I}(\theta)\right| \xrightarrow{\text{a.s}} 0. \tag{3.4.10}$$

□

Now we shall introduce some testing techniques using score function, namely Rao score Test; Wald test, and Likelihood Ratio (LR) test.

We first introduce score test, proposed by C. Rao, and the null hypothesis is $\mathscr{H}_0 : \theta = \theta_0$.

Proposition 1. *In the random sample $X_1,\cdots,X_n$ denote $U(\theta_0)$ as the score function, and $\mathscr{I}_1(\theta_0)$ is the Fisher information of a single variable evaluated at θ_0, then*

$$\frac{1}{\sqrt{n}}U(\theta_0) \xrightarrow{d} N(0,\mathscr{I}_1(\theta_0)) \tag{3.4.11}$$

Proof. Denote $\widehat{\theta}$ as the MLE of θ, then via a Taylor expansion we see that

$$0 = U(\widehat{\theta}) \approx U(\theta_0) + U'(\theta_0)(\widehat{\theta} - \theta_0), \tag{3.4.12}$$

thus

$$\frac{1}{\sqrt{n}}U(\theta_0) \approx -\frac{U'(\theta_0)}{\sqrt{n}}\left(\widehat{\theta} - \theta_0\right) = -\frac{U'(\theta_0)}{n}\cdot\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \tag{3.4.13}$$

where using the asymptotic normality of the MLE we have

$$\sqrt{n}\cdot\Big(\widehat{\theta}-\theta_0\Big)\xrightarrow{d}N\Big(0,\mathscr{I}_1^{-1}(\theta_0)\Big) \tag{3.4.14}$$

Also the empirical Fisher information converges to the actual Fisher information in probability, namely

$$\widehat{\mathscr{I}}(\theta_0)=-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log f(x,\theta)\xrightarrow{P}\mathscr{I}_1(\theta_0) \tag{3.4.15}$$

Combining the two results above, and using Slutzky's theorem, we have

$$\frac{1}{\sqrt{n}}U(\theta_0)\xrightarrow{d}N\Big(0,\mathscr{I}_1(\theta_0)\Big). \tag{3.4.16}$$

□

Definition 10. *We define the following Rao score statistic under $\mathscr{H}_0$:*

$$R_n=\frac{U(\theta_0)}{\sqrt{n\mathscr{I}_1(\theta_0)}}\xrightarrow{d}N(0,1). \tag{3.4.17}$$

which converges in distribution to a standard normal under $\mathscr{H}_0$. Under this test with significance level α, one shall reject $\mathscr{H}_0:\theta=\theta_0$ and in favour of $\mathscr{H}_1:\theta\neq\theta_0$ if $|R_n|>z_{\alpha/2}$; if $\mathscr{H}_1:\theta>\theta_0$ then we reject $\mathscr{H}_0$ if $R_n>z_\alpha$; if $\mathscr{H}_1:\theta<\theta_0$ then we reject $\mathscr{H}_0$ if $R_n<-z_\alpha$.

Another test is proposed by Abraham Wald, known as the *Wald Test*, which is similar to Rao score test. Again using the asymptotic normality of the MLE, and under the null hypothesis $\mathscr{H}_0:\theta=\theta_0$, we have that

$$\sqrt{n}\Big(\widehat{\theta}-\theta_0\Big)\xrightarrow{d}N\Big(0,\mathscr{I}_1^{-1}(\theta_0)\Big) \tag{3.4.18}$$

Definition 11. *We define the following Wald statistic under $\mathscr{H}_0$:*

$$W_n=\sqrt{n\mathscr{I}_1(\theta_0)}\cdot\Big(\widehat{\theta}-\theta_0\Big)\xrightarrow{d}N(0,1). \tag{3.4.19}$$

Finally, we define the *likelihood ratio (LR) statistic* statistic:

Definition 12. *Suppose we wish to test $\mathscr{H}_0:\theta=\theta_0,\mathscr{H}_1:\theta\neq\theta_0$, then the LR statistic is defined as*

$$\lambda_n:=\frac{\sup_{\theta=\theta_0}L(x,\theta)}{\sup_{\theta\in\Theta}L(x,\theta)}. \tag{3.4.20}$$

A nice property follows, if the regularity conditions are satisfied:

Theorem 3. *Under regularity conditions, we have $-2\log\lambda_n\xrightarrow{d}\chi^2_{(1)}$.*

Proof. For simplicity, we use $\ell(x,\theta)$ to denote the log-likelihood function. We denote the MLE of θ by $\widehat{\theta}$, and by definition we have

$$-2\log\lambda_n=-2\Big(\ell(x,\theta_0)-\ell(x,\widehat{\theta})\Big), \tag{3.4.21}$$

then we perform a Taylor expansion of $\frac{\partial}{\partial\theta}\ell(x,\theta)\Big|_{\theta=\widehat{\theta}}$ around θ_0, and by Bartlett's identity, we will have

$$\frac{\partial}{\partial\theta}\ell(x,\theta)\Big|_{\theta=\widehat{\theta}} = \frac{\partial}{\partial\theta}\ell(x,\theta)\Big|_{\theta=\theta_0} + (\widehat{\theta}-\theta_0)\cdot\frac{\partial^2}{\partial\theta^2}\ell(x,\theta)\Big|_{\theta=\theta_0} + o(2) = 0. \tag{3.4.22}$$

which is,

$$\frac{\partial}{\partial\theta}\ell(x,\theta)\Big|_{\theta=\theta_0} = (\widehat{\theta}-\theta_0)\cdot\left(-\frac{\partial^2}{\partial\theta^2}\ell(x,\theta)\Big|_{\theta=\theta_0}\right) - o(2) \tag{3.4.23}$$

Another Taylor expansion of $\ell(x,\theta)\Big|_{\theta=\widehat{\theta}}$ around θ_0 will yield

$$\ell(x,\theta)\Big|_{\theta=\widehat{\theta}} = \ell(x,\theta)\Big|_{\theta=\theta_0} + (\widehat{\theta}-\theta_0)\cdot\frac{\partial}{\partial\theta}\ell(x,\theta)\Big|_{\theta=\theta_0} + \frac{1}{2}(\widehat{\theta}-\theta_0)^2\cdot\frac{\partial^2}{\partial\theta^2}\ell(x,\theta)\Big|_{\theta=\theta_0} + o(3). \tag{3.4.24}$$

So the original log-ratio function now becomes

$$-2\log\lambda_n = 2\left((\widehat{\theta}-\theta_0)\cdot\frac{\partial}{\partial\theta}\ell(x,\theta)\Big|_{\theta=\theta_0} + \frac{1}{2}(\widehat{\theta}-\theta_0)^2\cdot\frac{\partial^2}{\partial\theta^2}\ell(x,\theta)\Big|_{\theta=\theta_0} + o(3)\right). \tag{3.4.25}$$

We substitute $\frac{\partial}{\partial\theta}\ell(x,\theta)\Big|_{\theta=\theta_0}$ into the previous expression, and we have

$$-2\log\lambda_n = 2(\widehat{\theta}-\theta_0)^2\cdot\left(-\frac{\partial^2}{\partial\theta^2}\ell(x,\theta)\Big|_{\theta=\theta_0} + \frac{1}{2}\cdot\frac{\partial^2}{\partial\theta^2}\ell(x,\theta)\Big|_{\theta=\theta_0} + o(1)\right) \tag{3.4.26}$$

$$= 2n(\widehat{\theta}-\theta_0)^2\cdot\left(-\frac{1}{2n}\cdot\frac{\partial^2}{\partial\theta^2}\ell(x,\theta)\Big|_{\theta=\theta_0} + o\left(\frac{1}{n}\right)\right). \tag{3.4.27}$$

Then we use the asymptotic normality of MLE, under null hypothesis $\mathscr{H}_0$, combine with the fact that $(N(0,1))^2 \sim \chi^2_{(1)}$, as well as the convergence of empirical Fisher information, and Slutsky's theorem, we will have

$$-2\log\lambda_n \xrightarrow{d} \chi^2_{(1)}. \tag{3.4.28}$$

□

Hence, we may define the following test:

Definition 13. *Consider the test $\mathscr{H}_0 : \theta = \theta_0$ and $\mathscr{H}_1 : \theta \neq \theta_0$ in the random sample $X_1, \cdots, X_n \sim f(x,\theta)$, then under regularity conditions, we will reject $\mathscr{H}_0$ at significance level α if $-2\log\lambda_n > \chi^2_{1,\alpha/2}$ or $-2\log\lambda_n < \chi^2_{1,1-\alpha/2}$.*

One thing need to point out is that all three tests we have introduced are asymptomatically equivalent to some extent, that is, they reach the same decision with probability 1 as $n \to \infty$.

Proposition 2. *Under regularity conditions, we have the following relations among Rao score test R_n, Wald test W_n and LR test λ_n:*

$$R_n \xrightarrow{P} W_n \text{ and } W_n^2 \xrightarrow{P} -2\log\lambda_n. \tag{3.4.29}$$

3.5 Likelihood Inference on Multiple Parameters

We would like to extend our results from the previous section, now we will assume we have multiple parameters, and we write it as a $p \times 1$ vector $\boldsymbol{\theta}$ (where p is the number of parameters we have). Also we will assume the regularity conditions are hold.

Definition 14. *Suppose we have a random sample $X_1, \cdots, X_n \sim f(x, \boldsymbol{\theta})$ where we have p parameters. The score function of one of them is now defined as the gradient of the likelihood function:*

$$U_i(\boldsymbol{\theta}) = \nabla \ell_i(x, \boldsymbol{\theta}) = \left(\frac{\partial}{\partial \theta_1} \ell_i(x, \boldsymbol{\theta}), \frac{\partial}{\partial \theta_2} \ell_i(x, \boldsymbol{\theta}), \cdots, \frac{\partial}{\partial \theta_p} \ell_i(x, \boldsymbol{\theta}) \right). \tag{3.5.1}$$

The score of the sample is $U(\boldsymbol{\theta}) = \sum_{i=1}^{n} U_i(\boldsymbol{\theta})$.

Proposition 3. *The mean of the score is still zero, we have $\mathbb{E}[U_i(\boldsymbol{\theta})] = \mathbf{0}$.*

Proof. By definition, we have

$$\mathbb{E}[U_i(\boldsymbol{\theta})] = \mathbb{E}[\nabla \ell_i(X, \boldsymbol{\theta})] = \begin{bmatrix} \mathbb{E}\left[\frac{\partial}{\partial \theta_1} \ell_i(X, \boldsymbol{\theta})\right] \\ \mathbb{E}\left[\frac{\partial}{\partial \theta_2} \ell_i(X, \boldsymbol{\theta})\right] \\ \vdots \\ \mathbb{E}\left[\frac{\partial}{\partial \theta_p} \ell_i(X, \boldsymbol{\theta})\right] \end{bmatrix} \overset{\text{Theorem 1}}{=} \mathbf{0}. \tag{3.5.2}$$

□

Definition 15. *Similar to the second Bartlett's identity, the variance of $U_i(\boldsymbol{\theta})$ is now a $p \times p$ covariance matrix, given by*

$$\mathscr{I}_i(\boldsymbol{\theta}) = \left[\mathbb{E}\Big[-\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(X_i, \boldsymbol{\theta}) \Big]_{ij} \right], 1 \le i, j \le p. \tag{3.5.3}$$

The Fisher information of the whole sample is just $\mathscr{I}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathscr{I}_i(\boldsymbol{\theta})$. The empirical Fisher information is

$$\widehat{\mathscr{I}(\boldsymbol{\theta})} = \left[-\frac{\partial^2}{\partial \theta_i \partial \theta_j} \sum_{i=1}^{n} \log f(X_i, \boldsymbol{\theta}) \right]_{ij}, 1 \le i, j \le p. \tag{3.5.4}$$

Proposition 4. *The empirical Fisher information converges to the actual one almost surely:*

$$\widehat{\mathscr{I}(\boldsymbol{\theta})} = \left[-\frac{\partial^2}{\partial \theta_i \partial \theta_j} \sum_{i=1}^{n} \log f(X_i, \boldsymbol{\theta}) \right]_{ij} \overset{a.s}{\to} \mathscr{I}_i(\boldsymbol{\theta}) = \left[\mathbb{E}\Big[-\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(X_i, \boldsymbol{\theta}) \Big]_{ij} \right]. \tag{3.5.5}$$

Proof. The proof is the same idea as in Theorem 2.

□

The Rao, Wald, LR statistic in multi-variable case could be easily extended from the single variable case:

Definition 16. *When $\boldsymbol{\theta} \in \mathbb{R}^p$, we have:*

$$W_n := n \cdot \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)^T \cdot \mathscr{I}(\boldsymbol{\theta}_0) \cdot \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} \chi^2_{(p)}, \tag{3.5.6}$$

$$R_n := \frac{1}{n} \cdot \nabla \ell(\boldsymbol{\theta}_0)^T \cdot \mathscr{I}^{-1}(\boldsymbol{\theta}_0) \cdot \nabla \ell(\boldsymbol{\theta}_0) \xrightarrow{d} \chi^2_{(p)}, \tag{3.5.7}$$

$$\lambda_n := -2\left(\ell(\boldsymbol{\theta}_0) - \ell(\widehat{\boldsymbol{\theta}})\right) \xrightarrow{d} \chi^2_{(p)}. \tag{3.5.8}$$

under the null hypothesis $\mathscr{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\theta}}$ is the MLE.

Chapter 4

Non-Parametric Approach to Survival Data

4.1 Introduction

In the previous chapter, we constructed the likelihood function for samples with different types of censoring and truncation. Note that the key assumption was that we know the sample belongs to a parametric family $f(x,\theta)$. That is, we already know the sample has a certain distribution and only the parameter is unknown, like $Weibull(\alpha,\beta)$, $Exponential(\lambda)$, etc. In this chapter, we will study non-parametric family, where we do not know the distribution of the sample in advance: All we have is just a collection of observable data. We will try to build the likelihood function and estimate the distribution of the sample as well as key survival features.

4.2 Empirical Cumulative Distribution Function (ECDF)

Now suppose we have a random sample $X_1,\cdots,X_n$ with an unknown distribution F, without the consideration of any kind of censoring or truncation, we define a new distribution F_n by

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \chi(X_i \leq x) \tag{4.2.1}$$

where χ is the indicator function given by

$$\chi(X_i \leq x) = \begin{cases} 1 & X_i \leq x \\ 0 & \text{otherwise} \end{cases} \tag{4.2.2}$$

We call the distribution in (4.2.1) the Empirical Cumulative Distribution Function (ECDF), if we compare to the actual cumulative distribution function $F(x) \equiv \mathbb{P}(X \leq x)$, we may notice some similarities. We approximate the value of $F(x)$ by counting the number of individuals that lie in the interval $(-\infty,x)$, then of course as the sample sizes increases, $F_n(x)$ would become closer to $F(x)$. The similar technique is used in estimate the mean μ, which is approximated by the sample mean $\overline{X}_n = \dfrac{X_1+\cdots+X_n}{n}$ and we know the weak law of large number states that $\overline{X}_n \xrightarrow{P} \mu$. In fact by Glivenko-Cantelli theorem, we have

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow{a.s} 0 \tag{4.2.3}$$

which is even stronger than convergence in probability, indeed we have almost surely convergence. Furthermore, we can derive the expected value of $F_n(x)$ by

$$\begin{aligned}
\mathbb{E}(F_n(x)) &= \frac{1}{n}\cdot\sum_{i=1}^{n}\mathbb{E}[\chi(X_i\leq x)] && (4.2.4)\\
&= \frac{1}{n}\cdot\sum_{i=1}^{n}[1\cdot\mathbb{P}(\chi(X_i\leq x)=1)+0\cdot\mathbb{P}(\chi(X_i\leq x)=0] && (4.2.5)\\
&= \frac{1}{n}\cdot\sum_{i=1}^{n}\mathbb{P}(X_i\leq x) && (4.2.6)\\
&= \frac{1}{n}\cdot\sum_{i=1}^{n}F(x) && (4.2.7)\\
&= F(x) && (4.2.8)
\end{aligned}$$

which shows that the impirical estimator is unbiased, and we have the following relation:

$$\mathbb{P}(|F_n(x)-F(x)|\geq\varepsilon)\leq 2e^{-2n\varepsilon^2} \tag{4.2.9}$$

for any $\varepsilon>0, n\geq 1$, which is a variation of Chebyshev's Inequality. The variance of $F_n(x)$ can be computed by

$$\begin{aligned}
\mathbb{V}(F_n(x)) &= \mathbb{E}(F_n^2(x))-[\mathbb{E}(F_n(x))]^2 && (4.2.10)\\
&= \mathbb{E}\left(\frac{1}{n}\sum_{j=1}^{n}\chi(X_j\leq x)\right)^2-F^2(x) && (4.2.11)\\
&= \frac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\chi(X_i\leq x)\chi(X_j\leq x)\right)-F^2(x) && (4.2.12)\\
&= \frac{1}{n^2}\mathbb{E}\left(\sum_{i=j}\chi^2(X_i\leq x)+\sum_{i\neq j}\chi(X_i\leq x)\chi(X_j\leq x)\right)-F^2(x) && (4.2.13)\\
&= \frac{1}{n^2}\left(nF(x)+n(n-1)F^2(x)\right)-F^2(x) && (4.2.14)\\
&= \frac{1}{n}F(x)-\frac{1}{n}F^2(x). && (4.2.15)
\end{aligned}$$

Once we have the mean and the variance, by Central Limit Theorem, it follows naturally that

$$\frac{\sqrt{n}(F_n(x)-F(x))}{\sqrt{F(x)(1-F(x))}}\xrightarrow{d}N(0,1) \tag{4.2.16}$$

as $n\to\infty$. Since by definition we have $S(x)=1-F(x)$, so we also obtained the empirical survival function.

4.3 Kaplan-Meier Estimator for the Survival Function

Previously, for a continuous random variable X with density $f(x)$, and distribution $F(x)$, we define the survival function to be

$$S(x) = \mathbb{P}(X \geq x) = 1 - F(x) \tag{4.3.1}$$

Using the empirical version, if given the sample $X_1, \cdots, X_n$, we can derive the survival function as

$$S_n(x) = \frac{1}{n} \sum_{i=1}^{n} \chi(X_i > x) \tag{4.3.2}$$

Now since we are dealing samples with potential censoring, (4.3.2) might not be a good estimate for us since many observations would be censored and we are not able to find the exact X_i. Thus we will introduce an estimator Kaplan and Meier proposed:

Theorem 4. ***Kaplan-Meier Estimator***

Suppose we have observed the exact event time (ordered) $t_1, \cdots, t_n$ where D_j deaths are observed at t_j, N_j individuals are at risk (event still not occurred) at t_j^-, then the Kaplan-Meier (KM) estimator for the survival function $S(t)$ is given by

$$\widehat{S}(t) = \prod_{t_j \leq t} \left(1 - \frac{D_j}{N_j}\right) \tag{4.3.3}$$

To better understand KM estimator, here we propose a plug in approach to get the KM estimator: We will split observed failure times into sub-intervals: $[0, t_1), [t_1, t_2), \cdots, [t_n, +\infty)$, then if $t \in [0, t_1)$, it means that no deaths had been observed, in this case the survival function is just defined by constant 1. Then if $t \in [t_1, t_2)$, by definition we have

$$\begin{aligned} S(t) = \mathbb{P}(X > t) &:= \mathbb{P}(\text{Survived after } t_1\ ) && (4.3.4)\\ &= \frac{\text{Number of Individuals survived after } t_1}{\text{Number of individuals at risk at } t_1^-} && (4.3.5)\\ &= \frac{N_1 - D_1}{N_1}. && (4.3.6) \end{aligned}$$

Now we consider $t \in [t_2, t_3)$, then we have the conditional probability given by

$$\begin{aligned} S(t) = \mathbb{P}(X > t) &:= \mathbb{P}(\text{Survived after } t_2) && (4.3.7)\\ &= \mathbb{P}(\text{Survived after } t_2 \mid \text{Survived after } t_1) \cdot \mathbb{P}(\text{Survived after } t_1) && (4.3.8)\\ &= \frac{\text{Number of Individuals survived after } t_2}{\text{Number of individuals at risk at } t = t_2^-} \times \frac{N_1 - D_1}{N_1} && (4.3.9)\\ &= \frac{N_2 - D_2}{N_2} \times \frac{N_1 - D_1}{N_1} && (4.3.10) \end{aligned}$$

Thus, we may use this plug in approach recursively, and we will get the KM estimator:

$$\widehat{S}(t) = \prod_{t_j \leq t} \left(\frac{N_j - D_j}{N_j}\right) \tag{4.3.11}$$

which is of the same form of (4.3.3). Since our survival function is based on the terms of products, this is also known as the Product Limit (PL) Estimator.

One natural question arises: How good is KM estimator? It turns out that KM estimator is the one the maximizes the likelihood of S, i.e $\mathscr{L}(S)$ obtains a maximum at $\widehat{S}$.

Theorem 5. *The KM Estimator is the one that maximizes $\mathscr{L}(S)$.*

Proof. We begin by stating the relation between the survival function $S(t)$ and the hazard function $\lambda(t)$ in the discrete case.

Lemma 3. *The survival function is given by*

$$S(t) = \prod_{t_j \leq t} (1 - \lambda_j) \tag{4.3.12}$$

where

$$\lambda_j = \mathbb{P}(T = t_j | T \geq t_j) = \frac{f(t_j)}{S(t_j^-)} \tag{4.3.13}$$

is the discrete hazard function.

Proof. In the sense of the discrete setting, $S(t_j^-) = S(t_{j-1})$. Now, for each j, we can derive $S(t_j)$ in terms of conditional probability:

$$S(t_j) = \mathbb{P}(X > t_j | X \geq t_j) \cdot \mathbb{P}(X \geq t_j) + \mathbb{P}(X > t_j | X < t_j) \cdot \mathbb{P}(X < t_j) \tag{4.3.14}$$

where

$$\mathbb{P}(X > t_j | X \geq t_j) + \mathbb{P}(X = t_j | X \geq t_j) = 1 \tag{4.3.15}$$

So (4.3.14) becomes

$$S(t_j) = (1 - \lambda_j) S(t_j^-) = (1 - \lambda_j) S(t_{j-1}) \tag{4.3.16}$$

In this case, by applying (4.3.16) recursively for all $t_j \leq t$, we will arrive at

$$S(t) = \prod_{t_j \leq t} (1 - \lambda_j) \tag{4.3.17}$$

□

Now, suppose we have observations $t_1, t_2, \cdots, t_j$ where we have D_i deaths at each t_i, N_i are at risk (event time still not occurred) at t_i^-. Then $N_i - D_i$ are censored in $[t_i, t_{i+1})$. We may think of the hazard at t_j as a binomial distribution: Among N_j samples that are at risk, we have D_j deaths and $N_j - D_j$ survivals, they yields a probability of λ_j and $1 - \lambda_j$ respectively, thus there will be $\binom{N_j}{D_j}$ ways to choose those samples, and we take the product of all the point t_j, we obtain

$$\mathscr{L}(\boldsymbol{\lambda}) = \prod_{j=1}^{J} \binom{N_j}{D_j} \lambda_j^{D_j} (1 - \lambda_j)^{N_j - D_j}. \tag{4.3.18}$$

Hence it suffices to maximize $\lambda_j^{D_j} (1 - \lambda_j)^{N_j - D_j}$ for each λ_j, so we have

$$\frac{\partial \mathscr{L}(\boldsymbol{\lambda})}{\partial \lambda_j} \propto \lambda_j^{D_j - 1} (1 - \lambda_j)^{N_j - D_j - 1} \left[D_j (1 - \lambda_j) - \lambda_j (N_j - D_j) \right] \tag{4.3.19}$$

where by definition, when

$$D_j(1-\lambda_j)-\lambda_j(N_j-D_j)=0 \tag{4.3.20}$$

we will get the value that maximizes each λ_j, given by

$$\widehat{\lambda}_j=\frac{D_j}{N_j}. \tag{4.3.21}$$

So by (4.3.17), we replace each λ_j by $\widehat{\lambda}_j$, we will also maximize the survival function S, where

$$\widehat{S}(t)\equiv\prod_{t_j\leq t}(1-\widehat{\lambda}_j) \tag{4.3.22}$$

By (4.3.21), we can derive $\widehat{S}(t)$ as

$$\widehat{S}(t)=\prod_{t_j\leq t}\left(1-\frac{D_j}{N_j}\right) \tag{4.3.23}$$

□

It is important to note that KM estimator is also a statistic, so we are also interested in estimating its expectation and variance. M. Zhou (1988) proved that the bias of Kaplan-Meier estimator is exponentially small. Also several improvements of KM estimator based on specific sets of data were also introduced. As for the variance, the Greenwood's Formula is the most common one:

Theorem 6. ***Greenwood's Formula***

Let $\widehat{S}(t)$ to be the KM estimator, then

$$\mathbb{V}(\widehat{S}(t))=\widehat{S}^2(t)\cdot\sum_{t_j\leq t}\frac{D_j}{N_j(N_j-D_j)} \tag{4.3.24}$$

where D_j is the number of death observed at t_j, N_j is the number of individuals at risk at t_j^-, as in the KM estimator.

Proof. Since $\widehat{\lambda}_j=\dfrac{D_j}{N_j}$ is the maximum likelihood estimator for the hazard component λ_j, where in the discrete setting we can rewrite it in terms of empirical version:

$$\widehat{\lambda}_j=\frac{1}{N_j}\sum_{i=1}^{N_j}\chi\{X_i\leq t_j\} \tag{4.3.25}$$

Then by (4.2.8) and (4.2.15), it follows that $\mathbb{E}\widehat{\lambda}_j=\lambda_j$ and $\mathbb{V}\widehat{\lambda}_j=\dfrac{\lambda_j(1-\lambda_j)}{N_j}$, hence the central limit theorem states that

$$\sqrt{n}(\widehat{\lambda}_j-\lambda_j)\xrightarrow{d}N\left(0,\frac{\lambda_j(1-\lambda_j)}{N_j}\right) \tag{4.3.26}$$

By Delta-Method, we have

$$\sqrt{n}(\log(1-\widehat{\lambda}_j)-\log(1-\lambda_j))\xrightarrow{d}N\left(0,\frac{\lambda_j(1-\lambda_j)}{N_j}\cdot\left[-\frac{1}{1-\lambda_j}\right]^2\right) \tag{4.3.27}$$

Hence

$$\mathbb{V}[\log(1-\widehat{\lambda}_j)] = \frac{\lambda_j}{N_j(1-\lambda_j)} \tag{4.3.28}$$

Now we express $\widehat{S}(t)$ in terms of the product of $1-\lambda_j$, we have

$$\mathbb{V}[\log \widehat{S}(t)] = \sum_{t_j \le t} \frac{\lambda_j}{N_j(1-\lambda_j)} \tag{4.3.29}$$

Again, we apply Delta-Method on (4.3.29), we will have

$$\mathbb{V}[\widehat{S}(t)] = \widehat{S}^2(t) \cdot \sum_{t_j \le t} \frac{\lambda_j}{N_j(1-\lambda_j)} \tag{4.3.30}$$

Since $\lambda_j = \dfrac{D_j}{N_j}$, hence we have proved Greenwood's formula. □

We could use Greenwood's formula to get the approximate $100(1-\alpha)\%$ confidence interval:

Proposition 5. *For a fixed t, the approximate $100(1-\alpha)\%$ confidence interval of $S(t)$ is given by*

$$\widehat{S}(t) \pm z_{\alpha/2}\sqrt{\mathbb{V}[\widehat{S}(t)]} \tag{4.3.31}$$

Proof. Denote by $S(t)$, as the true value of the survival function at t, and $\widehat{S}(t)$ to be our estimate, then central limit theorem states that

$$\frac{\widehat{S}(t) - S(t)}{\sqrt{\dfrac{S(t)(1-S(t))}{n}}} \xrightarrow{d} (0,1). \tag{4.3.32}$$

When the sample is large, using Slutsky's theorem, we could replace $S(t)$ by $\widehat{S}(t)$ in the denominator, and hence we have

$$\frac{\widehat{S}(t) - S(t)}{\sqrt{\dfrac{\widehat{S}(t)(1-\widehat{S}(t))}{n}}} \xrightarrow{d} N(0,1), \tag{4.3.33}$$

then using the pivot quantity, given level α, we aim to find $\boldsymbol{L}$ and $\boldsymbol{U}$ such that $\mathbb{P}\{\boldsymbol{L} \le S(t) \le \boldsymbol{U}\} = 1-\alpha$, and hence we have the $100(1-\alpha)\%$ confidence interval given by

$$\widehat{S}(t) \pm z_{\alpha/2}\sqrt{\mathbb{V}[\widehat{S}(t)]}. \tag{4.3.34}$$

□

We may notice that there is an obvious drawback: the bound for our confidence interval may lie outside $[0,1]$. Based on that, we introduce the complimentary log-log transformation:

$$Z(t) = \log\Big(-\log \widehat{S}(t)\Big), \tag{4.3.35}$$

then again we applying delta-method on (4.3.29), we get

$$\mathbb{V}(\widehat{Z(t)}) = \frac{1}{(\log \widehat{S}(t))^2} \cdot \sum_{t_i \le t} \frac{D_i}{N_i(N_i - D_i)}. \tag{4.3.36}$$

where the $100(1-\alpha)\%$ confidence interval of $S(t)$ is now

$$Z(t) \pm z_{\alpha/2} \cdot \sqrt{\mathbb{V}(Z(t))}. \tag{4.3.37}$$

Replacing $Z(t)$, we will get the exponential Greenwood formula, which is attributed by Hosmer and Lemeshow (1999). It yields the following symmetric confidence interval:

$$S(t) \in \Big(\exp(-\exp(c_+(t))), \exp(-\exp(c_-(t))) \Big) \tag{4.3.38}$$

where

$$c_\pm(t) = \log(-\log \widehat{S}(t)) \pm z_{\alpha/2} \cdot \sqrt{\frac{1}{\Big(\log \widehat{S}(t)\Big)^2} \cdot \sum_{t_i \leq t} \frac{D_i}{N_i(N_i - D_i)}}. \tag{4.3.39}$$

In this case, the confidence interval is guaranteed to lie in $(0,1)$, which gives us a better estimate of $S(t)$. This estimate behaves well for sample sizes greater or equal to 25 with up to 50% of the observations being censored (Hosmer and Lemeshow, 1999).

4.4 Nelson-Aalen Estimator for the Cumulative Hazard Function

Nelson-Aalen estimator can be used to estimate the cumulative hazard function of the survival data. Denote $h(t)$ as the hazard function, we recall that the cumulative hazard function is defined as

$$\Lambda(t) = \int_0^t h(s)dx. \tag{4.4.1}$$

Previously, we have introduced the relation among $h(t)$, the survival function $S(t)$ and the density function $f(t)$, given by

$$h(t) = \frac{f(t)}{S(t)} = \frac{F'(x)}{1-F(x)}, S(t) = 1-F(t), \tag{4.4.2}$$

hence we have the equation

$$\Lambda(t) = \int_0^t h(x)dx = \int_0^t \frac{dF(x)}{S(x)}. \tag{4.4.3}$$

Given two independent random samples $X_1, \cdots, X_n$ and $C_1, \cdots, C_n$ where the first one represents the event time and the second one represents the censoring time, we already defined that $T_i = \min\{X_i, C_i\}$ and $\delta_i = \chi\{X_i \leq C_i\}$, then we define

$$H(t) = \mathbb{P}\{T_i \leq t\}, H^U(t) = \mathbb{P}\{T_i \leq t, \delta_i = 1\} \tag{4.4.4}$$

where $H^U(t)$ is a sub-distribution representing all the uncensored and observed data. Then we can rewrite the cumulative hazard function as

$$\Lambda(t) = \int_0^t \frac{dH^U(x)}{1-H(x)} \tag{4.4.5}$$

In the non-parametric setting, we replace $H(t)$ and $H^U(t)$ by the empirical distribution function

$$\widehat{H}_n(t) = \frac{1}{n}\sum_{i=1}^n \chi\{T_i \leq t\} \quad \text{and} \quad \widehat{H}_n^U(t) = \frac{1}{n}\sum_{i=1}^n \chi\{T_i \leq t, \delta_i = 1\} \tag{4.4.6}$$

So our estimate for the cumulative hazard function is given by (Läuter and Liero, 2004)

$$\widehat{\Lambda}(t) = \int_0^t \frac{d\widehat{H}_n^U(x)}{1-\widehat{H}_n(x^-)} \tag{4.4.7}$$

Another approach follows: If we denote $t_1 < t_2 < \cdots < t_n$ to be the time where we observed the event before t, also let D_j to be the number of death at time t_j, N_j is the number of individuals at risk just prior to t_j, then at each t_j the risk can be estimated by $\widehat{\lambda}(t_j) = \dfrac{D_j}{N_j}$, and thus we may derive $\widehat{\Lambda}(t)$ as

$$\widehat{\Lambda}(t) = \sum_{t_j \leq t} \frac{d_j}{n_j} \tag{4.4.8}$$

Where (4.4.8) is known as the Nelson-Aalen estimator. We see that it is a non-decreasing step function, with increments D_j/N_j at each observed failure time. If we recall that the Kaplan-Meier estimator takes the form

$$\widehat{S}(t) = \prod_{t_j \leq t}\left(1 - \frac{d_j}{n_j}\right), \tag{4.4.9}$$

by taking the logarithm of the Kaplan-Meier estimator yields

$$\log(\widehat{S}(t)) = \sum_{t_j \leq t} \log\left(1 - \frac{d_j}{n_j}\right). \tag{4.4.10}$$

from a well known Taylor expansion, we know that if $d_j << n_j$, usually treated as no ties in observed failures, we then have

$$\log(\widehat{S}(t)) = \sum_{t_j \leq t} -\frac{d_j}{n_j} \tag{4.4.11}$$

hence we have found the relationship between the Kaplan-Meier estimator and Nelson-Aalen estimator:

$$-\log(\widehat{S}(t)) = \widehat{\Lambda}(t). \tag{4.4.12}$$

Hence to estimate the variance of the Nelson-Aalen estimator, we have the form

$$\mathbb{V}(\widehat{\Lambda}(t)) = \mathbb{V}(-\log(\widehat{S}(t)) = \mathbb{V}\left(-\log\left(\prod_{t_j \leq t} 1 - \frac{d_j}{n_j}\right)\right), \tag{4.4.13}$$

by using a Taylor expansion, when $d_j << n_j$ (where we often treated as no ties in observed event time or in a large sample), we have

$$\mathbb{V}(\widehat{\Lambda}(t)) = \mathbb{V}\left(\sum_{t_j \leq t} \frac{d_j}{n_j}\right) = \sum_{t_j \leq t} \mathbb{V}\left(\frac{d_j}{n_j}\right), \text{ assuming independence.} \tag{4.4.14}$$

Then, at each time t_j consider the number of observed death d_j as a random variable D_j, then we have

$$\mathbb{V}(\widehat{\lambda}(t)) = \sum_{t_j \leq t} \frac{\mathbb{V}(D_j)}{n_j^2} \tag{4.4.15}$$

Commonly, D_j can be treated as a binomial distribution $D_j \sim Binomial\left(n_j, d_j/n_j\right)$, and so we have

$$\mathbb{V}(\widehat{\lambda}(t)) = \sum_{t_j \leq t} \frac{d_j(n_j - d_j)}{n_j^3}, \tag{4.4.16}$$

or assuming D_j forms a Poisson distribution with parameter $n_j h(t_j)$, where $h(t_j)$ represents the hazard rate at t_j, then we also have

$$\mathbb{V}(\widehat{\lambda}(t)) = \sum_{t_J \leq t} \frac{d_j}{n_j^2}. \tag{4.4.17}$$

Denote $\widehat{\sigma^2}$ as the variance of the Nelson-Aalen estimator, we are able to derive the $100(1-\alpha)\%$ confidence interval for $\Lambda(t)$:

$$\left(\widehat{\Lambda}(t) - z_{1-\alpha/2} \cdot \widehat{\sigma}(t) \ , \ \widehat{\Lambda}(t) + z_{1-\alpha/2} \cdot \widehat{\sigma}(t)\right) \tag{4.4.18}$$

With a log-log transformation we get the improved confidence interval as

$$\widehat{\Lambda}(t) \cdot \exp\left(\pm z_{1-\alpha/2} \cdot \widehat{\sigma}(t)/\widehat{\Lambda}(t)\right). \tag{4.4.19}$$

4.5 An Improvement: Aalen-Johansen Estimator

Aalen-Johansen Estimator is a generalization of Kaplen-Meier Estimator, but in a multi-dimensional form. When we have observed the time event T, some questions remain unanswered. "What is the actual cause of the death?" "Is the death of the individual related to the disease that we are actually interested in?". Hence we introduce some new variables, including different causes of death. Also, we can use a Markov process to describe the behavior of the survival data (Borgan, 2005): We define two states, namely alive and death, where we split death into several sub-states indicating the different causes of death, the transition from being alive to death at a given time t can be viewed as the hazard function evaluated at t. In general, there are multiple states, but we could summary into 3: healthy state, disease state and death state.

1. A Multirisk Model:

We first consider a simple case where we have one transient state (alive) and several absorbing states (death by cause $h, h = 1,2,\cdots,k$). The diagram below shows a simple case when $k = 3$.

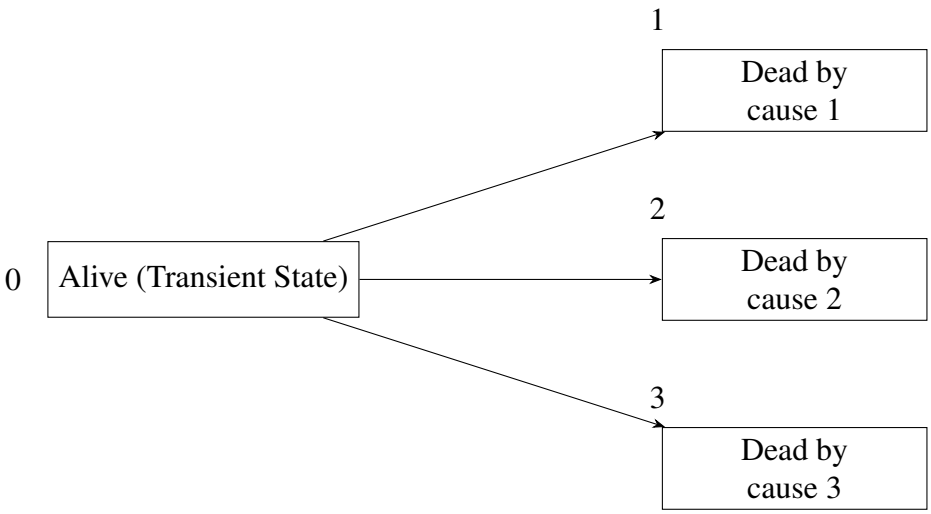


Figure 4.1: A multirisk model with $k = 3$, 0 represents the transient state (alive) and $1,2,3$ represent the absorbing states (dead by caush h).

Denote by $\alpha_{0h}(t)$, as the marginal risk function at time t by cause h. That is, $\alpha_{0h}(t)$ is the instantaneous risk of death at $t+\Delta t$ by cause h given that the individual is alive at t. We call $\alpha_{0h}(t)$ the *cause-specific hazard rate functions*. Recall that for a single parameter, the hazard function $h(t)$ and survival function $S(t)$ has the following relation:

$$S(x) = \exp\left\{-\int_0^x h(t)dt\right\} \tag{4.5.1}$$

Now, we define $\mathbb{P}_{00}(s,t)$, as the probability that an individual is alive at s will also be alive at t, $t > s$. Then the survival probability $\mathbb{P}_{00}(s,t)$ is the product of each marginal survival function $S_h(s,t)$:

$$\mathbb{P}_{00}(s,t) = \exp\left\{-\int_s^t \sum_{h=1}^{k} \alpha_{0h}(u)du\right\}. \tag{4.5.2}$$

Also, define $\mathbb{P}_{0h}(s,t)$, as the probability that the individual in state 0 (being alive) at s will in state h (dead by cause h) at t, and $\mathbb{P}_{0h}(s,t)$ is also denoted as *transition probability*. This can be computed by the cumulation of probability of death at any time t_0 in the interval (s,t), given by $\mathbb{P}_{00}(s,t_0)\alpha_{0h}(s,t_0)$, thus taking the integral yields

$$\mathbb{P}_{0h}(s,t) = \int_s^t \mathbb{P}_{00}(s,u)\alpha_{0h}(u)du. \tag{4.5.3}$$

Now let's consider the non-parametric estimators for those equations above. From Kaplan-Meier estimator, we can derive $\mathbb{P}_{00}(s,t)$ as

$$\widehat{\mathbb{P}}_{00}(s,t) = \prod_{s<t_j<t}\left(1-\frac{D_{0j}}{N_{0j}}\right) \tag{4.5.4}$$

where we denote D_{0h_j} as the number of death at time t_j due to h, i.e the number of transition from 0 to h at time t_j, and $D_{0j} = \sum_{h=1}^{k} D_{0h_j}$ as the total number of death at t_j, and N_{0j} be the number of individuals at risk just prior to time t_j (number of individuals at state 0). From here, the Nelson-Aalen estimator for the cumulative cause-specific hazard function $\widehat{A}_{0h}$ is given by

$$\widehat{A}_{0h}(t) = \sum_{t_j\le t}\frac{D_{0h_j}}{N_{0j}}, \tag{4.5.5}$$

the cumulative incidence can also be estimated by

$$\widehat{\mathbb{P}}_{0h}(s,t) = \sum_{s<t_j<t}\widehat{\mathbb{P}}_{00}(s,t_{j-1})\cdot\frac{D_{0h_j}}{N_{0j}}. \tag{4.5.6}$$

2. An Alive-Illness-Death Model:

To study the occurrence of a chronic disease as well as death in a homogeneous population, we may adopt the Markov illness-death model shown below, where we have 3 different states, 0 (healthy), 1 (illness) and 2 (dead):

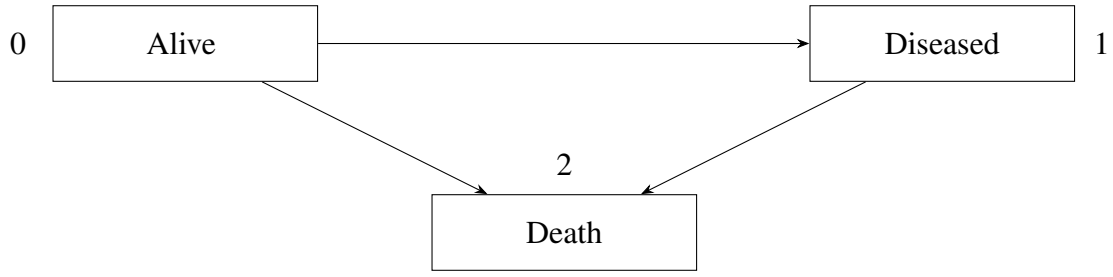


Figure 4.2: An alive-illness-death model without recovery

We define $\alpha_{01}(t), \alpha_{02}(t)$ and $\alpha_{12}(t)$ as the instantaneous risk of experiencing a transition from i to j at time $t+\Delta t$. We denote $\mathbb{P}_{00}(s,t)$, $\mathbb{P}_{01}(s,t)$, $\mathbb{P}_{11}(s,t)$, $\mathbb{P}_{02}(s,t)$, $\mathbb{P}_{12}(s,t)$ as the probability that the individual at state i at s will in state j at t, where in this model we still have

$$\mathbb{P}_{00}(s,t) = \exp\left\{-\int_s^t\Big(\alpha_{01}(u)+\alpha_{02}(u)\Big)du\right\}, \mathbb{P}_{11}(s,t) = \exp\left\{-\int_s^t\alpha_{12}(u)du\right\} \tag{4.5.7}$$

and

$$\mathbb{P}_{12}(s,t) = \int_s^t \mathbb{P}_{11}(s,u)\alpha_{12}(u)du. \tag{4.5.8}$$

The difference occurs in the transition probability. In our previous model, after reaching the absorbing state we would have reached the equilibrium state and no further transition is possible. Here if we would like to compute $\mathbb{P}_{01}(s,t)$, we would not only compute the transition probability from state 0 to state 1, but also we need to consider the effect of transiting to state 2. So the final calculation yields

$$\mathbb{P}_{01}(s,t) = \int_s^t \underbrace{\mathbb{P}_{00}(s,u)\cdot\alpha_{01}(u)}_{\text{transition from 0 to 1}}\cdot\underbrace{\mathbb{P}_{11}(u,t)}_{\text{stay in state 1}}\,du. \tag{4.5.9}$$

Similarly we can derive the expression for $\mathbb{P}_{02}(s,t)$. Where there are two cases: With direct transition from state 0 to state 2, and with an intermediate state 1, so we have the following expression:

$$\mathbb{P}_{02}(s,t) = \int_s^t \mathbb{P}_{00}(s,u)\cdot\alpha_{02}(u)du \tag{4.5.10}$$

$$+\int_{v=u}^{v=t}\int_{u=s}^{u=t}\underbrace{\Big(\mathbb{P}_{00}(s,u)\cdot\alpha_{01}(u)\cdot\mathbb{P}_{11}(u,v)\Big)}_{\text{transition from 0 to 1}}\cdot\underbrace{\Big(\mathbb{P}_{11}(v,t)\cdot\alpha_{12}(v)\Big)}_{\text{transition from 1 to 2}}dudv \tag{4.5.11}$$

Where the expression is also known as the *Chapman-Kolmogorov equation*, where we have $\mathbb{P}_{02}(s,t) = \mathbb{P}_{01}(s,k)\mathbb{P}_{12}(k,t), s<k<t$. We may also derive the KM, NA estimator for the observed data. Suppose we have observed the distinct failure times $t_1 < t_2 < \cdots$. Here, we denote D_{01_j} as the number of individuals who get diseased at time t_j (in this example only 1 represents the disease state but it could be generalized), and D_{02_j} represents the number of individuals who die directly without getting diseased at time t_j, and D_{12_j} represents the number of individuals who die due to disease at time t_j (similarly in this example only 2 represents disease state but the idea could also be generalized). Then we denote $D_{0j} = D_{02_j} + D_{12_j}$ as the total number of observed death at t_j. Also we use N_{0j}, (N_{1j}) as the total number of healthy (sick) individuals just before t_j (they are at risk of possible transition), then we may use KM estimator directly to get the estimate of $\mathbb{P}_{00}(s,t)$ and $\mathbb{P}_{11}(s,t)$:

$$\widehat{\mathbb{P}}_{00}(s,t) = \prod_{s<t_j\le t}\left(1-\frac{D_{0j}}{N_{0j}}\right); \widehat{\mathbb{P}}_{11}(s,t) = \prod_{s<t_j\le t}\left(1-\frac{D_{12_j}}{N_{1j}}\right) \tag{4.5.12}$$

While $\mathbb{P}_{01}(s,t), \mathbb{P}_{12}(s,t)$ may be estimated by

$$\widehat{\mathbb{P}}_{01}(s,t) = \sum_{s<t_j\le t}\widehat{\mathbb{P}}_{00}(s,t_{j-1})\cdot\left(\frac{D_{01_j}}{N_{0j}}\right)\cdot\widehat{\mathbb{P}}_{11}(t_j,t); \widehat{\mathbb{P}}_{12}(s,t) = \sum_{s<t_j\le t}\widehat{\mathbb{P}}_{11}(s,t_{j-1})\cdot\frac{D_{12_j}}{N_{1j}}. \tag{4.5.13}$$

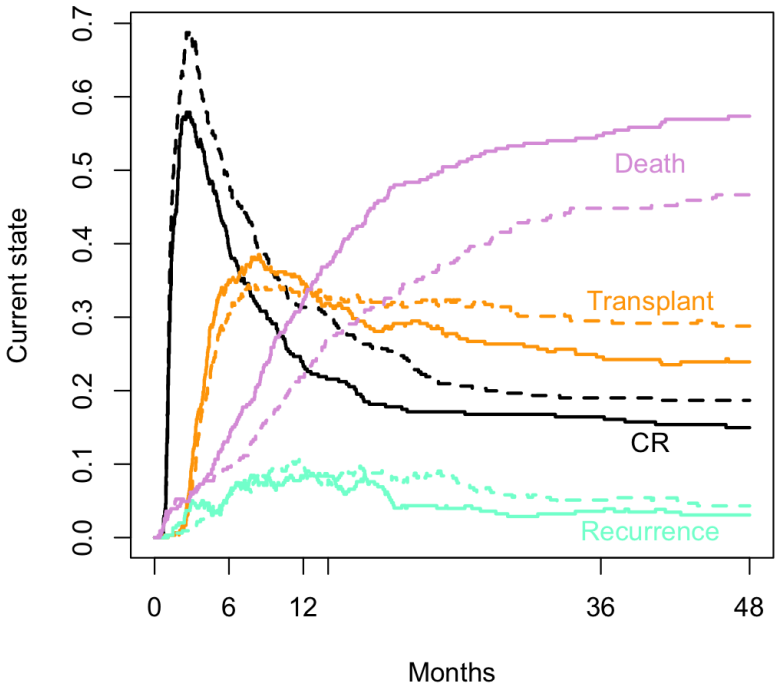
Finally, we have the following estimate of $\mathbb{P}_{02}(s,t)$:

$$\widehat{\mathbb{P}}_{02}(s,t) = \sum_{s<t_j\le t}\widehat{\mathbb{P}}_{00}(s,t_{j-1})\cdot\left(\frac{D_{02_j}}{N_{0j}}\right) \tag{4.5.14}$$

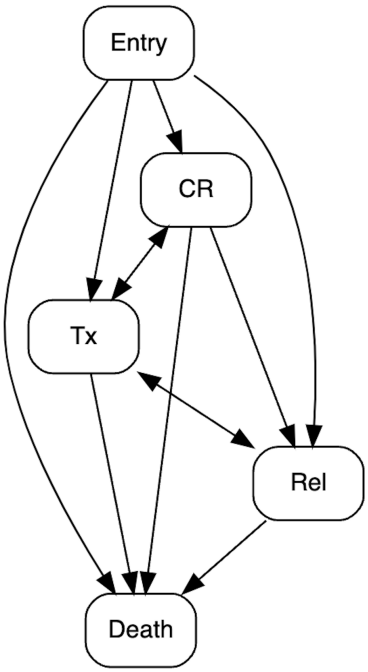
$$+\sum_{s<t_j<t_k\le t}\widehat{\mathbb{P}}_{00}(s,t_{j-1})\cdot\left(\frac{D_{01_j}}{N_{0j}}\right)\cdot\widehat{\mathbb{P}}_{11}(t_j,t_k)\cdot\widehat{\mathbb{P}}_{11}(t_{k+1},t)\cdot\left(\frac{D_{12_j}}{N_{1j}}\right). \tag{4.5.15}$$

3. The general Case of Multi-State Healthy-Disease-Death Model:

The real-life model is often considered to be more complicated, with multiple disease states and possible transition between different diseases as well as the recovery from several diseases. The `myeloid` package in **R** has a complete set of data of leukemia patients treated with 2 different treatments (A or B) and marked the important transitions of each patient. The graph (left) shows the percentage of patient with leukemia at different states at time t, and the graph (right) shows the canonical path of patients with leukemia, where entry represents the start of the disease, CR represents the complete response from treatment A or B; TX represents the hematologic stem cell transplant, Rel represents the patient get full remission from the disease, and death marks the death of the patient.



(a) The full multi-state curves for the two treatment arms A and B. Arm B is dashed in the picture.



(b) The full multi-state graph.

Then we consider a general case of the Markov process: Let $\mathscr{I} := \{0,1,\cdots,k\}$ be the state space ($k+1$ different states), where we now we denote $\alpha_{gh}(t)$ as the instantaneous risk of transition from state g to state h where $g \neq h$ at time $t+\Delta t$. Here those k different states may represent healthy, getting disease A, getting disease B, and death, etc. We also denote $\mathbb{P}_{gh}(s,t)$ to be the probability that an individual at state g at time s will be at state h at time t, then we will have a $(k+1)\times(k+1)$ transition matrix (Borgan, 2005)

$$\boldsymbol{P}(s,t) = \begin{bmatrix} \mathbb{P}_{00} & \mathbb{P}_{10} & \cdots & \mathbb{P}_{k0} \\ \mathbb{P}_{01} & \mathbb{P}_{11} & \cdots & \mathbb{P}_{k1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{P}_{0k} & \mathbb{P}_{1k} & \cdots & \mathbb{P}_{kk} \end{bmatrix} \tag{4.5.16}$$

By definition of the transition matrix, in jth column, we have

$$\sum_{n=0}^{k} \mathbb{P}_{jn}(s,t) = 1 \tag{4.5.17}$$

also, we denote by index 0, to be the "healthy state", and $1,2,\cdots,k-1$ to be the "sick state", and k to be the "death state". So this means that the last column of the matrix, would have

$$\mathbb{P}_{k0} = \cdots = \mathbb{P}_{k(k-1)} = 0, \mathbb{P}_{kk} \equiv 1. \tag{4.5.18}$$

Now suppose we have observed the transition time $t_1 < t_2 < \cdots$ between any two states, also let $g,h \in \mathscr{I}, g \neq h$, we denote D_{ghj} as the number of individuals with transition from state g to state h at observed time t_j, and $D_{gj} = \sum_{h \neq g} D_{ghj}$ as the total number of transitions out of state g at observed time t, N_{gj} be the number of individuals at state g just prior to time t_j. Finally we define the $(k+1)\times(k+1)$ matrix $\widehat{\boldsymbol{\alpha}}_j$ with entries (g,h) by

$$\widehat{\boldsymbol{\alpha}}_j(g,h) = \begin{cases} \dfrac{D_{ghj}}{N_{gj}} & g \neq h \\ \\ -\dfrac{D_{gj}}{N_{gj}} & g = h \end{cases} \tag{4.5.19}$$

then the Aalen Johansen estimator takes the form

$$\widehat{\boldsymbol{P}}(s,t) = \prod_{s<t_j<t} \left(\boldsymbol{I} + \widehat{\boldsymbol{\alpha}}_j\right) \tag{4.5.20}$$

where $\boldsymbol{I}$ is the $(k+1)\times(k+1)$ identity matrix.

4.6 Life-Tables: An Overview

A *life table* is a table which shows the survival rate of a certain population at a certain age or a certain time period. It represents the survivorship of people in a population, or the population's longevity. It usually has several components, where each column will represent a certain age or a time period of the population, and serval rows representing the number of survivors at age or time period x; the number of deaths between age or time period x and $x+1$; the survival rate between age or time period x and $x+1$; the life expectancy at age or time period x, and so on. For example, the table 3.1 below shows the life table of Canadian populations between age 60 and 80 from year 2021 to 2023.

Element	A	B	C	D
Age group	2021 to 2023	2021 to 2023	2021 to 2023	2021 to 2023
60 years	91,805	552	0.00601	24.95
61 years	91,253	598	0.00655	24.10
62 years	90,655	648	0.00715	23.26
63 years	90,007	703	0.00781	22.42
64 years	89,304	763	0.00854	21.59
65 years	88,541	828	0.00935	20.77
66 years	87,713	898	0.01024	19.96
67 years	86,815	975	0.01123	19.17
68 years	85,840	1,058	0.01233	18.38
69 years	84,782	1,149	0.01355	17.60
70 years	83,634	1,246	0.01490	16.84
71 years	82,387	1,352	0.01641	16.08
72 years	81,035	1,466	0.01809	15.34
73 years	79,570	1,588	0.01996	14.62
74 years	77,982	1,719	0.02204	13.90
75 years	76,263	1,858	0.02437	13.21
76 years	74,405	2,007	0.02697	12.52
77 years	72,398	2,163	0.02988	11.86
78 years	70,235	2,327	0.03313	11.21
79 years	67,908	2,498	0.03678	10.57
80 years	65,410	2,674	0.04087	9.96

Table 4.1: The Life Table of Canadian Population Between Age 60 to 80 from year 2021 to 2023. The elements in the first rows represent: (A) The number of survivors at age x; (B) The number of deaths between age x and $x+1$; (C) Death probability between age x and $x+1$; (D) Life expectancy (in years) at age x. *Source: Statistics Canada. Table 13-10-0114-01 Life expectancy and other elements of the complete life table, three-year estimates, Canada, all provinces except Prince Edward Island.*

There are two types of life tables, called *current life tables* and *cohort life tables*. A current life table measures the mortality of the population at a specific time period, while a cohort life table measures the mortality of the population at a specific age. That is, in a cohort life table all individuals from the population is believed to have the same date (period) of birth, like the one shown in table 3.1. While in a current life table, the ages of the individuals may very, and it would not be a great representation of the mortality behavior of the individuals at a certain age. Hence throughout the discussion we will focus on cohort life table and we will see how we can estimate key survival data from it.

The ideal case is that we start our observation at time $t = 0$ when the entire population is born, and we make track of the death of each individual, until the last individual dies. But like I have discussed in the very beginning, it is not always possible due to cost, so instead we use the idea of "censoring" and set our observation into discrete time periods, usually 1 year or 3 years depending on the nature of the study. Then we may use the difference of the population between year x and $x+1$ as the estimate of the number of deaths at year x. This idea is very similar to the concept of interval censoring which I have discussed earlier. Hence in a cohort life table, we would define the following features based on the observed data:

Survivors l_x***:*** This represents the number of individuals alive (at-risk) at exactly age x, while the initial population is set as l_0.

Theoretical deaths d_x***:*** This represents the theoretical number of deaths between ages x and $x+1$.

Conditional probability of death q_x***:*** This represents the conditional probability that an individual will die before age $x+1$ given that the individual is alive at age x, and we have $q_x = d_x/l_x$. Its compliment $p_x = 1 - q_x$ is the conditional probability that the individual will survive at age $x+1$ given the individual is alive at age x, it is given by $p_x = l_x/l_{x-1}$.

Total survival probability P_x***:*** This represents the unconditional probability that the population will survive beyond age x, it starts from the beginning and is given by $P_x = l_x/l_0$.

Years lived L_x***:*** This represents the cumulative time lived by the entire population between year x and $x+1$. All individuals alive at year x enters our study, either one year or the proportion of the period they lived, and in practice we always use 0.5 to denote the average "proportion". Thus we have the following estimate: $L_x = (l_x - d_x) + 0.5d_x$.

Total time lived T_x***:*** This represents the total time lived beyond age x by the all individuals at age x, which is given by $T_x = L_x + L_{x+1} + \cdots$.

Life expectation e_x***:*** This represents the mean of additional years lived by those individuals who are age x, and is given by $e_x = T_x/l_x$.

The survival function $S(x)$ in this case can be estimated by

$$\widehat{S}(t) = P_{x \le t} = \frac{l_1}{l_0} \times \frac{l_2}{l_1} \times \cdots \times \frac{l_x}{l_{x-1}} = \prod_{x \le t} \left(1 - \frac{d_x}{l_x}\right). \tag{4.6.1}$$

As we can see, it is the product of the conditional probability of survival, which is also a product estimator. The key difference between Kaplan-Meier and life-table method is that, in KM estimator, we have observed the exact time of occurrence t, while in the life-table method the population is interval censored.

4.7 The Log-Rank Test

At the very beginning of this paper, I introduced several examples about different types of censoring and truncation, now we go back to one of them: In 1963, Freireich et al report the results of a clinical trial of a drug 6-mercaptopurine (6-MP) versus a placebo in 42 children with acute leukemia. In the experiment,

21 children were given 6-MP treatment and 21 children were given a placebo. The trial was conducted by matching pairs of patients at a given hospital by remission status (complete or partial) and randomizing within the pair to either a 6-MP (group 1) or placebo maintenance therapy (group 2). Patients were followed until their leukemia returned (relapse) or until the end of the study (in months). The data we collected is shown in table 3.2 and figure 3.2 shows the KM / NA curves of those two groups:

$t_{(f)}$	d_{1f}	d_{2f}	n_{1f}	n_{2f}
1	0	2	21	21
2	0	2	21	19
3	0	1	21	17
3	0	2	21	16
5	0	2	21	14
6	3	0	21	12
7	1	0	17	12
8	0	4	16	12
10	1	0	15	8
11	0	2	13	8
12	0	2	12	6
13	1	0	12	4
15	0	1	11	4
16	1	0	11	3
17	0	1	10	3
22	1	1	7	2
23	1	1	6	1

Table 4.2: The remission data among the sample of 42 children. For each ordered failure time $t_{(f)}$ in the sample, we show the numbers of subjects d_{if} failing at that time, separately by group i, followed by the numbers of subjects n_{if} at risk at that time. Here 1 is the treatment group and 2 is the placebo group. *Source: Survival Analysis by David G. Kleinbaum and Mitchel Klein, chapter 2.*

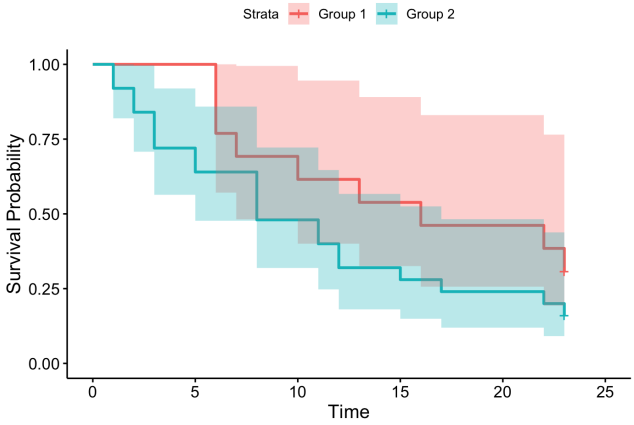


Figure 4.4: Kaplan–Meier survival curve with 95% confidence interval.

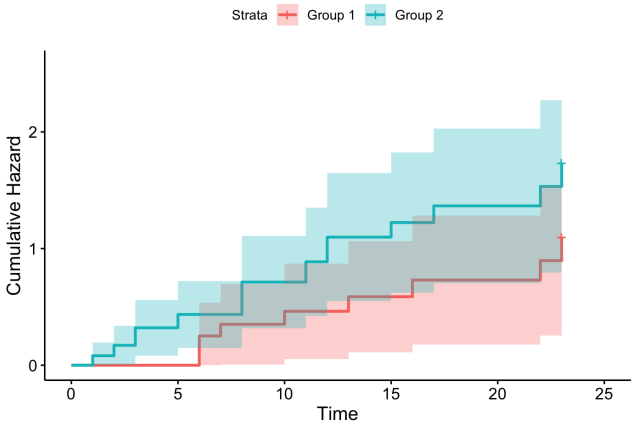


Figure 4.5: Nelson-Aalen hazard curve with 95% confidence interval.

We are now interested in testing whether the two KM curves are statistically equivalent, i.e to see if 6-MP actually plays an important rule in leukemia remission in the previous example.

The log-rank test is a large sample χ-square test that uses as its test criterion a statistic that provides an overall comparison of the KM curves being compared (Klein and Moeschberger, 2003). The categories for the log-rank statistic are defined by each of the ordered failure times for the entire set of data being analyzed.

We first define the expected number of fails for each group at each observed failure time, defined by the proportion of individuals at risk in this group times the total number of fails over both groups

$$e_{1f} = \left(\frac{n_{1f}}{n_{1f}+n_{2f}}\right) \times \left(d_{1f}+d_{2f}\right) \tag{4.7.1}$$

$$e_{2f} = \left(\frac{n_{2f}}{n_{1f}+n_{2f}}\right) \times \left(d_{1f}+d_{2f}\right) \tag{4.7.2}$$

Let $n_f = n_{1f}+n_{2f}$ and $d_f = d_{1f}+d_{2f}$, at time $t_{(f)}$, we obtain the following table:

	Group 1	Group 2	Total
Fails	d_{1f}	d_{2f}	d_f
Survivals	$n_{1f}-d_{1f}$	$n_{2f}-d_{2f}$	n_f-d_f
Total	n_{1f}	n_{2f}	n_f

Table 4.3: contingency table for all subjects in the risk set at time $t_{(f)}$:

Under the null hypothesis $\mathscr{H}_0$: the survival curve of group 1,2 are statistically equivalent, the random variable D_{if} forms a hyper-geometric distribution, with

$$\mathbb{E}[D_{if}] = e_{if} = \frac{n_{if}}{n_f}\cdot d_f \text{ and } \mathbb{V}[D_{if}] = v_{if} = n_{if}\cdot\frac{d_f}{n_f}\cdot\frac{n_f-d_f}{n_f}\cdot\frac{n_f-n_{if}}{n_f-1}, \tag{4.7.3}$$

also we define the total difference between observed fails and the expected fails in each group by

$$O_i - E_i = \sum_{f=1}^{n}\left(d_{if}-e_{if}\right), i = 1,2. \tag{4.7.4}$$

Definition 17. *The log-rank statistic is defined by*

$$Z^2 = \frac{(O_i-E_i)^2}{\mathbb{V}(O_i-E_i)} \tag{4.7.5}$$

for any group $i = 1,2$.

Theorem 7. *Under the null hypothesis* $\mathscr{H}_0$ *: the KM curve for two groups are statistically equivalent, i.e* $S_1(t) = S_2(t)$*, the log-rank statistic is approximately* χ^2 *distribution with one degree of freedom, that is,*

$$Z^2 = \frac{(O_i-E_i)^2}{\mathbb{V}(O_i-E_i)} \sim \chi^2_{(1)}. \tag{4.7.6}$$

Proof. The proof can be done by applying central limit theorem, where we have

$$\frac{\sum_f\left(D_{if}-e_{if}\right)}{\sqrt{\sum_f v_{if}}} \sim N(0,1). \tag{4.7.7}$$

□

We may also express the test statistic by

$$\frac{(O_1-E_1)^2}{E_1} + \frac{(O_2-E_2)^2}{E_2} \sim \chi^2_{(1)}. \tag{4.7.8}$$

To take the decision on the null hypothesis, we compare the test statistic with the chi-square value for a given significance level α, under the condition that the null hypothesis is true. In the case there are multiple groups, say k groups and we would like to test $\mathscr{H}_0 : S_1(t) = \cdots = S_k(t)$ and $\mathscr{H}_1$: At least two groups are different, then we may use the generalized log-rank test, given by

$$\sum_{i=1}^{k}\frac{(O_i-E_i)^2}{E_i} \sim \chi^2_{(k-1)}. \tag{4.7.9}$$

Chapter 5

Analyzing Survival Data Via Counting Processes

5.1 Introduction

In the last chapter, we discussed some famous non-parametric estimators for key survival features through an intuitive way. In this chapter we will re-visit those estimators from the view of stochastic process. It will provide a rather theoretical way to investigate survival data as a counting process. A detailed and rigorous derivation can be found in *Counting Processes and Survival Analysis* by Thomas R. Fleming, David P. Harrington.

5.2 Basic Martingale Theory

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space, a *stochastic process* is defined as follows:

Definition 18. *A stochastic process* $\{X(t) : t \in [0,\infty)\}$ *is a sequence of random variables indexed by time* $t \in [0,\infty)$*, defined on the probability space* $(\Omega, \mathscr{F}, \mathbb{P})$.

Definition 19. *A filtration* $\{\mathscr{F}_t, t \in [0,\infty)\}$ *is a non-decreasing family of sub-*σ*-algebras of* $\mathscr{F}$*, i.e* $\mathscr{F}_s \subset \mathscr{F}_t$ *if* $s < t$*. A stochastic process* $\{X(t), t \in [0,\infty)\}$ *is said to be adopted to a filtration* $\{\mathscr{F}_t, t \in [0,\infty)\}$ *if for each* t*,* $X(t)$ *is* $\mathscr{F}_t$*-measurable.*

We first define a predictable stochastic process:

Definition 20. *A stochastic process* $X(t)$ *adapted to a filtration* $\mathscr{F}_t$ *is predictable, if* $X(t)$ *is* $\mathscr{F}_{t-}$ *measurable. That is,* $X(t) = \mathbb{E}(X(t)|\mathscr{F}_{t-})$.

Definition 21. *A stochastic process* $M = \{M(t), t \in [0,\infty)\}$ *is a martingale with respect to the filtration* $\mathscr{F}_t$ *if:*

(i) $M(t)$ *is adopted to* $\mathscr{F}_t$*;*

(ii) $\mathbb{E}|M(t)| < \infty$ *for all* $t \in [0,\infty)$*;*

(iii) For any $0 < s \leq t$, $\mathbb{E}(M(t)|\mathscr{F}_s) = M(s)$.

Definition 22. *An adapted stochastic process $X = \{X(t), t \in [0,\infty)\}$ is called a sub-martingale if $\mathbb{E}(X(t)\Big|\mathscr{F}_s) \geq X(s), t > s$; Likely X is called a super-martingale if $\mathbb{E}(X(t)\Big|\mathscr{F}_s) \leq X(s), t > s$.*

Proposition 6. *Let $M(t)$ be a martingale, then if $M(0) = 0$, then $\mathbb{E}(M(t)) = 0$ for all $t \geq 0$.*

Proof. By definition, it is easy to see that for any $t \geq 0$, $\mathbb{E}(M(t)|\mathscr{F}_0) = M(0) = 0$. □

To measure the "jump" size at each point case, we define the jump as

$$dM(t) = M(t+\Delta t)^{-}) - M(t^{-}) \tag{5.2.1}$$

when $\Delta t \to 0$. Another way to define a martingale is by the "jump":

Proposition 7. *Let $M(t)$ be integrable and adapted to the filtration $\mathscr{F}_t$, then $M(t)$ is a martingale if and only if $\mathbb{E}(dM(t)|\mathscr{F}_{t-}) = 0$ for all t.*

Proof. Let $M(t)$ be a martingale, then by definition

$$\mathbb{E}(dM(t)|\mathscr{F}_{t-}) = \mathbb{E}(M(t+\Delta t)^{-}|\mathscr{F}_{t-}) - \mathbb{E}(M(t)|\mathscr{F}_{t-}) = M(t-) - M(t-) = 0. \tag{5.2.2}$$

Now assume $\mathbb{E}(dM(t)|\mathscr{F}_{t-}) = 0$ holds, let $0 < s \leq t$, then

$$\begin{aligned}
\mathbb{E}(M(t)|\mathscr{F}_s) &= \mathbb{E}(M(t) + M(s) - M(s)|\mathscr{F}_s) && (5.2.3)\\
&= M(s) + \mathbb{E}(M(t) - M(s)|\mathscr{F}_s) && (5.2.4)\\
&= M(s) + \mathbb{E}\left(\int_t^s dM(u)\Big|\mathscr{F}_s\right) && (5.2.5)\\
&= M(s) + \int_s^t \mathbb{E}(dM(u)|\mathscr{F}_s) && (5.2.6)\\
&= M(s) + \int_s^t \mathbb{E}(\mathbb{E}(dM(u)|\mathscr{F}_{u-})|\mathscr{F}_s) && (5.2.7)\\
&= M(s). && (5.2.8)
\end{aligned}$$

□

Definition 23. *Let $M(t)$ be a square-integrable martingale (i.e $\mathbb{E}(M^2(t)) < \infty$) adapted to the filtration $\mathscr{F}(t)$, the variation process of $M(t)$ is defined by $\langle M\rangle(t)$, such that $M^2(t) - \langle M\rangle(t)$ is a martingale.*

In fact, $\langle M\rangle(t)$ is unique, non-decreasing and $\langle M\rangle(0) = 0$, and it is a sub-martingale. It follows from Doob-Meyer decomposition theorem:

Theorem 8. *(Doob-Meyer) Any sub-martingale X can be decomposed uniquely as*

$$X = X^* + M \tag{5.2.9}$$

where X^* is a non-decreasing predictable stochastic process and M is a zero-mean martingale.

We shall omit the proof for continuous martingales, a short proof is done by Beiglböck, and Veliyev, 2011. But a proof for discrete martingale is straightforward:

Proof. For uniqueness, assume Y_n, W_n are both non-decreasing predictable stochastic processes with $Y_0 = W_0 = 0$, then we define $Z_n = Y_n - W_n$, it is clear that $Z_0 = 0$, also we have $Z_n = (X_n - W_n) - (X_n - Y_n)$, where by definition $X_n - W_n, X_n - Y_n$ are all martingales, and hence Z_n is a martingale and is also predictable, which means we have $\mathbb{E}(Z_{n+1}|\mathscr{F}_n) = Z_n$ (martingale) as well as $Z_{n+1} = \mathbb{E}(Z_{n+1}|\mathscr{F}_n)$ (predictable), then $Z_n \equiv 0$. To find such Y_n, we define

$$Y_n = \sum_{j=0}^{n-1} \Big(\mathbb{E}(X_{j+1} \Big| \mathscr{F}_j) - X_j \Big), \tag{5.2.10}$$

then using the fact that X is a sub-martingale and Y is predictable, we have

$$\begin{aligned}
\mathbb{E}(X_{n+1} - Y_{n+1} \Big| \mathscr{F}_n) &= \mathbb{E}(X_{n+1} \Big| \mathscr{F}_n) - Y_{n+1} && (5.2.11)\\
&= \mathbb{E}(X_{n+1} \Big| \mathscr{F}_n) - \left[\sum_{j=0}^{n-1} \Big(\mathbb{E}(X_{j+1} \Big| \mathscr{F}_j) - X_j \Big) + \mathbb{E}(X_{n+1} \Big| \mathscr{F}_n) - X_n \right] && (5.2.12)\\
&= \mathbb{E}(X_{n+1} \Big| \mathscr{F}_n) - Y_n - \mathbb{E}(X_{n+1} \Big| \mathscr{F}_n) + X_n && (5.2.13)\\
&= X_n - Y_n. && (5.2.14)
\end{aligned}$$

□

The variation process has many nice properties. Firstly if $M(t)$ is a zero-mean martingale, then $\mathbb{E}(M(t)) = 0$, and hence

$$\mathbb{V}(M(t)) = \mathbb{E}(M^2(t)) = \mathbb{E}\langle M \rangle(t), \tag{5.2.15}$$

which further implies, $\langle M \rangle(t)$ is a unbiased estimator of $\mathbb{V}(M(t))$. We shall further discuss some properties of square integrable matngales:

Lemma 4. *Let $M(t)$ be a square-integrable martingale adapted to $\mathscr{F}_t$, then*

$$\mathbb{E}(dM^2(t)|\mathscr{F}_{t-}) = \mathbb{E}((dM(t))^2|\mathscr{F}_{t-}) = \mathbb{V}(dM(t)|\mathscr{F}_{t-}). \tag{5.2.16}$$

Proposition 8. *Let $M(t)$ be a square integrable martingale, $K(t)$ is a bounded and predictable stochastic process with respect to $\mathscr{F}_t$, then:*

(i) $\int_0^t K(s)dM(s)$ *is also a square integrable martingale with respect to $\mathscr{F}_t$;*

(ii) $\left\langle \int_0^t K(s)dM(s) \right\rangle = \int_0^t K^2(s)d\langle M \rangle(s).$

Next we define local martingales. The purpose to do so is that, we may have some processes that satisfy martingale property on a locally domain but fail globally. Thus in order to satisfy the property globally we would define a stopping time, such that the process terminates at a certain time where the martingale property is preserved. i.e, we terminates the process at a time where its adeptness to the filtration is preserved.

Definition 24. *A stopping time τ is a random variable with respect to the filtration $\mathscr{F}_t$, such that $\{\tau < t\} \in \mathscr{F}_t$.*

Definition 25. *A sequence of stopping times $\{\tau_n : n \geq 1\}$ is called a localizing sequence if $\{\tau : n \geq 1\}$ is non-decreasing and $\lim_{n\to\infty} \tau_n = \infty$.*

Definition 26. *A process $X(t)$ is locally bounded if there exists a localizing sequence $\{\tau_n : n \geq 1\}$ such that the stopped process $X_{\tau_n}(t) = X(t \wedge \tau_n)$ is bounded for each n.*

Definition 27. *A process $X(t)$ is locally integrable if there exists a localizing sequence $\{\tau_n : n \geq 1\}$ such that $\mathbb{E}[X(t \wedge \tau_n)] < \infty$ for all t and n; A process $X(t)$ is called locally square integrable if there exists a localizing sequence such that $\mathbb{E}[X^2(t \wedge \tau_n)] < \infty$ for all t and n.*

Definition 28. *A stochastic process $M(t)$ is a local martingale with respect to a filtration $\mathscr{F}_t$ if it is adapted and continuous with left limits, and there exists a localizing sequence $\{\tau_n : n \geq 1\}$ such that $M(t \wedge \tau_n)$ is a martingale with respect to $\mathscr{F}_t$.*

5.3 Properties of Nelson-Aalen Estimator

As we have seen from the previous chapters, we are often interested in the number of deaths (events) at different time t, and we may use $N(t)$ to model the total number of deaths (events) from time 0 to t, where by its nature $N(t)$ is non-decreasing. Formally, $N(t)$ is also a counting process defined as follows:

Definition 29. *We say a stochastic process $N(t)$ is a counting process with respect to time t (either continuous or discrete) if it satisfies the followings:*

(i) $N(0) = 0$;
(ii) For $0 \leq s < t$, $N(t) - N(s)$, shows the number of events occurred in the interval $(s,t]$.

We can see that a counting process is also a sub-martingale, then Doob-Meyer decomposition theorem suggests that we may decompose such a counting process into the sum of a martingale M and a predictable process X^*. We shall use this relation to derive the Nelson-Aalen estimator.

In a random censoring model, assume for each individual $i = 1, \cdots, n$. Let T_i represents the true event time and C_i represents the censoring time. An important assumption is that T_i, C_i are mutually independent. Define $X_i = \min\{T_i, C_i\}$ as the observation, $\delta_i = \chi\{T_i \leq C_i\}$, and when $\delta_i = 1$ it represents the true event time was observed while $\delta_i = 0$ indicates the individual is censored. An obvious counting process is defined to be $N(t) = \sum N_i(t) = \sum_i \chi\{X_i \leq t, \delta_i = 1\}$, which counts the number of individuals whose event times were observed before or at time t. Furthermore, let $Y(t) = \sum_i Y_i(t) = \sum_i \chi\{X_i \geq t\}$ as the at-risk processes at time t, which counts the number of individuals at risk prior to time t. Then by defining the hazard rate function at t to be $\lambda(t)$, we may use $Y_i(t) \cdot \lambda(t)$ as the "expected death" at t, hence by integrating $Y(s)\lambda(s)$ from 0 to t, we will get the cumulative number of events from 0 to t, which is a non-decreasing process. Also note that we have $d\Lambda(t) = \lambda(t)dt$ where $\Lambda(t)$ is the cumulative risk.

Theorem 9. *Given $N(t) = \sum_i \chi\{X_i \leq t, \delta_i = 1\}, Y_i(t) = \sum_i \chi\{X_i \geq t\}$, and denote $\Lambda(t)$ as the cumulative hazard function, then*

$$M(t) = N(t) - \int_0^t Y(s)d\Lambda(s) \tag{5.3.1}$$

is a martingale.

Proof. By summation property of a martingale, it suffices to show that

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)d\Lambda(s) \tag{5.3.2}$$

is a martingale. It is easy to show $M_i(t)$ is measurable, and also by triangle inequality, we have

$$\mathbb{E}(|M(t)|) \leq \mathbb{E}(N_i(t)) + \mathbb{E}\left[\int_0^t Y_i(s)d\Lambda(s))ds\right] \tag{5.3.3}$$

$$= \mathbb{E}(N_i(t)) + \int_0^t \mathbb{P}(X_i \geq s)d\Lambda(s) \tag{5.3.4}$$

$$\leq 1 + \int_0^t \mathbb{P}(X_i \geq s)d\Lambda(s) \tag{5.3.5}$$

$$= 1 + \int_0^t e^{-\Lambda(s)}d\Lambda(s) \tag{5.3.6}$$

$$\leq 2, \tag{5.3.7}$$

which means M_i is bounded. Finally to show it is a martingale it suffices to show that $\mathbb{E}(dM_i(t)|\mathscr{F}_{t-}) = 0$. We have $dM(t) = dN(t) - Y(t)d\Lambda(t)$, so

$$\begin{aligned}\mathbb{E}(dM_i(t)|\mathscr{F}_{t-}) &= \mathbb{E}(dN_i(t) - Y(t)d\Lambda(t)|\mathscr{F}_{t-}) && (5.3.8)\\ &= \mathbb{E}(N_i(t+dt)^- - N_i(t)^-|\mathscr{F}_{t-}) - \mathbb{E}(Y_i(t)\lambda(t)dt|\mathscr{F}_{t-}) && (5.3.9)\\ &= \mathbb{P}(t \le X_i < t+dt, \delta_i = 1|Y_i(t)) - Y_i(t)\lambda(t)dt && (5.3.10)\\ & && (5.3.11)\end{aligned}$$

If $Y_i(t) = 0$, then it means the individual has already experienced the event (or censored) before time t, so we have

$$\mathbb{P}(t \le X_i < t+dt, \delta_i = 1|X_i < t) = 0. \tag{5.3.12}$$

If $Y_i(t) = 1$, which means the individual is still at risk at t, then

$$\mathbb{P}(t \le X_i < t+dt, \delta_i = 1|X_i \ge t) = \lambda(t)dt \tag{5.3.13}$$

Hence combining the two cases give us

$$\mathbb{P}(t \le X_i < t+dt, \delta_i = 1|Y_i(t)) = Y_i(t)\lambda(t)dt, \tag{5.3.14}$$

and hence $\mathbb{E}(dM_i(t)|\mathscr{F}_{t-}) = 0$, which means $M(t)$ is a martingale.

□

Since $M(t)$ is now a martingale, we may write $dM(t) = dN(t) - Y(t)\lambda(t)dt$, we divide both sides by $Y(t)$, where we ignore the case when $Y(t) = 0$. The rigorous treatment when $Y(s) = 0$ can be found in Fleming and Harrington, 2005. In this case we have

$$\lambda(t)dt = \frac{dN(t)}{Y(t)} - \frac{dM(t)}{Y(t)}, \tag{5.3.15}$$

by taking integrals on both sides, we have

$$\Lambda(t) = \int_0^t \frac{dN(s)}{Y(s)} - \int_0^t \frac{dM(s)}{Y(s)}. \tag{5.3.16}$$

We define the Nelson-Aalen estimator of $\Lambda(t)$ is given by

$$\widehat{\Lambda}(t) = \int_0^t \frac{dN(s)}{Y(s)} \tag{5.3.17}$$

where it takes the form of a Riemann-Stieltjes integral, where $dN(t) = N((t+\Delta t)^-) - N(t^-)$ is the increment over a small time period and $Y(s)$ is the risk process (number of individuals at risk), and it is equivalent to the summation form we obtained in the previous chapter. Based on the Nelson-Aalen estimator (5.3.17), we shall discuss some of its properties:

(i) Unbiasedness: *We claim that the Nelson-Aalen estimator is almost unbiased.*

If we look at the difference between the true cumulative hazard and the estimation by Nelson-Aalen , we will see that

$$\widehat{\Lambda}(t) - \Lambda(t) = \int_0^t \frac{dM(s)}{Y(s)}, \tag{5.3.18}$$

since martingale property is preserved under integration, and the integral on the right hand side is a summation form of zero-mean martingales, so we have

$$\mathbb{E}(\widehat{\Lambda}(t)-\Lambda(t))=\mathbb{E}\left[\int_0^t \frac{dM(s)}{Y(s)}\right]=0, \tag{5.3.19}$$

hence Nelson-Aalen estimator is almost unbiased (since we ignored the case when $Y(s)=0$, if without this consideration we would have an unbiased estimator). The biased case is well explained in Fleming and Harrington, 2005.

Theorem 10. *Define* $A(t)=\int_0^t Y(s)\lambda(s)ds$, *then* M^2-A *is a martingale.*

Proof. We first show M is square integrable. Since $N(t),Y(t),\Lambda(t)$ are all non-negative, then

$$M_i^2(t)\le N_i^2(t)+\left(\int_0^t Y_i(s)d\Lambda(s)\right)^2\le 1+\Lambda(t)^2<\infty, \tag{5.3.20}$$

then we have

$$\begin{aligned}
\mathbb{E}(dM^2(t)|\mathscr{F}_{t-}) &= \mathbb{V}(dM(t)|\mathscr{F}_{t-}) && (5.3.21)\\
&= \mathbb{V}(dN(t)-Y(t)d\Lambda(t)|\mathscr{F}_{t-}) && (5.3.22)\\
&= \mathbb{V}(dN(t)|\mathscr{F}_{t-}),\ \text{since } Y(s) \text{ is a constant conditioning on } \mathscr{F}_{t-} && (5.3.23)
\end{aligned}$$

where $dN(t)$ takes only 0 and 1 so we may view it as a Bernoulli random variable:

$$\begin{aligned}
\mathbb{V}(dN(t)|\mathscr{F}_{t-}) &= \mathbb{E}(dN(t)|\mathscr{F}_{t-})(1-(E(dN(t)|\mathscr{F}_{t-})) && (5.3.24)\\
&= Y(t)\lambda(t)dt(1-Y(t)\lambda(t)dt) && (5.3.25)\\
&= Y(t)\lambda(t). && (5.3.26)
\end{aligned}$$

We define $A(t)=\int_0^t Y(s)\lambda(s)ds$, then the above deviation shows that

$$\mathbb{E}(dM^2(t)-dA(t)\Big|\mathscr{F}_{t-})=0. \tag{5.3.27}$$

□

(ii) Variance: *The variance of the Nelson-Aalen estimator is given by*

$$\mathbb{V}(\widehat{\Lambda}(t))=\int_0^t \frac{1}{Y^2(s)}dN(s)=\sum_{t_j\le t}\frac{D_j}{N_j^2}. \tag{5.3.28}$$

From here, we may further derive the variance of the Nelson-Aalen estimator. We have

$$\mathbb{V}(\widehat{\Lambda}(t))=\mathbb{E}\langle\widehat{\Lambda}\rangle(t)=\mathbb{E}\langle\widehat{\Lambda}-\Lambda\rangle(t)=\mathbb{E}\left(\int_0^t \frac{1}{Y^2(s)}d\langle M\rangle(s)\right), \tag{5.3.29}$$

using the fact that $\langle M\rangle=A$, we further have

$$\mathbb{E}\left(\int_0^t \frac{1}{Y^2(s)}d\langle M\rangle(s)\right)=\mathbb{E}\left(\int_0^t \frac{1}{Y^2(s)}dA(s)\right), \tag{5.3.30}$$

also note that $dA(s) = Y(s)\lambda(s)ds = Y(s)d\Lambda(s)$, so combine the results above we have

$$\mathbb{V}(\widehat{\Lambda}(t)) = \mathbb{E}\left(\int_0^t \frac{1}{Y(s)} d\Lambda(s)\right). \tag{5.3.31}$$

Now, by replacing $d\Lambda(s)$ by its estimate $d\widehat{\Lambda}(s)$, we are interested in the difference:

$$\int_0^t \frac{1}{Y(s)} d\widehat{\Lambda}(s) - \int_0^t \frac{1}{Y(s)} d\Lambda(s) = \int_0^t \frac{1}{Y^2(s)} dN(s) - \int_0^t \frac{1}{Y(s)} d\Lambda(s) \equiv \int_0^t \frac{1}{Y^2(s)} dM(s). \tag{5.3.32}$$

Then the mean of the above term is zero (we ignore the case when $Y(s) = 0$) and hence we have the estimate of the variance given by

$$\mathbb{V}(\widehat{\Lambda}(t)) = \int_0^t \frac{1}{Y^2(s)} dN(s), \tag{5.3.33}$$

which again is a Riemann-Stieltjes integral, and it is equivalent to the form $\sum_{t_j \le t} D_j / Y_j^2$ we got in the previous chapter.

(iii) Consistency: *We claim that the Nelson-Aalen estimator is consistent.*

It is first easy to see that $\widehat{\Lambda}(t) \xrightarrow{P} \Lambda$ as $n \to \infty$, since by Chebyshev's inequality we have

$$\mathbb{P}(|\widehat{\Lambda}(t) - \Lambda(t)| > \varepsilon) < \frac{\mathbb{V}(\widehat{\Lambda}(t))}{\varepsilon^2} \to 0, n \to \infty. \tag{5.3.34}$$

In fact, we are able to reach uniform consistency. It will be introduced in the next section with Kaplan-Meier estimator and Lenglart inequality.

5.4 Properties of Kaplan-Meier Estimator

In this section, we will adapt the same set up as the previous chapter, and we already know that Kaplan-Meier estimator takes the form

$$\widehat{S}(t) = \prod_{t_j \le t} \left(1 - \frac{D_j}{N_j}\right) = \prod_{s \le t} \left(1 - \frac{dN(s)}{Y(s)}\right) \tag{5.4.1}$$

where the first expression is the one we have established in the previous chapter, D_j denotes the number of deaths at t_j and N_j denotes the number of individuals at risk prior to t_j. The second expression is the one obtained by counting process where $Y(s)$ is the risk process and dN is the increment of deaths we have observed. We may further derive the expression of $\widehat{S}(t)$:

Proposition 9. $\widehat{S}(t) = 1 - \int_0^t \widehat{S}(s^-) d\widehat{\Lambda}(s)$.

Proof. We begin by investigating the differential $d\widehat{S}(t) = \widehat{S}((t+dt)^-) - \widehat{S}(t^-)$, where

$$d\widehat{S}(t) = \prod_{s \le (t+dt)^-} \left(1 - \frac{dN(s)}{Y(s)}\right) - \prod_{s \le t^-} \left(1 - \frac{dN(s)}{Y(s)}\right) \tag{5.4.2}$$

$$= \prod_{t- \le s \le (t+dt)^-} \left(1 - \frac{dN(s)}{Y(s)}\right) \cdot \prod_{s \le t^-} \left(1 - \frac{dN(s)}{Y(s)}\right) - \prod_{s < t^-} \left(1 - \frac{dN(s)}{Y(s)}\right) \tag{5.4.3}$$

$$\tag{5.4.4}$$

By letting $dt \to 0$, the above equation becomes

$$d\widehat{S}(t) = \left(1 - \frac{dN(t)}{Y(t)}\right) \cdot \prod_{s \le t^-} \left(1 - \frac{dN(s)}{Y(s)}\right) - \prod_{s < t^-} \left(1 - \frac{dN(s)}{Y(s)}\right) \tag{5.4.5}$$

$$= -\frac{dN(t)}{Y(t)} \widehat{S}(t^-) \tag{5.4.6}$$

Then integrating from 0 to t gives

$$\int_0^t d\widehat{S}(u) = \widehat{S}(t) - 1 = \int_0^t -\frac{dN(u)}{Y(u)} \widehat{S}(u^-), \tag{5.4.7}$$

which gives

$$\widehat{S}(t) = 1 - \int_0^t \widehat{S}(u^-) \frac{dN(u)}{Y(u)} = 1 - \int_0^t \widehat{S}(u^-) d\widehat{\Lambda}(u). \tag{5.4.8}$$

□

Proposition 10. *If $S(t) > 0$, then*

$$\frac{\widehat{S}(t)}{S(t)} = 1 - \int_0^t \frac{\widehat{S}(u^-)}{S(u)} \left(\frac{dN(u)}{Y(u)} - d\Lambda(u)\right). \tag{5.4.9}$$

In order to prove this proposition, we will adapt the results from two technical lemmas introduced below:

Lemma 5. *Let U, V be right continuous and of bounded variations on $(0,t]$, then*

$$U(t)V(t) = U(0)V(0) + \int_0^t U(s^-) dV(s) + \int_0^t V(s) dU(s). \tag{5.4.10}$$

Lemma 6. *For a right continuous function $W(s)$ we have $d(W(s)^{-1}) = -\dfrac{dW(s)}{W(s)W(s^-)}$.*

Proof. With lemma 5,6 introduced, we let $U(s) = \widehat{S}(s), W(s) = S(s), V(s) = S(s)^{-1}$ and the desired result will follow. □

We now introduce the properties of $\widehat{S}(t)$:

(i) Unbiasedness: *We claim that the Kaplan-Meier estimator is almost unbiased.*

First ignore the small possibility that $Y(s) = 0$, we multiply $S(t)$ on both sides and we will get

$$\widehat{S}(t) = S(t) - S(t) \int_0^t \frac{\widehat{S}(u^-)}{S(u)} \left(\frac{dN(u)}{Y(u)} - d\Lambda(u)\right), \tag{5.4.11}$$

and by (5.3.15) we know that

$$d\Lambda(u) = \lambda(u)du = \frac{dN(u)}{Y(u)} - \frac{dM(u)}{Y(u)} \tag{5.4.12}$$

hence we have

$$\widehat{S}(t) - S(t) = S(t) \int_0^t \frac{\widehat{S}(u^-)}{S(u)} \cdot \frac{dM(u)}{Y(u)}, \tag{5.4.13}$$

finally by proposition 11, we know the right hand side is again a martingale, and by definition it has mean zero, which means $\mathbb{E}(\widehat{S}(t) - S(t)) = 0$. Hence we claim that the Kaplan-Meier estimator is almost unbiased (We ignored the case when $Y(s) = 0$. A rigorous treatment can also be found in Fleming and Harrington, 2005, and Zhou (1988) also showed that the true bias is exponentially small.

(ii) Uniform Consistency:

To show the uniform consistency, we shall introduce Lenglart inequality which will help us establish the result:

Lemma 7. *(Lenglart inequality) Let M be a local square-integrable martingale. Suppose H is a predictable and locally bounded process, then for any finite stopping time τ, and any $\varepsilon, \eta > 0$, we have*

$$\mathbb{P}\left(\sup_{0 \le s \le \tau} \left| \int_0^s H(u) dM(u) \right| > \varepsilon \right) \le \frac{\eta}{\varepsilon^2} + \mathbb{P}\left(\int_0^\tau H^2(s) d\langle M \rangle(s) > \eta \right). \tag{5.4.14}$$

Proposition 11. *The Kaplan-Meier estimator is uniformly consistent. If $u \in (0, \infty]$ is such that $Y(s) \to \infty, n \to \infty$, for any $s \le u$, then* $\sup_{0 \le s \le u} |\widehat{S}(s) - S(s)| \to 0, n \to \infty$.

Proof. Let $Z(t) = \int_0^t \frac{\widehat{S}(u^-)}{S(u)} \frac{dM(u)}{Y(u)}$, using the fact that $S(t) \le 1$, we have $|\widehat{S}(t) - S(t)| \le |Z(t)|$. Hence for any ε, η, we have

$$\mathbb{P}\left(\sup_{0 \le s \le u} |\widehat{S}(t) - S(t)| > \sqrt{\varepsilon} \right) \le \mathbb{P}\left(\sup_{0 \le s \le u} Z^2(t) > \varepsilon \right) \le \frac{\eta}{\varepsilon} + \mathbb{P}\left(\int_0^u \frac{\widehat{S}^2(s^-)}{S^2(s)} \frac{d\Lambda(s)}{Y(s)} > \eta \right) \tag{5.4.15}$$

$$\le \frac{\eta}{\varepsilon} + \mathbb{P}\left(\frac{\Lambda(u)}{S^2(u) Y(u)} > \eta \right), \tag{5.4.16}$$

with assumption that $Y(u) \to \infty$, we have established the desired result. □

We may also establish the uniform consistency for Nelson-Aalen estimator:

Proposition 12. *The Nelson-Aalen estimator is uniformly consistent. If $u \in (0, \infty]$ is such that $Y(s) \to \infty, n \to \infty$, for any $s \le u$, then* $\sup_{0 \le s \le u} |\widehat{\Lambda}(s) - \Lambda(s)| \to 0, n \to \infty$.

Proof. Let $Z(t) = \widehat{\Lambda}(t) - \Lambda(t) = \int_0^t \frac{dM(s)}{Y(s)}$, then for any ε, η, we have

$$\mathbb{P}\left(\sup_{0 \le s \le u} |\widehat{\Lambda}(s) - \Lambda(s)| > \sqrt{\varepsilon} \right) = \mathbb{P}\left(\sup_{0 \le s \le u} Z^2(s) > \varepsilon \right) \le \frac{\eta}{\varepsilon} + \mathbb{P}\left(\int_0^u \frac{d\Lambda(s)}{Y(s)} > \eta \right) \tag{5.4.17}$$

$$\le \frac{\eta}{\varepsilon} + \mathbb{P}\left(\frac{\Lambda(u)}{Y(u)} > \eta \right), \tag{5.4.18}$$

and since $Y(u) \to \infty$, we have established the desired result. □

5.5 Kernel Smoothed Estimation

In this section, we will introduce some smoothing methods to improve the estimation of the hazard rate. In our previous discussion, the hazard rate is a non-decreasing and right continuous step function. We shall use smoothing technique to make our estimate as a continuous function which also provides a more stable and interpretable representation of the hazard rate. There are several smoothing techniques, and details can be found in Wang [7]. We mainly introduce kernel method in this section.

Definition 30. *The kernel function $K(t)$ is a non-negative function, with $K(u) = K(-u)$ for all u and $\int_{-\infty}^{\infty} K(u)du = 1$.*

Some typical choice of kernel includes the Gaussian kernel, Epanechnikov kernel and uniform kernel. Another important property of a kernel is the effect of bandwidth. We define $K_h(t) = \frac{1}{h}K\left(\frac{t}{h}\right)$, and the bandwidth h will determine the behavior of the kernel. A typical kernel smoothed estimate of a density function $f(t)$ takes the form

$$\widetilde{f}(t) = \sum_{i=1}^{n} w(X_i)\frac{1}{h}K\left(\frac{t-X_i}{h}\right) \tag{5.5.1}$$

where X_i are observed data, K is the kernel and $w(X_i)$ is the weight of each X_i. A typical choice of $w(X_i)$ is simply take each to be $\frac{1}{n}$ and the resulting kernel estimate is

$$\widetilde{f}(t) = \frac{1}{hn}\sum_{i=1}^{n} K\left(\frac{t-X_i}{h}\right). \tag{5.5.2}$$

To estimate the hazard rate, we set the weight $w(X_i)$ as the estimated hazard, and adapt the notion from previous sections, we define the kernel-smoothed hazard estimator as

Definition 31. *The kernel-smoothed hazard estimator is defined by*

$$\widetilde{\lambda}(t) = \int_0^{\infty} K_h(t-s)\frac{dN(s)}{Y(s)}. \tag{5.5.3}$$

The behavior of the kernel smoothed estimator $\widetilde{\lambda}$ shall depend on the value of bandwidth h, and in this section we will derive an optimal bandwidth via mean integrated square error (MISE). To establish the result, we will make the following assumptions:

A1 The true hazard density $\lambda(t)$ is at least twice continuously differentiable and square integrable. Its second derivative is also square integrable;

A2 The kernel K is bounded with finite moment and compactly supported;

A3 The bandwidth h only depends on n and as $n \to \infty$, we have $h \to 0$ and $nh \to \infty$.

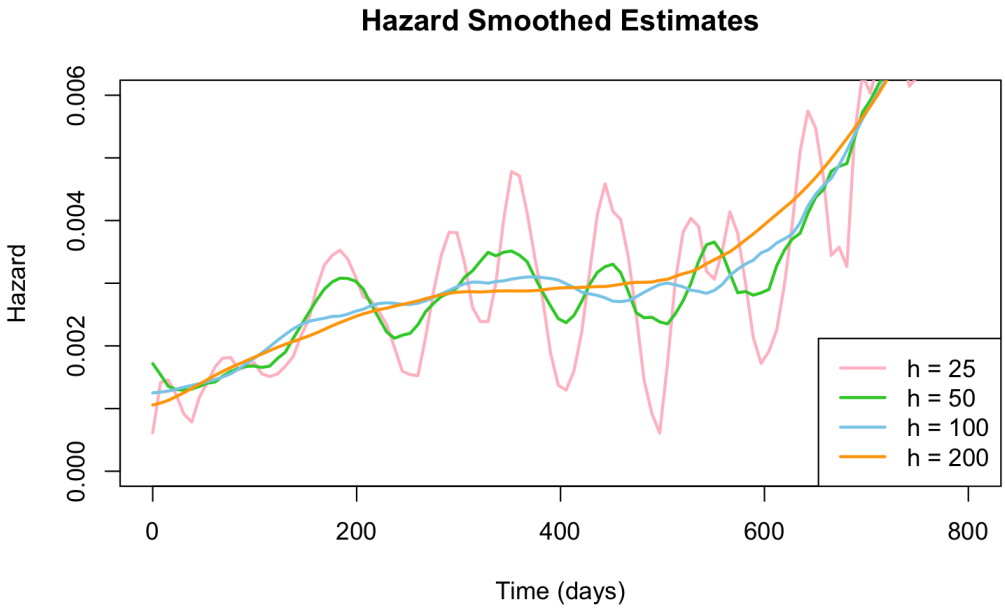


Figure 5.1: The smoothed hazard curve using Epanechnikov kernel with different bandwidth. Datas are from `lung` package in **R** where it studies the survival time of patients with lung cancer.

While below is the hazard estimate using Nelson-Aalen estimator:

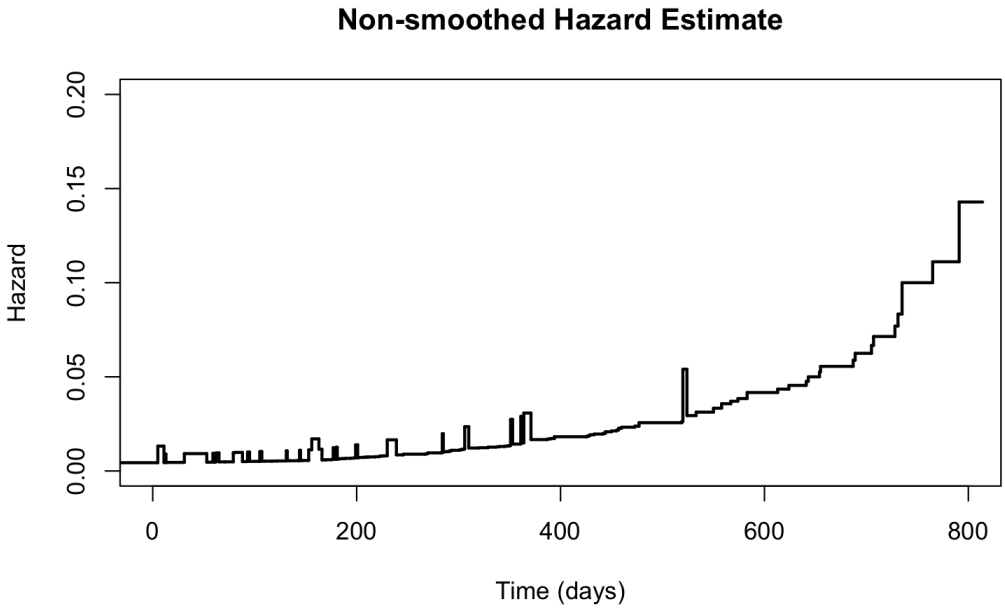


Figure 5.2: Non-Smoothed hazard estimate using Nelson-Aalen estimator

As we can see, a larger h will result in a smoother estimate, but will then lose local informations and increase the bias but lower the variance. While a small h can better describe the hazard rate and will have a lower bias but higher variance. So our goal now is to find the optimal bandwidth. One way

optimize h is first by defining the mean integrated squared error (MISE), and then we minimize MISE. Another famous approach to find optimal bandwidth is by cross-validation criterion (LCV) and will not be introduced here.

Definition 32. *Let $\widetilde{\lambda}(t)$ be a kernel estimate of $\lambda(t)$, the mean integrated squared error (MISE) is defined as*

$$MISE(h) = \mathbb{E}\left[\int_0^\infty (\widetilde{\lambda}(t) - \lambda(t))^2 dt\right] \tag{5.5.4}$$

We may further write MISE as a sum of an integrated square bias term and an integrated variance term:

$$\text{MISE}(h) = \int_0^\infty \left(\mathbb{E}\widetilde{\lambda}(t) - \lambda(t)\right)^2 dt + \int_0^\infty \mathbb{V}(\widetilde{\lambda}(t))dt, \tag{5.5.5}$$

and the following two lemmas will give an estimate of each of them:

Lemma 8. *The bias of the estimate $\widetilde{\lambda}(t)$ is given by*

$$Bias[\widetilde{\lambda}(t)] = \frac{h^2\lambda''(t)}{2}\int_{-\infty}^\infty u^2 K(u)du + O(h^3) \tag{5.5.6}$$

Proof. Note that the Nelson-Aalen estimator is unbiased (we ignore the case when $Y(s) = 0$), hence in the expectation if we replace the estimate by true hazard, we will get rid of the expectation. We define $\lambda(s) = 0$ when $s < 0$ and write

$$\mathbb{E}[\widetilde{\lambda}(t)] = \int_{-\infty}^\infty K_h(t-s)\lambda(s)ds = \int_{-\infty}^\infty \frac{1}{h}K\left(\frac{t-s}{h}\right)\lambda(s)ds, \tag{5.5.7}$$

then we let $u = (t-s)/h$, so we have $ds = -hdu$ and the integral becomes

$$\mathbb{E}[\widetilde{\lambda}(t)] = \int_{\infty}^{-\infty} \frac{1}{h}K(u)\lambda(t-hu)(-hdu) = \int_{-\infty}^\infty K(u)\lambda(t-hu)du. \tag{5.5.8}$$

Then we perform a Taylor expansion on λ, and assume λ is at least twice continuously differentiable, we will have

$$\lambda(t-hu) = \lambda(t) - hu\lambda'(t) + \frac{h^2u^2}{2}\lambda''(t) - O(h^3), \tag{5.5.9}$$

hence we have

$$\mathbb{E}[\widetilde{\lambda}(t)] = \int_{-\infty}^\infty K(u)\cdot\left(\lambda(t) - hu\lambda'(t) + \frac{h^2u^2}{2}\lambda''(t) - O(h^3)\right)du \tag{5.5.10}$$

$$= \lambda(t)\int_{\infty}^\infty K(u)du - h\lambda'(t)\int_{-\infty}^\infty uK(u)du + \frac{h^2\lambda''(t)}{2}\int_{\infty}^\infty u^2K(u)du - O(h^3) \tag{5.5.11}$$

$$= \lambda(t) + \frac{h^2\lambda''(t)}{2}\int_{-\infty}^\infty u^2K(u)du - O(h^3). \tag{5.5.12}$$

We ingore the $O(h^3)$ term, and by definition, we have

$$\text{Bias}[\widetilde{\lambda}(t)] = |\lambda(t) - \mathbb{E}[\widetilde{\lambda}(t)]| = \frac{h^2\lambda''(t)}{2}\int_{-\infty}^\infty u^2K(u)du + O(h^3) \propto h^2. \tag{5.5.13}$$

□

Lemma 9. *The variance of the kernel-smoothed hazard estimator is given by*

$$\mathbb{V}(\widetilde{\lambda}(t)) = \frac{1}{h}\frac{\lambda(t)}{Y(t)}\int_0^\infty K^2(u)du. \tag{5.5.14}$$

Proof. From Doob-Meyer decomposition, we may write $dN(s) = dM(s) + Y(s)\lambda(s)ds$ where M is a zero mean martingale. Then we have

$$\widetilde{\lambda}(t) = \int_0^t K_h(t-s)\frac{dM(s)}{Y(s)} + \int_0^t K_h(t-s)\frac{\cancel{Y(s)}\lambda(s)ds}{\cancel{Y(s)}}. \tag{5.5.15}$$

We see that the variance of the second term is 0, and use the property of variation process, we have

$$\mathbb{V}(\widetilde{\lambda}(t)) = \mathbb{E}\left\{\int_0^t K_h^2(t-s)\frac{d\langle M\rangle(s)}{Y^2(s)}\right\}. \tag{5.5.16}$$

We know that $\langle M\rangle(u) = A(u) = \int_0^s Y(u)\lambda(u)du$, hence we have $d\langle M\rangle(s) = Y(s)\lambda(s)ds$, and hence we have

$$\mathbb{V}(\widetilde{\lambda}(t)) = \mathbb{E}\left\{\int_0^\infty K_h^2(t-s)\frac{\lambda(s)ds}{Y(s)}\right\} = \int_{-\infty}^\infty \frac{1}{h}K^2\left(\frac{t-s}{h}\right)\frac{\lambda(s)}{Y(s)}ds. \tag{5.5.17}$$

where we define $\lambda(s) = 0$ when $s < 0$. Then by substutiting $u = (t-s)/h$ we will have

$$\mathbb{V}(\widetilde{\lambda}(t)) = \frac{1}{h}\int_{-\infty}^\infty K^2(u)\frac{\lambda(t-hu)}{Y(t-hu)}du \approx \frac{1}{h}\cdot\frac{\lambda(t)}{Y(t)}\int_{-\infty}^\infty K^2(u)du \propto h^{-1}. \tag{5.5.18}$$

for small bandwidth h. □

Then, we ignore all higher order terms and plug in our estimates, the MISE is approximately given by

$$\text{MISE}(h) \approx \text{AMISE}(h) = \int_0^\infty \frac{h^4(\lambda''(t))^2}{4}\left(\int_{-\infty}^\infty u^2K(u)du\right)^2 dt + \int_0^\infty \frac{\lambda(t)}{hY(t)}\int_{-\infty}^\infty K^2(u)dudt, \tag{5.5.19}$$

where AMISE is the asymptotic MISE. The above term is a function of h, and we see that $\text{MISE}(h) \propto h^4 + h^{-1}$, we may write it as $\text{MISE}(h) = C_1h^4 + C_2/h$ where C_1, C_2 are independent of h, hence by taking derivative, we find the optimal bandwidth with minimal MISE is $h_{\text{opt}} = \left(\frac{C_2}{4C_1}\right)^{1/5}$. We define

$$\mu(K) = \int_{-\infty}^\infty u^2K(u)du \quad R(K) = \int_{-\infty}^\infty K^2(u)du, \tag{5.5.20}$$

then we have

$$h_{\text{opt}} = \left[\frac{R(K)\cdot\int_0^\infty \frac{\lambda(t)}{Y(t)}dt}{\mu^2(K)\cdot\int_0^\infty(\lambda''(t))^2)}\right]^{1/5}. \tag{5.5.21}$$

In a large sample, we may replace $Y(t)$ by $nS(t)$. In this case we arrive at

$$h_{\text{opt}} = \left[\frac{R(K)\cdot\int_0^\infty \frac{\lambda(t)}{S(t)}dt}{n\mu^2(K)\cdot\int_0^\infty(\lambda''(t))^2}\right]^{1/5} = O(n^{1/5}). \tag{5.5.22}$$

Now we obtain the order of optimal bandwidth, but from here it is not possible to compute the exact value using the formula above since it involves unknown $\lambda(t)$. A plug in approach is possible by assuming the density of $\lambda(t)$ forms a certain distribution. Another way is to estimate $R(\lambda^{''}(t))$, which can be derived from a pilot kernel K_0 (Catherine. R. Loader [8]):

$$\lambda^{''}(t)dt = \frac{w_0(X_i)}{h^3}\sum_{i=1}^{n} K_0^{''}\left(\frac{X_i - t}{h}\right) \tag{5.5.23}$$

leading to

$$\int_0^{\infty}(\lambda^{''}(t))dt = \frac{w_0^2(X_i)}{h^6}\sum_{i=1}^{n}\sum_{j=1}^{n}\int_{-\infty}^{\infty} K^{''}\left(\frac{X_i - t}{h}\right)K^{''}\left(\frac{X_j - t}{h}\right)dt. \tag{5.5.24}$$

Although the pilot bandwidth also vary, the most common solution to the pilot bandwidth problem is through an "assumed" relation between the pilot bandwidth and original bandwidth (Loader, 1999 , and Gasser, Kneip and Kohler, 1991 (GKK)) ; Sheather and Jones (1991) (SJPI) both provided different approaches to solve pilot bandwidth problem. Finally $\lambda(t), S(t)$ on the numerator may be estimated by the standard method.

In this section, we assume a fixed bandwidth for simplicity. But this would lead to boundary effects since the support of the kernel may exceed the available range of data, and Müller, Wang, (1994) proposed a varying kernels and bandwidths method.

Chapter 6

Non-Parametric Hypothesis Testing

6.1 Introduction

Previously we have studied some famous non-parametric estimators and their asymptotic properties. In this chapter we will further study the testing methods for non-parametric inferences. The functions we will be testing are hazard functions and survival functions, that is we shall verify the Nelson-Aalen estimator and Kaplan-Meier estimator. In parametric inference, the testing methods had been studied in chapter 2, where we introduced Wald test, Rao test and Likelihood Ratio test. In this chapter, we will introduce some testing methods for non-parametric inference.

6.2 One Sample Testing

One sample test can be viewed as verifing the statement that the hazard function we obtained via Nelson-Aalen estimator is accurate. In this case we will have the null hypothesis $\mathscr{H}_0 : \lambda_0(t)$ is the accurate hazard function for all $t < \tau$ and the alternative hypothesis to be $\lambda_0(t)$ is not the accurate estimate of hazard function for $t < \tau$. Recall the Nelson-Aalen estimator states that the cumulative hazard function can be estimated by

$$\Lambda(t) = \sum_{t_i \leq t} \frac{d_i}{T(t_i)} \tag{6.2.1}$$

where d_i is the number of death at time t_i and $Y(t_i)$ is the total number of individuals at risk prior to t_i. Under the null hypothesis, the expected hazard rate at t_i is equal to $\lambda_0(t_i)$. So the expected cumulative hazard rate at τ under the null hypothesis is given by

$$E(\tau) = \int_0^\tau \lambda_0(t)dt \tag{6.2.2}$$

and the cumulative hazard rate at τ using Nelson-Aalen estimator is given by

$$O(\tau) = \sum_{t_i \leq \tau} \frac{d_i}{Y(t_i)} \tag{6.2.3}$$

and here we made the assumption that $Y(t) \neq 0$. A more rigorous approach is to multiply the whole thing by a weight function $W(t)$ and it assigns value 0 whenever $Y(t) = 0$. A test statistic $Z(\tau)$ is

defined by taking the difference between the observed and expected cumulative hazard rates under the null hypothesis:

$$Z(\tau) = O(\tau) - E(\tau) = \sum_{t_i \leq \tau} W(t_i) \frac{d_i}{Y(t_i)} - \int_0^\tau W(t) \lambda_0(t) dt. \tag{6.2.4}$$

We first make an claim on the variance of the statistic $Z(\tau)$:

Lemma 10. *Under the null hypothesis, the variance of $Z(\tau)$ is given by*

$$\mathbb{V}[Z(\tau)] = \int_0^\tau W^2(t) \frac{\lambda_0(t)}{Y(t)} dt. \tag{6.2.5}$$

Proof. When the null hypothesis is true, then by the derivation from counting processes, we know that

$$Z(\tau) = \sum_{t_i \leq \tau} W(t_i) \frac{d_i}{Y(t_i)} - \int_0^\tau W(t) \lambda_0(t) dt \tag{6.2.6}$$

is a martingale. If we adopt the notion from the previous chapter, we can rewrite the above test statistic as

$$Z(\tau) = M(\tau) = \int_0^\tau W(t) \left(\frac{dN(t)}{Y(t)} - \lambda_0(t) dt \right) \tag{6.2.7}$$

where we have

$$dM(t) = dN(t) - Y(t) \lambda_0(t) \tag{6.2.8}$$

so now we have

$$Z(\tau) = \int_0^\tau W(t) \frac{dM(t)}{Y(t)} dt. \tag{6.2.9}$$

To compute the variance, we will use the properties of variation process:

$$\mathbb{V}[Z(\tau)] = \mathbb{E} \left\langle \int_0^\tau W(t) \frac{dM(t)}{Y(t)} dt \right\rangle = \mathbb{E} \left(\int_0^\tau \frac{W^2(t)}{Y^2(t)} d\langle M \rangle(t) \right) \tag{6.2.10}$$

where $d\langle M \rangle$ can be replaced by $d\Lambda(t) = Y(t)\lambda_0(t)$, so hence we have

$$\mathbb{V}[Z(\tau)] = \mathbb{E} \left(\int_0^\tau W^2(t) \frac{\lambda_0(t)}{Y(t)} dt \right) = \int_0^\tau W^2(t) \frac{\lambda_0(t)}{Y(t)} dt. \tag{6.2.11}$$

□

Under the null hypothesis, one may conduct the testing based on the following theorem:

Theorem 11. *Under the null hypothesis, the statistic satisfies $\dfrac{Z^2(\tau)}{\mathbb{V}[Z(\tau)]} \sim \chi^2_{(1)}$.*

Proof. Since we know Z is also a mean-zero martingale, then the martingale central limit theorem states that

$$\frac{Z(\tau) - \mathbb{E}[Z(\tau)]}{\sqrt{\mathbb{V}[Z(\tau)]}} \xrightarrow{d} N(0,1), \tag{6.2.12}$$

and we use the fact that the square of a standard normal distribution results in a $\chi^2_{(1)}$ distribution. □

We may then compute the desired test statistic and compare it to a standard χ^2 table. Depending on the alternative hypothesis $\mathscr{H}_1$, the rejection region of $\mathscr{H}_0$ is given by the following table:

Alternative Hypothesis $\mathscr{H}_1$	Rejection Region of $\mathscr{H}_0$ Given Significance Level α
$\mathscr{H}_1 : \lambda(t) \neq \lambda_0(t)$	$\gamma > \chi^2_{\alpha/2}$ or $\gamma < \chi^2_{1-\alpha/2}$
$\mathscr{H}_1 : \lambda(t) > \lambda_0(t)$	$\gamma > \chi^2_{\alpha}$
$\mathscr{H}_1 : \lambda(t) < \lambda_0(t)$	$\gamma < \chi^2_{1-\alpha}$

Also the choice of the weight might give us different tests results. There are many popular choice of weight function $W(t)$ and each has its purpose. One choice is to make $W(t) = 1$ which assigns unit weight to all observed failures, and when $W(t) = Y(t)$, it is the case of an one-sample log-rank test. Lastly Fleming and Harrington (1982) introduced the weight function of the form $W(t) = S_0(t)^p[1-S_0(t)]^q$ with $p, q \geq 0$ to be chosen, where under null hypothesis $S_0(t) = \exp(-\Lambda_0(t))$ is the "hypothesized" survival function. By adjusting the value of p, q, one can adjust the weight of each observed failure time based on their time of arrvival. If $p > q$, it means we assign more weight to those early deaths, and $p < q$ means we assign more weight to those late deaths.

6.3 Theory of Log-Rank Test

More often, our experiment will have a treatment group as well as a placebo group. Our goal is to see if the treatment is effective, and it will involve comparing different hazard rates. In this section we will introduce hypothesis testing methods for two or more groups, denoted by $1, 2, \cdots, k$, our null hypothesis will be

$$\mathscr{H}_0 : \lambda_1(t) = \lambda_2(t) = \cdots = \lambda_k(t) \quad \text{for } t \leq \tau \tag{6.3.1}$$

and the alternative hypothesis is

$$\mathscr{H}_1 : \text{ at least one of } \lambda_k(t) \text{ is different from the other for some } t \leq \tau. \tag{6.3.2}$$

The first test we will introduce is log-rank test where we breifly mentioned in the end of Chapter 4. Here we do not assign a weight function $W(t)$ and all observed deaths are considered to have equal and unit weight. We later will introduce weighted log-rank test and many other tests, depending on the choice of our weight function.

We first consider a testing of two samples: Denote by sample 1 (treatment group) and sample 2 (placebo group), and our null hypothesis is $\mathscr{H}_0 : \lambda_1(t) = \lambda_2(t)$. We denote D_{ij}, Y_{ij} as the number of deaths and number of individuals at risk at time t_j in group i, respectively. and we have $D_j = D_{ij} + D_{2j}, Y_j = Y_{ij} + Y_{2j}$. We addtionally define an indicator variable $Z(X_i)$, $i = 1, \cdots, n$ where n is the whole sample size, and $Z(X_i) = 1$ if the individual belongs to the treatment group and $Z(X_i) = 0$ otherwise (placebo group). By adapting the notion from counting process, we can derive D_{1j} and Y_{1j} as

$$D_{ij} = \sum_{i=1}^{n} Z(X_i) dN_i(t_j), Y_{1j} = \sum_{i=1}^{n} Z(X_i) Y_i(t_j) \tag{6.3.3}$$

where $dN_i(t_j)$ denotes the increment of counting process for individual i at time t_j, $Y_i(t_j)$ is the at risk process for individual i at time t_j. The numerator of the log-rank statistic can be expressed by (see Chapter 4.7)

$$\mathscr{U} = \sum_{t_j} \left(D_{1j} - \frac{D_j Y_{1j}}{Y_j} \right), \tag{6.3.4}$$

which is the difference between observed death in treatment group and the total number of deaths multiplied by the proportion of individuals in treatment group. We then denote $\bar{Z}(t) = \dfrac{\sum_{i=1}^{n} Z(X_i) Y_i(t)}{\sum_{i=1}^{n} Y_i(t)}$ as the proportion of individuals at risk that belong to the treatment group at time t, so we have

$$\bar{Z}(t)\sum_{i=1}^{n} Y_i(t) = \sum_{i=1}^{n} Z(X_i)Y_i(t) \tag{6.3.5}$$

and we can further write the numerator of the log-rank statistic as

$$\mathscr{U} = \sum_{t_j}\sum_{i=1}^{n}(Z_i - \bar{Z}(t_j))dN_i(t_j) \tag{6.3.6}$$

which can be further written in terms of a stochastic intergal:

$$\mathscr{U} = \sum_{i=1}^{n}\int_0^{\infty}(Z_i - \bar{Z}(s))dN_i(s) \tag{6.3.7}$$

where we take the time up to ∞ and recall that $M_i(t) = N_i(t) - \int_0^t Y_i(s)\lambda_i(s)ds$ is a martingale, so we have

$$\mathscr{U} = \sum_{i=1}^{n}\int_0^{\infty}(Z_i - \bar{Z}(s))dM_i(s) + \sum_{i=1}^{n}\int_0^{\infty}(Z_i - \bar{Z}(s))Y_i(s)\lambda_i(s)ds \tag{6.3.8}$$

and under null hypothesis, the second intergal in the equation above is

$$\begin{aligned}
\sum_{i=1}^{n}\int_0^{\infty}(Z_i - \bar{Z}(s))Y_i(s)\lambda_i(s)ds &= \int_0^{\infty}\sum_{i=1}^{n}(Z_i - \bar{Z}(s))Y_i(s)\lambda_i(s)ds && (6.3.9)\\
&= \int_0^{\infty}\left(\sum_{i=1}^{n} Z_iY_i(s) - \bar{Z}\sum_{i=1}^{n} Y_i(s)\right)\lambda_i(s)ds && (6.3.10)\\
&= 0. && (6.3.11)
\end{aligned}$$

and hence

$$\mathscr{U} = \sum_{i=1}^{n}\int_0^{\infty}(Z_i - \bar{Z}(s))dM_i(s). \tag{6.3.12}$$

Proposition 13. *Define*

$$U(t) = \sum_{i=1}^{n}\int_0^{\infty}(Z_i - \bar{Z}(s))dM_i(s), \tag{6.3.13}$$

we claim that $U(t)$ is a martingale with filtration $\mathscr{F}_t = \sigma\{N_i(s), Y_i(s), Z_i\}$.

Proof. First it is easy to see that the function

$$\int_0^t (Z_i - \bar{Z}(s))dM_i(s) = \int_0^t (Z_i - \bar{Z}(s))dN_i(s) - \int_0^t (Z_i - \bar{Z}_i(s))Y_i(s)\lambda_i(s)ds \tag{6.3.14}$$

is $\mathscr{F}_t$ measurable, and the sum of measureable functions are measurable as well. Then we see that

$$\begin{aligned}
\left|\int_0^t (Z_i - \bar{Z}(s))dM_i(s)\right| &\le \int_0^t |Z_i - \bar{Z}(s)|\cdot|dM_i(s)| && (6.3.15)\\
&\le \int_0^t |dM_i(s)| && (6.3.16)\\
&\le \int_0^t |dN_i(s)| + \int_0^t |Y_i(s)\lambda_i(s)ds| && (6.3.17)\\
&\le 2 && (6.3.18)
\end{aligned}$$

which means $U(t)$ is integrable, and by a similiar argument in the proof of theorem 9, we can also show that $\mathbb{E}|U(t)|$ is bounded. Lastly we have

$$\mathbb{E}[dU(t)|\mathscr{F}_{t-}] = \mathbb{E}\left[\sum_{I=1}^{n}(Z_i - \bar{Z}(t))dM_i(t)\middle|\mathscr{F}_{t-}\right] = \sum_{i=1}^{n}(Z_i - \bar{Z}(t)) \cdot \mathbb{E}[dM_i(t)|\mathscr{F}_{t-}] = 0, \tag{6.3.19}$$

and thus $U(t)$ is a martingale. □

Since $U(0) = 0$, so we can conclude that $U(t)$ is a mean-zero martingale for all finite t. When we extend t to ∞, we will use dominated convergence theorem to establish the property for all t.

Theorem 12. *Let $\{X_k\}_{k=1}^{\infty}$ be a sequence of random variables such that $X_k \xrightarrow{a.s} X$ and there exists intergable random variable Y with $|X_k| \leq Y$ almost surely, then X is integrable and $\lim_{k\to\infty}\mathbb{E}[X_k] = \mathbb{E}[X]$.*

The formal proof of dominated convergence theorem relies heavily on the basic theory of Lebesgue integration, we will omit the proof here, but it can be found in Chapter 2.6 of [27]. by applying dominated convergence theorem to $U(t)$, we simply have

$$\mathbb{E}[\mathscr{U}] = \lim_{t\to\infty}\mathbb{E}[U(t)] = 0, \tag{6.3.20}$$

which means, the log-rank statistic has mean zero for all $t \geq 0$. Now we would like to study the variance of $\mathscr{U}(t)$, which is the denominator of the log-rank statistic, and since we already know that $\mathscr{U}(t)$ is a martingale so its variance can be easily computed via variation process. We again first consider finite time $U(t)$ and then apply dominated convergence theorem. By definition, $\langle U\rangle(t)$ is such that $U^2(t) - \langle U\rangle(t)$ is a martingale, and we have

$$\mathbb{E}(dU^2(t)|\mathscr{F}_{t-}) = \mathbb{V}(dU(t)|\mathscr{F}_{t-}) = \mathbb{E}[\langle dU(t)\rangle] = \mathbb{E}\left[\sum_{i=1}^{n}(Z_i - \bar{Z}(t))^2 d\langle M_i\rangle(t)\right] \tag{6.3.21}$$

which means

$$\langle U\rangle(t) = \sum_{i=1}^{n}\int_0^t (Z_i - \bar{Z}(s))^2 d\langle M_i\rangle(t). \tag{6.3.22}$$

We may furthermore simplify $d\langle M_i\rangle(t)$ as $d\langle M_i\rangle(t) = Y_i(t)\lambda_i(t)dt$, and hence the variance of the log-rank statistic can be written as

$$\mathbb{V}(U(t)) = \mathbb{E}\left[\sum_{i=1}^{n}\int_0^t (Z_i - \bar{Z}(s))^2 Y_i(s)\lambda_i(s)ds\right] \tag{6.3.23}$$

where we replace $\lambda_i(s)ds$ by Nelson-Aalen estimator under the null hypothesis $\mathscr{H}_0 : \lambda_1(t) = \lambda_2(t) = \lambda(t)$, and have the following estimate:

$$\mathbb{V}(U(t)) = \sum_{i=1}^{n}\int_0^t (Z_i - \bar{Z}(s))^2 Y_i(s)\frac{dN(s)}{Y(s)} = \sum_{t_j\leq t}\frac{Y_{1j}Y_{2j}}{Y_j^2}D_j. \tag{6.3.24}$$

As we set $t \to \infty$, using dominated convergence theorem will yield

$$\mathbb{V}(\mathscr{U}) = \mathbb{E}\mathscr{U} = \lim_{t\to\infty}\mathbb{E}U^2(t) = \sum_{i=1}^{n}\mathbb{E}\left[\int_0^{\infty}(Z_i - \bar{Z}(s))^2 Y_i(s)\lambda(s)ds\right] \tag{6.3.25}$$

and its estimate is also given by

$$\mathbb{V}(U(t)) = \sum_{i=1}^{n} \int_0^t (Z_i - \bar{Z}(s))^2 Y_i(s) \frac{dN(s)}{Y(s)} = \sum_{t_j \le t} \frac{Y_{1j} Y_{2j}}{Y_j^2} D_j. \tag{6.3.26}$$

So finally the complete log-rank statistic is defined by

$$\mathbf{Z}_{\text{log-rank}} = \frac{U^2(t)}{\mathbb{V}(U(t))} \tag{6.3.27}$$

and using martingale central limit theorem, it can be shown that $\mathbf{Z}_{\text{log-rank}} \sim \chi^2_{(1)}$ when our test involves two samples.

Now we consider the case when a weight function $W(t)$ is given. In this case we have a weighted log-rank test, and now the numerator of the weighted log-rank statistic is simply

$$U_W(t) = \sum_{t_j \le t} W(t_j) \left(D_{1j} - \frac{D_j Y_{1j}}{Y_j} \right), \tag{6.3.28}$$

and its variance is

$$\mathbb{V}U_W(t) = \mathbb{E}\left[\sum_{I=1}^{n} \int_0^{\infty} W^2(s)(Z_i - \bar{Z}(s))^2 Y_i(s) \lambda(s) ds \right] \tag{6.3.29}$$

which can be estimated by

$$\mathbb{V}U_W(t) = \sum_{t_J \le t} \frac{W(t_j)^2 Y_{1j} Y_{2j}}{Y_j^2} D_j. \tag{6.3.30}$$

In multi-sample case, the idea can be generalized, and we have $\mathbf{Z}_{\text{log-rank}} \sim \chi^2_{(k-1)}$ when we have k samples. Now let's consider the case when we have more than 2 groups, in this case we denote Y_{ij} as the number of individuals in group i at risk at time t_j, and D_j be the total number of observed death at t_j. For group i, the numerator of the weighted log-rank statistics is

$$U_i(t) = \sum_{t_j \le t} W(t_j) \left(D_{ij} - \frac{D_j Y_{ij}}{Y_j} \right) \tag{6.3.31}$$

which variance can be estimated by

$$\mathbb{V}U_i(t) = \sum_{t_j \le t} W(t_j)^2 \frac{Y_{ij}}{Y_i} \left(1 - \frac{Y_{ij}}{Y_i} \right) \left(\frac{Y_i - D_i}{Y_i - 1} \right) D_i \tag{6.3.32}$$

and the covariance of $U_k(t)$, $U_h(t)$ $(k \neq h)$, can be estimated by

$$\mathbf{Cov}(U_k(t), U_h(t)) = -\sum_{t_j \le t} W(t_j)^2 \frac{Y_{jk}}{Y_j} \frac{Y_{jh}}{Y_j} \left(\frac{Y_j - D_j}{Y_j - 1} \right) D_i \tag{6.3.33}$$

and the covariance is simply $\mathbb{V}(U_j)$ whenever $j = k$. Since

$$\sum_{i=1}^{k} U_i(t) = \sum_{i=1}^{k} \sum_{t_j \le t} W(t_j) \left(D_{ij} - \frac{D_j Y_{ij}}{Y_j} \right) \tag{6.3.34}$$

$$= \sum_{t_j \le t} W(t_j) \left(D_j - D_j \right) \tag{6.3.35}$$

$$= 0. \tag{6.3.36}$$

So the vector $\boldsymbol{U} = \begin{pmatrix} U_1 & \cdots & U_k \end{pmatrix}$ is linearly dependent. In this case, we construct the test statistic by selecting any $k-1$ of those U_i's, and we denote $\mathbf{V}$ as the corresponding $k-1 \times k-1$ covariance matrix, then the statistic given by

$$\begin{pmatrix} U_1(t) & \cdots & U_{k-1}(t) \end{pmatrix} \mathbf{V}^{-1} \begin{pmatrix} U_1(t) & \cdots & U_{k-1}(t) \end{pmatrix}^T \tag{6.3.37}$$

has a $\chi^2_{(k-1)}$ distribution under the null hypothesis. There are many popular ways to design the weight function, including the regular log-rank test where $W(t) = 1$; the Wilcoxon test where $W(t) = Y(t)$; Tarone-Ware test where $W(s) = \sqrt{Y(s)}$ and Fleming-Harrington test where $W(t) = S(t)^p(1-S(t))^q$. Below is a graph indicating the relative weights for different tests with respect to time:

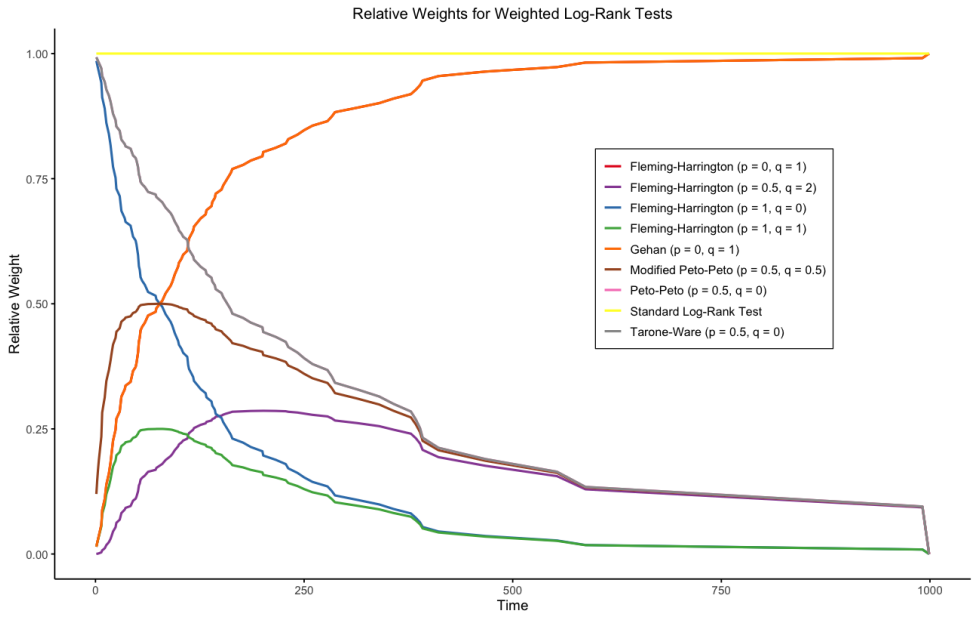


Figure 6.1: The relative weight functions (with different parameters) with respect to time, where sample is generated from `veteran` package in **R**, a sample consists of 137 patients with lung cancer and two groups: standard treatment group (1) and test drug group (2).

We perform a log-rank test (the weight function is set to be 1) on the sample:

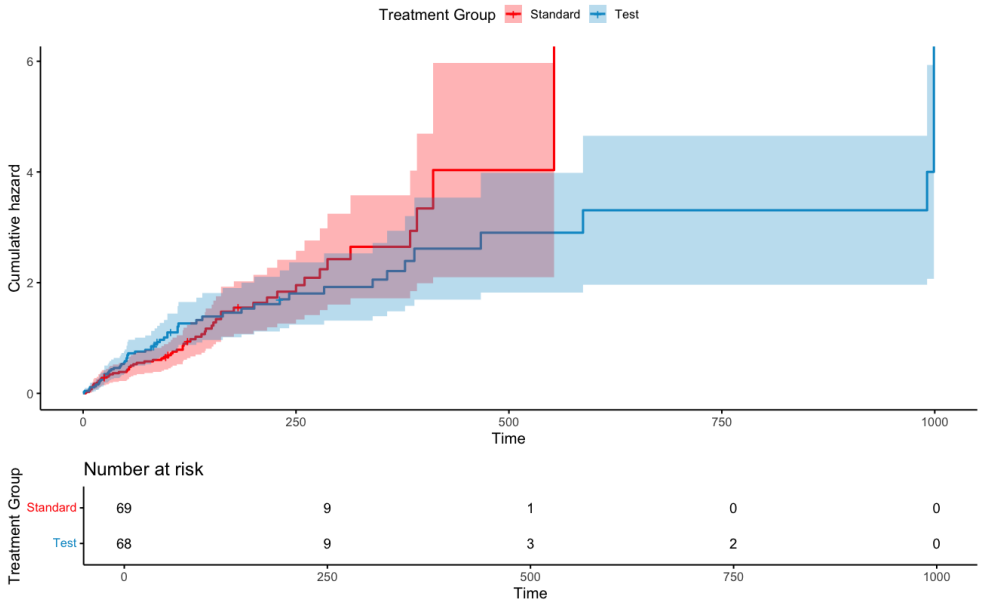


Figure 6.2: We have plotted the Nelson-Aalen estimators as well as the confidence intervals for both groups. The chart below will be the data we use to construct the log-rank test.

Based on our data, the following log-rank test can be constructed:

Group	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
Standard (trt = 1)	69	45	61.1	4.24	9.77
Test (trt = 2)	68	55	38.9	6.70	9.77

Log-Rank Test Statistic: $\chi^2_{(1)} = 9.8$, $p = 0.0017$

Table 6.1: Log-Rank Test Results for Treatment Groups

and we reject the null hypothesis based on our result.

6.4 Some Other Tests

6.4.1 Tests for Trend

In this section, we are interested in an ordered hazard rate of different groups, and we will set the null hypothesis to be $\mathscr{H}_0 : \lambda_1(t) = \cdots = \lambda_k(t)$ for all $t \leq \tau$ against $\mathscr{H}_1 : \lambda_1(t) \leq \lambda_2(t) \cdots \leq \lambda_k(t)$ for $t \leq \tau$. It is very useful when we shall determine which treatment is the best among potential treatments and we are able to derive an ordered hazard rate so we can see the effectiveness of each treatment. We shall adapt the notion from the previous chapter, and denote $\mathbf{V}$ as the full $k \times k$ covariance matrix this time. Suppose we have a sequence of increasing scores $a_1 < \cdots < a_k$, then we compute the following statistic:

$$Z = \frac{\sum_{j=1}^{k} a_j U_j(t)}{\sqrt{\sum_{j=1}^{k}\sum_{g=1}^{k} a_j a_g \mathbf{Cov}(U_j(t), U_g(t))}} \tag{6.4.1}$$

which asymptotically has a standard normal distriution under the null hypothesis. The test statistic Z arises from the martingale central limit theorem. Under $\mathscr{H}_0$ U_j are all mean-zero martingales and we have

$$\frac{\sum_{j=1}^{k} a_j U_j(t) - \mathbb{E}\left[\sum_{j=1}^{k} a_j U_j(t)\right]}{\sqrt{\mathbf{Var}\left(\sum_{j=1}^{k} a_j U_j(t)\right)}} \xrightarrow{d} N(0,1). \tag{6.4.2}$$

Here we examine a study of 90 patients with larynx cancer (Moeschberger and Klein, 2022, Section 1.8). Patients were followed after the first treatment until their death or the end of study. The cancer had 4 different stages based on primary tumor, nodal involvement and distant metastasis grading. For those four stages (Stage I,II,III,VI), we assign scores $a_j = j, j = 1,2,3,4$ and we wish to test the alternative $\mathscr{H}_1 : \lambda_1(t) < \lambda_2(t) < \lambda_3(t) < \lambda_4(t)$ or equivalently $S_1(t) > S_2(t) > S_3(t) > S_4(t)$. The data set is available in the appendix, and after some computations the test statistics is given by $Z = 3.72$ with p-value less than 0.0001, which rejects the null hypothesis and we are in favour of $\mathscr{H}_1$. The survival curve based on Kaplan-Meier estimator is shown in the figure below for better illustration:

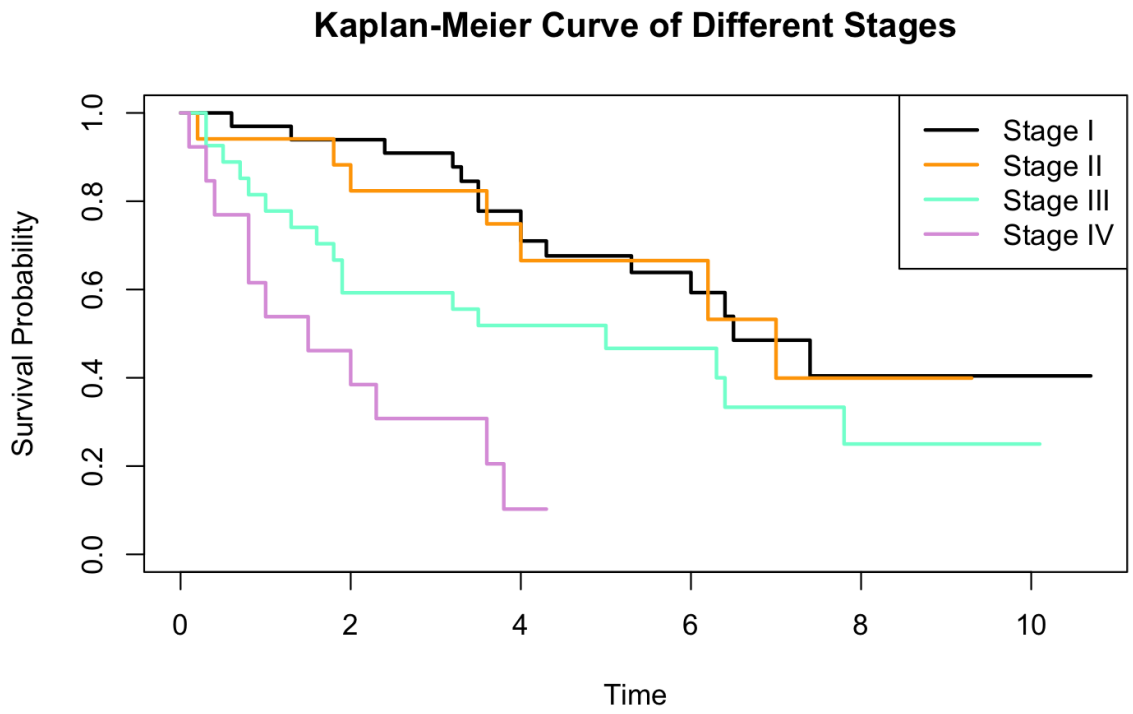


Figure 6.3: The Kaplan-Meier survival curves by different cancer stage

6.4.2 Renyi Type Tests

Renyi-Type tests works well if we would like to detect the crossing of hazards in different groups. In a two-sample example, suppose we wish to test $\mathscr{H}_0 : \lambda_1(t) = \lambda_2(t)$ versus $\mathscr{H}_1 : \lambda(t) \neq \lambda_2(t)$ where $t \leq \tau$, as we have previously shown that, denote $t_1 < t_2 < \cdots < t_D \leq \tau$ as the observed event times, Y_{ij} (D_{ij}) to

be the number of individuals at risk (death) at t_i in group $j = \{1,2\}$, $Y_i = Y_{i1} + Y_{i2}, D_i = D_{i1} + D_{i2}$. Let $W(t)$ as the weight function chosen and we showed that the numerator of the weighted log-rank statistic is given by

$$Z(t) = \sum_{t_j \leq t} W(t_j) \left[D_{j1} - Y_{j1} \left(\frac{D_j}{Y_j} \right) \right] \tag{6.4.3}$$

and the variance of $Z(t)$ is given by

$$\mathbb{V}(Z(t)) = \sum_{t_j \leq t} W^2(t_j) \left(\frac{Y_{j1}}{Y_j} \right) \left(\frac{Y_{j2}}{Y_j} \right) \left(\frac{Y_j - D_j}{Y_j - 1} \right) D_j \tag{6.4.4}$$

the entire weight log-rank statistic has an asymptotically standard normal distribution:

$$\frac{Z(\tau)}{\sqrt{\mathbb{V}(Z(\tau))}} \sim N(0,1), \tag{6.4.5}$$

or its square has an asymptotically $\chi^2_{(1)}$ distribution. When the hazard rates in two group cross each other, then in the numerator $Z(t)$ the early possitive differences between the two hazard rates are canceled out by later differences in the rates with opposite signs (Klein and Moeschberger, Chapter 7.6, 2003) so we may result in a smaller statistic and then might falsely accept $\mathscr{H}_0$ (Type II error). Hence an alternative test statistic proposed by Gill (1980) is given by

$$Q = \frac{\sup\{|Z(t)|, t \leq \tau\}}{\sqrt{\mathbb{V}(Z(\tau))}}, \tag{6.4.6}$$

under the null hypothesis, the distribution of Q can be approximated by the distribution of $\sup(|B(x)|, x \in [0,1])$ where $B(x)$ is the standard Brownian motion process. The figure is a simulated path of $B(x)$:

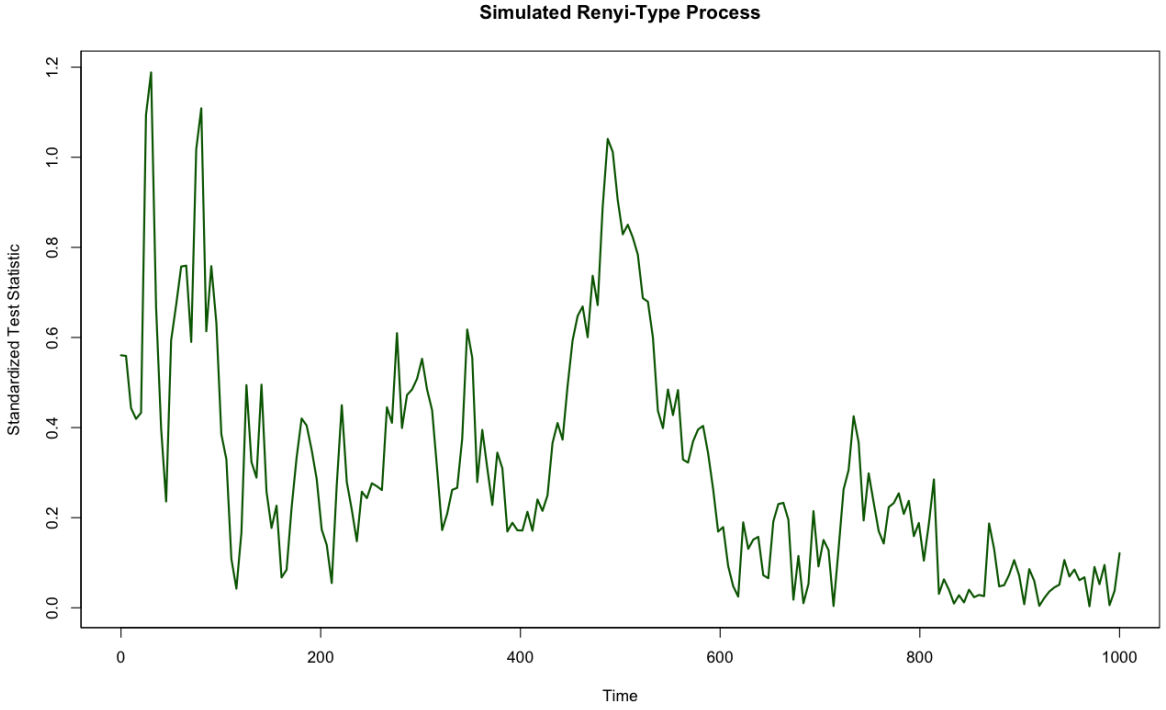


Figure 6.4: The simulated standard Brownian motion $B(t)$

Billingsly (1968) showed that for a standard Brownian motion $B(t)$, we have

$$\mathbb{P}(\sup|B(t)| > y) = 1 - \frac{4}{\pi}\sum_{k=0}^{\infty}\frac{(-1)^k}{2k+1}\exp\left(-\frac{\pi^2(2k+1)^2}{8y^2}\right) \tag{6.4.7}$$

Here we examine a study of a clinical trial of chemotherapy against chemotherapy combined with radiotherapy in the treatment of locally unresectable gastric cancer. (Klein and Moeschberger, Chapter 7.6, 2003). The original study was conducted by the Gastrointestinal Tumor Study Group (1982). In this trial, forty-five patients were randomized to each of the two arms and followed for about eight years. The data obtained can be found in the appendix. We will test the hypothesis that the survival rate of the two groups is the same using weight function $W(t) = 1$. Under this, the Reyni-type test statistic is given by $Q = 2.20$, by the definition, it has a p-value

$$\mathbb{P}(\sup|B(t)| > 2.20) \approx 0.053 \tag{6.4.8}$$

which concludes that we do not reject $\mathscr{H}_0$ when we take $\alpha = 0.05$ as the significance level. Below we plotted the survival curve for two groups:

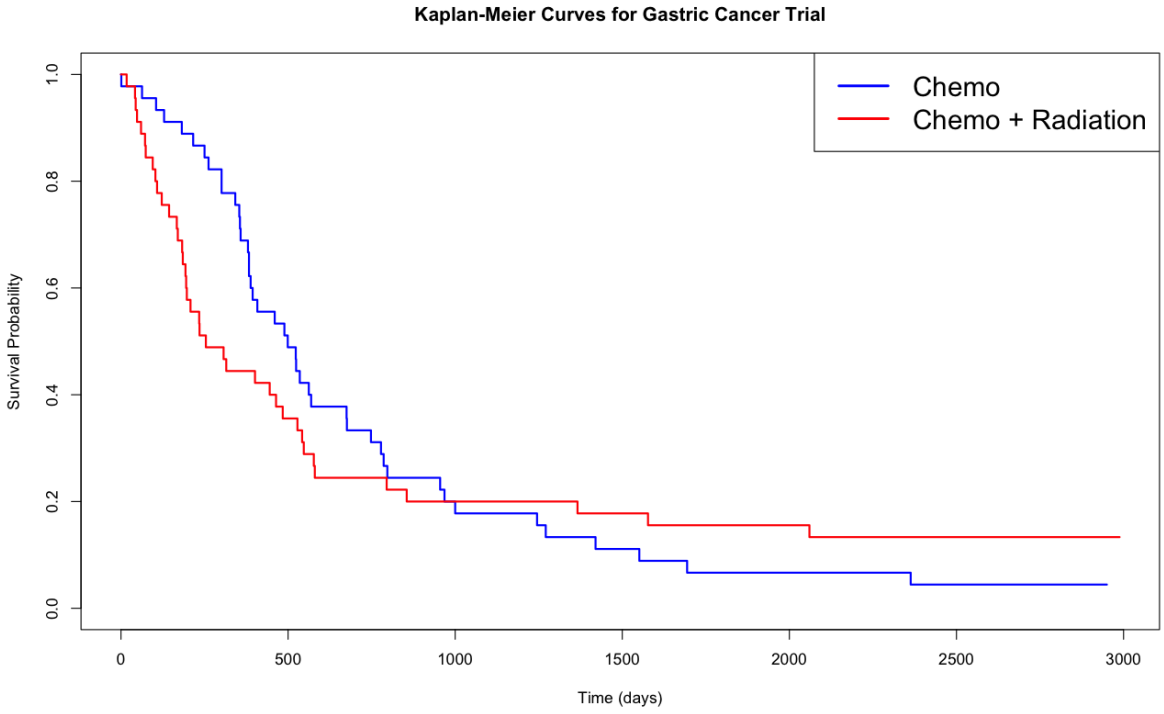


Figure 6.5: The Kaplan-Meier survival curves by different cancer treatment

6.4.3 Cramer-Von Mises Type Tests

Cremer-von Mises test is based on the integral of squared difference between two empirical survival functions. With potential right-censoring, we will then design a test based on the integral of the squared difference between the two estimated cumulative hazard functions using Nelson-Aalen estimator (Klein and Moeschberger, Chapter 7.7, 2003). We recall cumulative hazard function in group $i = \{1,2\}$ is estimated by

$$\widehat{\Lambda}_i(t) = \sum_{t_j \le t}\frac{d_{ij}}{Y_{ij}} \tag{6.4.9}$$

which variance is estimated by

$$\mathbb{V}(\widehat{\Lambda}_i(t)) = \sigma_i^2(t) = \sum_{t_j \le t} \frac{d_{ij}}{Y_{ij}(Y_{ij}-1)}. \tag{6.4.10}$$

and so the estimated variance of $\widehat{\Lambda}_1(t) - \widehat{\Lambda}_2(t)$ is $\sigma^2(t) = \sigma_1^2(t) + \sigma_2^2(t)$. The Cramer-von Mises test statistic is given by

$$Q = \left(\frac{1}{\sigma^2(\tau)}\right)^2 \cdot \int_0^\tau (\Lambda_1(t) - \Lambda_2(t))^2 d\sigma^2(t) \tag{6.4.11}$$

which can be estimated by

$$Q = \left(\frac{1}{\sigma^2(\tau)}\right)^2 \cdot \sum_{t_j \le \tau} (\widehat{\Lambda}_1(t_j) - \widehat{\Lambda}_2(t_j))^2 (\sigma^2(t_j) - \sigma^2(t_{j-1})) \tag{6.4.12}$$

It can be shown that asymptotically the distribution of Q is

$$Q \sim \int_0^1 (B(x))^2 dx \tag{6.4.13}$$

where $B(x)$ is a standard Brownian motion process.

Chapter 7

Cox's Semi-parametric Regression Models

7.1 Introduction

In this chapter, we would like to discuss different regression models for time to event survival data. Recall in chapter 2 we discussed several likelihood inference for parametric family, where we define $T_i = \min\{X_i, C_i\}$ as the observed time, where C_i is the censoring time and X_i is the true event time. Assume X_i forms a certain distribution $X_i \sim f(x,\theta)$, we derived in chapter 2 that the likelihood function in this case is

$$L(\theta) = K \prod_{\text{censored}} S(X_i) \cdot \prod_{\text{Uncensored}} f(X_i) \tag{7.1.1}$$

where K is a constant. Then we know the log-likelihood is given by

$$\ell(\theta) \propto \sum_{\text{Censored}} \log S(X_i) + \sum_{\text{Uncensored}} \log f(X_i). \tag{7.1.2}$$

For example, if we assume $X_i \sim Exponential(\theta)$, then we have $f(x) = \theta e^{-\theta x}$, $S(x) = 1 - F(x) = e^{-\theta x}$ and we then have

$$\ell(\theta) \propto \sum_{\text{censored}} \Big(-\theta x_i \Big) + \sum_{\text{uncensored}} \Big(\log(\theta) - \theta x_i \Big). \tag{7.1.3}$$

Then the partial derivative is given by

$$\frac{\partial \ell(\theta)}{\partial \theta} = - \sum_{\text{Censored}} x_i + \sum_{\text{Uncensored}} \Big(\frac{1}{\theta} - x_i \Big) \tag{7.1.4}$$

Hence we get the MLE of θ given by

$$\widehat{\theta}_{\text{MLE}} = \frac{U}{\sum_{i=1}^{n} X_i} \tag{7.1.5}$$

where U denotes the number of uncensored observation.

Hence from this example we see that it is straightforward to model survival data when we assumed we have a parametric inference with single parameter. However in reality, we will usually have more information on each individual, also known as the covariates. For example in the study of lung cancer, we not only have the observed time $T_i = \min\{X_i, C_i\}$ for each individual, we may also have the covariates

such as that person's smoking habits, gender, age, etc. If we denote all covariates as a column vector $\boldsymbol{X}$, then the survival function is the conditional probability given $\boldsymbol{X}$, denote as

$$S(t|\boldsymbol{x}) = \mathbb{P}\{T > t|\boldsymbol{X} = \boldsymbol{x}\}. \tag{7.1.6}$$

So the purpose of this chapter is to introduce some of the methods to model survival data with different covariates.

7.2 Proportional Hazards Model and Time Independent Fixed Covariates

If we consider possible time-independent covariates row vector $\boldsymbol{Z} = (Z_1, Z_2, \cdots, Z_k)$ that might effect the hazard rate, Cox (1972) introduced a semi-parametric model for the hazard function:

$$\lambda(t|\boldsymbol{Z}) = \lambda_0(t)c(\boldsymbol{\beta}^T\boldsymbol{Z}) \tag{7.2.1}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_k)$ is a parameter row vector which is of our interest. $\lambda_0(t)$ is the baseline hazard. The vector product $\boldsymbol{\beta}^T\boldsymbol{Z} = \beta_1 Z_1 + \cdots + \beta_k Z_k$ describes a linear combination of different covariates and it is a known function with chosen parameters. Since we do not make any assumption on $\lambda_0(t)$, it is a semi-parametric approach to survival data. One popular model is by taking the exponential

$$\lambda(t|\boldsymbol{Z}) = \lambda_0(t)\exp(\boldsymbol{\beta}^T\boldsymbol{Z}). \tag{7.2.2}$$

If we do not assume anything on covariates, by making $\boldsymbol{Z} = \boldsymbol{0}$, we see that the Cox model simply becomes $\lambda_0(t)$ and this is the (base) hazard rate we were estimating in the previous two chapters. One advantage of Cox's model is that if we choose two different covariates $\boldsymbol{Z}_1, \boldsymbol{Z}_2$, the ratio of the two hazard rate is a constant, given by

$$\frac{\lambda(t|\boldsymbol{Z}_1)}{\lambda(t|\boldsymbol{Z}_2)} = \exp(\boldsymbol{\beta}^T(\boldsymbol{Z}_1 - \boldsymbol{Z}_2)). \tag{7.2.3}$$

Which means in Cox's model the exponential part is independent of time and thus remains a constant under the ratio. We may use the same notion to derive the survival function. Recall the relation that $S(t) = e^{-\Lambda(t)}$ where $\Lambda(t)$ is the cumulative hazard function, we have

$$S(t|\boldsymbol{Z}) = e^{-\Lambda(t)} = \exp\Big(-\Lambda_0(t)\exp(\boldsymbol{\beta}^T\boldsymbol{Z})\Big). \tag{7.2.4}$$

One way to deal with covariates is by assigning different values to different gtoups, in a study of 863 kidney transplant patients ([Klein, Moeschberger]), two different covariates were introduced: The patients' gender (male and female) and race (white and black). Then we have 4 total different covariates, namely black male; black female; white male; white female. Then a covariate vector $\boldsymbol{Z} = (Z_1, Z_2, Z_3, Z_4)$ can be used to describe the above feature, by assigning 1 if it matches the group and 0 otherwise. Under this case a black female would have the covariate $\boldsymbol{Z} = (0,1,0,0)$. The drawback is that we still de not have access to the baseline hazard. Another way is to define a reference group where it represents the baseline hazard. So we may set $Z_1 = 1$ if we have a black male and 0 otherwise; $Z_2 = 1$ if we have a white male and 0 otherwise; $Z_3 = 1$ if we have a black female and 0 otherwise. Then white female is considered as the reference group and its covariate vector is simply $\boldsymbol{Z} = \boldsymbol{0}$ and its hazard rate in Cox model is equal to the base hazard $\lambda_0(t)$, and we have

$$\lambda(t) = \lambda_0(t)\exp\Big(\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3\Big). \tag{7.2.5}$$

We may solve the parameters via numerical methods introduced in the next section, it turns out that we have $\beta_1 = 0.1596, \beta_2 = 0.2484, \beta_3 = 0.6567$. Since the ratio of two hazard in Cox's model is independent of time, so the advantage of defining a reference group is that we can take the ratio of it with non-reference groups and see that relative hazard rate, and we can see the performace compared to the reference group. For example we have

$$\frac{\lambda(t|Z_1)}{\lambda_0(t)} = 1.17, \frac{\lambda(t|Z_2)}{\lambda_0(t)} = 1.28, \frac{\lambda(t|Z_3)}{\lambda_0(t)} = 1.93. \tag{7.2.6}$$

Thus the relative hazard risks for black male, white male and black female compared to white female are given by $1.17, 1.28, 1.93$ respectively. One can also use this method to test if a certain treatment is effictive, by defining the group with placebo as the reference group and see the relative hazard risk for treatment group. If the hazard rate is greater than 1, it means the treatment is somehow not effective, if less than 1 it means the treatment is somehow effective.

7.3 Partial Likelihoods and Numerical Estimates for Distinct Time Event

We will assume no ties exist in this section. In our sample of size n, we have the triple written as $(T_i, \delta_i, \boldsymbol{Z}_i)$ where we denote $\boldsymbol{Z}_i$ as the covariate for ith individual. Suppose we have observed the distinct ordered event time $t_1 < \cdots < t_m$, and we are interested in the death at each t_i. Among all individuals still at risk, we like to investigate the proability that the death at t_i is exactly due to the cause of covariate $\boldsymbol{Z}_i$. If we define $R(t_i)$ as the set of individuals at risk prior to t_i, then the conditional probability of one death at t_i due to $\boldsymbol{Z}_i$ given that one death (with all possible covariates) is observed is now given by

$$\mathbb{P}(\text{death at } t_i \text{ due to } \boldsymbol{Z}_i | \text{one observed death at } t_i) = \frac{\exp\left(\boldsymbol{\beta}^T \boldsymbol{Z}_i\right)}{\sum_{j \in R(t_i)} \exp\left(\boldsymbol{\beta}^T \boldsymbol{Z}_j\right)} \tag{7.3.1}$$

Assuming independence, we are interested in the joint event of death at t_i due to $\boldsymbol{Z}_i$ for all observation. By multiplying corresponding conditional probability over all observed deaths, we get the Cox-partial likelihood function:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{\exp\left(\boldsymbol{\beta}^T \boldsymbol{Z}_i\right)}{\sum_{j \in R(t_i)} \exp\left(\boldsymbol{\beta}^T \boldsymbol{Z}_j\right)} \tag{7.3.2}$$

which can also be viewed as the probability of a specific partition of those time events. We may also derive the Cox partial likelihood as follows:

We first consider the likelihood with potential censoring, in chapter 2 we derived the likelihood function of the following form:

$$L(\theta) = \prod_{i=1}^{n} f(t_i|\theta)^{\delta_i} S(t_i|\theta)^{1-\delta_i} \tag{7.3.3}$$

where n is the total number of individuals and δ_i is the indicator whether the individual is censored or not. Using the fact that $f(t) = \lambda(t)S(t)$, and we replace our parameter θ by the parameter vector $\boldsymbol{\beta}$, and

conditioning on the covariate $\boldsymbol{Z}_i$ at each time t_i, we have

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \lambda(t_i|\boldsymbol{Z}_i)^{\delta_i} S(t_i|\boldsymbol{Z}_i) \tag{7.3.4}$$

then we use Cox's likelihood formula and we have

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \lambda_0(t_i)^{\delta_i} \Big[\exp\Big(\boldsymbol{\beta}^T \boldsymbol{Z}_i\Big)\Big]^{\delta_i} \exp\Big(-\Lambda_0(t_i)\exp(\boldsymbol{\beta}^T \boldsymbol{Z}_i)\Big). \tag{7.3.5}$$

At this time we fix $\boldsymbol{\beta}$ first and consider to maximize the likelihood as a function of the baseline hazard $\lambda_0(t)$, and we define a new subset $j = 1, \cdots, D$ as all the uncensored observations (distinct observed events), we have

$$L(\lambda_0(t)) = \left[\prod_{j=1}^{D} \lambda_0(t_j)\exp\Big(\boldsymbol{\beta}^T \boldsymbol{Z}_j\Big)\right] \cdot \exp\left\{-\sum_{i=1}^{n} \Lambda_0(t_i)\exp\Big(\boldsymbol{\beta}^T \boldsymbol{Z}_i\Big)\right\} \tag{7.3.6}$$

and the log-likelihood takes the form proportional to

$$\ell(\lambda_0(t)) \propto \sum_{j=1}^{D} \left(\log\lambda_0(t_j) - \int_0^{t_j} \lambda_0(s)ds \cdot \exp\Big(\boldsymbol{\beta}^T \boldsymbol{Z}_j\Big)\right) - \sum_{t_i \neq t_j} \int_0^{t_i} \lambda_0(s)ds \cdot \exp\Big(\boldsymbol{\beta}^T \boldsymbol{Z}_i\Big). \tag{7.3.7}$$

Since the integral of the hazard function is always positive, so in order to maximize $\ell(\lambda_0(t))$ we can make $\lambda_0(t) = 0$ except for t_js and this will eliminate the integral. We use $\lambda_0(t_j)$ as the hazard rate at those discrete observed time (also the MLE), and we substitute into the likelihood function, we will have the multivariable function given by

$$L(\lambda_0(t_1), \cdots, \lambda_0(t_j)) \propto \prod_{j=1}^{D} \lambda_0(t_j)\exp\left[-\lambda_0(t_j) \cdot \sum_{j \in R(t_j)} \exp\Big(\boldsymbol{\beta}^T \boldsymbol{Z}_j\Big)\right], \tag{7.3.8}$$

so for each $\lambda_0(t_j)$ its MLE is given by

$$\widehat{\lambda_0}(t_j) = \frac{1}{\sum_{k \in R(t_j)} \exp\Big(\boldsymbol{\beta}^T \boldsymbol{Z}_k\Big)} \tag{7.3.9}$$

and we sum over all observed time t_j gives the Breslow's estimator for the baseline cumulative hazard rate:

$$\widehat{\Lambda}(t) = \sum_{t_j \leq t} \frac{1}{\sum_{k \in R(t_j)} \exp\Big(\boldsymbol{\beta}^T \boldsymbol{Z}_k\Big)}. \tag{7.3.10}$$

Now the Cox's partial likelihood function can be obtained by replacing the hazard rate by its MLE in the likelihood function above.

We now provide several ways to estimate the parameter. First the log-partial likelihood function is now given by

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{m} \left[\boldsymbol{\beta}^T \boldsymbol{Z}_i - \log\left(\sum_{j \in R(t_i)} \exp\Big(\boldsymbol{\beta}^T \boldsymbol{Z}_j\Big)\right)\right]. \tag{7.3.11}$$

In order to derive the MLE of $\boldsymbol{\beta}$, we first note that $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_k)$, we first derive the partial derivative of $\ell(\boldsymbol{\beta})$ with respect to parameter $\beta_h, h = 1, \cdots, k$, and let Z_{ih} as the hth covariate of the ith individual, we have

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_h} = \sum_{i=1}^{m} \left[Z_{ih} - \frac{\sum_{j \in R(t_i)} Z_{jh} \cdot \exp\left(\boldsymbol{\beta}^T \boldsymbol{Z}_j\right)}{\sum_{j \in R(t_i)} \exp\left(\boldsymbol{\beta}^T \boldsymbol{Z}_j\right)} \right]. \tag{7.3.12}$$

The second derivative is given by

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_h^2} = -\sum_{i=1}^{m} \left[\frac{\sum_{j \in R(t_i)} Z_{jh}^2 \cdot \exp(\boldsymbol{\beta}^T \boldsymbol{Z}_j)}{\sum_{j \in R(t_i)} \exp(\boldsymbol{\beta}^t \boldsymbol{Z}_j)} - \left(\frac{\sum_{j \in R(t_i)} Z_{jh} \exp(\boldsymbol{\beta}^T \boldsymbol{Z}_j)}{\sum_{j \in R(t_i)} \exp(\boldsymbol{\beta}^T \boldsymbol{Z}_j)} \right)^2 \right]. \tag{7.3.13}$$

Then the partial maximum likelihood estimates (PMLE) can be found by solving the system of equations

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_h} = 0, h = 1, \cdots, k. \tag{7.3.14}$$

and to ensure the estimate $\widehat{\boldsymbol{\beta}}$ we have is indeed the PMLE, the Hessian matrix $\boldsymbol{I}$ defined by

$$\boldsymbol{I}_{gh} = \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_g \beta_h} \tag{7.3.15}$$

should be positive definite. Several methods can be used to maximize the PMLE, Klein and Moeschberger [1] proposed three different numerical methods: method of steepest ascent; multi-variate Newton-Raphson method and Marquardt's method.

Method of Steepest Ascent: This method is also known as Gradient method. The idea is to first pick our initial guess $\boldsymbol{x}_0$ and set it to be the maximum (an educated guess), then we compute the gradient at this point. We then move in the direction of the gradient at this point by a distance α_0, and then pick $\boldsymbol{x}_0 + \alpha_0 \nabla f(\boldsymbol{x}_0)$ as the updated maximum $\boldsymbol{x}_1$, and then we apply this procedure recursively until we have reached some accurancy level. Now our task would be choose the best α_k for all steps. If we consider this procedure as a map (Robert Foote, 1996) [2] T and define $T(\boldsymbol{x}_k) = \boldsymbol{x}_k + \alpha(\boldsymbol{x}_k) \nabla f(\boldsymbol{x}_k)$, then we have $\boldsymbol{x}_{k+1} = T(\boldsymbol{x}_k)$, and if $f(\widetilde{\boldsymbol{x}})$ is the maximum we will have $T(\widetilde{\boldsymbol{x}}) = \widetilde{\boldsymbol{x}}$ which is a fixed point. At point $\boldsymbol{x}_k$, we now assume $\boldsymbol{x}_k$ is constant and we are moving in the direction of $\nabla f(\boldsymbol{x}_k)$, so the line function is given by

$$g(t_k) = f(\boldsymbol{x}_k + t_k \nabla f(\boldsymbol{x}_k)). \tag{7.3.16}$$

Now we only have a function with one variable t, and to find its maximum, we perform a second order Taylor expansion around $t_k = 0$

$$g(t) = g(0) + t_k g'(0) + \frac{1}{2} t_k^2 g''(0), \tag{7.3.17}$$

then by computing $g'(t_k)$ and solve for $g'(t_k) = 0, g''(t_k) < 0$, we have the maximizer t_k^* given by

$$t_k^* = -\frac{g'(0)}{g''(0)}. \tag{7.3.18}$$

Hence we let

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + t_k^* \nabla f(\boldsymbol{x}_k) \tag{7.3.19}$$

as the updated maximizer (eduated guess) of the function and we repeat this process. Note that using Chain rule we may compute

$$g'(0) = \frac{d}{dt}\bigg|_{t_k=0} f\Big(\boldsymbol{x}_k + t_k \nabla f(\boldsymbol{x}_k)\Big) = \nabla f(\boldsymbol{x}_k)^2 = ||\nabla f(\boldsymbol{x}_k)||^2 \tag{7.3.20}$$

and

$$g''(0) = \frac{d^2}{dt^2}\bigg|_{t_k=0} f\Big(\boldsymbol{x}_k + t_k \nabla f(\boldsymbol{x}_k)\Big) = \nabla f(\boldsymbol{x}_k) \cdot (\boldsymbol{H}(\boldsymbol{x}_k)\nabla f(\boldsymbol{x}_k)) \tag{7.3.21}$$

where $\boldsymbol{H}$ is the Hessian matrix. So the overall procedure is

$$T(\boldsymbol{x}_k) = \boldsymbol{x}_k - \frac{||\nabla f(\boldsymbol{x}_k)||^2}{\nabla f(\boldsymbol{x}_k) \cdot (\boldsymbol{H}(\boldsymbol{x}_k)\nabla f(\boldsymbol{x}_k))} \cdot \nabla f(\boldsymbol{x}_k). \tag{7.3.22}$$

Multi-variate Newton-Raphson Method: Similiar to the previous method, we will first start with an educated guess of the maximum $f(\boldsymbol{x}_0)$ and then it recursively apply the following method:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \boldsymbol{H}(\boldsymbol{x}_k)^{-1} \nabla f(\boldsymbol{x}_k) \tag{7.3.23}$$

where $\boldsymbol{H}(\boldsymbol{x}_k)$ is the Hessian matrix.

Marquardt's Method: This method comprimise between the method of steepest ascent and Newton-Raphson method. where we introduce a constant γ and an initial guess $\boldsymbol{x}_0$ of where the maximum is attained. Then it recursively applies the following method:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \boldsymbol{S}_k \Big(\boldsymbol{S}_k \boldsymbol{H}(\boldsymbol{x}_k) \boldsymbol{S}_k + \gamma \boldsymbol{I}\Big)^{-1} \boldsymbol{S}_k \nabla f(\boldsymbol{x}_k) \tag{7.3.24}$$

where $\boldsymbol{H}$ is the Hessian matrix and $\boldsymbol{S}_k = \text{diag}\Big(\Big|\boldsymbol{H}_{11}(\boldsymbol{x}_k)\Big|^{-1/2}, \cdots, \Big|\boldsymbol{H}_{kk}(\boldsymbol{x}_k)\Big|^{-1/2}\Big)$, and $\boldsymbol{I}$ is the $k \times k$ identity matrix. We can see that when $\gamma \to \infty$ it is the method of steepest ascent and when $\gamma \to 0$ it is Newton-Raphson's method.

Several features can be used to decide whether we have obtained a good estimate using the methods above, one is to check the source of convergence, including $|f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k)| < \varepsilon$ for pre-determined ε.

Below is an example to illustrate the numerial estimation using three methods presented for a group of uncensored observations. The example is taken from Klein and Moeschberger, Example A.1, A.2. In a Weibull model with two parameters α and β, and the hazard function takes the form $\lambda(t) = \beta t^{\beta-1} e^{-\alpha t^\beta}$ we have observed the following uncensored observation:

$$\text{Data: } 2.57, 0.58, 0.82, 1.02, 0.78, 0.46, 1.04, 0.43, 0.69, 1.37 \tag{7.3.25}$$

The table below shows the numeical estimates of α, β with 5 iterations using three different methods:

Method	$\widehat{\alpha}$ after 5 iterations	$\widehat{\beta}$ after 5 iterations
Method of Steepest Ascent	0.839	1.792
Newton-Raphson	0.832	1.796
Marquardt	0.845	1.777

7.4 Partial Likelihoods with Tied Events

Often, due to the way times are recorded, ties between event times are found in the data (Klein and Moeschberger, chpter 8.4), so in this chapter we investigate several ways to compute the partial likelihoods with tied events. We define $t_1 < \cdots < t_D$ be the distinct ordered event times. Let d_i be the number of ties (observed deaths) at t_i, denote $\mathscr{D}_i$ as the set of individuals who die at time i, $\mathscr{R}_i$ be the total number of individuals at risk prior to t_i. Denote $\boldsymbol{S}_i$ as the sum of all covariates of individuals who die at t_i, ie $\boldsymbol{S}_i = \sum_{j\in\mathscr{D}_i} \boldsymbol{Z}_j$.

In this section we introduce three methods to compute partial likelihoods when ties are present and we will all assume the true hazard rate satisfies Cox's proportional model. We first consider a method proposed by Cox (1972), the likelihood is the product of all conditional probabilities of the specific d_i individual $I_{i1}, \cdots, I_{id_i}$ who fail at t_i given d_i individual fail at t_i from $\mathscr{R}_i$, which yields

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{D} \mathbb{P}\Big(I_{i1}, \cdots, I_{id_i} \text{ fail at } t_i \Big| \ d_i \text{ individuals fail at } t_i \text{ from } \mathscr{R}_i \Big). \tag{7.4.1}$$

If we denote $S(i,d_i)$ as the set of all set of possible d_i individuals that may fail at t_i, then by using the law of total probability we have

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\mathbb{P}\Big(I_{i1}, \cdots, I_{id_i} \text{ fail at } t_i \Big| \ d_i \text{ individuals fail at } t_i \text{ from } \mathscr{R}_i \Big)}{\sum_{\ell\in S(i,d_i)} \mathbb{P}\Big(\ell_{i1}, \cdots, \ell_{id_i} \text{ fail at } t_i \Big| \ d_i \text{ individuals fail at } t_i \text{ from } \mathscr{R}_i \Big)} \tag{7.4.2}$$

which simplifies to

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp\Big(\boldsymbol{\beta}^T \boldsymbol{Z}_{I_{i1}}\Big) \cdots \exp\Big(\boldsymbol{\beta}^T \boldsymbol{Z}_{I_{id_i}}\Big)}{\sum_{\ell\in S(i,d_i)} \exp\Big(\boldsymbol{\beta}^T \boldsymbol{Z}_{\ell_{i1}}\Big) \cdots \exp\Big(\boldsymbol{\beta}^T \boldsymbol{Z}_{\ell_{id_i}}\Big)} \tag{7.4.3}$$

using the notation introduced above, and let $\boldsymbol{S}_{i\ell}$ be the sum of covariates of individuals in ℓ in $S(i,d_i)$, Cox's partial likelihood with ties finally becomes

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp\Big(\boldsymbol{\beta}^T \boldsymbol{S}_i\Big)}{\sum_{\ell\in S(i,d_i)} \exp\Big(\boldsymbol{\beta}^T \boldsymbol{S}_{i\ell}\Big)}. \tag{7.4.4}$$

Breslow's (1974) method is almost identical to Cox's partial likelihood, Breslow suggested replacing the denominator by the one in original Cox partial likelihood but raise to the power by d_i, which gives the following formula:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp\Big(\boldsymbol{\beta}^T \boldsymbol{S}_i\Big)}{\Big[\sum_{j\in\mathscr{R}_i} \exp\Big(\boldsymbol{\beta}^T \boldsymbol{Z}_j\Big)\Big]^{d_i}}. \tag{7.4.5}$$

Efron (1977) suggests a partial likelihood of the form

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{D} \left[\frac{\exp\Big(\boldsymbol{\beta}^T \boldsymbol{S}_i\Big)}{\prod_{j=1}^{d_i} \left(\sum_{k\in\mathscr{R}_i} \exp\Big(\boldsymbol{\beta}^T \boldsymbol{Z}_k\Big) - \frac{j-1}{d_i} \sum_{k\in\mathscr{D}_i} \exp\Big(\boldsymbol{\beta}^T \boldsymbol{Z}_k\Big) \right)} \right]. \tag{7.4.6}$$

All three methods would work quite well if the number of ties is small. A simple example can be used to illustrate this: Suppose we have a group of 4 individuals $\{1,2,3,4\}$ and 1,2 fail at time t, and let $\boldsymbol{Z}_i$ denote the covariates for i, then the partial likelihood by using three methods are:

Method	Estimate
Breslow's Method	$L(\boldsymbol{\beta}) = \dfrac{\exp\left(\boldsymbol{\beta}^T(\boldsymbol{Z}_1+\boldsymbol{Z}_2)\right)}{\left[\exp\left(\boldsymbol{\beta}^T\boldsymbol{Z}_1\right)+\exp\left(\boldsymbol{\beta}^T\boldsymbol{Z}_2\right)+\exp\left(\boldsymbol{\beta}^T\boldsymbol{Z}_3\right)+\exp\left(\boldsymbol{\beta}^T\boldsymbol{Z}_4\right)\right]^2}$
Cox's Method	$L(\boldsymbol{\beta}) = \dfrac{\exp\left(\boldsymbol{\beta}^T(\boldsymbol{Z}_1+\boldsymbol{Z}_2)\right)}{\sum_{i>j,i,j\in\{1,2,3,4\}}\exp\left(\boldsymbol{\beta}^T(\boldsymbol{Z}_i+\boldsymbol{Z}_j)\right)}$
Efron's Method	$L(\boldsymbol{\beta}) = \dfrac{\exp\left(\boldsymbol{\beta}^T(\boldsymbol{Z}_1+\boldsymbol{Z}_2)\right)}{\prod_{j=1}^2\left[\sum_{k=1}^4\exp\left(\boldsymbol{\beta}^T\boldsymbol{Z}_k\right)-\frac{j-1}{2}\sum_{k=1}^2\exp\left(\boldsymbol{\beta}^T\boldsymbol{Z}_k\right)\right]}$

7.5 Testing the Covariates

In the previous two sections, we introduced different methods to estimate the parameters $\boldsymbol{\beta}$, after the estimation one is then interested in testing the accuracy of them. Since we have a semi-parameter inference, and the covariates form a known distribution up to some unknown parameters, we may use the testing methods in parametric inference introduced in chapter 2. In this section we shall use hypothesis testing to test some of the parameters β. We first will review the testing methods introduced in chapter 2, namely multi-variable Rao test, Wald test and LR test. We denote $\ell(\boldsymbol{\beta})$ as the log-likelihood function, $\mathscr{I}_1(\boldsymbol{\beta})$ as the Fisher information matrix for one sample, $\widehat{\boldsymbol{\beta}}$ as the MLE of $\boldsymbol{\beta}$, suppose the sample size is n and $\boldsymbol{\beta}\in\mathbb{R}^p$, the diagram below illustrates the formulas for those tests:

Method	Null Hypothesis	Testing Quantity Under Null Hypothesis
Wald Test	$\mathscr{H}_0:\boldsymbol{\beta}=\boldsymbol{\beta}_0$	$\gamma = n\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)^T\cdot\mathscr{I}_1(\boldsymbol{\beta}_0)\cdot\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)\xrightarrow{d}\chi^2_{(p)}$
Rao Test	$\mathscr{H}_0:\boldsymbol{\beta}=\boldsymbol{\beta}_0$	$\frac{1}{n}\nabla\ell(\boldsymbol{\beta}_0)^T\cdot\mathscr{I}_1(\boldsymbol{\beta}_0)^{-1}\cdot\nabla\ell(\boldsymbol{\beta}_0)\xrightarrow{d}\chi^2_{(p)}$
Likelihood Ratio Test	$\mathscr{H}_l:\boldsymbol{\beta}=\boldsymbol{\beta}_0$	$-2\left(\ell(\boldsymbol{\beta}_0)-\ell(\widehat{\boldsymbol{\beta}})\right)\xrightarrow{d}\chi^2_{(p)}$

All three tests behave asymptotically the same. Depending on significance level α and alternative hypothesis $\mathscr{H}_1$, one can compute the rejection region given by the table below

Alternative Hypothesis $\mathscr{H}_1$	Rejection Region of $\mathscr{H}_0$ Given Significance Level α
$\mathscr{H}_1:\boldsymbol{\beta}\neq\boldsymbol{\beta}_0$	$\gamma>\chi^2_{\alpha/2,p}$ or $\gamma<\chi^2_{1-\alpha/2,p}$
$\mathscr{H}_1:\boldsymbol{\beta}>\boldsymbol{\beta}_0$	$\gamma>\chi^2_{\alpha,p}$
$\mathscr{H}_1:\boldsymbol{\beta}<\boldsymbol{\beta}_0$	$\gamma<\chi^2_{1-\alpha,p}$

More often, we are interested in testing a subset of $\boldsymbol{\beta}$, we partition the parameter in a way that $\boldsymbol{\beta}=\begin{pmatrix}\boldsymbol{\beta}_1^T & \boldsymbol{\beta}_2^T\end{pmatrix}^T$ where $\boldsymbol{\beta}_1$ is the parameter vector of interest, and denote $\widehat{\boldsymbol{\beta}}=\begin{pmatrix}\widehat{\boldsymbol{\beta}_1}^T & \widehat{\boldsymbol{\beta}_2}^T\end{pmatrix}^T$ as the MLE, the Fisher information can also be partitioned in such a way:

$$\mathscr{I}=\begin{pmatrix}\mathscr{I}_{11} & \mathscr{I}_{12}\\ \mathscr{I}_{21} & \mathscr{I}_{22}\end{pmatrix} \tag{7.5.1}$$

then local tests can be done by replacing $\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}, \mathscr{I}$ with $\boldsymbol{\beta}_1, \widehat{\boldsymbol{\beta}}_1, \mathscr{I}_{11}$ respectively.

We investigate one example from Klein and Moeschberger (Chapter 8.5): In a study of 90 male patients with larynx cancer, the researcher determined a test model based on 4 covariates Z_1, Z_2, Z_3, Z_4 that could effect the survival rate of the patient. Z_1, Z_2, Z_3 represents different stages of the cancer (Stage (ii), (iii), (iv) respectively and the stage (i) is used for reference group with its hazard equal to baseline hazard) and equals to 1 if the patient is in that stage, 0 otherwise, and Z_4 is the patient's age, we will use the actual age of each patient to represent Z_4. One hypothesis is that the stage of the cancer will not effect the patient's survival rate, that is, the only covariate that will effect the survival rate is the age. So in this case we make our null hypothesis to be $\mathscr{H}_0 : \beta_1 = \beta_2 = \beta_3 = 0$ while the alternative hypothesis is that $\mathscr{H}_1$: at least one of $\beta_1, \beta_2, \beta_3$ is non-zero. The data is also available from `larynx` package in **R** and may be used to compute test statistic directly. The full dataset is in the Appendix. Under the null hypothesis, the Cox proportional hazard model becomes

$$\lambda(t) = \lambda_0(t) \exp\left(\beta_4 Z_4\right) \tag{7.5.2}$$

where Z_4 is the only covariate we shall consider. Then using the methods we introduced in section 3 and 4, by constructing the partial likelihood, the MLE of β_4 is given by $\widehat{\beta}_4 = 0.023$, after knowing the partial likelihood, the score and the full Fisher information matrix are given by

$$\nabla \ell(0,0,0,0.023) = \begin{pmatrix} -2.448 \\ 3.0583 \\ 7.4400 \\ 0 \end{pmatrix}, \mathscr{I}(0,0,0,0.023) = \begin{pmatrix} 7.637 & -2.608 & -0.699 & -24.730 \\ -2.608 & 9.994 & -0.979 & -8.429 \\ -0.699 & -0.979 & 3.174 & 11.306 \\ -24.730 & -8.429 & 11.306 & 4775.716 \end{pmatrix} \tag{7.5.3}$$

since we are only interested in testing the first 3 parameters, then the Rao test statistic is given by

$$\chi^2_{\text{Rao}} = \begin{pmatrix} -2.448 & 3.0583 & 7.4400 \end{pmatrix} \begin{pmatrix} 7.637 & -2.608 & -0.699 \\ -2.608 & 9.994 & -0.979 \\ -24.730 & -8.429 & 11.306 \end{pmatrix}^{-1} \begin{pmatrix} -2.448 \\ 3.0583 \\ 7.4400 \end{pmatrix} = 20.577 \tag{7.5.4}$$

which yields a p-value of 0.0015. We may also perform LR test and Wald test, and the results show that

$$\chi^2_{\text{LR}} = 15.454, p = 0.0015 \quad \chi^2_{\text{Wald}} = 17.63, p = 0.0005 \tag{7.5.5}$$

so above suggests rejecting the null hypothesis, meaning the the stage of the cancer will effect the patients' survival rate. Now we further we want to know how different stages will effect the patients' survival rate. Since stage (i) is chosen to be the reference group, so we may compute the ratio of the survival rate with other groups. If a patient is in stage 2, then the relative risk is simply $\lambda_2(t) = \exp\left(\beta_1\right)$. By testing $\mathscr{H}_0 : \beta_1 = 0$, we will know if there is a difference in the survival rate in stage 1 and 2. We may perform similiar tests on other covariates compared to stage (i), and we obtain the following table:

Covariates	MLE	Wald Test Statistic	p-Value	Relative Risk to Stage (i)
Z_1	0.1386	0.09	0.7644	1.15
Z_2	0.6383	3.21	0.0730	1.89
Z_3	1.6931	16.08	0.0001	5.44
Z_4	0.0189	1.76	0.1847	1.02

From this table we would see the relative risk compared to stage (i), and we can conclude that the higher the stage, the higher the risk. The KM plot of patients in different stages may also illustrate this:

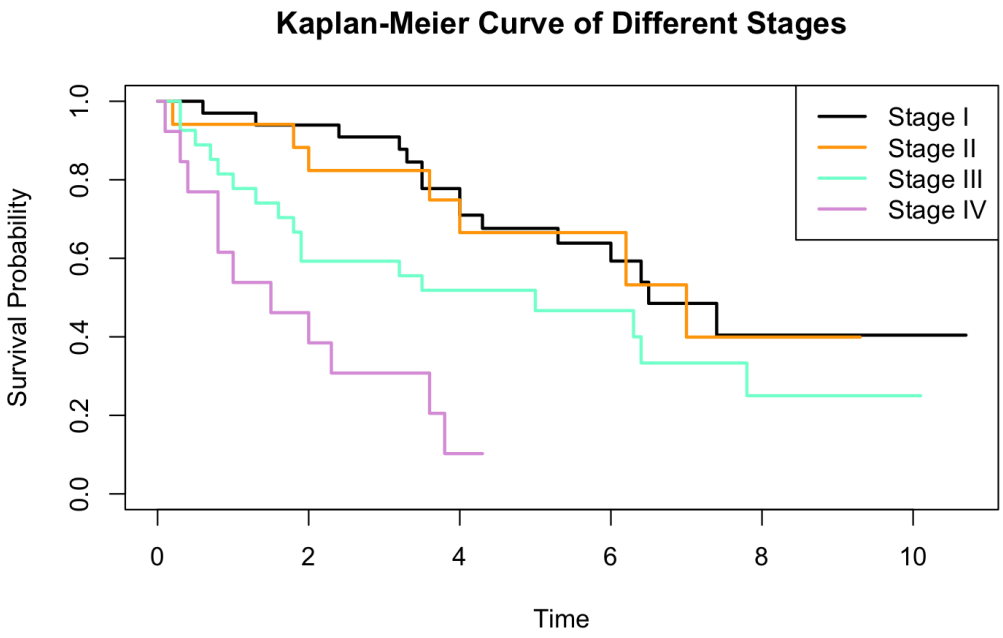


Figure 7.1: The Kaplan-Meier survival curve for patients in different stages

While the next figure shows the Kaplan-Meier curve based on the age group:

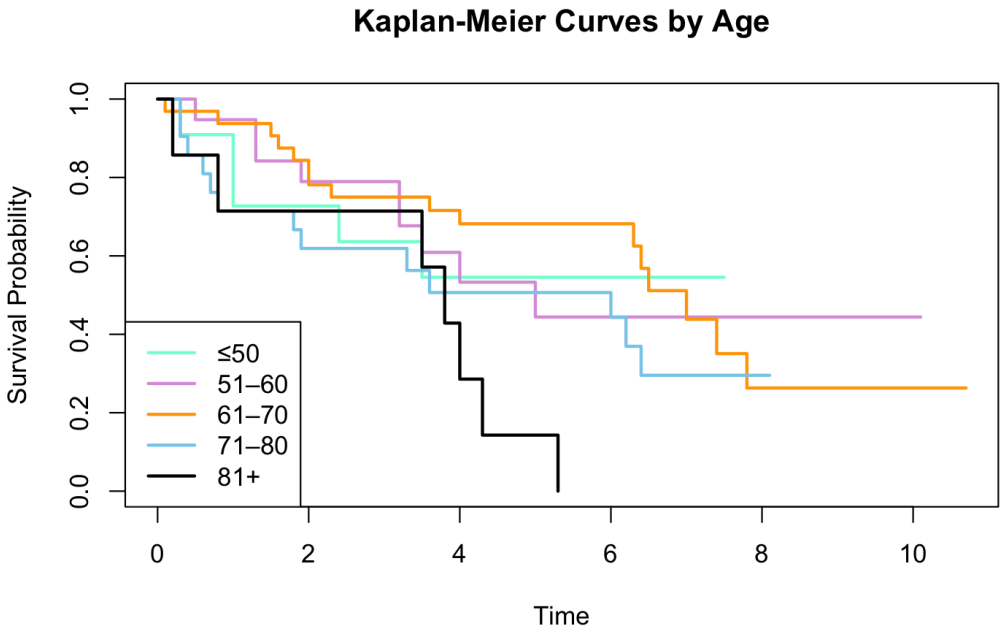


Figure 7.2: The Kaplan-Meier survival curve for patients in different age groups

Bibliography

[1] Arnaud Doucet. *Wald, Rao and Likelihood Ratio Tests*. Stat 461–561 Lecture Notes, University of British Columbia, 2008.

[2] Thomas R. Fleming and David P. Harrington. *Counting Processes and Survival Analysis*. 2005.

[3] Yen-Chi Chen. *Introduction to Non-Parametric Statistics, Lecture 5*. University of Washington, 2018.

[4] D. R. Cox. *Regression Models and Life Tables*. Imperial College London, 1972.

[5] E. L. Kaplan and P. Meier. *Non-parametric Estimation from Incomplete Observations*. 1958.

[6] Henning Läuter and Hannelore Liero. *Nonparametric Estimation and Testing in Survival Models*. University of Potsdam, 2004.

[7] Ørnulf Borgan. *Aalen–Johansen Estimator*. University of Oslo, 2005.

[8] Ørnulf Borgan. *Three Contributions to the Encyclopedia of Biostatistics*. 1997.

[9] Patrick Breheny. *Survival Data Analysis, BIOS:7210/STAT:7570/IGPI:7210 Lecture Notes*. University of Iowa, 2019.

[10] Frank E. Harrell Jr. *Regression Modeling Strategies*. 2001.

[11] David G. Kleinbaum and Mitchel Klein. *Survival Analysis: A Self-Learning Text*.

[12] Odd O. Aalen, Ørnulf Borgan, and Håkon K. Gjessing. *Survival and Event History Analysis*. 2008.

[13] Stanley Sawyer. *The Greenwood and Exponential Greenwood Confidence Intervals in Survival Analysis*. Washington University in St. Louis, 2003.

[14] Steve Selvin. *Survival Analysis for Epidemiologic and Medical Research*. Cambridge University Press, 2010.

[15] Terry Therneau. *A Package for Survival Analysis in R*. 2024.

[16] Mathias Beiglboeck, Walter Schachermayer, and Bezirgen Veliyev. *A Short Proof of the Doob-Meyer Theorem*. 2010.

[17] Odd O. Aalen, Ørnulf Borgan, and Håkon K. Gjessing. *Survival and Event History Analysis, Section 2.2.2*. 2008.

[18] Yi Li. *Lecture Notes on Survival Analysis: A Counting Processes Approach*. University of Michigan.

[19] Menggang Yu. *Introduction to Counting Processes in Survival Analysis*. University of Wisconsin.

[20] M. Zhou. *Two-Sided Bias Bound of the Kaplan–Meier Estimator*. Massachusetts Institute of Technology, 1988.

[21] Jane-Ling Wang. *Smoothing Hazard Rates*. University of California, Davis, 2003.

[22] Catherine R. Loader. *Bandwidth Selection: Classical or Plug-in?* Lucent Technologies, 1999.

[23] Hans-Georg Müller and Jane-Ling Wang. *Hazard Rate Estimation Under Random Censoring with Varying Kernels and Bandwidths*. University of California, Davis, 1994.

[24] John P. Klein and Melvin L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*, 2003.

[25] Robert Foote. *Steepest Descent and Ascent*. Wabash College, 1996.

[26] Jiajun Zhang. *An Introduction to Measure Theory and Real Analysis*. Based on MATH 454 taught by Prof. Linan Chen, McGill University, 2024.

Chapter 8

Appendix

Table 8.1: Larynx Cancer Patient Data

Stage	Time	Age	Year	Death (0 if censored)	Stage	Time	Age	Year	Death (0 if censored)
1	0.6	77	76	1	2	6.2	74	72	1
1	1.3	53	71	1	2	7.0	62	73	1
1	2.4	45	71	1	2	7.5	50	73	0
1	2.5	57	78	0	2	7.6	53	73	0
1	3.2	58	74	1	2	9.3	61	71	0
1	3.2	51	77	0	3	0.3	49	72	1
1	3.3	76	74	1	3	0.3	71	76	1
1	3.3	63	77	0	3	0.5	57	74	1
1	3.5	43	71	1	3	0.7	79	77	1
1	3.5	60	73	1	3	0.8	82	74	1
1	4.0	52	71	1	3	1.0	49	76	1
1	4.0	63	76	1	3	1.3	60	76	1
1	4.3	86	74	1	3	1.6	64	72	1
1	4.5	48	76	0	3	1.8	74	71	1
1	4.5	68	76	0	3	1.9	72	74	1
1	5.3	81	72	1	3	1.9	53	74	1
1	5.5	70	75	0	3	3.2	54	75	1
1	5.9	58	75	0	3	3.5	81	74	1
1	5.9	47	75	0	3	3.7	52	77	0
1	6.0	75	73	1	3	4.5	66	76	0
1	6.1	77	75	0	3	4.8	54	76	0
1	6.2	64	75	0	3	4.8	63	76	0
1	6.4	77	72	1	3	5.0	59	73	1
1	6.5	67	70	1	3	5.0	49	76	0
1	6.5	79	74	0	3	5.1	69	76	0
1	6.7	61	74	0	3	6.3	70	72	1
1	7.0	66	74	0	3	6.4	65	72	1
1	7.4	68	71	1	3	6.5	65	74	0
1	7.4	73	73	0	3	7.8	68	72	1
1	8.1	56	73	0	3	8.0	78	73	0
1	8.1	73	73	0	3	9.3	69	71	0
1	9.6	58	71	0	3	10.1	51	71	0
1	10.7	68	70	0	4	0.1	65	72	1
2	0.2	86	74	1	4	0.3	71	76	1
2	1.8	64	77	1	4	0.4	76	77	1
2	2.0	63	75	1	4	0.8	65	76	1
2	2.2	71	78	0	4	0.8	78	77	1
2	2.6	67	78	0	4	1.0	41	77	1
2	3.3	51	77	0	4	1.5	68	73	1
2	3.6	70	77	1	4	2.0	69	76	1
2	3.6	72	77	0	4	2.3	62	71	1
2	4.0	81	71	1	4	2.9	74	78	0
2	4.3	47	76	0	4	3.6	71	75	1
2	4.3	64	76	0	4	3.8	84	74	1
2	5.0	66	76	0	4	4.3	48	76	0