Lecture Notes on Survival Analysis: A Counting Process Approach

Yi Li
Biostatistics Department
University of Michigan

Contents

Preface

Survival analysis plays a pivotal role in statistical modeling, particularly in medical, biological, and reliability studies where time-to-event data is fundamental. Over the past few decades, the development of rigorous mathematical frameworks, such as counting processes and martingales, has significantly advanced the field. This note, based on several lectures delivered as part of the Michigan Biostatistics survival analysis class, provides a comprehensive yet accessible treatment of survival analysis through the lens of modern statistical methodology, focusing on counting processes and their applications to survival models.

Unlike many existing books on survival analysis, this note takes a deeper dive into the underlying stochastic structures that govern survival data. By leveraging counting processes and martingale theory, we provide a robust theoretical foundation for a various topics, such as nonparametric estimation and comparison of survival functions, Cox models, and competing risk processes. A distinguishing feature of this note is its balance between theoretical rigor and practical implementation. While theoretical derivations are presented in a detailed and structured manner, the note also incorporates applied examples, demonstrating how these advanced methods can be utilized in real-world scenarios. Additionally, it connects classical techniques with their modern extensions, bridging the gap between foundational concepts and cutting-edge applications. By equipping readers with the tools necessary to analyze and interpret survival data within a rigorous stochastic framework, the note aims to foster a more profound appreciation of the mathematical principles that underpin survival analysis.

This note is intended for graduate students, researchers, and practitioners who seek a deeper understanding of survival analysis. A working knowledge of probability theory and statistical inference will be beneficial, though essential concepts are introduced as needed. While measure theory can enhance understanding, this note does not assume prior knowledge of it. I have intentionally refrained from introducing the definitions of σ-algebras and conditional expectations with respect to σ-algebras. Those interested in these topics can refer to a separate note I wrote while teaching measure theory in Harvard Biostatistics https://public.websites.umich.edu/~yili/yinote.pdf. Additionally, I have tried to keep technicalities to a minimum to prevent confusion with complex concepts, and to make the material self-contained. The list of references has been kept brief, with a more comprehensive compilation to be provided at a later stage. I welcome feedback and discussions from readers.

1 Stochastic Processes

We consider $(\Omega, \mathcal{F}, P)$, a probability space, where Ω is the sample space containing all the possible outcomes (each element of Ω is called a sample point), $\mathcal{F}$ is a collection of subsets of Ω (called a σ-algebra) so that each element of $\mathcal{F}$ is an event, and P measures the probability of each event. In the context of stochastic processes, we introduce $\{\mathcal{F}_t\}_{t\geq 0}$ as a time-indexed filtration, meaning a non-decreasing family of sub-σ-algebras of $\mathcal{F}$. That is, $\mathcal{F}_t$ is a σ-algebra containing history up to time t, i.e., the collection of all the information up to t, such that information increases over time, or $\mathcal{F}_s \subseteq \mathcal{F}_t \subset \mathcal{F}$ if $s \leq t$; see a graphical illustration in Figure 1. The definition of a filtration generated by a series of events can be found in Chung (1974) and Fleming & Harrington (2013).

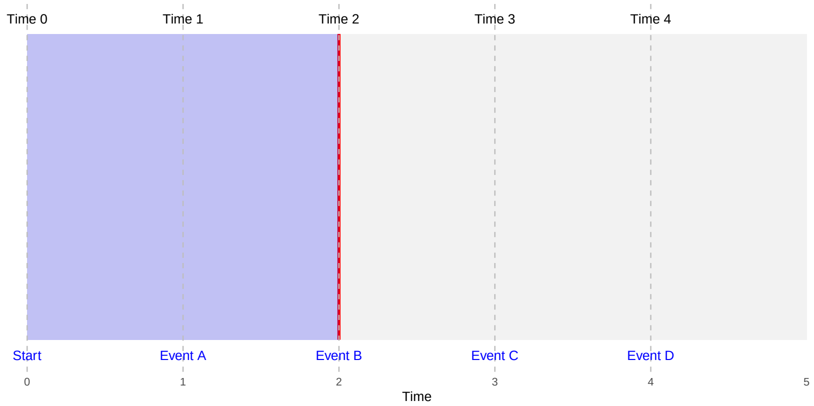


Figure 1: Illustration of a filtration: the progression of information over time, with accumulated information shaded in blue and future information shaded in gray. The red line at $t = 2$ marks the current time.

A stochastic process $\{X(t, \omega)\}_{t\geq 0}$ is a collection of random variables indexed by time t, defined on a $(\Omega, \mathcal{F}, P)$, where $\omega \in \Omega$ represents a sample point in the sample space. The process is essentially a bivariate map from the time domain and the sample space to a subset of the real line. The evolution of the process over time is governed by a specified probability law.

A stochastic process $\{X(t, \omega)\}_{t\geq 0}$ is said to be **adapted** to a filtration $\{\mathcal{F}_t\}_{t\geq 0}$ if, for each t, the random variable $X(t, \omega)$ is $\mathcal{F}_t$-measurable. This means that for any $b < \infty$, the event

$$\{\omega : X(t, \omega) < b\} \in \mathcal{F}_t, \quad \forall t \geq 0.$$

Intuitively, this implies that at any time t, the value of $X(t, \omega)$ is completely determined by the information contained in $\mathcal{F}_t$. Given a specific ω, the function $X(t, \omega)$ as a function of t is called the **sample path** or realization of the stochastic process. Figure 4 shows 5 sample paths of a stochastic process (more specifically, Brownian Motion) corresponding to 5 different ω's. For notational convenience, we often drop ω and write $X(t)$ instead of $X(t, \omega)$. Even with this simplification, it is important to keep in mind that the randomness of the process is induced by ω. When the context is clear, we may use $X(t)$ to refer to the entire process $\{X(t)\}_{t\geq 0}$, though Wikipedia recommends against it.

Finally, in contexts involving multiple subjects, we may introduce subscripts, such as $\mathcal{F}_{i,t}$ and $X_i(t)$, to denote individual-specific filtrations and processes for a particular subject i.

1.1 Counting processes in the survival context

Building on the defined probability spaces and filtrations, we turn to an application of stochastic processes in survival analysis. All the random variables introduced below are defined in a common probability space $(\Omega, \mathcal{F}, P)$. As a special case of stochastic processes, we define counting processes in a survival context with a homogeneous population consisting of n independent subjects indexed by $i = 1, 2, \ldots, n$. For each subject, define the following random variables on $(\Omega, \mathcal{F}, P)$:

- $T_i : \Omega \to [0, \infty)$ denotes the true underlying continuous survival time for subject i.
- $C_i : \Omega \to [0, \infty)$ represents the potential continuous censoring time for subject i, assumed to be independent of T_i.

Our primary interest is in the distribution of T_i. However, since T_i is subject to right censoring, we do not observe T_i directly. Instead, we observe

$$X_i = \min(T_i, C_i) \overset{\text{def}}{=} T_i \wedge C_i,$$

which represents the observed survival time, along with the event indicator

$$\Delta_i = I(T_i \leq C_i),$$

where $\Delta_i = 1$ indicates that the event occurred, and $\Delta_i = 0$ indicates that the observation was censored. Here, $I(\cdot)$ denotes the indicator function. We aim to use the observed data, based on some conditions, to estimate the distributions of T_i's.

Assume that the T_i's are independently and identically distributed (i.i.d.). For any $t > 0$, define the hazard function to be

$$\lambda(t) = \lim_{dt \to 0^+} \frac{1}{dt} P(t \leq T_i < t + dt \mid T_i \geq t),$$

which is the instantaneous failure rate, given survival up to t. Also, define the cumulative hazard function

$$\Lambda(t) = \int_0^t \lambda(s)\, ds,$$

which is of our main interest. It measures the total accumulated risk of an event occurring up to time t and is directly linked to the survival function

$$S(t) \overset{def}{=} P(T_i > t)$$

via $S(t) = \exp\{-\Lambda(t)\}$, meaning that higher cumulative hazard values correspond to lower survival probabilities. We assume that $\Lambda(t) < \infty$ for all $t < \infty$, implying that the survival time T_i is unbounded. For simplicity, we also assume that the censoring times C_i, $i = 1, \ldots, n$, are i.i.d. Based on the observed i.i.d. data $\{(X_i, \Delta_i)\}_{i=1}^n$, we can estimate $\Lambda(t)$ via the Nelson-Aalen estimator.

A key assumption ensuring the validity of the Nelson-Aalen estimator is independent censoring. Unfortunately, based solely on the observed data $\{(X_i, \Delta_i)\}_{i=1}^n$, this assumption is not statistically testable (Tsiatis 2006); one must rely on subject matter knowledge to justify it. When censoring is dependent, the Nelson-Aalen estimator may yield biased results. Survival analysis with dependent censoring is still an actively research area; for various methods to address dependent censoring, see Tsiatis (2006). Throughout this note, we assume independent censoring, as is commonly done in the literature.

Moreover, instead of the traditional approach that directly works with random variables, we utilize counting processes, which offer a more elegant, flexible, and theoretically robust framework for survival analysis and event-time modeling. Their deep connection to martingales, stochastic integrals, and intensity functions makes them a powerful tool for both theoretical and applied research.

Definition 1.1. *Define the at-risk process for subject i, $Y_i(t)$, as $Y_i(t) = I(X_i \geq t)$, where $I(\cdot)$ is the indicator function.*

The at-risk process seamlessly integrates censoring mechanisms and keeps track of whether an individual is still at risk of experiencing the event at time t; see Figure 2.

Definition 1.2. *Define $N_i^*(t) = I(T_i \leq t)$, indicating whether death occurs prior to or at t.*

This notion uses counting processes to conceptualize event occurrence. However, with censoring, T_i is not always observable, thus one cannot directly analyze $N_i^*(t)$. Instead, a counting process is defined to handle censored observations, accounting for censoring in the process. In particular, we introduce a modified counting process such that an event (e.g., death) is observed to occur only when the individual is still at risk.

Definition 1.3. *Define the observable increment as*

$$dN_i(s) = Y_i(s)dN_i^*(s), \tag{1.1}$$

where $dN_i^*(s) = N_i^*((s+ds)^-) - N_i^*(s^-)$.

The inclusion of $Y_i(s)$ in (1.1) ensures that event counting is restricted to an individual who is still at risk at time s. As N_i^* is right-continuous, it follows that $N_i^*((s+ds)^-) = N_i^*(s^+) = N_i^*(s)$, which implies that $dN_i^*(s) = N_i^*(s) - N_i^*(s^-)$, corresponding to the conventional definition of an increment. However, more broadly, for a stochastic process X, we define the jump size at t as $X((t+dt)^-) - X(t^-)$ rather than $X(t) - X(t^-)$ to account for a subtle distinction discussed immediately before Proposition 2.5.

Definition 1.4. *Define the observable counting process $N_i(t)$ by*

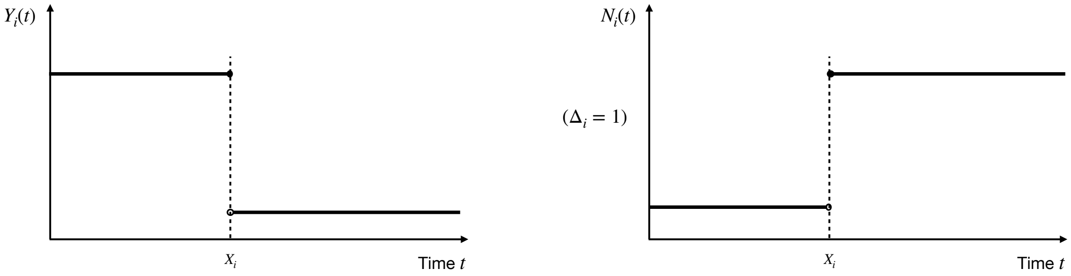
$$N_i(t) = \int_0^t dN_i(s) = \int_0^t Y_i(s)dN_i^*(s).$$

Here (and hereafter) the Stieltjes integral (Carter et al. 2000) of the type of $\int_u^t k(s)dN_i(s)$, where N_i is right continuous, is the sum of the values of k at the jump times of $N_i(\cdot)$ in the interval $(u,t]$; see Figure 2. Later on, we will also frequently use the notion of $\int_u^t k(s)dM_i(s)$, where $M_i(s) = N_i(s) - \int_0^s Y_i(u)d\Lambda(u)$ is a (local) martingale (defined later). Then,

$$\int_u^t k(s)dM_i(s) \overset{def}{=} \int_u^t k(s)dN_i(s) - \int_u^t k(s)Y_i(s)d\Lambda(s)$$

where the second integral is indeed a Stieltjes integral with respect to the deterministic process $\Lambda(\cdot)$, inheriting well-defined properties from integrals of random processes with respect to nondecreasing and deterministic functions.

Obviously, $N_i(t)$ is the observable version of $N_i^*(t)$. It can be shown that $N_i(t) = I(X_i \le t, \Delta_i = 1)$, indicating whether patient i is observed to die by time t, while $dN_i(s)$ is whether patient i is observed to die at time s. In the following, we summarize the properties and interpretations for $N_i(t)$, $dN_i(t)$ and $Y_i(t)$.



(a) $Y_i(t)$

(b) $N_i(t)$

Figure 2: Example figures for $Y_i(t)$ and $N_i(t)$.

1.2 Summary of counting processes and the important results

1. $Y_i(t) = Y_i(t^-) = I(X_i \ge t)$ is left continuous, flagging whether patient i is at risk at time t.
2. $\sum_{i=1}^n Y_i(t) \overset{def}{=} Y(t)$ is also left continuous, counting the number of patients at risk at t among these n patients.
3. $N_i(t) = N_i(t^+) = I(X_i \le t, \Delta_i = 1)$ is right continuous, flagging whether patient i is observed to die at and before time t.
4. $\sum_{i=1}^n N_i(t) \overset{def}{=} N(t)$ is right continuous, counting how many patients observed to die at and before time t among the n patients.
5. $dN_i(t) = \Delta_i \cdot I(X_i = t)$, indicating whether patient i is observed to die at time t.
6. $\sum_{i=1}^n dN_i(t) \overset{def}{=} dN(t)$ denotes the number of patients die at time t among these n patients.

Looking ahead, we enumerate important results which will be detailed later. We first define the expectation. Let $G : \Omega \to \mathbb{R}$ be a random variable defined on $(\Omega, \mathcal{F}, P)$. The expectation of G is defined as $\mathbb{E}[G] = \int_\Omega G(\omega)\, dP(\omega)$. In particular, if G has a probability density function f_G on $\mathbb{R}$, then $\mathbb{E}[G] = \int_\mathbb{R} x\, f_G(x)\, dx$.

1. Define $dM_i(t) = dN_i(t) - Y_i(t)d\Lambda(t),$ where $M_i(t)$ is a martingale process (defined later). Then $\mathbb{E}dM_i(t) = 0$. So $dM_i(t)$ is like the residual term in regression models.
2. $\text{Cov}(dM_i(t), dM_i(s)) = 0, \ s \neq t$.
3. $\mathbb{E}M_i(t) = 0,$ where $M_i(t) = N_i(t) - \int_0^t Y_i(s)d\Lambda(s)$.
4. $\mathbb{E}N_i(t) = \mathbb{E}A_i(t),$ where $A_i(t) = \int_0^t Y_i(s)d\Lambda(s) = \Lambda(t \wedge X_i)$.
5. $\text{Var}(M_i(t)) = \mathbb{E}A_i(t)$.

1.3 Nelson-Aalen estimator: representation and derivation via counting processes

Given the data $(X_i, \Delta_i), i = 1, \ldots, n$, we extract a total of $n_d \, (\leq n)$ distinct observed failure time points, $t_1 < t_2 \ldots < t_{n_d}$. Let D_j, Y_j denote the number of observed deaths and subjects at risk at t_j respectively. The Nelson-Aalen estimator of the cumulative hazard can be expressed as

$$\widehat{\Lambda}(t) = \sum_{t_j \leq t} \frac{D_j}{Y_j}$$

which can be succinctly expressed in terms of counting process as well

$$\widehat{\Lambda}(t) = \sum_{t_j \leq t} \frac{dN(t_j)}{Y(t_j)} = \int_0^t \frac{dN(s)}{Y(s)}.$$

This follows from the definition of the Stieltjes integral, noting that

$$D_j = dN(t_j) \quad \text{and} \quad Y_j = Y(t_j),$$

with $N(t)$ only exhibiting jumps at $t_1, t_2, \ldots$.

We can also derive the Nelson-Aalen estimator (Aalen 1978) using the fact of $\mathbb{E}dM_i(t) = 0$; see Section 1.2. Indeed, the first moment estimation gives

$$\sum_{i=1}^n dM_i(t) = \sum_{i=1}^n dN_i(t) - \sum_{i=1}^n Y_i(t)d\Lambda(t) = 0,$$

which gives

$$d\widehat{\Lambda}(t) = \frac{\sum_{i=1}^n dN_i(t)}{\sum_{i=1}^n Y_i(t)} = \frac{dN(t)}{Y(t)},$$

implying the Nelson-Aalen estimator of $\Lambda(t)$ with $0 < t < \infty$:

$$\widehat{\Lambda}(t) = \int_0^t \frac{dN(s)}{Y(s)}. \tag{1.2}$$

The Nelson-Aalen estimator yields a right continuous step function; see Figure 3. In the following,

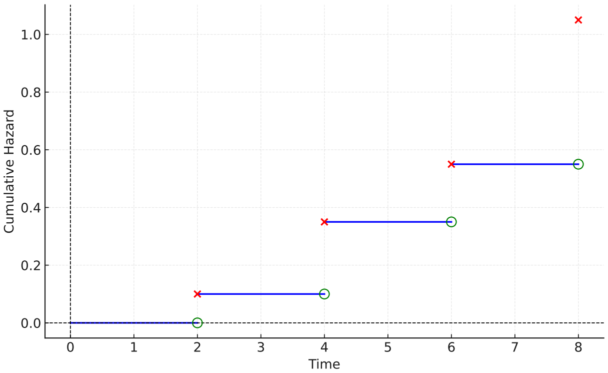


Figure 3: Example plot of the Nelson-Aalen estimator: the left-hand limits are marked using empty green circles, while the cumulative hazard at event times is marked with solid red dots. The plot shows right-continuous nature of the estimated function.

we will derive the variance of $\widehat{\Lambda}(t)$ and other statistical properties of $\widehat{\Lambda}(t)$, such as ubiasedness, consistency and asymptotic normality, using the notion of counting processes and martingales.

We will first introduce the concept of a martingale, a modern probability tool that can significantly simplify theoretical analysis. In stochastic processes, martingales represent systems with the 'fair game' property, where there is no net gain or loss on average over time. They model randomness in a way that future values depend only on the present and not the past history. We will show that counting processes can be decomposed into a martingale component (capturing randomness) and a predictable compensator component (representing deterministic trends). This decomposition, as illustrated below, plays a crucial role in deriving estimators and testing hypotheses in survival analysis.

2 Martingales and the Properties

We present the definition of martingales and their properties and discuss how to construct martingales.

Definition 2.1 (Definition of Martingale)**.** *$M(t)$ is a martingale with respect to $\mathcal{F}_t$ if*

1. *Adaptedness: $M(t)$ is adapted to $\mathcal{F}_t$. More precisely, we should say $\{M(t)\}_{t\geq 0}$ is adapted or measurable to a filtration $\{\mathcal{F}_t\}_{t\geq 0}$; we use measurable and adapted interchangeably later.*
2. *Integrability: $\mathbb{E}|M(t)| < \infty$ for all $0 < t < \infty$.*
3. *Martingale ('fair game') property: $\mathbb{E}\big(M(t)|\mathcal{F}_s\big) = M(s)$ for any $0 \leq s \leq t$, where the conditional expectation with respect to a filtration is given in Chung (1974). Notationwise, here and all "=" below should be interpreted as "=" holds almost surely.*

Measurability with respect to a filtration ensures the process depends only on current and past information, aligning with the flow of time. Integrability maintains mathematical consistency by avoiding infinite values. The martingale property, which is central to the definition of martingales, reflects their 'fair game' nature, where future values, given present knowledge, are not systematically predictable. This property is critical for various applications, including stochastic processes, financial (stock price) modeling, and survival analysis Also, whenever we mention a martingale, the underlying filtration, e.g., $\mathcal{F}_t$, should be specified, in which case we may say $M(t)$ is an "$\mathcal{F}_t$-martingale."

Property 1 (Properties of martingale)**.** *Let $M(t)$ be a martingale with respect to $\mathcal{F}_t$.*

1. *If $M(0) = 0$, then $\mathbb{E}M(t) = 0$.*

Proof. $\mathbb{E}M(t) = \mathbb{E}(\mathbb{E}\{M(t)|\mathcal{F}_0\}) = E(M(0)) = 0$. □

2. *(uncorrelated increments under disjoint intervals) If $0 \leq v \leq u \leq s \leq t$, then $\mathbb{E}\big((M(t) - M(s))(M(u) - M(v))\big) = 0$. Also, $\mathbb{E}(dM(v)dM(s)) = 0$ or $Cov(dM(v), dM(s)) = 0$ for $v \neq s$.*

Proof. This follows as $\mathbb{E}\big((M(t)-M(s))(M(u)-M(v))\big) = \mathbb{E}\{\mathbb{E}\big((M(t)-M(s))(M(u)-M(v))\big)|\mathcal{F}_s\}$. □

Definition 2.2 (Predictable)**.** *$X(t)$ is predictable with respect to $\mathcal{F}_t$ if $X(t)$ is determined by $\mathcal{F}_{t^-}$, i.e., a predictable process is one whose behavior at t is determined by the information over $[0, t)$. That is, for a predictable process $X(t)$, if given $\mathcal{F}_{t^-}$, we know the exact value of $X(t)$.*

Example 2.3 (Example)**.** *A left continuous (and measurable) process is predictable. That is, if $X(t) = X(t^-)$, $X(t)$ is predictable.*

Example 2.4. *$Y_i(t)$ is predictable as it is left continuous. But $N_i(t)$ is not predictable. This is because, even with the survival information up to t but not including t, $N_i(t)$ may or may not have a jump at t so it is not predictable.*

Define the jumpsize of $M(t)$ at t as

$$dM(t) = M((t + dt)^-) - M(t^-), \tag{2.1}$$

i.e., the jumpsize of $M(\cdot)$ from t to $t+dt$. We use $M((t+dt)^-)$ instead of $M(t+dt^-)$ or $M(t+dt)$ or $M(t)$ for a subtle reason. This way, we can write, aligning with the Itô stochastic integral (Øksendal 2003), that

$$\int_s^t dM(u) = \sum_{s<u\leq t} dM(u) = \sum_{s<u\leq t} M((u + du)^-) - M(u^-) = M(t) - M(s).$$

This facilitates the proof of the following important proposition for the infinitesimal characterization of a martingale.

Proposition 2.5. *Suppose $M(t)$ is integrable and is determined by $\mathcal{F}_t$, then $M(t)$ is a martingale if and only if*

$$\mathbb{E}\big(dM(t)|\mathcal{F}_{t^-}\big) = 0.$$

Proof. "⇒:" If $M(t)$ is a martingale, then it holds that

$$\mathbb{E}\big(M((t+dt)^-) - M(t^-) \mid \mathcal{F}_{t^-}\big) = M(t^-) - M(t^-) = 0.$$

Here, the first equality is by the definition of martingale and $(t+dt)^- \geq t^-$. By the definition in (2.1), we have that

$$\mathbb{E}\big(dM(t)|\mathcal{F}_{t^-}\big) = \mathbb{E}\big(M((t+dt)^-) - M(t^-) \mid \mathcal{F}_{t^-}\big) = 0.$$

"⇐:" We need to show that $\mathbb{E}\big(M(t)|\mathcal{F}_s\big) = M(s)$ for any $s \leq t$. We have

$$\begin{aligned}
\mathbb{E}\{M(t)|\mathcal{F}_s\} &= \mathbb{E}\{M(s) + M(t) - M(s)|\mathcal{F}_s\} \\
&= M(s) + \mathbb{E}\{\int_s^t dM(u)|\mathcal{F}_s\} \\
&= M(s) + \int_s^t \big[\mathbb{E}(dM(u)|\mathcal{F}_s)\big] \\
&= M(s) + \int_s^t \mathbb{E}\big\{\big[\mathbb{E}(dM(u)|\mathcal{F}_{u^-})\big]\big|\mathcal{F}_s)\big\} \\
&= M(s).
\end{aligned}$$

Here, as noted before, $\int_s^t dM(u) = \sum_{s<u\leq t} M((u+du)^-) - M(u^-) = M(t) - M(s)$. Also, the second equality is by that $M(s)$ is measurable with respect to $\mathcal{F}_s$, the third equality is by Fubini's Theorem and the fourth equality is by the iterated expectation theorem (Chung 1974):

$$\mathbb{E}\left\{\mathbb{E}\left(G \mid \mathcal{F}_s\right) \mid \mathcal{F}_t\right\} = \mathbb{E}\left\{\mathbb{E}\left(G \mid \mathcal{F}_t\right) \mid \mathcal{F}_s\right\} = \mathbb{E}\left(G \mid \mathcal{F}_s\right)$$

if $\mathcal{F}_s \subset \mathcal{F}_t$, where G denotes a generic random variable. The last equality comes from the condition that $\mathbb{E}(dM(u)|\mathcal{F}_{u^-}) = 0$. □

Then, if $M(t)$ is a martingale, immediately

$$\mathbb{E}\big(dM(t)\big) = \mathbb{E}(\mathbb{E}\big(dM(t)|\mathcal{F}_{t^-}\big)) = 0.$$

This local characterization of martingales is valuable as it provides an explicit method for computing the compensator of a right-continuous but non-predictable process, which we will utilize in our subsequent development.

An example of a continuous Martingale is the Brownian motion (Mörters & Peres 2010), often denoted as $B(t), t \geq 0$, satisfying (i) $B(0) = 0$; (ii) Independent increments: The changes in $B(t)$ over non-overlapping time intervals are independent; (iii) Gaussian increments: For any $0 \leq s < t$, the increment $B(t) - B(s) \sim N(0, t-s)$; (iv) Continuous paths: $B(t)$ is continuous in t; (v) Martingale property: $\mathbb{E}[B(t)|B(u), 0 \leq u \leq s] = B(s)$, for $s \leq t$. Figure 4 shows sample paths of a Brownian motion. Brownian motion has been widely applied across various fields: in physics, it describes the movement of particles suspended in a fluid, as first observed by Robert Brown (Brown 1828); in finance, it serves as the foundation for modeling stock prices and option pricing within the Black-Scholes framework (Black & Scholes 1973); in engineering, it is utilized for signal processing, noise modeling, and control systems (Franklin et al. 2023); and in biology, it helps explain the random motion of molecules and cellular processes (Berg 2023).

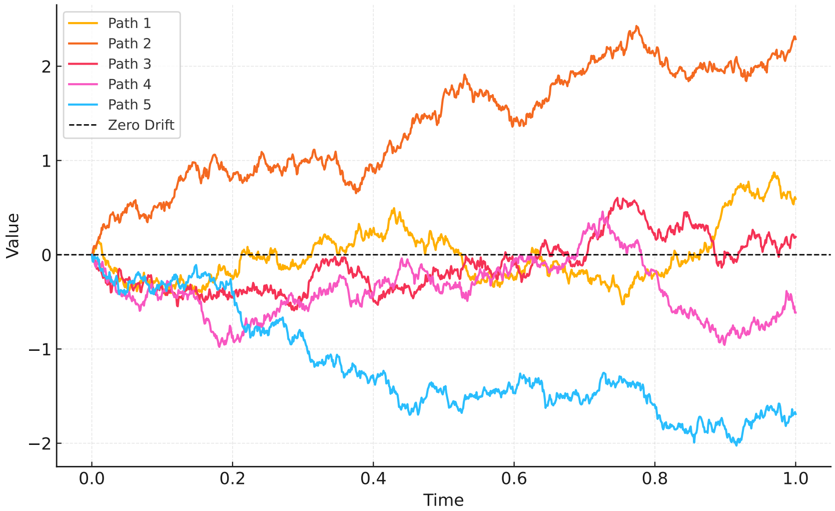


Figure 4: Sample paths of a Brownian motion: the plot highlights the unpredictable nature of the paths while maintaining no overall trend (zero drift); 5 sample paths are shown, with each being a realized trajectory of the process over time.

Now we apply the martingale theory to the survival process. First, for subject i, define $\mathcal{F}_{i,t} = \sigma\{N_i(s), Y_i(s), 0 \leq s \leq t\}$ the σ-algebra generated by the events in the bracket. Basically, this is the survival information for subject i up to (including) t; see a formal definition regarding a σ-algebra generated by events in Chung (1974). We will show

$$M_i(t) = N_i(t) - \int_0^t Y_i(s) d\Lambda(s)$$

is a martingale process with respect to $\mathcal{F}_{i,t}$. Here, $A_i(t) \overset{def}{=} \int_0^t Y_i(s) d\Lambda(s)$ is called the compensator of $N_i(t)$, i.e., the "predictable" part of $N_i(t)$. While it is natural to think that many predictable processes are left-continuous, not all predictable processes have this property. In fact, $A_i(t)$ is right continuous because Λ is right continuous by definition. On the other hand, if Λ has a discontinuity at t, $A_i(t)$ will not be left continuous at t. In this case, $A_i(t)$ is still predictable. However, if $\Lambda(t)$ is continuous or even differentiable with a derivative $\lambda(t)$ (as we will assume in the following), $A_i(t)$ is continuous everywhere and we can simply write $A_i(t) = \int_0^t Y_i(s)\lambda(s)ds$.

Proposition 2.6. *Assume T_i, C_i are independent and T_i is continuous. If $P(X_i \geq t) > 0$, then*

$$\lim_{dt \to 0^+} \frac{1}{dt} P(t \leq X_i < t + dt,\ \Delta_i = 1 \mid X_i \geq t) = \lambda(t).$$

Proof. We consider

$$P(t \leq X_i < t + dt, \Delta_i = 1 \mid X_i \geq t) = \frac{P(t \leq X_i < t + dt, \Delta_i = 1)}{P(X_i \geq t)},$$

while expanding the numerator,

$$\begin{aligned} P(t \leq X_i < t+dt, \Delta_i = 1) &= P(t \leq T_i < t+dt, T_i \leq C_i) \\ &= P(t \leq T_i < t+dt) P(T_i \leq C_i \mid T_i \in [t, t+dt]). \end{aligned}$$

Since T_i is continuous, its density function gives:

$$\lim_{dt \to 0+} \frac{1}{dt} P(t \leq T_i < t+dt) = f(t).$$

Then we consider

$$\begin{aligned} & P(T_i \leq C_i \mid T_i \in [t, t+dt)) \\ = & \frac{P(T_i \leq C_i, T_i \in [t, t+dt))}{P(T_i \in [t, t+dt))} \\ = & \frac{\int_t^{t+dt} f(s) P(C_i \geq s) ds}{\int_t^{t+dt} f(s) ds}, \end{aligned}$$

where the last equality comes from the independence of C_i with T_i. Let $dt \to 0^+$ and apply L'Hôpital's rule, we have

$$\lim_{dt \to 0^+} P(T_i \leq C_i \mid T_i \in [t, t+dt)) = \frac{f(t) P(C_i \geq t)}{f(t)} = P(C_i \geq t) = S_C(t^-),$$

where $S_C(t)$ is the survival function of C_i or $S_C(t) = P(C_i > t)$. Thus, the numerator (after dividing by dt and taking the limit) becomes:

$$\begin{aligned} & \lim_{dt \to 0+} \frac{1}{dt} P(t \leq T_i < t+dt, T_i \leq C_i) \\ = & \lim_{dt \to 0+} \frac{1}{dt} P(t \leq T_i < t+dt) \lim_{dt \to 0+} P(T_i \leq C_i \mid T_i \in [t, t+dt)) \\ = & f(t) S_C(t-). \end{aligned}$$

For the denominator:

$$P(X_i \geq t) = P(T_i \geq t) P(C_i \geq t) = S(t) S_C(t-).$$

Hence,

$$\lim_{dt \to 0^+} \frac{1}{dt} P(t \leq X_i < t+dt,\ \Delta_i = 1 \mid X_i \geq t) = \frac{f(t) S_C(t-)}{S(t) S_C(t-)} = \frac{f(t)}{S(t)} = \lambda(t).$$

□

The result indicates that under independent censoring, the observed data can be used to estimate the hazard function, so methods like the Nelson-Aalen or Kaplan-Meier estimators can be applied. Moreover, the result serves as a key step in developing the martingale framework for survival analysis, as shown below.

Proposition 2.7. *$M_i(t) = N_i(t) - \int_0^t Y_i(s) d\Lambda(s)$ is a martingale process with respect to $\mathcal{F}_{i,t}$.*

Proof. First, we show $M_i(t)$ is measurable with respect to $\mathcal{F}_{i,t}$. In fact, $N_i(t)$ is measurable with respect to $\mathcal{F}_{i,t}$. Also, $Y_i(s)$, when $s \leq t$, is measurable with respect to $\mathcal{F}_{i,s}$ and hence $\mathcal{F}_{i,t}$. This follows because of the increasing property of filtrations, i.e., $\mathcal{F}_{i,s} \subset \mathcal{F}_{i,t}$ when $s \leq t$. Therefore, $\int_0^t Y_i(s)\lambda(s)ds$ is measurable with respect to $\mathcal{F}_{i,t}$, leading to $N_i(t) - \int_0^t Y_i(s)d\Lambda(s)$ is measurable with respect to $\mathcal{F}_{i,t}$ as the countable summation of measurable functions is measurable.

Second, we note that $M_i(t)$ is integrable for any $t < \infty$. This follows as

$$\begin{aligned}
\mathbb{E}|M_i(t)| &< \mathbb{E}N_i(t) + \int_0^t P(X_i \geq s)d\Lambda(s) \\
&\leq 1 + \int_0^t P(T_i \geq s)d\Lambda(s) \\
&= 1 + \int_0^t e^{-\Lambda(s)}d\Lambda(s) \\
&= 1 + 1 - S(t) < 2. \qquad (2.2)
\end{aligned}$$

Now with Proposition 2.5, it only remains for us to prove

$$\mathbb{E}\big(dM_i(t)|\mathcal{F}_{i,t^-}\big) = 0.$$

In fact, it holds that

$$\begin{aligned}
\mathbb{E}\big(dM_i(t)|\mathcal{F}_{i,t^-}\big) &= \mathbb{E}\big(dN_i(t) - Y_i(t)d\Lambda(t)|\mathcal{F}_{i,t^-}\big) \\
&\overset{(2.1)}{=} \mathbb{E}\big(N_i((t+dt)^-) - N_i(t^-)|\mathcal{F}_{i,t^-}\big) - \mathbb{E}\big(Y_i(t)\lambda(t)dt|\mathcal{F}_{i,t^-}\big) \\
&= \mathbb{E}\big(I(t \leq X_i < t+dt, \Delta_i = 1)|Y_i(t)\big) - Y_i(t)\lambda(t)dt.
\end{aligned}$$

Here, $\mathbb{E}\big(Y_i(t)d\Lambda(t)|\mathcal{F}_{i,t^-}\big) = Y_i(t)d\Lambda(t)$ because $Y_i(t)$ is predictable. Also, the third equality follows because events of $X_i \geq t$ and $X_i < t$ are measurable with respect to $\mathcal{F}_{i,t^-}$, and you can compute the conditional expectation conditional on $Y_i(t) = I(X_i \geq t)$. More rigorous justifications can be found in Theorem 1.3.1 of Fleming & Harrington (2013). We can separately compute the expectation for the two cases depending on the value of $Y_i(t)$.

- Case 1: If $Y_i(t) = 0$ (the patient has already experienced an event or been censored before time t), then:

  $$\mathbb{E}\big[I(t \leq X_i < t+dt,\ \Delta_i = 1) \mid Y_i(t) = 0\big] = P(t \leq X_i < t+dt,\ \Delta_i = 1 \mid X_i < t) = 0$$

  This follows because patient who is no longer at risk at t^- will have no chance to be observed to die later.

- Case 2: If $Y_i(t) = 1$ (the patient is still at risk just before time t), then:

  $$\mathbb{E}\big[I(t \leq X_i < t+dt, \Delta_i = 1)|X_i \geq t\big] = P(t \leq X_i < t+dt,\ \Delta_i = 1 \mid X_i \geq t) = \lambda(t)dt$$

  which holds with independent censoring (Proposition 2.6).

Combining these two cases, we have:

$$\mathbb{E}\big[I(t \le X_i < t+dt, \Delta_i = 1)|Y_i(t)\big] = Y_i(t)\lambda(t)dt.$$

We hence have

$$\mathbb{E}\big(dM_i(t)|\mathcal{F}_{i,t^-}\big) = \mathbb{E}\big(I(t \le X_i < t+dt, \Delta_i = 1)|\mathcal{F}_{i,t^-}\big) - Y_i(t)\lambda(t)dt = 0.$$

That is, we have shown that $M_i(t) = N_i(t) - \int_0^t Y_i(s)\lambda(s)ds$ is a martingale process with respect to $\mathcal{F}_{i,t}$ or, succinctly, M_i is an $\mathcal{F}_{i,t}$-martingale. □

Recalling the definition of $\widehat{\Lambda}(t)$ in (1.2), we observe that

$$\widehat{\Lambda}(t) - \Lambda(t) = \int_0^t \frac{dN(s) - Y(s)d\Lambda(s)}{Y(s)} = \int_0^t \frac{dM(t)}{Y(s)}, \tag{2.3}$$

where we ignore the small possibility of $Y(s) = 0$. More rigorous treatment can be found in Theorem 3.2.1 of Fleming & Harrington (2013), which noted that $\int_0^t \frac{dN(s)}{Y(s)} = \int_0^t \frac{I(Y(s)>0)}{Y(s)} dN(s)$ with $0/0 \overset{def}{=} 0$. Here, $dM(s) = \sum_{i=1}^n dM_i(t)$. With i.i.d. data, $M(t) = \sum_{i=1}^n M_i(t)$ can be shown to be a martingale with respect to a richer σ-algebra $\mathcal{F}_t = \sigma\{N_i(s), Y_i(s), i = 1, ..., n, 0 \le s \le t\}$, the survival information of the whole population up to (including) t; see Exercise 1.11(a) of Fleming & Harrington (2013). Recall

$$N(t) = \sum_{i=1}^n N_i(t), \quad Y(t) = \sum_{i=1}^n Y_i(t),$$

the martingale $M(t)$ can be written as

$$M(t) = N(t) - \int_0^t Y(s)d\Lambda(s).$$

We define the compensator $A(t)$ of $N(t)$ as

$$A(t) = \sum_{i=1}^n A_i(t) = \int_0^t Y(s)d\Lambda(s).$$

The mean and variance for the martingale $M(t)$ can be written as

$$\begin{aligned} &\mathbb{E}M(t) = \mathbb{E}\{\mathbb{E}M(t)|\mathcal{F}_0\} = \mathbb{E}M(0) = 0; \\ &\text{Var}\, M(t) = \mathbb{E}M^2(t) - (\mathbb{E}M(t))^2 = \mathbb{E}M^2(t). \end{aligned}$$

To compute $\mathbb{E}M^2(t)$ (if it exists or the second moment of the process exists), we introduce the variation process of $M(t)$, denoted as $\langle M\rangle(t)$, which helps understand the behavior and characteristics of $M(t)$, particularly in terms of sample paths and changes over time.

2.1 Variation process and covariation process

Square integrable processes are important for analyzing predictable variation and martingale dynamics, and square integrability provides a framework to manage and quantify the variability of stochastic processes.

Definition 2.8. *A stochastic process $X(t)$ is said to be square integrable if for all t,*

$$\mathbb{E}[X^2(t)] < \infty.$$

This means that the second moment of $X(t)$ exists and is finite, which ensures the process does not exhibit unbounded variance over time. Square integrability is particularly important when studying martingales and their associated properties. A key concept related to square integrable martingales is the variation process, which quantifies the accumulation of variability in a martingale. This is formalized in the following definition.

Definition 2.9. *Suppose that $M(t)$ is a square integrable martingale with respect to $\mathcal{F}_t$, $\langle M\rangle(t)$ is called the variation process of $M(t)$ if*

$$M^2(t) - \langle M\rangle(t) \quad \textit{is a martingale with respect to } \mathcal{F}_t.$$

Here, $\langle M\rangle(t)$ is the unique predictable and non-decreasing process with $\langle M\rangle(0) = 0$.

In fact, the existence and uniqueness of such $\langle M\rangle(t)$ is the result of the famous Doob-Meyer decomposition theorem (Meyer 1963).

Theorem 2.10. *(Doob-Meyer decomposition) Let $\mathcal{X}(t)$ be a right continuous (with left hand limits) and integrable process, and is measurable with respect to $\mathcal{F}_t$, satisfying $\mathbb{E}\{\mathcal{X}(t)|\mathcal{F}_s\} \geq \mathcal{X}(s)$ whenever $s \leq t$. Then, there exists a unique, increasing, predictable process $\mathcal{A}(t)$ with respect to $\mathcal{F}_t$, and a right continuous (with left hand limits) martingale process $\mathcal{M}(t)$ with respect to $\mathcal{F}_t$ such that:*

$$\mathcal{X}(t) = \mathcal{M}(t) + \mathcal{A}(t), \quad \textit{for all } t \geq 0,$$

Such defined $\mathcal{X}(t)$ in the theorem is called submartingale in the literature (Meyer 1963). It is easy to verify that the counting process $N(t)$ and also $M^2(t) = (N(t) - A(t))^2$ are submartingales. As opposed to martingales which represent a "fair game," submartingale represent a "favorable game," meaning there is an expected upward tendency over time, given the current information. Readers may refer to a probability text book for the proof of Doob-Meyer decomposition. With the uniquely defined $\langle M\rangle(t)$ when $M(t) = N(t) - A(t)$ is a martingale, it is easy to see that

$$\operatorname{Var} M(t) = \mathbb{E}M^2(t) = \mathbb{E}\langle M\rangle(t). \tag{2.4}$$

Hence, (2.4) implies that $\langle M\rangle(t)$ is an unbiased estimator for $\operatorname{Var} M(t)$.

To compute $\langle M\rangle(t)$, we resort to the infinitesimal characterization of a martingale. We first compute the expectation $\mathbb{E}(dM^2(t)|\mathcal{F}_{t-})$ for a general martingale.

Lemma 2.11. *Suppose that $M(t)$ is a square integrable martingale with respect to $\mathcal{F}_t$, then*

$$\mathbb{E}\big\{dM^2(t)|\mathcal{F}_{t-}\big\} = \mathbb{E}\big\{[dM(t)]^2|\mathcal{F}_{t-}\big\} = \text{Var}\left(dM(t)|\mathcal{F}_{t-}\right).$$

Proof. Recall the definition of $dM^2(t)$ we have

$$dM^2(t) = M^2((t+dt)^-) - M^2(t^-).$$

By separating $M((t+dt)^-)$ into $M((t+dt)^-) - M(t^-) + M(t^-)$, we write $dM^2(t)$ as

$$\begin{aligned} dM^2(t) &= M^2((t+dt)^-) - M^2(t^-) \\ &= \big[M((t+dt)^-) - M(t^-) + M(t^-)\big]^2 - M^2(t^-) \\ &= \big[dM(t) + M(t^-)\big]^2 - M^2(t^-) \\ &= \big[dM(t)\big]^2 + 2M(t^-)dM(t). \end{aligned}$$

Here, the third equality comes from the definition $dM(t) = M((t+dt)^-) - M(t^-)$. Given $\mathcal{F}_{t^-}$, we have

$$\begin{aligned} \mathbb{E}\big\{dM^2(t)|\mathcal{F}_{t-}\big\} &= \mathbb{E}\big\{\big[dM(t)\big]^2 + 2M(t^-)dM(t)|\mathcal{F}_{t-}\big\} \\ &= \mathbb{E}\big\{\big[dM(t)\big]^2|\mathcal{F}_{t-}\big\} + 2M(t^-)\mathbb{E}\big\{dM(t)|\mathcal{F}_{t-}\big\} \\ &= \mathbb{E}\big\{\big[dM(t)\big]^2|\mathcal{F}_{t-}\big\}. \end{aligned}$$

Here, the second equality is by $M(t^-)$ is deterministic given $\mathcal{F}_{t^-}$ and the last equality applies Proposition 2.5 that when $M(t)$ is a martingale with respect to $\mathcal{F}_t$, then $\mathbb{E}\big(dM(t)|\mathcal{F}_{t-}\big) = 0$. $\mathbb{E}\big\{[dM(t)]^2|\mathcal{F}_{t-}\big\} = \text{Var}\left(dM(t)|\mathcal{F}_{t-}\right)$ directly comes from $\mathbb{E}\big(dM(t)|\mathcal{F}_{t-}\big) = 0$. □

By Lemma 2.11, we calculate the value of $\mathbb{E}\big\{dM^2(t)|\mathcal{F}_{t-}\big\}$ when $M(t) = N(t) - \int_0^t Y(s)d\Lambda(s)$. Before applying the lemma, we note that

$$M^2(t) \le 2N(t) + 2\left\{\int_0^t Y(s)d\Lambda(s)\right\}^2 \le 2 + \Lambda^2(t) < \infty,$$

and hence $M^2(t)$ is integrable and the Doob-Meyer decomposition would apply. Now recall $A(t) = \int_0^t Y(s)d\Lambda(s)$. Note that if we condition on $\mathcal{F}_{t^-}$, $Y(t)$ is constant, therefore we have

$$\begin{aligned} \mathbb{E}\big\{dM^2(t)|\mathcal{F}_{t-}\big\} &= \text{Var}\left\{dM(t)|\mathcal{F}_{t-}\right\} \\ &= \text{Var}\left\{dN(t) - Y(t)d\Lambda(t)|\mathcal{F}_{t-}\right\} \\ &= \text{Var}\left\{dN(t)|\mathcal{F}_{t-}\right\} \\ &= \mathbb{E}\big\{dN(t)|\mathcal{F}_{t-}\big\}\big(1 - \mathbb{E}\big\{dN(t)|\mathcal{F}_{t-}\big\}\big) \\ &= Y(t)\lambda(t)dt \cdot (1 - Y(t)\lambda(t)dt) \\ &= Y(t)\lambda(t)dt = dA(t). \end{aligned}$$

Here, the third equality comes from the statement above that $Y(t)$ is a constant conditional on $\mathcal{F}_{t^-}$, the fourth equality comes from the property of Bernoulli distribution and the second-to-last equality is by the property that $(dt)^2 = 0$ (Refer to the property of exterior derivative). We hence have

$$\mathbb{E}\big\{dM^2(t) - dA(t)|\mathcal{F}_{t-}\big\} = 0.$$

Again by Proposition 2.5, we have $M^2(t) - A(t)$ is a martingale with respect to $\mathcal{F}_t$; refer to Doob-Meyer Decomposition for the uniqueness of $A(t)$. We have that $\langle M\rangle(t) = A(t)$. As a remark, with respect to $\mathcal{F}_t$, we have $A(t)$ is a compensator for both $N(t)$ and $M^2(t)$. That is, $M(t) = N(t) - A(t)$ and $M^2(t) - A(t)$ are both martingales. We have the following properties.

Property 2. *Suppose that $M(t)$ is a square integrable martingale and $K(t)$ is a bounded predictable process with respect to $\mathcal{F}_t$, then*

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

   Proof. Homework. Hint: Let $Z(t) = \int_0^t K(s)dM(s)$. We apply the Ito isometry to obtain $\mathbb{E}Z^2(t) = \int_0^t K^2(s)d\langle M\rangle(s) < \infty$. We then apply the local charaterization proposition to $Z(t)$. □

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

   Proof. Homework. Hint: Compute $\mathbb{E}(dZ^2(t)|\mathcal{F}_{t-})$. □

We can also define the covariation process of two martingales.

Definition 2.12. *Suppose that $M_1(t)$ and $M_2(t)$ are two square integrable martingales with respect to $\mathcal{F}_t$, $\langle M_1, M_2\rangle(t)$ is called the variation process of $M_1(t)$ and $M_1(t)$ if*

$$M_1(t)M_2(t) - \langle M_1, M_2\rangle(t) \quad \textit{is a martingale with respect to } \mathcal{F}_t.$$

Here, $\langle M_1, M_2\rangle(t)$ is the unique predictable and right continuous process with $\langle M_1, M_2\rangle(0) = 0$.

The covariation process generalizes the notion of quadratic variation to two different martingales, describing their joint variability over time. If $M_1 = M_2$, then

$$\langle M_1, M_2\rangle(t) = \langle M_1\rangle(t) = \langle M_2\rangle(t),$$

which recovers the standard quadratic variation. In addition, it can be shown that

$$\langle M_1, M_2\rangle(t) = \frac{1}{4}\{\langle M_1 + M_2\rangle(t) - \langle M_1 - M_2\rangle(t)\}.$$

Hence, the existence and uniqueness of the covariation process come from the Doob-Meyer decomposition as both $M_1 + M_2$ and $M_1 - M_2$ are square integrable martingales.

Property 3. *Suppose that $M_1(t), M_2(t)$ are two square integrable martingales and $K_1(t), K_2(t)$ are two bounded predictable processes with respect to $\mathcal{F}_t$, then*

$$\langle\int_0^t K_1(s)dM_1(s), \int_0^t K_2(s)dM_2(s)\rangle = \int_0^t K_1(s)K_2(s)d\langle M_1, M_2\rangle(s)$$

Proof. Homework. □

2.2 Local martingales

As we have seen and shall see, martingales play a central role in the theory of survival analysis. However, many stochastic processes of interest fail to satisfy the strict requirements of martingales over their entire domain. Local martingales become useful, allowing us to generalize martingale properties while retaining their local behavior. The term "local" refers to behavior restricted to a finite interval or under certain conditions. By relaxing global constraints, local properties allow processes to exhibit martingale-like behavior locally, even if they fail to do so globally due to integrability or boundedness constraints. Local properties enhance scalability, enabling the analysis of a process in manageable segments that can be pieced together for a global understanding.

Specifically, local martingales enable the study of processes that locally satisfy martingale properties without requiring global adherence to strict conditions. They generalize martingales, capturing processes that behave like martingales on a local level and thereby broadening the applicability. As many of our applications require boundedness or integrability only on a local scale, the introduction of local martingales allows theorems to be applicable in more general contexts or in a broader range of situations.

To define local martingales, we first define a stopping time is a random variable τ such that the occurrence of the event "time τ" depends only on the information available up to time τ. Formally,

Definition 2.13. *τ is a stopping time with respect to a filtration $\{\mathcal{F}_t\}_{t\geq 0}$ if $\{\tau \leq t\} \in \mathcal{F}_t$ for all $t \geq 0$.*

Stopping times are used to "stop" a process at a random time, preserving its adaptedness to the filtration; see Figure 5 where a process is stopped at two stopping times.

Definition 2.14. *A sequence of stopping times $\{\tau_n\}_{n\geq 1}$ is called a localizing sequence if $\tau_n \uparrow \infty$ as $n \to \infty$ almost surely, and $\tau_n \leq \tau_{n+1}$ for all n.*

Localizing sequences allow us to break a process into intervals where it exhibits desirable properties, such as boundedness or integrability; again see Figure 5.

Definition 2.15. *A process $X(t)$ is locally bounded if there exists a localizing sequence $\{\tau_n\}$ such that the stopped process $X_{\tau_n}(t) = X(t \wedge \tau_n)$ is bounded for each n.*

For a non-random function, it is locally bounded if it is bounded on each finite interval $[0, s], s < \infty$. For example, $f(t) = t$ is locally bounded in $[0, \infty)$, but not bounded in the usual sense. For a stochastic process, local boundedness means that the boundedness would hold on intervals whose right end point is a random time determined by each sample path, which may be easier to satisfy than the original stochastic process.

Definition 2.16. *A process $X(t)$ is locally integrable if there exists a localizing sequence $\{\tau_n\}$ such that $\mathbb{E}[|X(t \wedge \tau_n)|] < \infty$ for all t and n. A process $X(t)$ is locally square integrable if there exists a localizing sequence $\{\tau_n\}$ such that $\mathbb{E}[X^2(t \wedge \tau_n)] < \infty$ for all t and n.*

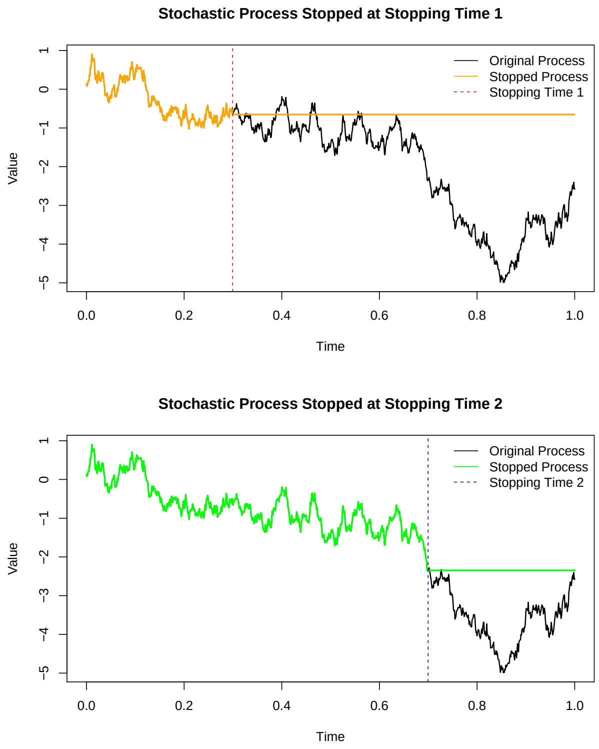


Figure 5: A sample path of Brownian Motion stopped at two stopping times.

Definition 2.17. *A stochastic process $M(t)$ is a local martingale with respect to a filtration $\{\mathcal{F}_t\}_{t\geq 0}$ if it is adapted and continuous with left limits, and if there exists a localizing sequence $\{\tau_n\}$ such that for each n, the stopped process $M_{\tau_n}(t) = M(t \wedge \tau_n)$ is a martingale respect to $\{\mathcal{F}_t\}_{t\geq 0}$.*

Local martingales generalize martingales by allowing the martingale property to hold on intervals defined by stopping times. Every martingale is a local martingale because the martingale property holds globally, and we can trivially choose $\tau_n = n$. However, not every local martingale is a martingale, as a local martingale may fail to satisfy the integrability condition required for martingales over the entire time domain. For example, Brownian motion $B(t)$ is a martingale with respect to $\mathcal{F}_s = \sigma\{B(u), 0 \leq u \leq s\}$, because $\mathbb{E}[B(t)|\mathcal{F}_s] = B_s$ for all $s \leq t$. On the other hand, the process $X(t) = B^3(t)$ is not a martingale, because $\mathbb{E}\{dB^3(s)|\mathcal{F}_{s-}\} = \mathbb{E}\{B^2(s^-)dB(s) + B(s^-)ds|\mathcal{F}_{s-}\} = B(s^-)ds$, which has a nonzero drift. However, it is a local martingale by taking $\tau_n = \sup\{s : |B(s)| \leq n\}$; the stopped process $B^3(t \wedge \tau_n)$, for any $n < \infty$, eliminates issues from the unbounded growth of $B(t)$, making the process a martingale within finite intervals. The following generalizes variation and covariation processes to local (square integrable) martingales.

Definition 2.18. *Suppose that $M(t)$ is a local square integrable martingale with respect to $\mathcal{F}_t$, $\langle M\rangle(t)$ is called the variation process of $M(t)$ if*

$$M^2(t) - \langle M\rangle(t) \quad \textit{is a local martingale with respect to } \mathcal{F}_t.$$

Here, $\langle M\rangle(t)$ is the unique predictable and non-decreasing process with $\langle M\rangle(0) = 0$.

If $\{\tau_n\}_{n=1}^{\infty}$ is the localizing sequence such that the stopped process $M_{\tau_n}(t) = M(t \wedge \tau_n)$ is a square integrable martingale with respect to $\mathcal{F}_t$ for each n, then it follows that, for any t

$$\langle M\rangle(t) = \lim_{n\to\infty} \langle M_{\tau_n}\rangle(t),$$

where the limit is taken pointwise. The uniqueness of $\langle M\rangle(t)$ is established by Theorem 2.2.3 in Fleming & Harrington (2013). Similarly, we can define the covariation process for two local martingales.

Definition 2.19. *Suppose that $M_1(t)$ and $M_2(t)$ are two local square integrable martingales with respect to $\mathcal{F}_t$, $\langle M_1, M_2\rangle(t)$ is called the variation process of $M_1(t)$ and $M_1(t)$ if*

$$M_1(t)M_2(t) - \langle M_1, M_2\rangle(t) \quad \textit{is a local martingale with respect to } \mathcal{F}_t.$$

Here, $\langle M_1, M_2\rangle(t)$ is the unique predictable and right continuous process with $\langle M_1, M_2\rangle(0) = 0$.

We can extend **Properties 2 and 3** to local square integrable martingales.

Property 4. *Let $M(t)$ be a local square integrable martingale and $K(t)$ be a locally bounded predictable process with respect to $\mathcal{F}_t$, then*

1. *$\int_0^t K(s)dM(s)$ is also a local square integrable martingale with respect to $\mathcal{F}_t$.*
2. *$\langle \int_0^t K(s)dM(s)\rangle = \int_0^t K^2(s)d\langle M\rangle(s)$.*
3. *Suppose that $M_1(t), M_2(t)$ are two local square integrable martingales and $K_1(t), K_2(t)$ are two locally bounded predictable processes with respect to $\mathcal{F}_t$, then*

$$\langle \int_0^t K_1(s)dM_1(s), \int_0^t K_2(s)dM_2(s)\rangle = \int_0^t K_1(s)K_2(s)d\langle M_1, M_2\rangle(s).$$

3 Properties of the Nelson-Aalen Estimator: a Martingale Approach

We apply **Property 2** of a martingale to obtain the following results for the Nelson-Aalen estimator.

3.1 Unbiasedness

First, we have that $\widehat{\Lambda}(t) - \Lambda(t) = \int_0^t \frac{dM(s)}{Y(s)}$ is a martingale with respect to $\mathcal{F}_t$, because $Y(t)$ in (2.3) is left continuous and hence is predictable with respect to $\mathcal{F}_t$. In addition, it is bounded above from 1. (Again we ignore the small possibility of $Y(t) = 0$.) Therefore $\mathbb{E}\widehat{\Lambda}(t) - \Lambda(t) = 0$. That is, $\mathbb{E}\widehat{\Lambda}(t) = \Lambda(t)$.

3.2 Variance of the Nelson-Aalen estimate

From (2.4),

$$\operatorname{Var}\widehat{\Lambda}(t) = \mathbb{E}\langle\widehat{\Lambda} - \Lambda\rangle(t) = \mathbb{E}\int_0^t \frac{d\langle M\rangle(s)}{Y^2(s)} = \mathbb{E}\int_0^t \frac{dA(s)}{Y^2(s)} = \mathbb{E}\int_0^t \frac{d\Lambda(s)}{Y(s)}$$

This follows by taking $K(s) = 1/Y(s)$ in **Property 2** and also using the fact that $\langle M\rangle(t) = A(t)$.

This means that $\int_0^t \frac{d\Lambda(s)}{Y(s)}$ is an unbiased "estimator" of $\operatorname{Var}\widehat{\Lambda}(t)$. However, because $\Lambda(s)$ is unknown we replace it by its estimate and assess

$$\int_0^t \frac{d\widehat{\Lambda}(s)}{Y(s)} - \int_0^t \frac{d\Lambda(s)}{Y(s)} = \int_0^t \frac{dN(s)}{Y^2(s)} - \int_0^t \frac{d\Lambda(s)}{Y(s)} = \int_0^t \frac{dM(s)}{Y^2(s)},$$

which is a martingale with respect to $\mathcal{F}_t$ by taking $K(s) = 1/Y^2(s)$ in **Property 2** and ignoring the small possibility of $Y(t) = 0$. Hence,

$$\mathbb{E}\left\{\int_0^t \frac{d\widehat{\Lambda}(s)}{Y(s)} - \int_0^t \frac{d\Lambda(s)}{Y(s)}\right\} = \mathbb{E}\int_0^t \frac{dM(s)}{Y^2(s)} = 0,$$

leading to

$$\int_0^t \frac{dN(s)}{Y^2(s)} = \sum_{t_j \le t} \frac{D_j}{Y_j^2}$$

is an unbiased estimator of $\operatorname{Var}\widehat{\Lambda}(t)$. Hence, we have justified the empirical variance formula for the Nelson-Aalen estimator.

3.3 Consistency

As $n \to \infty$, heuristically, we expect $\mathbb{E}\int_0^t \frac{d\Lambda(s)}{Y(s)} \to 0$ under appropriate conditions (e.g., at the tail of t), which implies $\operatorname{Var}(\widehat{\Lambda}(t)) \to 0$. This suggests that the Markov inequality can be applied to show that, for any $\varepsilon > 0$,

$$\begin{aligned} P\big(|\widehat{\Lambda}(t) - \Lambda(t)| \ge \varepsilon\big) &\le \frac{\mathbb{E}|\widehat{\Lambda}(t) - \Lambda(t)|^2}{\varepsilon^2} \\ &= \frac{\operatorname{Var}(\widehat{\Lambda}(t))}{\varepsilon^2} \to 0. \end{aligned}$$

Hence, $\widehat{\Lambda}(t) \xrightarrow{\text{p}} \Lambda(t)$.

Moreover, by using a valuable tool in the context of stochastic processes, the Lenglart inequality (introduced later as Lemma 4.3), we can rigorously demonstrate that within an interval where subjects have a non-zero probability of being at risk at the end of the interval, uniform consistency holds at every time point within the interval; see Proposition 4.5.

3.4 Asymptotic normality

We show that $\sqrt{n}(\widehat{\Lambda}(t)-\Lambda(t))$ asymptotically follows a normal distribution. To establish this result, we employ the Martingale Central Limit Theorem (MCLT) (Brown 1971). For simplicity and to avoid introducing unnecessary new concepts, we present a simplified version of the theorem.

Theorem 3.1. *(Martingale Central Limit Theorem) Let $\{M_i(t)\}_{i=1}^n$ be independent martingales with respect to $\mathcal{F}_t$, each with a predictable quadratic variation process $\langle M_i \rangle(t)$. Consider*

$$U^n(t) = \sum_{i=1}^n \int_0^t H_i^n(s) dM_i(s).$$

Suppose:

1. *The integrands $H_i^n(t)$ are locally bounded and predictable.*
2. *Quadratic variation: there exists $\alpha(t) > 0$ such that $\langle U^n \rangle(t) \to \alpha(t)$ in probability.*
3. *Big jump tightness: for any $\epsilon > 0$, $\langle U_\epsilon^n \rangle(t) \to 0$ in probability, where*

$$U_\epsilon^n(t) = \sum_{i=1}^n \int_0^t H_i^n(s) I(H_i^n(s) \geq \epsilon) dM_i(s).$$

Then

$$U^n(t) \xrightarrow{d} N(0, \alpha(t)).$$

The condition of "big jump tightness", similar to the Lindeberg condition in the classical CLT, is a stability condition ensuring that when n gets large, there cannot be too many big size jumps. The local boundedness as mentioned in the theorem is as defined in Definition 2.15. In particular, recall that $H(t)$ is locally bounded if there exists a localizing sequence of $\{\tau_n\}_{n=1}^\infty$, such that the stopped process $H^{\tau_n}(t) = H(t \wedge \tau_n)$ is bounded for every n.

We are ready to present the asymptotic normality of the Nelson-Aalen estimator and prove it by showing that it satisfies every condition of MCLT.

Proposition 3.2. *Suppose at a $t > 0$ with $S(t)S_c(t) > 0$, i.e. $P(X_i > t) > 0$ or there is nonzero probability for subjects to be at risk at t. Then*

$$\sqrt{n}(\widehat{\Lambda}(t) - \Lambda(t)) \xrightarrow{d} \mathcal{N}(0, \sigma^2(t)),$$

where $\sigma^2(t) = \int_0^t \frac{d\Lambda(s)}{S(s)S_c(s)}$, with $S(s) = P(T_i \geq s)$ and $S_c(s) = P(C_i \geq s)$.

Proof. Define

$$U^n(t) = \sqrt{n}(\widehat{\Lambda}(t) - \Lambda(t)) = \sqrt{n} \int_0^t \frac{1}{Y(s)} dM(s) = \sum_{i=1}^n \int_0^t \frac{\sqrt{n}}{Y(s)} dM_i(s),$$

where $M_i(t), i = 1, \ldots, n$ are independent martingales. First it is obvious that

$$H_i^n(s) \overset{def}{=} \frac{\sqrt{n}}{Y(s)}, i = 1, \ldots, n$$

is predicable. Take $\tau_n = \sup_{s>0}\{Y(s) \geq 1\} \wedge n$, which leads to $\{\tau_n \leq t\} \in \mathcal{F}_t$ for any $t > 0$ and n, and hence they are stopping times. This follows as (i) when $t \geq n$, $\{\tau_n \leq t\} = \Omega \in \mathcal{F}_t$; (ii) when $t < n$, $\{\tau_n \leq t\} = \{Y(t) = 0\} \in \mathcal{F}_t$. Also, $\tau_n \to \infty$ as $n \to \infty$, because the support of each X_i is $(0, \infty)$. Then the stopped process $H_i^n(\tau_n \wedge t) \leq \sqrt{n}$, and is hence locally bounded.

We next consider the quadratic variation. It follows

$$\langle U^n \rangle(t) = \int_0^t \frac{n}{Y^2(s)} d\langle M \rangle(s) = \int_0^t \frac{\lambda(s)}{Y(s)/n} ds.$$

We may show

$$\frac{Y(s)}{n} \to S(s)S_c(s)$$

in probability uniformly over $[0, t]$ by using empirical process arguments (Shorack & Wellner 1986, van der Vaart & Wellner 1996). Also, because, $\frac{1}{S(s)S_c(s)} \leq \frac{1}{S(t)S_c(t)} < \infty$ when $s \leq t$,

$$\frac{n}{Y(s)} \to \frac{1}{S(s)S_c(s)}$$

in probability uniformly over $[0, t]$ as well. Hence,

$$\int_0^t \frac{\lambda(s)}{Y(s)/n} ds - \int_0^t \frac{\lambda(s)}{S(s)S_c(s)} ds = \int_0^t \left(\frac{n}{Y(s)} - \frac{1}{S(s)S_c(s)} \right) \lambda(s) ds \to 0$$

in probability. Hence,

$$\langle U_n \rangle(t) = \int_0^t \frac{n}{Y^2(s)} d\langle M \rangle(s) \to \sigma^2(t) = \int_0^t \frac{\lambda(s)ds}{S(s)S_c(s)}$$

in probability. Here, $\sigma^2(t)$ corresponds to $\alpha(t)$ in the statement of MCLT.

We then study the big jump control by considering

$$\langle U_\epsilon^n \rangle(t) = \int_0^t \frac{n}{Y^2(s)} I\left(\frac{\sqrt{n}}{Y(s)} \geq \epsilon\right) d\langle M \rangle(s) = \int_0^t \frac{n}{Y(s)} I\left(\frac{\sqrt{n}}{Y(s)} \geq \epsilon\right) \lambda(s) ds$$

As $Y(s)$ is non-increasing,

$$I\left(\frac{\sqrt{n}}{Y(s)} \geq \epsilon\right) \leq I\left(\frac{\sqrt{n}}{Y(t)} \geq \epsilon\right)$$

when $s \leq t$. Hence,

$$\langle U_\epsilon^n \rangle(t) \leq \frac{n}{Y(t)} I\left(\frac{\sqrt{n}}{Y(t)} \geq \epsilon\right) \Lambda(t).$$

We first show $\frac{n}{Y(t)} I\left(\frac{\sqrt{n}}{Y(t)} \geq \epsilon\right) \xrightarrow{p} 0$. To see this, we need to prove for any $\epsilon' > 0$

$$P\left\{\frac{n}{Y(t)} I\left(\frac{\sqrt{n}}{Y(t)} \geq \epsilon\right) > \epsilon'\right\} \to 0. \tag{3.1}$$

As the event of $\{\frac{n}{Y(t)} I\left(\frac{\sqrt{n}}{Y(t)} \geq \epsilon\right) \geq \epsilon'\}$ is equal to $\{\frac{n}{Y(t)} \geq \epsilon'\} \cap \{\frac{\sqrt{n}}{Y(t)} \geq \epsilon\}$, it leads to

$$P\left\{\frac{n}{Y(t)} I\left(\frac{\sqrt{n}}{Y(t)} \geq \epsilon\right) \geq \epsilon'\right\} < P\left\{\frac{\sqrt{n}}{Y(t)} \geq \epsilon\right\} = P\{Y(t) \leq \sqrt{n}/\epsilon\}$$

With the Law of Large Numbers, $Y(t)/n \xrightarrow{p} S(t)Sc(t) > 0$. So for any $\delta > 0$, $P(|Y(t)/n - S(t)S_c(t)| > \delta) \to 0$. In particular, taking $\delta = S(t)Sc(t)/2$, we have $P(Y(t) < nS(t)S_c(t)/2) \to 0$. Now for a given ϵ, when n is sufficiently large, i.e. when $n > \frac{4}{\epsilon^2 S^2(t) S_c^2(t)}$, it follows that

$$\sqrt{n}/\epsilon \leq nS(t)S_c(t)/2,$$

implying

$$P(Y(t) \leq \sqrt{n}/\epsilon) < P(Y(t) \leq nS(t)S_c(t)/2) \to 0.$$

Hence (3.1) holds. As $\Lambda(t) < \infty$ when $t < \infty$, it follows $\langle U_\epsilon^n \rangle(t) \xrightarrow{p} 0$.

With all the conditions satisfied for the Martingale CLT, we thus obtain

$$\sqrt{n}(\widehat{\Lambda}(t) - \Lambda(t)) \xrightarrow{d} N(0, \sigma^2(t)).$$

□

The normality results justify the use of confidence intervals based on the normal distribution. To implement it, we first recall that $P(X_i \geq s) = P(T_i \geq s, C_i \geq s) = S(s)S_c(s)$ given T_i and C_i are independent, and hence $Y(s)/n = \frac{1}{n}\sum_{i=1}^n Y_i(s) \approx S(s)S_c(s)$ by the Law of Large Numbers. So, an estimator for $\sigma^2(t)$ is

$$\int_0^t \frac{d\widehat{\Lambda}(s)}{Y(s)/n} = n\int_0^t \frac{dN(s)}{Y^2(s)} = n\sum_{t_j \leq t} \frac{D_j}{Y_j^2}.$$

3.5 Numerical example: Nelson-Aalen estimator

Consider a toy study with 3 individuals, with the observed survival times (X_i) and event indicators (Δ_i):

Individual	X_i (Time)	Δ_i (Event)
1	2	1
2	3	0
3	4	1

We estimate $\Lambda(t)$ at 4 and give the variance of the estimate. Distinct event times are $t_1 = 2$ and $t_2 = 4$. At each event time, we calculate the number of individuals at risk $(Y(t_j))$ and the number of observed events $(dN(t_j))$:

t_j	$Y(t_j)$	$dN(t_j)$
2	3	1
4	1	1

The Nelson-Aalen estimator is calculated as:

$$\widehat{\Lambda}(t) = \sum_{t_j \leq t} \frac{dN(t_j)}{Y(t_j)}.$$

In particular, at $t = 4$:

$$\widehat{\Lambda}(4) = \frac{1}{3} + \frac{1}{1} \approx 1.333.$$

The variance of the Nelson-Aalen estimator is given by:

$$\text{Var}(\widehat{\Lambda}(t)) = \sum_{t_j \leq t} \frac{dN(t_j)}{Y(t_j)^2}.$$

Using the data:

$$\text{Var}(\widehat{\Lambda}(4)) = \frac{1}{3^2} + \frac{1}{1^2} \approx 1.111.$$

Thus, the estimated variance at $t = 4$ is approximately 1.111.

4 The Kaplan-Meier Estimator

The **Kaplan-Meier estimator**, or **product-limit estimator**, is a non-parametric method for estimating the *survival function* from time-to-event data, commonly used in medical research, reliability engineering, and social sciences. Introduced by *Edward Kaplan* and *Paul Meier* in 1958, it provides a stepwise survival curve that adjusts at observed event times, effectively handling censored data without assuming a specific probability distribution.

Let T be a non-negative random variable representing the survival time, and recall the survival function is defined as:

$$S(t) = P(T > t),$$

which gives the probability that the event has not occurred by time t. The derivation of the estimate of $S(t)$ is best understood when T only takes discrete values. Specifically, assume we observe survival times from n individuals. Define:

- $t_1 < t_2 < \cdots < t_m$: ordered, distinct event times.
- D_j: number of individuals experiencing the event at time t_j.

- Y_j: number of individuals at risk just before time t_j.

Given a t such that $t_{j-1} \leq t < t_j$, we use the conditional probability rule:

$$S(t) = P(T > t) = P(T > t, T > t_{j-1}, \ldots, T > t_1) = P(T > t | T > t_{j-1}) P(T > t_{j-1} | T > t_{j-2}) \cdots P(T > t_1) \tag{4.1}$$

Each conditional probability can be estimated as:

$$P(T > t_j | T > t_{j-1}) = 1 - P(T \leq t_j | T \geq t_j) = 1 - P(T = t_j | T \geq t_j) = 1 - P(X = t_j, \Delta = 1 | X \geq t_j)$$

where $X = T \wedge C, \Delta = I(T \leq C)$, the last equality comes from the independence of T and C, and the last probability can be estimated by

$$\widehat{P}(X = t_j, \Delta = 1 | X \geq t_j) = \frac{D_j}{Y_j}.$$

Thus, based on (4.1), the Kaplan-Meier estimator is given by:

$$\widehat{S}(t) = \prod_{t_j \leq t} \left(1 - \frac{D_j}{Y_j}\right) = \prod_{s \leq t} \left(1 - \frac{dN(s)}{Y(s)}\right).$$

Figure 6 illustrates that the Kaplan-Meier estimator is a step function that is right-continuous with left-hand limits. The survival probability remains constant until the next failure time point, at which it drops.

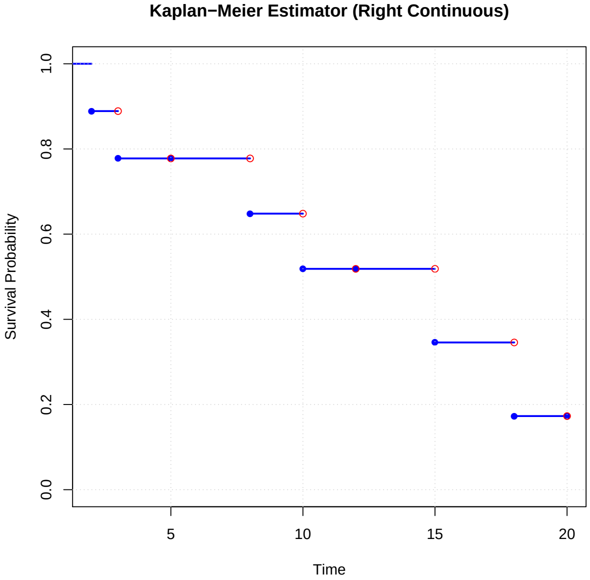


Figure 6: Illustration of the Kaplan-Meier estimator; the plot demonstrates the step-function nature of the estimator, where each step corresponds to an event (death), and censored observations do not lower the steps. Right-continuous survival probabilities are marked with filled blue circles, while left-discontinuities at the end of each step are indicated by open red circles.

Consider the differential

$$\begin{aligned} d\widehat{S}(t) &= \widehat{S}((t+dt)^-) - \widehat{S}(t^-) \\ &= \prod_{s\le (t+dt)^-}\left(1-\frac{dN(s)}{Y(s)}\right) - \prod_{s\le t^-}\left(1-\frac{dN(s)}{Y(s)}\right) \\ &= \prod_{t\le s\le (t+dt)^-}\left(1-\frac{dN(s)}{Y(s)}\right)\prod_{s<t}\left(1-\frac{dN(s)}{Y(s)}\right) - \prod_{s<t}\left(1-\frac{dN(s)}{Y(s)}\right) \\ &= \left(1-\frac{dN(t)}{Y(t)}\right)\prod_{s<t}\left(1-\frac{dN(s)}{Y(s)}\right) - \prod_{s<t}\left(1-\frac{dN(s)}{Y(s)}\right) \\ &= \left(-\frac{dN(t)}{Y(t)}\right)\widehat{S}(t^-) \end{aligned}$$

where the second to last equality holds as $dt \to 0$. Integrating from 0 to t:

$$\widehat{S}(t) = \widehat{S}(0) - \int_0^t \widehat{S}(s^-)\frac{dN(s)}{Y(s)} = 1 - \int_0^t \widehat{S}(s^-)d\widehat{\Lambda}(s) \tag{4.2}$$

as $\widehat{S}(0) = 1$ and $\frac{dN(s)}{Y(s)} = d\widehat{\Lambda}(s)$.

Lemma 4.1. *Let U, V be right-continuous functions of locally bounded variation on $(0, t]$. Then*

$$U(t)V(t) = U(0)V(0) + \int_0^t U(s-)dV(s) + \int_0^t V(s)dU(s). \tag{4.3}$$

Proof. When $s \le t$,

$$U(s) - U(0) = \int_0^s dU(x) = \int_0^t I(x \le s)dU(x),$$

while

$$U(s^-) - U(0) = \int_0^t I(x < s)dU(x).$$

Then

$$\begin{aligned} &(U(t) - U(0))(V(t) - V(0)) \\ =\ & \int_0^t\int_0^t dU(x)dV(y) \\ =\ & \int_0^t\int_0^t I(x<y)dU(x)dV(y) + \int_0^t\int_0^t I(y\le x)dV(y)dU(x) \\ =\ & \int_0^t (U(y^-) - U(0))dV(y) + \int_0^t (V(x) - V(0))dU(x). \end{aligned}$$

The results follow immediately.

□

Proposition 4.2. *If $S(t) > 0$, then*

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

Proof. For a right continuous function $W(s)$,

$$\begin{aligned} d(W(s)^{-1}) &= W((s+ds)^-)^{-1} - W(s^-)^{-1} \\ &= -\frac{W((s+ds)^-) - W(s^-)}{W((s+ds)^-)W(s-)} \\ &= -\frac{dW(s)}{W(s)W(s-)}, \end{aligned}$$

where the last equality holds because of the right continuity. Setting $U(s) = \widehat{S}(s)$, $W(s) = S(s)$, and $V(s) = S(s)^{-1}$, and using (4.3) we obtain:

$$\widehat{S}(t)S(t)^{-1} = \widehat{S}(0)S(0)^{-1} + \int_0^t \widehat{S}(s-)dV(s) + \int_0^t S^{-1}(s)dU(s). \tag{4.4}$$

Then the results follow by using (4.2). □

4.1 Unbiasedness

The proposition established that, when $S(t) > 0$,

$$\widehat{S}(t) - S(t) = -S(t) \int_0^t \frac{\widehat{S}(s-)}{S(s)} \frac{dM(s)}{Y(s)}. \tag{4.5}$$

Immediately, because $\frac{\widehat{S}(s-)}{Y(s)}$ is measurable with respect to $\mathcal{F}_s$ and bounded, hence $\int_0^t \frac{\widehat{S}(s-)}{S(s)} \frac{dM(s)}{Y(s)}$ is a martingale with respect to $\mathcal{F}_t$ by **Property 2**, and has expectation 0. We therefore have

$$\mathbb{E}\widehat{S}(t) = S(t).$$

For simplicity, we have ignored the small possibility (which goes to 0 as $n \to \infty$) that $Y_(s) = 0$ for some $s \in [0, t]$, in which case, Fleming & Harrington (2013) shows the bias goes to 0 at an exponential rate as $n \to \infty$.

4.2 Uniform consistency

We state a version of the Lenglart inequality which will be useful for showing the consistency of the Kaplan-Meier estimator.

Lemma 4.3. *(The Lenglart inequality) Let N be a counting process, and $M = N - A$ the corresponding local square-integrable martingale. Suppose H is an adapted left-continuous process with right-hand limits or, more generally, a predictable and locally bounded process. Then for any finite* *stopping time τ, and any $\varepsilon, \eta > 0$,*

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

Proof. See the proof of Corollary 3.4.1 in Fleming & Harrington (2013). □

Proposition 4.4. *Let T be a failure time random variable with continuous survival function $S(s) = P(T > s)$ and cumulative hazard function $\Lambda(s) = -\int_0^s \frac{dS(v)}{S(v)}$. If $u \in (0, \infty]$ is such that*

$$Y(s) \to \infty \quad \textit{in probability as } n \to \infty,$$

for any $s \le u$, then

$$\sup_{0\le s\le u} |\widehat{S}(s) - S(s)| \to 0 \quad \textit{as } n \to \infty,$$

where $\widehat{S}$ is the Kaplan-Meier estimator.

Proof. Denote by

$$Z(t) = \int_0^t \frac{\widehat{S}(s-)}{S(s)}\frac{dM(s)}{Y(s)}.$$

As $|\widehat{S}(t) - S(t)| \le |Z(t)|$, for any $\epsilon > 0$, we only need to quantify $P\left\{\sup_{0\le s\le u} Z^2(s) > \varepsilon\right\}$. In fact, by the Lenglart inequality, for any $\eta > 0$, it holds that

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

$$< \frac{\eta}{\varepsilon} + P\left\{\frac{\Lambda(u)}{S^2(u)Y(u)} > \eta\right\}.$$

Since $Y(u) \to \infty$ in probability as $n \to \infty$, the second term on the right-hand side above converges to zero as $n \to \infty$ for any $\eta > 0$. Since η and ϵ are arbitrary, the uniformly convergence holds. □

The empirical distribution function is a consistent estimator of an underlying arbitrary cumulative distribution function, uniformly over the entire real line by Theorem 5.5.1 of Chung (1974). The Kaplan-Meier estimator coincides with the empirical cumulative distribution function in the absence of censoring. However, it is unreasonable to expect uniform consistency over the entire real line, regardless of censoring or failure time distributions. For instance, if $P(C_i > t) = 0$ but $S(t) > 0$ for some time t, then there will never be items at risk at or after t, making the survival probability unestimable beyond t. Nevertheless, the theorem guarantees that if there is a nonzero probability of subjects being at risk at a given time, the Kaplan-Meier estimator provides a uniformly consistent estimate of the survival curve up to that point as the sample size increases.

Finally, as we noted before, we can easily prove the uniform consistency of the Nelson-Aalen estimator by using the Lenglart inequality as well.

Proposition 4.5. *Let $\widehat{\Lambda}(t)$ be the Nelson-Aalen estimator, as defined in (1.2), for the cumulative hazard function $\Lambda(t)$. If $u \in (0,\infty]$ is such that*

$$Y(s) \to \infty \quad \textit{in probability as } n \to \infty,$$

for any $s \leq u$, then

$$\sup_{0\leq s\leq u} |\widehat{\Lambda}(s) - \Lambda(s)| \to 0 \quad \textit{as } n \to \infty.$$

Proof. Denote by

$$Z(t) = \widehat{\Lambda}(t) - \Lambda(t) = \int_0^t \frac{dM(s)}{Y(s)}.$$

For any $\epsilon > 0$, we bound $P\left\{\sup_{0\leq s\leq u} Z^2(s) > \varepsilon\right\}$. In fact, by the Lenglart inequality, for any $\eta > 0$, it holds that

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

$$< \frac{\eta}{\varepsilon} + P\left\{\frac{\Lambda(u)}{Y(u)} > \eta\right\}.$$

Since $Y(u) \to \infty$ in probability as $n \to \infty$, the second term on the right-hand side above converges to zero as $n \to \infty$ for any $\eta > 0$. As η and ϵ are arbitrary, the uniformly convergence holds. □

4.3 Weak convergence over a time interval

We now establish the weak convergence of $\widehat{S}$ on a time interval. We state a lemma without proof.

Lemma 4.6. *Denote by $\mathcal{I} = \{t : \pi(t) = P(X_i > t) > 0\}$. Over $\mathcal{I}$, define a process*

$$U^{(n)}(\cdot) = \int_0^{\cdot} H^{(n)}(s)dM(s),$$

where $H^{(n)}$ is a locally bounded predictable process and $M(t) = N(t) - \int_0^t Y(s)d\Lambda(s)$. If there exists a nonnegative function h such that, for any $t \in \mathcal{I}$,

$$\sup_{0\leq s\leq t} \left|\{H^{(n)}(s)\}^2 Y(s) - h(s)\right| \to 0 \quad \textit{as } n \to \infty,$$

then

$$U^{(n)}(\cdot) \xrightarrow{d} Z(\cdot) \quad \textit{in } D[0,t], \quad t \in I,$$

as $n \to \infty$, where $Z(\cdot)$ is a zero-mean Gaussian process with independent increments and variance function of $v(t) = \int_0^t h(s)d\Lambda(s)$, i.e., $Z(t) = B(v(t))$ with $B(\cdot)$ being the Brownian motion, and $D[0,t]$ is the space of functions on $[0,t]$ which are right-continuous with finite left-hand limits.

Proof. See Anderson & Gill (1982) and Fleming & Harrington (2013).

□

Proposition 4.7. *Suppose*

$$\sup_{t\in[0,\infty)} \left| \frac{Y(t)}{n} - \pi(t) \right| \to 0 \tag{4.6}$$

and that B is Brownian Motion. Then, as $n \to \infty$, for any $t \in \mathcal{I}$:

1. $\sqrt{n}(\widehat{S}(\cdot) - S(\cdot)) \xrightarrow{d} S(\cdot)B(v(\cdot))$ *on $D[0,t]$, where $v(t) = \int_0^t \pi^{-1}(s)d\Lambda(s)$, and*

$$Cov\big[S(s)B(v(s)), S(t)B(v(t))\big] = S(s)S(t)v(s \wedge t).$$

2. *Let*

$$\widehat{v}(t) = n \int_0^t \frac{dN(s)}{Y^2(s)}.$$

Then

$$\sup_{0\le s\le t} |\widehat{v}(s) - v(s)| \to 0$$

in probability.

3. $$\sqrt{n}\frac{\widehat{S}(\cdot) - S(\cdot)}{\widehat{S}(\cdot)} \xrightarrow{d} B(v(\cdot))$$

on $D[0,t]$.

4. *Uniform bounds for the above convergence apply, e.g.,*

$$\sup_{0\le s\le t} \left(\frac{n}{\widehat{v}(s)}\right)^{1/2} \left| \frac{\widehat{S}(s) - S(s)}{\widehat{S}(s)} \right| \xrightarrow{d} \sup_{0\le s\le 1} |B(s)| \tag{4.7}$$

Proof. With the notion of (4.5), we can obtain the asymptotic distribution of the process

$$\left\{ \int_0^s H^n(u)dM(u) : 0 < s < t \right\},$$

where

$$H^{(n)}(s) = \sqrt{n}\frac{\widehat{S}(s-)}{S(s)Y(s)},$$

and

$$M(s) = N(s) - \int_0^s Y(u)d\Lambda(u).$$

To apply Lemma 4.6, we set $h(t) = \pi^{-1}(t)$, and show

$$\sup_{0\le s\le t} \left| n\frac{S^2(s-)}{S^2(s)Y(s)} - \pi^{-1}(t) \right| \to 0$$

in probability. In fact,

$$\begin{aligned}
& \sup_{0\le s\le t}\left|n\frac{\widehat{S}^2(s-)}{S^2(s)Y(s)}-\pi^{-1}(t)\right| \\
= & \sup_{0\le s\le t}\left|n\frac{\widehat{S}^2(s-)}{S^2(s)Y(s)}-\frac{1}{\pi(s)}\frac{\widehat{S}^2(s-)}{S^2(s)}+\frac{1}{\pi(s)}\frac{\widehat{S}^2(s-)}{S^2(s)}-\pi^{-1}(s)\right| \\
\le & \sup_{0\le s\le t}\frac{\widehat{S}^2(s-)}{S^2(s)}\left|\frac{n}{Y(s)}-\frac{1}{\pi(s)}\right|+\sup_{0\le s\le t}\frac{1}{\pi(s)S^2(s)}|\widehat{S}^2(s-)-S^2(s)| \\
\le & \frac{1}{S^2(t)}\sup_{0\le s\le t}\left|\frac{n}{Y(s)}-\frac{1}{\pi(s)}\right|+\frac{2}{\pi(t)S^2(t)}\sup_{0\le s\le t}|\widehat{S}(s-)-S(s)|
\end{aligned}$$

Hence, the claim is satisfied by the assumption of (4.6) and the uniform consistency of the Kaplan-Meier estimator as established in Proposition 4.4.

We then study

$$\begin{aligned}
\widehat{v}(s)-v(s) &= \int_0^s n\frac{dN(u)}{Y^2(u)}-\int_0^s\frac{d\Lambda(u)}{\pi(u)} \\
&= \int_0^s n\frac{dN(u)}{Y^2(u)}-\int_0^s n\frac{d\Lambda(u)}{Y(u)}+\int_0^s n\frac{d\Lambda(u)}{Y(u)}-\int_0^s\frac{d\Lambda(u)}{\pi(u)} \\
&= \int_0^s n\frac{dM(u)}{Y^2(u)}+\int_0^s\left(\frac{n}{Y(u)}-\frac{1}{\pi(u)}\right)d\Lambda(u) \qquad (4.8)
\end{aligned}$$

Consider the uniform convergence of the first term of (4.8). For any $\epsilon,\eta>0$, the Lenglart inequality gives

$$\begin{aligned}
P\left(\sup_{0\le s\le t}\left|\int_0^s n\frac{dM(u)}{Y^2(u)}\right|^2>\epsilon\right) &< \frac{\eta}{\epsilon}+P\left(\int_0^t\frac{n^2}{Y^4(s)}d\langle M\rangle(s)>\eta\right) \\
&< \frac{\eta}{\epsilon}+P\left(\int_0^t\frac{n^2}{Y^3(s)}d\Lambda(s)>\eta\right) \\
&< \frac{\eta}{\epsilon}+P\left(\frac{n^2\Lambda(t)}{Y^2(t)}\frac{1}{Y(t)}>\eta\right)
\end{aligned}$$

As $Y(t)/n\to\pi(t)>0$ in probability and, hence, $Y(t)\to\infty$ and $\frac{n^2}{Y^2(t)}\to\frac{1}{\pi^2(t)}>0$ in probability as $n\to\infty$, therefore, the probability of the event of

$$\frac{n^2\Lambda(t)}{Y^2(t)}\frac{1}{Y(t)}>\eta$$

goes to 0. Because η is arbitrary, hence,

$$P\left(\sup_{0\le s\le t}\left|\int_0^s n\frac{dM(u)}{Y^2(u)}\right|^2>\epsilon\right)\to 0.$$

On the other hand, the uniform convergence of

$$\frac{n}{Y(u)} - \frac{1}{\pi(u)}$$

over $[0, t]$ implies the uniform convergence to 0 of the second term of (4.8).

Part 3 of the theorem comes from Part 1 and the uniform convergence of $\widehat{S}$. Part 4 comes from the Continuous Mapping Theorem. □

4.4 Confidence intervals versus confidence bands

The weak convergence results allow the construction of confidence intervals and bands for S. For large n, point-wise $(1 - \alpha) \times 100/\%$ approximate confidence intervals for survival estimates at s can be given by:

$$\left[\widehat{S}(s) \pm z_{1-\alpha/2} \widehat{S}(s) \sqrt{\frac{\widehat{v}(s)}{n}} \right],$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2) \times 100$ percentile of the standard normal distribution, e.g. $z_{0.975} \approx 1.96$. On the other hand, the confidence band, for $s \in [0, t]$, can be obtained as:

$$\left[\widehat{S}(s) \pm c \widehat{S}(s) \sqrt{\frac{\widehat{v}(s)}{n}} \right], \tag{4.9}$$

where c is a constant chosen to ensure the desired simultaneous coverage probability (e.g., 95%) using the weak convergence result of (4.7). Specifically, c is chosen based on the distribution of the supremum of $B(t))$ over $[0, 1]$ to ensure the simultaneous coverage probability meets the desired level.

To see why (4.9) is the desired confidence band, we aim to show that (4.9) covers $S(s)$ over $[0, t]$ with probability (approximately) at least $1 - \alpha$, or more explicitly,

$$P\left(S(s) \in \left[\widehat{S}(s) \left(1 - c_{1-\alpha} \left(\frac{\widehat{v}(s)}{n} \right)^{1/2} \right), \widehat{S}(s) \left(1 + c_{1-\alpha} \left(\frac{\widehat{v}(s)}{n} \right)^{1/2} \right) \right], \forall s \in [0, t] \right) \geq 1 - \alpha. \tag{4.10}$$

From Proposition 4.7 (Part 4), we have:

$$P\left(\sup_{0 \leq s \leq t} \left(\frac{n}{\widehat{v}(s)} \right)^{1/2} \left| \frac{\widehat{S}(s) - S(s)}{\widehat{S}(s)} \right| \leq c_{1-\alpha} \right) \to P\left(\sup_{0 \leq s \leq 1} |B(s)| \leq c_{1-\alpha} \right) = 1 - \alpha. \tag{4.11}$$

Or, when n is sufficiently large,

$$P\left(\sup_{0 \leq s \leq t} \left(\frac{n}{\widehat{v}(s)} \right)^{1/2} \left| \frac{\widehat{S}(s) - S(s)}{\widehat{S}(s)} \right| \leq c_{1-\alpha} \right) \approx 1 - \alpha.$$

This means

$$P\left(\left(\frac{n}{\widehat{v}(s)}\right)^{1/2}\left|\frac{\widehat{S}(s)-S(s)}{\widehat{S}(s)}\right| \leq c_{1-\alpha}, \forall s\in[0,t]\right) \geq P\left(\sup_{0\leq s\leq t}\left(\frac{n}{\widehat{v}(s)}\right)^{1/2}\left|\frac{\widehat{S}(s)-S(s)}{\widehat{S}(s)}\right| \leq c_{1-\alpha}\right) \approx 1-\alpha,$$

meaning that with probability (approximately) at least $1-\alpha$,

$$\left|\frac{\widehat{S}(s)-S(s)}{\widehat{S}(s)}\right| \leq c_{1-\alpha}\left(\frac{\widehat{v}(s)}{n}\right)^{1/2}, \quad \forall s\in[0,t].$$

Rearranging the bound, this means (4.10) holds or (4.9) covers $S(s)$ for all $s \in [0,t]$ with probability (approximately) at least $1-\alpha$.

To select $c_{1-\alpha}$, one can apply the formula derived by Billingsley (2013):

$$P\left(\sup_{0\leq s\leq 1}|B(s)| \leq y\right) = 4\sum_{k=0}^{\infty}\frac{(-1)^k}{\pi(2k+1)}\exp\left(-\frac{\pi^2(2k+1)^2}{8y^2}\right).$$

Now, if we set

$$4\sum_{k=0}^{\infty}\frac{(-1)^k}{\pi(2k+1)}\exp\left(-\frac{\pi^2(2k+1)^2}{8c^2}\right) = 0.95$$

and solve for c numerically, we find the c satisfying

$$P\left(\sup_{0\leq s\leq 1}|B(s)| \leq c_{0.95}\right) = 0.95$$

is $c_{0.95} \approx 2.241$. Note that $c_{0.95} > z_{0.975} \approx 1.96$, meaning that confidence bands are wider than pointwise confidence intervals. This makes sense because confidence bands account for **simultaneous uncertainty** across an entire range of values, whereas pointwise confidence intervals only provide coverage at individual points. In particular, when estimating a survival function $S(t)$, a **pointwise** 95% confidence interval means that at a fixed t, the true $S(t)$ will fall within the interval 95% of the time if we repeat the procedure. However, a **confidence band** at the same confidence level ensures that the **entire** survival function lies within the band with 95% probability. Since this requires controlling the error across all t values, the resulting confidence bands are necessarily wider than the pointwise confidence intervals; see, for example, Figure 7.

Finally, perhaps a more straightforward approach to identify $c_{1-\alpha}$ is to determine the distribution of the supremum of $B(t)$ via simulations:

1. (simulate Brownian motion): We simulate n independent realizations of a Brownian motion process $B(t)$ over the interval $[0,1]$.
2. (calculate the supremum): For each simulated path, compute the supremum (maximum value) of the Brownian motion over the interval $[0,1]$.
3. (quantile selection): Based on the desired simultaneous coverage probability $1-\alpha$ (e.g., 95%), we determine the quantile from the distribution of the supremum values. This quantile will be our constant $c_{1-\alpha}$.

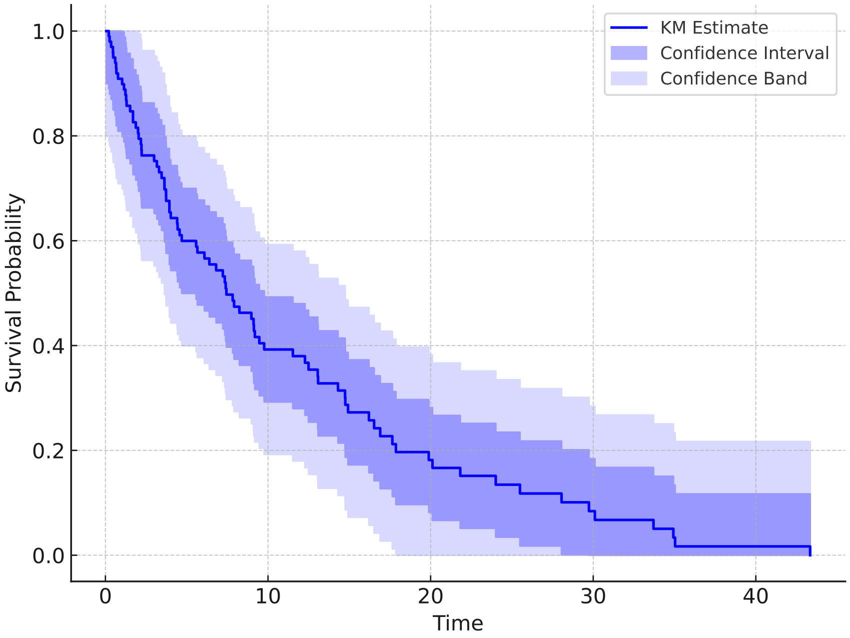


Figure 7: Kaplan-Meier survival curve with confidence intervals and confidence bands. The confidence intervals (shaded darker) are narrower, compared to confidence bands (lighter shading) which account for greater uncertainty across all time points.

We typically run simulations with a large number of sample paths (e.g., 100,000 simulations, each with 10,000 equally spaced time points over the interval $[0, 1]$), and compute the supremum of the absolute values for each path. The 95% quantile of the supremum values gives us the value of $c_{0.95}$. Our simulation returned a value of 2.243, which closely matches the theoretical value of 2.241. The following R code simulates Brownian motion paths, calculates the supremum for each path, and computes the 95% quantile (or any other desired quantile) of the supremum distribution.

```
# Set parameters
n_simulations <- 100000  # Number of Brownian motion paths
n_points <- 10000        # Number of time points in each path (for interval [0,1])
alpha <- 0.05            # Significance level for 95% quantile

# Function to simulate one Brownian motion path
simulate_brownian_motion <- function(n_points) {
  dt <- 1 / (n_points - 1)
  # Simulate Brownian motion increments (increment is normally distributed)
  increments <- sqrt(dt) * rnorm(n_points - 1)
  B_t <- c(0, cumsum(increments))  # Brownian motion with starting value 0
  return(B_t)
}

# Simulate multiple Brownian motion paths and calculate supremums
supremums <- numeric(n_simulations)  # Pre-allocate memory for supremum values
for (i in 1:n_simulations) {
  B_t <- simulate_brownian_motion(n_points)
  supremums[i] <- max(abs(B_t))  # Calculate the supremum (max absolute value)
}

# Calculate the quantile for the supremum distribution
c <- quantile(supremums, 1 - alpha)

# Output the result
cat("The␣value␣of␣c␣for␣the␣95%␣quantile␣is:", c, "\n")
```

5 Nonparametric Comparison of Hazard Functions

It is often necessary to compare the effectiveness of two treatment arms—typically an experimental treatment group and a control (placebo) group. The comparisons help assess whether a new treatment significantly improves survival outcomes compared to standard care or no treatment. More specifically, we aim to evaluate differences in survival distributions and determine whether the treatment has a statistically significant impact on the time to event (e.g., death or disease progression). For independent individuals $i = 1, \ldots, n$, we introduce a binary indicator Z_i, where $Z_i = 0$ if individual i belongs to the placebo arm and $Z_i = 1$ if they are in the treatment arm. We extract the distinct observed failure time points, $t_1 < t_2 < \ldots < t_{n_d}$, from these n subjects. Let D_{kj} and Y_{kj} denote, respectively, the number of observed deaths and the number of subjects at risk at time t_j for group $k = 0$ (placebo group) and 1 (treatment group). The total number of observed deaths and the total number of subjects at risk at time t_j across both groups are given by $D_j = D_{0j} + D_{1j}, \quad Y_j = Y_{0j} + Y_{1j}$. Since Z_i indicates treatment assignment, we can express the number of deaths and subjects at risk in the treatment group as

$$D_{1j} = \sum_{i=1}^{n} Z_i dN_i(t_j), \quad Y_{1j} = \sum_{i=1}^{n} Z_i Y_i(t_j),$$

where $dN_i(t_j)$ represents the increment in the counting process for individual i at time t_j, and $Y_i(t_j)$ is the at-risk indicator for individual i at time t_j.

5.1 The log-rank test

The log-rank test (Mantel 1966) is to test the null hypothesis, $H_0 : \lambda_0(t) = \lambda_1(t)$ for all $t > 0$, where $\lambda_k(t)$ is the hazard function for patients in group $k = 0, 1$, or plainly, both treatment arms have the same hazard function. That is, under H_0, the true survival time T_i, regardless of the value of Z_i, has a hazard of $\lambda_0(t) = \lambda_1(t) \overset{def}{=} \lambda(t)$.

We have shown that the numerator of the log-rank test statistic is

$$U_L = \sum_{j=1}^{n_d} (D_{1j} - D_j Y_{1j} / Y_j)$$

which can be expressed, by using the counting process notation, as

$$U_L = \sum_{j=1}^{n_d}\sum_{i=1}^{n}(Z_i - \bar{Z}(t_j))dN_i(t_j) = \sum_{i=1}^{n}\int_0^\infty (Z_i - \bar{Z}(s))dN_i(s)$$

where $\bar{Z}(s) = \frac{\sum_{i=1}^n Z_i Y_i(s)}{\sum_{i=1}^n Y_i(s)}$, or the proportion of the at-risk population at s from group 1. We now derive the variance of U_L under the null hypothesis of H_0. Indeed, if H_0 holds, some algebra will yield

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

where $M_i(t) = N_i(t) - \int_0^t Y_i(s)\lambda(s)ds$ is a martingale with respect to

$$\mathcal{F}_t = \sigma\{N_i(s), Y_i(s), Z_i, i = 1, ..., n, 0 \leq s \leq t\}.$$

Compared to the filtration defined in Section 2, this updated filtration is enriched by incorporating the treatment assignment information, Z_i.

Now consider the process

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

We note $U_L = U(\infty)$ or $U_L = \lim_{t\to\infty} U(t)$, where the limit is defined for each sample point in the sample space. We first establish the property of $U(t)$ before studying U_L.

Proposition 5.1. *$U(t)$ is a martingale with respect to $\mathcal{F}_t$.*

Proof. To prove that $U(t)$ is a martingale, we proceed in three steps: (1) show that $U(t)$ is adapted, (2) demonstrate that $U(t)$ is integrable, and (3) verify that it satisfies the martingale property using the infinitesimal characterization.

Step 1: we first show that $U(t)$ is adapted. Recall $M_i(t) = N_i(t) - \int_0^t Y_i(s)\lambda(s)ds$. We then consider the adaptness of

$$\int_0^t (Z_i - \bar{Z}(s))\, dM_i(s) = \sum_{s\leq t}(Z_i - \bar{Z}(s))\,(N_i((s+ds)^-) - N_i(s^-)) - \int_0^t (Z_i - \bar{Z}(s))Y_i(s)d\Lambda(s).$$

For any $s \leq t$, the term Z_i is adapted to $\mathcal{F}_t$, while the term $\bar{Z}(s) = \frac{\sum_{i=1}^n Z_i Y_i(s)}{\sum_{i=1}^n Y_i(s)}$ depends on $Y_i(s)$, which is measurable with respect to $\mathcal{F}_s$ and therefore with respect to $\mathcal{F}_t$. Hence, $\int_0^t (Z_i - \bar{Z}(s))Y_i(s)d\Lambda(s)$ is measurable with respect to $\mathcal{F}_t$. This is natural as the trajectory of the integrand is deterministic given $\mathcal{F}_t$. Moreover, $\bar{Z}(s^-), N_i((s+ds)^-), N_i(s^-)$ are all measurable with respect to to $\mathcal{F}_{(s+ds)^-}$ and hence with respect to $\mathcal{F}_t$. Therefore, so is the countable summation, $\sum_{s\leq t}(Z_i - \bar{Z}(s))\,(N_i((s+ds)^-) - N_i(s^-))$, with respect to to $\mathcal{F}_t$. In this case as N_i can jump at most once, the sum is either 0 or one term. As each stochastic integral $\int_0^t (Z_i - \bar{Z}(s))\, dM_i(s)$ is adapted to $\mathcal{F}_t$, their sum $U(t)$ is also adapted, i.e., $U(t)$ is $\mathcal{F}_t$-adapted. Here we have repeatedly used the fact that the countable summation of measurable functions is also measurable.

Step 2: we show that $U(t)$ is integrable by showing $\mathbb{E}[|U(t)|] < \infty$ for any $t < \infty$. Indeed, as $|Z_i - \bar{Z}(s)| < 2$, it follows

$$|\int_0^t (Z_i - \bar{Z}(s))dM_i(s)| \leq 2\int_0^t |dM_i(s)| \leq 2N_i(t) + 2\int_0^t Y_i(s)d\Lambda(s)$$

whose expectation is less than 4 by (2.2). Hence, $\mathbb{E}[|U(t)|] < 4n < \infty$ for any $t < \infty$, concluding that $U(t)$ is integrable.

Step 3: we verify the "fair game" property by using the infinitesimal characterization and showing

$$\mathbb{E}[dU(t) \mid \mathcal{F}_{t^-}] = 0,$$

where $\mathcal{F}_{t^-}$ is the filtration just before time t.

In fact, by definition,

$$dU(t) = \sum_{i=1}^{n} (Z_i - \bar{Z}(t))\, dM_i(t).$$

Since $Z_i - \bar{Z}(t)$ is predictable and $M_i(t)$ is a martingale, the increments $dM_i(t)$ satisfy $\mathbb{E}[dM_i(t) \mid \mathcal{F}_{t^-}] = 0$. Substituting $dU(t)$ into the conditional expectation:

$$\mathbb{E}[dU(t) \mid \mathcal{F}_{t^-}] = \mathbb{E}\left[\sum_{i=1}^{n} (Z_i - \bar{Z}(t))\, dM_i(t) \mid \mathcal{F}_{t^-}\right].$$

Since $Z_i - \bar{Z}(t)$ is predictable with respect to $\mathcal{F}_t$ and therefore adapted to $\mathcal{F}_{t^-}$, it follows that:

$$\mathbb{E}[dU(t) \mid \mathcal{F}_{t^-}] = \sum_{i=1}^{n} (Z_i - \bar{Z}(t)) \cdot \mathbb{E}[dM_i(t) \mid \mathcal{F}_{t^-}] = 0.$$

Thus, $U(t)$ satisfies the martingale property. Combining this with its adaptedness and integrability, we conclude that $U(t)$ is a martingale. □

Immediately, we can conclude $\mathbb{E}U(t) = 0$ for all t because $U(0) = 0$ and $U(t)$ is a martingale. We then prove the log-rank test statistic has mean 0 under the null by using the dominated convergence theorem (DCT) stated below.

Theorem 5.2. *(Dominated convergence theorem) Let $(X_k)_{k\geq 1}$ be a sequence of random variables such that (i) $X_k \xrightarrow{a.s.} X$ (almost surely) as $k \to \infty$; (ii) there exists an integrable random variable Y (i.e., $\mathbb{E}[|Y|] < \infty$) such that $|X_k| \leq Y$ for all $k \geq 1$ almost surely. Then, X is integrable (i.e., $\mathbb{E}[|X|] < \infty$), and*

$$\lim_{k\to\infty} \mathbb{E}[X_k] = \mathbb{E}[X].$$

The proof can be found in the probability text books. Applying DCT, we have the result of the unbiasedness of the log-rank test.

Proposition 5.3. $\mathbb{E}U_L = 0$. *That is, the expectation of the log-rank test statistic is 0 under the null hypothesis.*

Proof. Consider any sequence of $t_k \to \infty$ as $k \to \infty$ and the sequence of random variables, $\{U(t_k)\}_{k\geq 1}$. To apply DCT, we verify its required conditions:

(i) (pointwise convergence) By definition, $U(t_k) \to U_L$ pointwise and hence almost surely.

(ii)(integrable dominance) For each $i = \{1, \dots, n\}$ and any $t_k < \infty$, it follows that

$$\left| \int_0^{t_k} (Z_i - \bar{Z}(s))\, dM_i(s) \right| \leq 2 \int_0^{\infty} |dM_i(s)| \overset{def}{=} G_i.$$

Similar to what we have shown in (2.2), we can show $\mathbb{E}\int_0^\infty |dM_i(s)| < 4$ and hence $\mathbb{E}G_i \leq 8$. Therefore, we have identified a dominating random variable $G \overset{def}{=} \sum_{i=1}^n G_i$ such that $\mathbb{E}G \leq 8n$ is uniformly bounded (with respect to k) and

$$|U(t_k)| = \left| \sum_{i=1}^n \int_0^{t_k} (Z_i - \bar{Z}(s))\, dM_i(s) \right| \leq \sum_{i=1}^n \left| \int_0^{t_k} (Z_i - \bar{Z}(s))\, dM_i(s) \right| \leq \sum_{i=1}^n G_i = G$$

for any k. Then DCT implies that

$$\mathbb{E}[U_L] = \mathbb{E}\left[\lim_{k\to\infty} U(t_k)\right] = \lim_{k\to\infty} \mathbb{E}[U(t_k)] = 0.$$

□

We now study the variance of $U(t)$. Because $|Z_i - \bar{Z}(s)| \leq 2$, $\int_0^t (Z_i - \bar{Z}(s))dM_i(s)$ is square integrable by **Property 2**. So $U(t)$ is square integrable as

$$U^2(t) \leq 2\sum_{i=1}^n \left\{ \int_0^t (Z_i - \bar{Z}(s))dM_i(s) \right\}^2.$$

Further, as subjects are independent, we have the following result.

Lemma 5.4. $\mathbb{E}(dM_i(s)dM_j(s)|\mathcal{F}_{s^-}) = 0$ *when* $i \neq j$.

Proof. Exercise. □

With this, the following gives the variation process of $U(t)$.

Proposition 5.5. $\langle U\rangle(t) = \sum_{i=1}^n \int_0^t (Z_i - \bar{Z}(s))^2 Y_i(s) d\Lambda(s)$,

Proof. Applying Lemmas 2.11 and 8.6, we can show

$$\mathbb{E}(dU^2(t)|\mathcal{F}_{t^-}) = \sum_{i=1}^n (Z_i - \bar{Z}(t))^2 d\langle M_i\rangle(t),$$

where $d\langle M_i\rangle(t) = Y_i(t)d\Lambda(t)$. □

With $\langle U\rangle(t) = \sum_{i=1}^n \int_0^t (Z_i - \bar{Z}(s))^2 Y_i(s)d\Lambda(s)$, it follows

$$\operatorname{Var}(U(t)) = \mathbb{E}\sum_{i=1}^n \int_0^t (Z_i - \bar{Z}(s))^2 Y_i(s)d\Lambda(s)$$

which can be estimated by

$$\sum_{i=1}^n \int_0^t (Z_i - \bar{Z}(s))^2 Y_i(s)d\widehat{\Lambda}(s) = \int_0^t \sum_{i=1}^n (Z_i - \bar{Z}(s))^2 Y_i(s)\frac{dN(s)}{Y(s)}.$$

This can be shown to be equal to

$$\sum_{j:t_j\le t} \frac{Y_{1j}Y_{0j}}{Y_j^2} D_j,$$

which follows because $\sum_{i=1}^n (Z_i - \bar{Z}(s))^2 Y_i(s) = \sum_{i=1}^n Z_i Y_i(s) - (\sum_{i=1}^n Z_i Y_i(s))^2/Y(s)$, which is equal to $\frac{Y_{1j}Y_{0j}}{Y_j}$ when evaluated at t_j.

Considering $t \to \infty$ in $U(t)$, applying DCT may give

Proposition 5.6. *Suppose* $\mathbb{E}\Lambda^2(X_i) < \infty, i = 1, \ldots, n$. *Then*

$$\operatorname{Var} U_L = \sum_{i=1}^n \mathbb{E}\int_0^\infty (Z_i - \bar{Z}(s))^2 Y_i(s)d\Lambda(s).$$

Proof. As $\mathbb{E}U(t) = \mathbb{E}U_L = 0$, we only need to consider any sequence of $t_k \to \infty$ as $k \to \infty$ and the sequence of random variables, $\{U(t_k)\}_{k\ge1}$, and show

$$\mathbb{E}U_L^2 = \lim_{k\to\infty} \mathbb{E}U^2(t_k).$$

This follows by applying DCT. In particular, we note

$$\begin{aligned} |U(t_k)| &= \left|\sum_{i=1}^n \int_0^{t_k} (Z_i - \bar{Z}(s))\, dM_i(s)\right| \\ &\le \sum_{i=1}^n \left|\int_0^{t_k} (Z_i - \bar{Z}(s))\,(dN_i(s) - Y_i(s)d\Lambda(s))\right| \\ &< 2\sum_{i=1}^n (N_i(t_k) + \Lambda(X_i)) < 2\sum_{i=1}^n (1 + \Lambda(X_i)). \end{aligned}$$

Therefore, $U^2(t_k) \le 16\sum_{i=1}^n (1 + \Lambda^2(X_i)) \overset{def}{=} G$. Given the condition of $\mathbb{E}\Lambda^2(X_i) < \infty, i = 1, \ldots, n$, it follows that G is integrable. Hence we can apply DCT and conclude

$$\operatorname{Var} U_L = \mathbb{E}U_L^2 = \lim_{k\to\infty} \mathbb{E}U^2(t_k) = \lim_{k\to\infty} \operatorname{Var} U(t_k) = \sum_{i=1}^n \mathbb{E}\int_0^\infty (Z_i - \bar{Z}(s))^2 Y_i(s)d\Lambda(s).$$

□

We impose a sufficient condition $\mathbb{E}\Lambda^2(X_i) < \infty$, i.e., the transformed observed survival time $\Lambda(X_i)$ has finite variability, to prevent excessively heavy tails in the distribution of X_i.

Thus, we may estimate $\mathrm{Var}(U_L)$ with

$$\sum_{j=1}^{n_d} \frac{Y_{1j}Y_{0j}}{Y_j^2} D_j,$$

which justifies the use of the variance formula for the log-rank test. Note in the counting process, we do not allow ties, and hence $D_j = 1$ for all j.

5.2 The weighted log-rank test

The log-rank test is most powerful under the proportional hazards assumption, which assumes a constant hazard ratio. When this assumption fails, the weighted log-rank test extends the method by emphasizing early, middle, or late differences with tailored weight functions, enhancing sensitivity to time-specific survival differences. We consider the weighted log-rank test (Harrington & Fleming 1982), in the form of

$$U_W = \sum_{j=1}^{n_d} W_j(D_{1j} - D_j Y_{1j}/Y_j).$$

Suppose $W(s)$ is predictable and $W(t_j) = W_j$. Then we can express U_W, by using the counting process notation, as

$$\begin{aligned} U_W &= \sum_{j=1}^{n_d}\sum_{i=1}^{n} W_j(Z_i - \bar{Z}(t_j))dN_i(t_j) \\ &= \sum_{i=1}^{n}\int_0^\infty W(s)(Z_i - \bar{Z}(s))dN_i(s) \\ &= \sum_{i=1}^{n}\int_0^\infty W(s)(Z_i - \bar{Z}(s))dM_i(s). \end{aligned}$$

Then similarly, under the null, we can show $\sum_{i=1}^n \int_0^t W(s)(Z_i - \bar{Z}(s))dM_i(s)$ is a martingale with respect to $\mathcal{F}_t$ as both $W(s)$ and $\bar{Z}(s)$ are predictable. Further, under the null, it follows that $\mathbb{E}U_W = 0$ and

$$\mathrm{Var}\, U_W = \mathbb{E}\sum_{i=1}^{n}\int_0^\infty W^2(s)(Z_i - \bar{Z}(s))^2 Y_i(s) d\Lambda(s),$$

which can be estimated by

$$\sum_{j=1}^{n_d} \frac{W_j^2 Y_{1j}Y_{0j}}{Y_j^2} D_j.$$

In particular, for the Wilcoxon test (Peto & Peto 1972), where $W(s) = Y(s)$, and hence $W_j = Y_j$, the variance can be estimated by $\sum_{j=1}^{n_d} Y_{1j} Y_{0j} D_j$.

While similar in nature, the log-rank and Wilcoxon tests are suited to different application scenarios. The log-rank test is most sensitive to differences in hazard functions that are proportional over time, performing well when the hazard ratio remains constant and the proportional hazards assumption holds. The Wilcoxon test, as a special case of the weighted log-rank test, applies weights based on the number of individuals at risk, placing greater emphasis on earlier time periods. This makes it more sensitive to survival differences that occur early in the study and potentially more effective in detecting deviations from proportional hazards, such as when treatment effects diminish or intensify over time. More broadly, by selecting different weights, the weighted log-rank test can be adapted to emphasize specific time periods, allowing it to detect survival differences that align with particular patterns, including early effects, late effects, or proportional hazards throughout the follow-up period. This makes them particularly valuable in clinical trials where treatment effects vary over time, such as delayed benefits in immunotherapy.

5.3 Numerical example: Log-Rank and Wilcoxon tests

Consider a study with two groups, Group 0 (placebo) and Group 1 (treatment). The observed survival times (X_i) and event indicators (Δ_i) are as follows (with no ties in failure times):

Individual	Group	X_i (Time)	Δ_i (Event)
1	0	2	1
2	0	5	1
3	1	3	1
4	1	6	1

As the first step, we extract failure times. Here, the distinct observed failure times are $t_1 = 2$, $t_2 = 3$, $t_3 = 5$, and $t_4 = 6$. For each t_j, calculate the number of subjects at risk (Y_{kj}) and the number of events (D_{kj}) for each group ($k = 0, 1$):

t_j	Y_{0j}	D_{0j}	Y_{1j}	D_{1j}	$Y_j = Y_{0j} + Y_{1j}$	$D_j = D_{0j} + D_{1j}$
2	2	1	2	0	4	1
3	1	0	2	1	3	1
5	1	1	1	0	2	1
6	0	0	1	1	1	1

We then calculate the log-rank test statistic. The numerator of the log-rank test is:

$$U_L = \sum_{j=1}^{n_d} \left(D_{1j} - D_j \frac{Y_{1j}}{Y_j} \right).$$

For each t_j:

t_j	D_{1j}	$D_j \frac{Y_{1j}}{Y_j}$	$D_{1j} - D_j \frac{Y_{1j}}{Y_j}$
2	0	$1 \cdot \frac{2}{4} = 0.5$	$0 - 0.5 = -0.5$
3	1	$1 \cdot \frac{2}{3} \approx 0.67$	$1 - 0.67 \approx 0.33$
5	0	$1 \cdot \frac{1}{2} = 0.5$	$0 - 0.5 = -0.5$
6	1	$1 \cdot \frac{1}{1} = 1$	$1 - 1 = 0$

Hence, the numerator is:

$$U_L = (-0.5) + 0.33 + (-0.5) + 0 = -0.67$$

and the estimate of its variance is:

$$\widehat{\text{Var}}(U_L) = \sum_{j=1}^{n_d} \frac{Y_{1j} Y_{0j}}{Y_j^2} D_j,$$

where for each t_j:

t_j	$\frac{Y_{1j} Y_{0j}}{Y_j^2}$	D_j	$\frac{Y_{1j} Y_{0j}}{Y_j^2} D_j$
2	$\frac{2 \cdot 2}{4^2} = 0.25$	1	0.25
3	$\frac{2 \cdot 1}{3^2} \approx 0.22$	1	0.22
5	$\frac{1 \cdot 1}{2^2} = 0.25$	1	0.25
6	$\frac{1 \cdot 0}{1^2} = 0$	1	0

So

$$\widehat{\text{Var}}(U_L) = 0.25 + 0.22 + 0.25 + 0 = 0.72.$$

The test statistic is:

$$Z = \frac{U_L}{\sqrt{\widehat{\text{Var}}(U_L)}} = \frac{-0.67}{\sqrt{0.72}} \approx -0.79.$$

Finally, we calculate the Wilcoxon test statistic by noting the Wilcoxon test weights events by the number of individuals at risk ($W_j = Y_j$). That is, the numerator is:

$$U_W = \sum_{j=1}^{n_d} W_j \left(D_{1j} - D_j \frac{Y_{1j}}{Y_j} \right),$$

where for each t_j:

t_j	W_j	$W_j \cdot D_j \frac{Y_{1j}}{Y_j}$	$W_j \cdot \left(D_{1j} - D_j \frac{Y_{1j}}{Y_j} \right)$
2	4	$4 \cdot 0.5 = 2$	$4 \cdot (-0.5) = -2$
3	3	$3 \cdot 0.67 = 2.0$	$3 \cdot 0.33 = 1.0$
5	2	$2 \cdot 0.5 = 1$	$2 \cdot (-0.5) = -1$
6	1	$1 \cdot 1 = 1$	$1 \cdot 0 = 0$

So the numerator is:

$$U_W = (-2) + 1 + (-1) + 0 = -2.$$

On the other hand, the variance of the Wilcoxon test is:

$$\widehat{\mathrm{Var}}(U_W) = \sum_{j=1}^{n_d} Y_{1j} Y_{0j} D_j,$$

where at each t_j:

t_j	$Y_{1j}Y_{0j}$	D_j	$Y_{1j}Y_{0j}D_j$
2	$16 \cdot 0.25 = 4$	1	4
3	$9 \cdot 0.22 = 2$	1	2
5	$4 \cdot 0.25 = 1$	1	1
6	$1 \cdot 0 = 0$	1	0

Using this table gives

$$\widehat{\mathrm{Var}}(U_W) = 4 + 2 + 1 + 0 = 7,$$

and, therefore, the test statistic is:

$$Z = \frac{U_W}{\sqrt{\widehat{\mathrm{Var}}(U_W)}} = \frac{-2}{\sqrt{7}} \approx -0.76.$$

6 Kernel-Smoothed Hazard Estimator

Estimating the hazard function provides valuable insights into the instantaneous rate of failure (or event rate) over time, which is essential for understanding and predicting survival outcomes. The shape of the hazard function can reveal important information about the underlying dynamics of the event, such as whether the risk of failure increases or decreases as time progresses. In contrast to the Nelson-Aalen estimate, which produces a stepwise hazard estimate with spikes at the observed failure times and zero elsewhere, smooth estimates of the hazard function help reduce variance and noise. This leads to a more stable, continuous, and interpretable representation of the event risk over time. We consider the kernel-smoothed hazard estimator defined by

$$\widetilde{\lambda}(t) = \sum_{j=1}^{n_d} K_h(t - t_j) \frac{D_j}{Y_j} = \int_0^\infty K_h(t - s)\, \frac{dN(s)}{Y(s)},$$

where $K_h(u) = \frac{1}{h} K\left(\frac{u}{h}\right)$, and we recall that $N(s)$ and $Y(s)$ are the aggregated counting and at-risk processes, respectively. The kernel function K is central to the smoothing process with these properties (Wand & Jones 1995):

- Normalization: $K(u) \geq 0$ for all u and $\int_{-\infty}^{\infty} K(u)\, du = 1$.
- Symmetry: $K(u) = K(-u)$ so that the weighting is balanced around zero.
- Choice of Kernel: Typical kernels include:
  - Gaussian: $K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$.
  - Epanechnikov: $K(u) = \frac{3}{4}(1 - u^2) I(|u| \leq 1)$.
  - Uniform: $K(u) = \frac{1}{2} I(|u| \leq 1)$.

- Effect of Bandwidth: The bandwidth h determines the window (see Figure 8) over which the data are smoothed, controlling the trade-off between bias and variance; a smaller h leads to less smoothing (lower bias, higher variance), while a larger h leads to more smoothing (higher bias, lower variance).

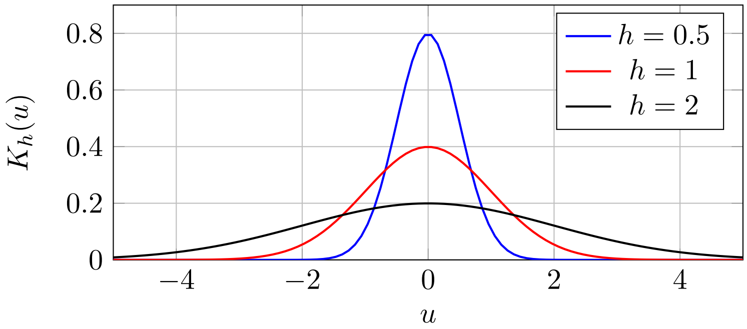


Figure 8: The effect of the bandwidth h on the kernel function. A larger h yields a wider, flatter kernel.

We next discuss how to estimate the variance of $\widetilde{\lambda}(t)$. Recall we use the Doob–Meyer decomposition for the counting process, obtaining

$$dN(s) = Y(s)\lambda(s)\,ds + dM(s),$$

where $\lambda(s)$ is the true hazard function and $dM(s)$ is a martingale increment satisfying

$$\mathbb{E}\big[dM(s) \mid \mathcal{F}_{s^-}\big] = 0.$$

Thus, we can write

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

That is,

$$\widetilde{\lambda}(t) = \int_0^\infty K_h(t-s)\lambda(s)\,ds + \epsilon(t),$$

where $\epsilon(t) = \int_0^\infty K_h(t-s)\,\frac{dM(s)}{Y(s)}$. Hence, $\widetilde{\lambda}(t)$ is decomposed into two terms. The first one is deterministic, while the second term, $\epsilon(t)$, is a stochastic integral with respect to the martingale $M(s)$ and captures the variance of $\widetilde{\lambda}(t)$, which is given by

$$\mathrm{Var}\big(\epsilon(t)\big) = \int_0^\infty K_h^2(t-s)\,\frac{d\langle M\rangle(s)}{Y^2(s)}.$$

Under the Doob–Meyer decomposition, the predictable variation process of $M(s)$ is

$$\langle M\rangle(s) = \int_0^s Y(u)\lambda(u)\,du,$$

so that

$$d\langle M\rangle(s) = Y(s)\lambda(s)\,ds.$$

Substituting, we obtain

$$\operatorname{Var}\big(\epsilon(t)\big) = \mathbb{E}\int_0^\infty K_h^2(t-s)\,\frac{Y(s)\lambda(s)\,ds}{Y^2(s)} = \mathbb{E}\int_0^\infty K_h^2(t-s)\,\frac{d\Lambda(s)}{Y(s)}.$$

Using the Nelson-Aalen estimator, we can estimate the variance with

$$\int_0^\infty K_h^2(t-s)\,\frac{dN(s)}{Y^2(s)} = \sum_{j=1}^{n_d} K_h^2(t-t_j)\frac{D_j}{Y_j^2}$$

Figure 9 illustrates how the hazard estimate depends on the bandwidth of the kernel function: For $h = 0.1$, the estimated hazard function is highly variable, showing sharp fluctuations due to overfitting; for moderate bandwidths ($h = 0.5, 1.0$), the estimate smooths out, providing a more stable representation of the hazard; for a large bandwidth ($h = 2.0$), the hazard function is overly smoothed, losing important details. Thus, it is critical to find an h that strikes a balance between bias and variation.

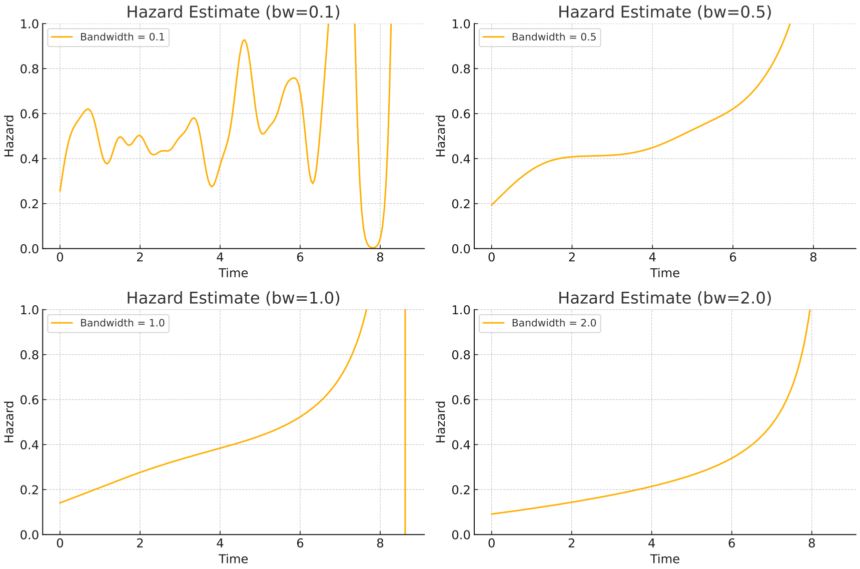


Figure 9: Kernel hazard function estimates for different bandwidths.

6.1 Derivation of the optimal bandwidth

Recall that we assume (T_i, C_i) are i.i.d. for $i = 1, \ldots, n$ and so are (X_i, Δ_i). For simplicity, we consider points t away from 0 (in particular, $t > h$). To facilitate the later development, we extend the integration limits to $(-\infty, \infty)$ as $\lambda(s) = 0$ when $s < 0$:

$$\mathbb{E}\big[\widetilde{\lambda}(t)\big] = \int_0^\infty K_h(t-s)\lambda(s)\,ds = \int_{-\infty}^\infty K_h(t-s)\lambda(s)\,ds.$$

Substitute $K_h(t-s) = \frac{1}{h}K\Big(\frac{t-s}{h}\Big)$ and change variables by letting

$$u = \frac{t-s}{h} \quad \Longrightarrow \quad s = t - h\,u, \quad ds = -h\,du.$$

Then,

$$\begin{aligned}\mathbb{E}\big[\widetilde{\lambda}(t)\big] &= \int_{-\infty}^\infty \frac{1}{h}K(u)\,\lambda(t-h\,u)\,(h\,du)\\ &= \int_{-\infty}^\infty K(u)\,\lambda(t-h\,u)\,du.\end{aligned}$$

Assume that $\lambda(t)$ is twice continuously differentiable and expand $\lambda(t - h\,u)$ in a Taylor series about t:

$$\lambda(t-h\,u) = \lambda(t) - h\,u\,\lambda'(t) + \frac{h^2u^2}{2}\lambda''(t) + o(h^2).$$

Thus,

$$\begin{aligned}\mathbb{E}\big[\widetilde{\lambda}(t)\big] &= \int_{-\infty}^\infty K(u)\left[\lambda(t) - h\,u\,\lambda'(t) + \frac{h^2u^2}{2}\lambda''(t) + o(h^2)\right]du\\ &= \lambda(t)\int_{-\infty}^\infty K(u)\,du - h\,\lambda'(t)\int_{-\infty}^\infty u\,K(u)\,du + \frac{h^2\lambda''(t)}{2}\int_{-\infty}^\infty u^2\,K(u)\,du + o(h^2).\end{aligned}$$

Since K satisfies normalization and is symmetric about zero, we have

$$\int_{-\infty}^\infty K(u)\,du = 1 \quad \text{and} \quad \int_{-\infty}^\infty u\,K(u)\,du = 0.$$

So K can be regarded as a probability density function. Define the second moment of K as

$$\mu_2(K) = \int_{-\infty}^\infty u^2\,K(u)\,du.$$

Thus,

$$\mathbb{E}\big[\widetilde{\lambda}(t)\big] = \lambda(t) + \frac{h^2\lambda''(t)}{2}\mu_2(K) + o(h^2).$$

The bias of the estimator is

$$\mathrm{Bias}\big[\widetilde{\lambda}(t)\big] = \mathbb{E}\big[\widetilde{\lambda}(t)\big] - \lambda(t) = \frac{h^2}{2}\mu_2(K)\lambda''(t) + o(h^2).$$

Neglecting the $o(h^2)$ term, the leading term in the bias is given by

$$\mathrm{Bias}\big[\widetilde{\lambda}(t)\big] \approx \frac{h^2}{2}\mu_2(K)\lambda''(t),$$

and the squared bias is

$$\text{Bias}^2[\widetilde{\lambda}(t)] \approx \frac{h^4}{4}\mu_2(K)^2\big(\lambda''(t)\big)^2.$$

Next, we recall the variance of $\widetilde{\lambda}(t)$ can be approximated by

$$\text{Var}\big(\widetilde{\lambda}(t)\big) \approx \int_0^\infty K_h^2(t-s)\,\frac{\lambda(s)}{Y(s)}\,ds = \int_{-\infty}^\infty K_h^2(t-s)\,\frac{\lambda(s)}{Y(s)}\,ds,$$

where the last equality holds because $\lambda(s) = 0$ when $s < 0$. Changing variables with $u = (t-s)/h$ so that $ds = -h\,du$ and using

$$K_h(t-s) = \frac{1}{h}K\left(\frac{t-s}{h}\right) = \frac{1}{h}K(u),$$

we have

$$\text{Var}\big(\widetilde{\lambda}(t)\big) \approx \frac{1}{h}\int_{-\infty}^\infty K^2(u)\,\frac{\lambda(t-hu)}{Y(t-hu)}\,du.$$

With h small, we approximate

$$\lambda(t-hu) \approx \lambda(t), \quad Y(t-hu) \approx Y(t),$$

because of the smoothness of $\lambda(t)$ and the left continuity of $Y(t)$. Thus,

$$\text{Var}\big(\widetilde{\lambda}(t)\big) \approx \frac{\lambda(t)}{h\,Y(t)}\int_{-\infty}^\infty K^2(u)\,du,$$

and defining

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

we obtain

$$\text{Var}\big(\widetilde{\lambda}(t)\big) \approx \frac{\lambda(t)R(K)}{h\,Y(t)}.$$

When the sample size n is large, one may approximate $Y(t) \approx n\,S_X(t)$, where $S_X(t) = P(X_i > t)$ is the survival function of the observed survival time.

We choose h to minimize the integrated mean squared error (IMSE), defined as

$$\text{IMSE}(h) = \int_0^\infty \mathbb{E}\Big[(\widetilde{\lambda}(t) - \lambda(t))^2\Big]\,dt.$$

Neglecting higher-order terms, this decomposes into the integrated squared bias and integrated variance:

$$\text{IMSE}(h) \approx \int_0^\infty \frac{h^4}{4}\mu_2(K)^2\big(\lambda''(t)\big)^2\,dt + \int_0^\infty \frac{\lambda(t)R(K)}{h\,Y(t)}\,dt.$$

For simplicity, denote

$$A = \frac{1}{4}\mu_2(K)^2\int_0^\infty \big(\lambda''(t)\big)^2 dt,$$

and, assuming $Y(t) \approx n\, S_X(t)$,

$$B = \frac{R(K)}{n} \int_0^\infty \frac{\lambda(t)}{S_X(t)} dt.$$

Then, we can write

$$\mathrm{IMSE}(h) \approx A\, h^4 + \frac{B}{h}.$$

To find the optimal h, we minimize IMSE(h) by differentiating it with respect to h and setting the derivative to 0: $\frac{d}{dh}\left[A\, h^4 + \frac{B}{h}\right] = 4A\, h^3 - \frac{B}{h^2} = 0$. Solving it for h yields: $h_{\text{opt}} = \left(\frac{B}{4A}\right)^{1/5}$. Substituting back the definitions of A and B, we have

$$h_{\text{opt}} = \left(\frac{R(K) \int_0^\infty \frac{\lambda(t)}{S_X(t)} dt}{n\, \mu_2(K)^2 \int_0^\infty \left(\lambda''(t)\right)^2 dt} \right)^{1/5} = O(n^{-1/5}).$$

Below Figure 10 shows bias, variance and IMSE and how optimal h can be obtained (with constants chosen as 1 for illustrative purposes).

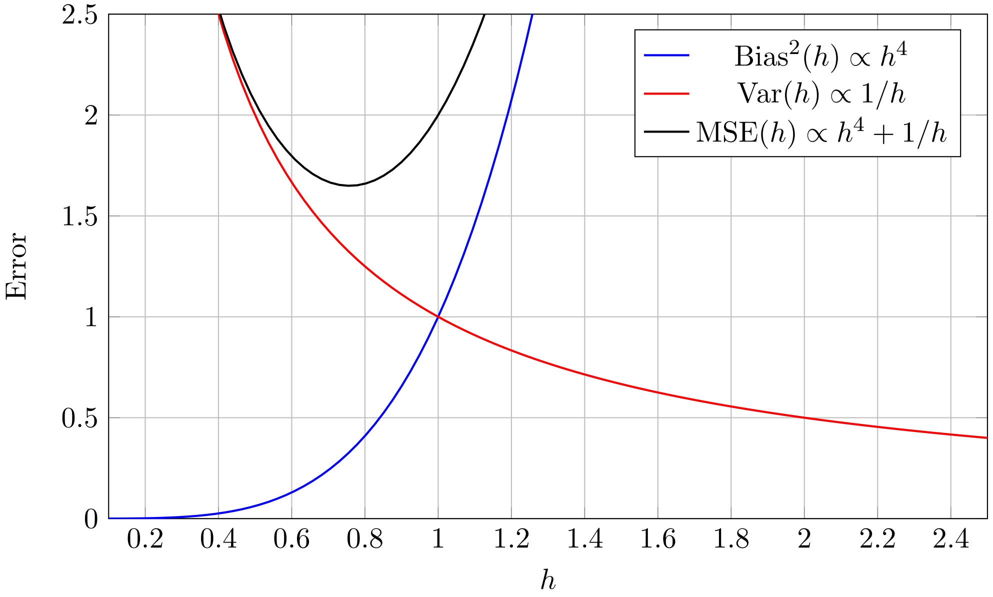


Figure 10: Squared bias, variance, and MSE versus the bandwidth h.

7 Cox Proportional Hazards Models

The Cox proportional hazards model (Cox 1972) is a widely used survival analysis method that estimates covariate effects on the hazard function without assuming a parametric baseline hazard, ensuring flexibility across fields like medicine and epidemiology. Its key innovation, partial likelihood, enables efficient estimation of regression coefficients without modeling the baseline hazard, focusing on relative hazard ratios (Tsiatis 2006).

We define the hazard function for individual i with a vector of covariates $Z_i \in \mathbb{R}^p$, at time t, as

$$\lambda_i(t) = \lim_{dt \to 0^+} \frac{1}{dt} P(t \le T_i < t + dt \mid T_i \ge t, Z_i).$$

The Cox proportional hazards model stipulates that

$$\lambda_i(t) = \lambda_0(t) e^{Z_i^\top \beta} \tag{7.1}$$

where $\lambda_0(t)$ represents the baseline hazard and β is the vector of regression coefficients which is to be estimated based on the observed (X_i, Δ_i, Z_i), for $i = 1, \dots, n$. The term "proportional hazards" indicates that covariates have a multiplicative effect on the baseline hazard, e.g., if one group has a hazard that is twice that of another group at any time point, it remains twice as high at all times. This assumption allows the model to estimate the coefficients without specifying the exact form of the baseline hazard function. For simplicity in theoretical derivations, we assume that Z_i belongs to a compact subset of $\mathbb{R}^p$ and that the true parameter value, β_0, also lies within a compact subset of $\mathbb{R}^p$. Additionally, we assume Z_i is time-independent, though our results can be extended to the time-dependent case. The filtration considered hereafter has been extended to incorporate covariate information. Specifically, for $t > 0$, we define

$$\mathcal{F}_t = \sigma\{N_i(s), Y_i(s), Z_i, 1 \le i \le n, 0 \le s \le t\}.$$

Proposition 7.1. *Suppose that $Z_i, i = 1, \dots, n$, are bounded in $\mathbb{R}^p$ and so is the true parameter value, β_0. We define a right continuous process:*

$$M_i(t) = N_i(t) - \int_0^t Y_i(s) e^{Z_i^\top \beta_0} \lambda_0(s) ds. \tag{7.2}$$

Then it is a square integrable martingale with respect to $\mathcal{F}_t$, with the variation process of

$$\langle M_i \rangle(t) = A_i(t) \overset{def}{=} \int_0^t Y_i(s) e^{Z_i^\top \beta_0} \lambda_0(s) ds.$$

Proof. Homework. □

We estimate β_0 based on the observed data. Let us extract the observed failure time points: $t_1 < t_2 < \dots < t_{n_d}$ from the data. Assuming no ties at the failure time points, we let (k) provide the case label for the patient failing at t_k (thus $T_{(k)} = t_k$), so the covariates associated with the n_d failures are $Z_{(1)}, \dots, Z_{(n_d)}$, and in particular, $\lambda_{(k)}(t) = \lambda_0(t) \exp(Z_{(k)}^\top \beta)$. For an individual who experiences an event at time t_k, the probability of this event occurring, given that one event occurs at this time point, is:

$$P(\text{the observed failure at } t_k | \text{one failure at } t_k \text{ among } R(t_k)) = \frac{\lambda_{(k)}(t_k)}{\sum_{j \in R(t_k)} \lambda_j(t_k)}, \tag{7.3}$$

where at each time t, the risk set $R(t)$ consists of individuals who are at risk of experiencing the event just before time t. It follows that the overall partial likelihood is the product of the individual likelihood contributions for all observed events:

$$L(\beta) = \prod_{k=1}^{n_d} \frac{\lambda_{(k)}(t_k)}{\sum_{j \in R(t_k)} \lambda_j(t_k)} = \prod_{i \in D} \frac{\lambda_i(X_i)}{\sum_{j \in R(X_i)} \lambda_j(X_i)} = \prod_{i \in D} \frac{e^{Z_i^\top \beta}}{\sum_{j \in R(X_i)} e^{Z_j^\top \beta}}, \tag{7.4}$$

where D is the set of the labels of individuals who were observed to have failed. We derive (7.3) and (7.4) in detail later.

Taking the natural logarithm of the partial likelihood yields the log partial likelihood:

$$\begin{aligned} \ell(\beta) &= \sum_{i \in D} \left(Z_i^\top \beta - \log \sum_{j \in R(X_i)} \exp(Z_j^\top \beta) \right) \\ &= \sum_{i=1}^{n} \Delta_i \left(Z_i^\top \beta - \log \sum_{j=1}^{n} Y_j(X_i) \exp(Z_j^\top \beta) \right) \\ &= \sum_{i=1}^{n} \int_0^\infty \left(Z_i^\top \beta - \log \sum_{j=1}^{n} Y_j(t) \exp(Z_j^\top \beta) \right) dN_i(t). \end{aligned}$$

We estimate β_0, the truth, by maximizing $\ell(\beta)$, and the resulting estimator, denoted by $\widehat{\beta}$, is called the maximum partial likelihood estimator (MPLE). This structure of the partial likelihood allows for the estimation of regression coefficients while circumventing the need to specify the baseline hazard function.

7.1 Derivation of partial likelihood

We consider a sequential conditioning argument as done in Fleming & Harrington (2013). Suppose $(A_1, B_1), (A_2, B_2), \ldots, (A_K, B_K)$ is a collection of pairs of events. Applying the recursive formula of conditional probability, the likelihood of all $2K$ events is:

$$\begin{aligned} P\{A_K B_K A_{K-1} B_{K-1} \ldots A_1 B_1\} &= \prod_{k=2}^{K} P\{A_k \mid B_k A_{k-1} B_{k-1} \ldots A_1 B_1\} P(A_1 | B_1) \\ &\quad \times \prod_{k=2}^{K} P\{B_k \mid A_{k-1} B_{k-1} \ldots A_1 B_1\} P\{B_1\}. \end{aligned}$$

The first two terms would form a partial likelihood for a parameter, if ignoring the last two.

Let us apply this to the observed data, (X_i, Δ_i, Z_i), for $i = 1, \ldots, n$. Let B_k be the event describing (i) the observed censoring times within the intervals $[t_{k-1}, t_k)$ for $k = 1, \ldots, n_d + 1$ (with $t_0 = 0$ and $t_{n_d+1} = \infty$), along with their associated case labels; and (ii) the fact that a failure has been observed at t_k. If A_k is the event specifying the label k of the case failing at t_k, then the observed data are equivalent to the event $B_1 A_1 \ldots B_{n_d} A_{n_d} B_{n_d+1}$ and the likelihood of the data will be:

$$P(B_1 A_1 \ldots B_{n_d} A_{n_d} B_{n_d+1}).$$

Fleming & Harrington (2013) reasoned that since the censoring times do not provide additional information about the failure distribution, it is plausible to assume that the events B_k contain little information about the regression parameter β. Therefore, a reasonable partial likelihood for β will be:

$$\prod_{k=2}^{n_d} P\{A_k \mid B_k A_{k-1} B_{k-1} \dots A_1 B_1\} P(A_1|B_1).$$

We next show that

$$P\{A_k \mid B_k A_{k-1} B_{k-1} \dots A_1 B_1\} = \frac{\lambda_{(k)}(t_k)}{\sum_{j \in R(t_k)} \lambda_j(t_k)},$$

which corresponds to (7.3). To prove this, we note that $B_k = B_k^{(\mathrm{i})} B_k^{(\mathrm{ii})}$, where $B_k^{(\mathrm{i})}$ = the observed times of censoring in the interval $[t_{k-1}, t_k)$ and the case labels associated with these censored times, and B_k^{ii} = a failure at t_k. (As no death would happen at ∞, we define $B_{n_d+1} = B_{n_d+1}^{(\mathrm{i})}$.)

$$\begin{aligned} P\{A_k \mid B_k A_{k-1} B_{k-1} \dots A_1 B_1\} &= \frac{P\{A_k B_k^{(\mathrm{ii})} \mid B_k^{(\mathrm{i})} A_{k-1} B_{k-1} \dots A_1 B_1\}}{P\{B_k^{(\mathrm{ii})} \mid B_k^{(\mathrm{i})} A_{k-1} B_{k-1} \dots A_1 B_1\}} \\ &= \frac{P\{A_k, \text{a failure at } t_k \mid B_k^{(\mathrm{i})} A_{k-1} B_{k-1} \dots A_1 B_1\}}{P\{\text{a failure at } t_k \mid B_k^{(\mathrm{i})} A_{k-1} B_{k-1} \dots A_1 B_1\}} \\ &= \frac{P\{A_k \mid B_k^{(\mathrm{i})} A_{k-1} B_{k-1} \dots A_1 B_1\}}{P\{\text{a failure at } t_k \mid B_k^{(\mathrm{i})} A_{k-1} B_{k-1} \dots A_1 B_1\}}. \end{aligned}$$

Note that the event of $B_k^{(\mathrm{i})} A_{k-1} B_{k-1} \dots A_1 B_1$ describes the risk set at t_k. Hence,

$$P\{A_k \mid B_k^{(\mathrm{i})} A_{k-1} B_{k-1} \dots A_1 B_1\} = P\{A_k \mid \text{individual } (k) \text{ is at risk at } t_k\} = \lambda_{(k)}(t_k).$$

The first equality holds because individual (k) is independent of the other individuals, and the second equality holds because the probability that individual (k) experiences a failure at t_k, given he/she is still at risk, is equal to his hazard function. On the other hand, the denominator of the conditional probability consists of the sum of hazard functions over all individuals in the risk set $R(t_k)$, as any of them could experience the failure:

$$P\{\text{a failure at } t_k \mid B_k^{(\mathrm{i})} A_{k-1} B_{k-1} \dots A_1 B_1\} = \sum_{j \in R(t_k)} \lambda_j(t_k).$$

Thus,

$$P\{A_k \mid B_k A_{k-1} B_{k-1} \dots A_1 B_1\} = \frac{\lambda_{(k)}(t_k)}{\sum_{j \in R(t_k)} \lambda_j(t_k)}.$$

Therefore,

$$\prod_{k=2}^{n_d} P\{A_k \mid B_k A_{k-1} B_{k-1} \dots A_1 B_1\} P(A_1|B_1) = \prod_{k=1}^{n_d} \frac{\lambda_{(k)}(t_k)}{\sum_{j \in R(t_k)} \lambda_j(t_k)}$$

which corresponds to (7.4).

7.2 Maximum partial likelihood estimator (MPLE)

Introduce

$$S^{(k)}(\beta, t) = \frac{1}{n}\sum_{j=1}^{n} Y_j(t) Z_j^{\otimes k} e^{Z_j^\top \beta}, k = 0, 1, 2.$$

Here, $Z_j^{\otimes 0} = 1, Z_j^{\otimes 1} = Z_j, Z_j^{\otimes 2} = Z_j Z_j^\top$.

To facilitate the theory, we consider the log partial likelihood as

$$\ell(\beta, \tau) = \sum_{i=1}^{n} \int_0^\tau \left(Z_i^\top \beta - \log S^{(0)}(\beta, t) \right) dN_i(t), \tag{7.5}$$

where $\tau < \infty$ such that $\Lambda_0(\tau) < \infty$. Similarly, the score function is

$$U(\beta, \tau) = \sum_{i=1}^{n} \int_0^\tau \left(Z_i - \bar{Z}(\beta, t) \right) dN_i(t), \tag{7.6}$$

where

$$\bar{Z}(\beta, t) = \frac{S^{(1)}(\beta, t)}{S^{(0)}(\beta, t)}. \tag{7.7}$$

We introduce the parameter τ in the log-likelihood and score functions to highlight their dependence on using information up to τ. The choice of τ instead of ∞ as the upper limit of integration helps prevent divergence and ensures that the integral is restricted to a finite observation period. This avoids unrealistic assumptions about unobserved or censored times. In practice, τ is often chosen as the maximum observation period in the study.

The maximum partial likelihood estimator (MPLE) of β, denoted by $\widehat{\beta}$, is obtained by maximizing (7.5) or equivalently solving the score equation $U(\beta, \tau) = 0$. Under mild regularity conditions, $\ell(\beta, \tau)$ is a concave function, ensuring a unique maximizer. We analyze the Hessian matrix of (7.5):

$$H(\beta) = -\int_0^\tau \left[\frac{S^{(2)}(\beta, t)}{S^{(0)}(\beta, t)} - \left\{ \frac{S^{(1)}(\beta, t)}{S^{(0)}(\beta, t)} \right\}^{\otimes 2} \right] dN(t) \overset{def}{=} -\mathcal{I}(\beta), \tag{7.8}$$

where we have also defined the observed information matrix $\mathcal{I}(\beta)$. Hence, concavity holds if $H(\beta)$ is negative definite or $\mathcal{I}(\beta)$ is positive definite, which requires the integrand to be non-degenerate. For example, if the number of individuals at risk, $S^{(0)}(\beta, t)$, is too small at any time t, the integrand may become unstable and the Hessian may become singular. We may require the sample size to be sufficiently large.

Furthermore, even if the log partial likelihood function $\ell(\beta)$ is concave, perfect separation by covariates (or linear combinations of covariates) can lead to situations where $\ell(\beta)$ does not have a finite solution. In such cases, as the parameter estimates approach infinity, the likelihood may diverge, making it impossible to obtain meaningful estimates for the coefficients.

7.3 Consistency of MPLE

If we can ensure that problematic scenarios are avoided, maximizing equation (7.5) can produce a unique and finite estimate $\widehat{\beta}$. In this context, we examine the asymptotic properties of the estimator $\widehat{\beta}$ and introduce additional sufficient regularity conditions that ensure the validity of maximum likelihood estimation. Throughout this discussion, $||\cdot||$ denotes the maximum absolute value of the elements of a vector or matrix, while $|\cdot|$ represents the Euclidean norm for a vector (and, trivially, the absolute value for a scalar).

(C.1) There exists an open and convex neighborhood $\mathcal{B}$ of $\beta_0 \in \mathbb{R}^p$ and, respectively, scalar, vector, and matrix functions, $s^{(0)}, s^{(1)}, s^{(2)}$ such that

$$\sup_{t\in[0,\tau],\beta\in\mathcal{B}} ||S^{(k)}(\beta,t) - s^{(k)}(\beta,t)|| \to 0$$

in probability.

(C.2) In the same $\mathcal{B}$, it holds that, for any $\beta \in \mathcal{B}$ and $t \in [0,\tau]$,

$$s^{(1)}(\beta,t) = \frac{\partial}{\partial\beta}s^{(0)}(\beta,t),\ s^{(2)}(\beta,t) = \frac{\partial}{\partial\beta}s^{(1)}(\beta,t) = \frac{\partial^2}{\partial\beta\partial\beta^\top}s^{(0)}(\beta,t).$$

We also assume each element of $s^{(k)}(\beta,t), k = 0,1,2$ is bounded, and in addition, $s^{(0)}(\beta,t)$ is bounded away from 0 in $\mathcal{B} \times [0,\tau]$. In addition, for each k and $t \in [0,\tau]$, $s^{(k)}(\beta,t)$ is equicontinuou at β_0. That is, for any $\epsilon > 0$, there exists a $\delta > 0$ such that $||\beta - \beta_0|| < \delta$ implies $||s^{(k)}(\beta,t) - s^{(k)}(\beta_0,t)|| < \epsilon$ for all $t \in [0,\tau]$.

(C.3) Define

$$v(\beta,t) = \frac{s^{(2)}(\beta,t)}{s^{(0)}(\beta,t)} - \left(\frac{s^{(1)}(\beta,t)}{s^{(0)}(\beta,t)}\right)^{\otimes 2}. \tag{7.9}$$

We assume

$$\Sigma(\beta_0,\tau) \overset{def}{=} \int_0^\tau v(\beta_0,s)s^{(0)}(\beta_0,s)\lambda_0(s)ds \tag{7.10}$$

is positive definite.

Condition (C.1) establishes an asymptotic stability requirement for the functions $S^{(k)}$. The first part of Condition (C.2) ensures the interchangeability of differentiation and limits. The other parts of Condition (C.2), along with Condition (2.3), specifies regularity conditions analogous to those commonly encountered in standard asymptotic likelihood theory. These conditions can be verified in certain specific cases and may be relaxed under alternative circumstances; see Fleming & Harrington (2013).

We state two useful lemmas.

Lemma 7.2. *Let $\mathcal{B}$ be an open convex subset of $\mathbb{R}^p$, and let $F_n, n = 1,2,\ldots$ be a sequence of random concave functions on $\mathcal{B}$ and f a real-valued function on $\mathcal{B}$ such that for all $\beta \in \mathcal{B}$,*

$$\lim_{n\to\infty} F_n(\beta) = f(\beta) \quad \textit{in probability.}$$

Then:

1. *The function f is concave.*
2. *For all compact subsets A of $\mathcal{B}$,*

$$\sup_{\beta\in A} |F_n(\beta) - f(\beta)| \to 0 \quad \textit{in probability, as } n \to \infty.$$

3. *If F_n has a unique maximum at β_n and f has one at β, then $\beta_n \to \beta_0$ in probability.*

Proof. See Anderson & Gill (1982). □

Proposition 7.3. *Let $\widehat{\beta}$ denote the MPLE of β maximizing (7.5), and β_0 the true value of β in (7.5). Then*

$$\lim_{n\to\infty} \widehat{\beta} = \beta_0 \quad \textit{in probability, i.e., } \widehat{\beta} \textit{ is consistent.}$$

Proof. We first establish convergence of the log partial likelihood by expressing a closely related term as a martingale. Let $X_n(\beta, \cdot)$ denote the process which, at time t, is the difference in log partial likelihoods over $[0, t)$, evaluated at an arbitrary β and the true value β_0.

$$X_n(\beta, t) = n^{-1} \left\{\ell(\beta, t) - \ell(\beta_0, t)\right\},$$

where

$$\ell(\beta, t) = \sum_{i=1}^{n} \int_0^t \left(Z_i^\top \beta - \log \sum_{j=1}^{n} Y_j(t) \exp(Z_j^\top \beta) \right) dN_i(t).$$

Hence,

$$X_n(\beta, t) = n^{-1} \sum_{i=1}^{n} \int_0^t \left[(\beta - \beta_0)^\top Z_i - \log \frac{S^{(0)}(\beta, s)}{S^{(0)}(\beta_0, s)} \right] dN_i(s),$$

Define

$$A_n(\beta, t) = n^{-1} \sum_{i=1}^{n} \int_0^t \left[(\beta - \beta_0)^\top Z_i - \log \frac{S^{(0)}(\beta, s)}{S^{(0)}(\beta_0, s)} \right] Y_i(s) \exp(Z_j^\top \beta_0) \lambda_0(s) ds.$$

Hence,

$$X_n(\beta, t) - A_n(\beta, t) = n^{-1} \sum_{i=1}^{n} \int_0^t \left[(\beta - \beta_0)^\top Z_i - \log \frac{S^{(0)}(\beta, s)}{S^{(0)}(\beta_0, s)} \right] dM_i(s),$$

where $M_i(t) = N_i(t) - \int_0^t Y_i(s) e^{Z_i^\top \beta_0} \lambda_0(s) ds$ is a square integrable martingale with respect to $\mathcal{F}_t$, with the variation process of $d\langle M_i \rangle(s) = Y_i(s) e^{Z_i^\top \beta_0} \lambda_0(s) ds$.

Take $\tau_{n,i} = n \wedge \sup \left\{ s : \left| (\beta - \beta_0)^\top Z_i - \log \frac{S^{(0)}(\beta,s)}{S^{(0)}(\beta_0,s)} \right| \le n \right\}$, then $\int_0^{t\wedge\tau_{n,i}} \left[(\beta - \beta_0)^\top Z_i - \log \frac{S^{(0)}(\beta,s)}{S^{(0)}(\beta_0,s)} \right] dM_i(s)$ is a square integrable martingale by **Property 2**. Therefore, for any given $\beta \in \mathcal{B}$, the process $X_n(\beta, \cdot) - A_n(\beta, \cdot)$ is a local square integrable martingale with the predictable variation process (by

Property 4):

$$\begin{aligned}
\langle X_n(\beta,\cdot) - A_n(\beta,\cdot)\rangle(t) &= n^{-2}\sum_{i=1}^{n}\int_0^\tau \left[(\beta-\beta_0)^\top Z_i - \log\frac{S^{(0)}(\beta,s)}{S^{(0)}(\beta_0,s)}\right]^2 d\langle M_i\rangle(s) \\
&= n^{-2}\sum_{i=1}^{n}\int_0^\tau \left[(\beta-\beta_0)^\top Z_i - \log\frac{S^{(0)}(\beta,s)}{S^{(0)}(\beta_0,s)}\right]^2 Y_i(s)e^{Z_i^\top\beta_0}\lambda_0(s)ds \\
&= n^{-1}\int_0^\tau \Bigg[(\beta-\beta_0)^\top S^{(2)}(\beta_0,s)(\beta-\beta_0) - 2(\beta-\beta_0)^\top S^{(1)}(\beta_0,s)\log\frac{S^{(0)}(\beta,s)}{S^{(0)}(\beta_0,s)} \\
&\quad + \left\{\log\frac{S^{(0)}(\beta,s)}{S^{(0)}(\beta_0,s)}\right\}^2 S^{(0)}(\beta_0,s)\Bigg]\lambda_0(s)ds
\end{aligned}$$

which converges to 0 in probability as $n\to\infty$ using Conditions (C.1) and (C.2) (so that the integral is bounded in probability).

Applying the Lenglart inequality to the process of $\{(X_n(\beta,t)-A_n(\beta,t))^2\}_{t\in[0,\tau]}$, we have

$$X_n(\beta,t) - A_n(\beta,t) \to 0 \quad \text{in probability}$$

uniformly over $[0,\tau]$. In particular,

$$X_n(\beta,\tau) - A_n(\beta,\tau) \to 0 \quad \text{in probability.}$$

Since under Condition (C.1), $A_n(\beta,\tau)$ converges to $A(\beta,\tau)$ for all $\beta\in\mathcal{B}$, where

$$A(\beta,\tau) = \int_0^\tau \left[(\beta-\beta_0)^\top s^{(1)}(\beta_0,s) - \log\frac{s^{(0)}(\beta,s)}{s^{(0)}(\beta_0,s)}s^{(0)}(\beta_0,s)\right]\lambda_0(s)\,ds,$$

it follows that $X_n(\beta,\tau)$ must also converge in probability to the same limit, as long as $\beta\in B$.

As $X_n(\beta,\tau)$ is a concave function of β with a unique maximum, and that $A(\beta,\tau)$ has a unique maximum at $\beta=\beta_0$ under Conditions (C.2) and (C.3), the theorem follows by Lemma 7.2. □

7.4 Asymptotic normality of MPLE

To show the asymptotic normality, we add a new condition.

(C.4) There exists a $\delta>0$ so that

$$\sup_{1\le i\le n, t\in[0,\tau]} n^{-1/2}|Z_i|Y_i(t)I(\beta_0^\top Z_i > -\delta|Z_i|)\to 0$$

in probability.

This condition, which is important for verifying the Lindeberg condition for the martingale central limit theorem holds trivially if the covariates Z_i are bounded. We present several useful lemmas.

Lemma 7.4. *Given any real numbers a and b, and $\varepsilon > 0$, then*

$$(a-b)^2 I(|a-b| \geq \varepsilon) \leq 4a^2 I(|a| \geq \varepsilon/2) + 4b^2 I(|b| \geq \varepsilon/2).$$

Proof. First, with $(a-b)^2 \leq 2a^2 + 2b^2$, multiplying both sides by $I(|a-b| > \varepsilon)$, we obtain $(a-b)^2 I(|a-b| \geq \varepsilon) \leq (2a^2 + 2b^2) I(|a-b| \geq \varepsilon)$. As $|a-b| \geq \varepsilon$ implies $|a| \geq \varepsilon/2$ or $|b| \geq \varepsilon/2$; otherwise, it would lead to $|a-b| < \varepsilon$. Hence,

$$I(|a-b| \geq \varepsilon) \leq I(|a| \geq \varepsilon/2) + I(|b| \geq \varepsilon/2).$$

Therefore,

$$(a-b)^2 I(|a-b| \geq \varepsilon) \leq (2a^2 + 2b^2)\left(I(|a| \geq \varepsilon/2) + I(|b| \geq \varepsilon/2)\right).$$

Expanding the right-hand side:

$$(a-b)^2 I(|a-b| > \varepsilon) \leq 2a^2 I(|a| > \varepsilon/2) + 2a^2 I(|b| > \varepsilon/2) + 2b^2 I(|a| > \varepsilon/2) + 2b^2 I(|b| > \varepsilon/2).$$

Now we show

$$a^2 I(|b| \geq \varepsilon/2) + b^2 I(|a| \geq \varepsilon/2) \leq b^2 I(|b| \geq \varepsilon/2) + a^2 I(|a| \geq \varepsilon/2).$$

Rewriting the inequality, we want to show:

$$a^2 I(|b| \geq \varepsilon/2) - a^2 I(|a| \geq \varepsilon/2) \leq b^2 I(|b| \geq \varepsilon/2) - b^2 I(|a| \geq \varepsilon/2). \tag{7.11}$$

For this, we consider 3 possible cases.

Case 1: $I(|a| \geq \varepsilon/2) = I(|b| \geq \varepsilon/2)$, in which case, the both sides of (7.11) are equal to 0, so the inequality holds trivially.

Case 2: $I(|b| \geq \varepsilon/2) = 1$ and $I(|a| \geq \varepsilon/2) = 0$, so (7.11) would simplify to: $a^2 \leq b^2$. On the other hand, this case means $|b| \geq \varepsilon/2$ and $|a| < \varepsilon/2$, so $a^2 \leq b^2$ will hold.

Case 3: $I(|b| \geq \varepsilon/2) = 0$ and $I(|a| \geq \varepsilon/2) = 1$, so (7.11) would simplify to: $b^2 \leq a^2$. In fact, this case means $|b| < \varepsilon/2$ and $|a| \geq \varepsilon/2$, so the inequality does hold.

Hence, after exhausting the 3 possible cases, we have shown (7.11) holds. Thus the lemma holds. □

We next present a simplified version of multivariate martingale central limit theorem, which will be used for showing the weak convergence of score functions.

Lemma 7.5. *Consider a sequence of p-variate local square integrable martingales, $(U_1^n, \ldots, U_p^n)$, where, for $l = 1, \ldots, p$ and for $0 \leq t \leq \tau$,*

$$U_l^n(t) = \sum_{i=1}^{n} \int_0^t H_{i,l}^n(s) dM_i(s)$$

where $M_i(s)$ is as defined in (7.2) and $H_{i,l}^n(t)$ is locally bounded and predictable with respect to $\mathcal{F}_t$. Also define

$$U_{l,\epsilon}^n(t) = \sum_{i=1}^n \int_0^t H_{i,l}^n(s) I(|H_{i,l}^n(s)| \geq \epsilon) dM_i(s)$$

Suppose for each $l, l' \in \{1, \dots, p\}$ and for all $t > 0$, the covariation process

$$\langle U_l^n, U_{l'}^n \rangle(t) \to C_{ll'}(t), \tag{7.12}$$

in probability and

$$\langle U_{l,\epsilon}^n \rangle(t) \to 0, \tag{7.13}$$

in probability. Then $\{U_1^n(\tau), \dots, U_p^n(\tau)\}$ converges weakly and jointly to a multivariate normal distrbution with mean 0 and a $p \times p$ variance-covariance matrix, whose (l, l')-th entry is $C_{ll'}(\tau)$.

We next consider an extension of the DCT theorem.

Lemma 7.6. *Suppose $|X_n| \leq Y$ a.s., and $\mathbb{E}(Y) < \infty$, and $X_n \to X$ in probability. Then $\mathbb{E}(X_n) \to \mathbb{E}(X)$.*

Proof. We prove by contradiction. Suppose $\mathbb{E}(X_n) \not\to \mathbb{E}(X)$. Then there exists an $\epsilon_0 > 0$ such that there is a sequence n_k such that

$$|\mathbb{E}(X_{n_k}) - \mathbb{E}(X)| \geq \epsilon_0. \tag{7.14}$$

As $X_{n_k} \to X$, there exists a further subsequence n_{k_j} such that $X_{n_{k_j}} \to X$ almost surely. Then DCT leads to

$$\mathbb{E}(X_{n_{k_j}}) \to \mathbb{E}(X),$$

which however contradicts to (7.14). □

We apply the multivariate martingale central limit theorem to obtain the next proposition, which is critical for obtaining the asymptotic normality results. We will prove it in detail.

Proposition 7.7. *Suppose Conditions (C.1)-(C.4) hold and the dimension of Z_i is p.*

(Part 1) Define the normalized vector score

$$n^{-1/2} U(\beta_0, \tau) = n^{-1/2} \sum_{i=1}^n \int_0^\tau \{Z_i - \bar{Z}(\beta_0, s)\} dN_i(s).$$

Then it converges weakly to a multivariate Gaussian distribution with mean 0 and a $p \times p$ variance-covariance matrix $\Sigma(\beta_0, \tau)$, whose l, l'-th entry is:

$$\int_0^\tau v(\beta_0, s)_{ll'} s^{(0)}(\beta_0, s) \lambda_0(s) ds.$$

(Part 2) If $\widehat{\beta}$ is a consistent estimator of β_0, then

$$||n^{-1}\mathcal{I}(\widehat{\beta}) - \Sigma(\beta_0, \tau)|| \to 0$$

in probability as $n \to \infty$. In addition, define $\widehat{\beta}(u) = \beta_0 + u(\widehat{\beta} - \beta_0)$ for $u \in [0,1]$. Then

$$||n^{-1}\mathcal{I}(\widehat{\beta}(u)) - \Sigma(\beta_0, \tau)|| \to 0$$

in probability uniformly with respect to $u \in [0,1]$.

Proof. (Part 1) Define

$$U^n(\beta_0, t) = n^{-1/2} \sum_{i=1}^n \int_0^t \{Z_i - \bar{Z}(\beta_0, s)\} dN_i(s),$$

which is equal to

$$U^n(\beta_0, t) = n^{-1/2} \sum_{i=1}^n \int_0^t \{Z_i - \bar{Z}(\beta_0, s)\} dM_i(s),$$

where $M_i(t) = N_i(t) - \int_0^t Y_i(s) e^{\beta_0^\top Z_i} \lambda_0(s) ds$.

Here, $U^n(\beta_0, t)$ is a p-variate process, with the l-th component written as:

$$U_l^n(\beta_0, t) = \sum_{i=1}^n \int_0^t n^{-1/2} \{Z_{i,l} - \bar{Z}_l(\beta_0, s)\} dM_i(s),$$

where $Z_{i,l}$ is the l-th component of Z_i and

$$\bar{Z}_l(\beta_0, x) = \frac{\sum_{j=1}^n Y_j(x) Z_{j,l} e^{\beta_0^\top Z_j}}{\sum_{i=1}^n Y_i(x) e^{\beta_0^\top Z_i}}.$$

Define $H_{i,l}^n(s) = n^{-1/2}(Z_{i,l} - \bar{Z}_l(\beta_0, s))$, so that

$$U_l^n(\beta_0, t) = \sum_{i=1}^n \int_0^t H_{i,l}^n(s) dM_i(s).$$

Since $H_{i,l}^n(s)$ is locally bounded (considering, for example, a localizing sequence, $\tau_n = n \wedge \sup\{s : |\bar{Z}_l(\beta_0, s)| \leq n\}$) and predictable, applying **Property 4** yields that $U_l^n(\beta_0, t)$ is a local square integrable martingale, and

$$\begin{aligned} \langle U_l^n(\beta_0, t), U_{l'}^n(\beta_0, t) \rangle &= \sum_{i=1}^n \int_0^t H_{i,l}^n(s) H_{i,l}^n(s) d\langle M_i \rangle(s) \\ &= \frac{1}{n} \sum_{i=1}^n \int_0^t (Z_{i,l} - \bar{Z}_l(\beta_0, s))(Z_{i,l'} - \bar{Z}_{l'}(\beta_0, s)) Y_i(s) e^{Z_i^\top \beta_0} \lambda_0(s) ds, \end{aligned}$$

which converges in probability to

$$\int_0^t v(\beta_0, s)_{ll'} s^{(0)}(\beta_0, s) \lambda_0(s) ds,$$

for all $t \in [0, \tau]$, under Conditions (C.1) and (C.2). Here, $v(\beta, s)$ is defined in (7.9).

We next verify the Lindeberg condition. For any $\epsilon > 0$, define, for all l and t, that

$$U^n_{l,\epsilon}(\beta_0, t) = \sum_{i=1}^n \int_0^t H^n_{i,l}(s) I(|H^n_{i,l}(s)| \geq \epsilon) dM_i(s),$$

which, by **Property 4**, is a local square integrable martingale with the variation process given by

$$\langle U^n_{l,\epsilon}(\beta_0, t)\rangle = \sum_{i=1}^n \int_0^t \{H^n_{i,l}(s)\}^2 I(|H^n_{i,l}(s)| \geq \epsilon) d\langle M_i\rangle(s) = \sum_{i=1}^n \int_0^t \{H^n_{i,l}(s)\}^2 I(|H^n_{i,l}(s)| \geq \epsilon) Y_i(s) e^{\beta_0^\top Z_i} \lambda_0(s) ds.$$

Applying Lemma 7.4, the last integral is bounded by

$$\begin{aligned} & \frac{4}{n} \sum_{i=1}^n \int_0^t Z^2_{i,l} I(n^{-1/2}|Z_{i,l}| \geq \epsilon/2) Y_i(s) e^{\beta_0^\top Z_i} \lambda_0(s) ds \\ + \quad & \frac{4}{n} \sum_{i=1}^n \int_0^t \bar{Z}^2_l(\beta_0, s) I(n^{-1/2}|\bar{Z}_l(\beta_0, s)| \geq \epsilon/2) Y_i(s) e^{\beta_0^\top Z_i} \lambda_0(s) ds. \end{aligned} \tag{7.15}$$

The second term of (7.15) can be written as

$$4 \int_0^t \bar{Z}^2_l(\beta_0, s) I(n^{-1/2}|\bar{Z}_l(\beta_0, s)| \geq \epsilon/2) S^{(0)}(\beta_0, s) \lambda_0(s) ds.$$

With Conditions (C.1) and (C.2), it follows that when n is large, $P((n^{-1/2}|\bar{Z}_l(\beta_0, s)| \geq \epsilon/2) \to 0$ uniformly in s. That is $I(n^{-1/2}|\bar{Z}_l(\beta_0, s)| \geq \epsilon/2) = 0$ with probability going to 1 uniformly in s. Hence, the second term of (7.15) converges to 0 in probability.

For the first term, we consider two cases: given the δ defined in (C.4), we consider the events of $\{\beta_0^\top Z_i > -\delta|Z_i|\}$ and $\{\beta_0^\top Z_i \leq -\delta|Z_i|\}$ separately. That is, we consider

$$\frac{4}{n} \sum_{i=1}^n \int_0^t Z^2_{i,l} I(n^{-1/2}|Z_{i,l}| \geq \epsilon/2, \beta_0^\top Z_i > -\delta|Z_i|) Y_i(s) e^{\beta_0^\top Z_i} \lambda_0(s) ds \tag{7.16}$$

$$+ \quad \frac{4}{n} \sum_{i=1}^n \int_0^t Z^2_{i,l} I(n^{-1/2}|Z_{i,l}| \geq \epsilon/2, \beta_0^\top Z_i \leq -\delta|Z_i|) Y_i(s) e^{\beta_0^\top Z_i} \lambda_0(s) ds. \tag{7.17}$$

On the other hand, Condition (C.4) implies that, for any $\epsilon' > 0$, there exists an N, such that when $n > N$, the event that $n^{-1/2}|Z_{i,l}| \geq \epsilon/2$ when $\beta_0^\top Z_i \leq -\delta|Z_i|$ and $Y_i(s) = 1$ for all $s \in [0, \tau]$ and $i = 1, \ldots, n,$, will happen with probability $< \epsilon'$. This means the probability of $I(n^{-1/2}|Z_{i,l}| \geq \epsilon/2, \beta_0^\top Z_i \leq -\delta|Z_i|) Y_i(s) = 0$ will converge to 1 uniformly for all $s \in [0, \tau]$ and $i = 1, \ldots, n$, uniformly. Hence, (7.16) converges to 0 in probability.

Studying the integrand of (7.17), we note that

$$Z^2_{i,l} I(n^{-1/2}|Z_{i,l}| \geq \epsilon/2, \beta_0^\top Z_i \leq -\delta|Z_i|) Y_i(s) e^{\beta_0^\top Z_i} \leq I(n^{-1/2}|Z_{i,l}| \geq \epsilon/2) Z^2_{i,l} e^{-\delta|Z_i|}.$$

We first consider the case that $n^{-1/2}|Z_{i,l}| \geq \epsilon/2$, and note $Z_{i,l}^2 e^{-\delta|Z_i|} \leq Z_{i,l}^2 e^{-\delta|Z_{i,l}|}$. Because $x^2 e^{-\delta x} \to 0$ when $\delta > 0$ as $x \to \infty$. Hence for any $\eta > 0$, there exists an N_0 such that when $n > N$, $Z_{i,l}^2 e^{-\delta|Z_{i,l}|} < \eta$. On the other hand, if $n^{-1/2}|Z_{i,l}| < \epsilon/2$, $I(n^{-1/2}|Z_{i,l}| \geq \epsilon/2, \beta_0^\top Z_i \leq -\delta|Z_i|)Y_i(s)e^{\beta_0^\top Z_i} = 0 < \eta$ holds trivially. As this N_0 does not depend on i, hence when $n > N_0$,

$$I(n^{-1/2}|Z_{i,l}| \geq \epsilon/2, \beta_0^\top Z_i \leq -\delta|Z_i|)Y_i(s)e^{\beta_0^\top Z_i} < \eta$$

for all i. Therefore, (7.17) is bounded by $\eta \int_0^\tau \lambda_0(s)ds$ which can be arbitrarily small, and must converge to 0 in probability. Hence, with all conditions of the (multivariate) Martingale CLT are satisfied, the first part of the results holds.

(Part 2) We consider

$$\begin{aligned}
& \left\| \frac{1}{n}\mathcal{I}(\widehat{\beta}) - \Sigma(\beta_0, \tau) \right\| \\
\leq & \left\| \int_0^\tau (V(\widehat{\beta}, s) - v(\widehat{\beta}, s)) \frac{1}{n} dN(s) \right\| && (7.18) \\
+ & \left\| \int_0^\tau (v(\widehat{\beta}, s) - v(\beta_0, s)) \frac{1}{n} dN(s) \right\| && (7.19) \\
+ & \left\| \int_0^\tau v(\beta_0, s) \frac{1}{n} \sum_{i=1}^n dM_i(s) \right\| && (7.20) \\
+ & \left\| \int_0^\tau v(\beta_0, s)(S^{(0)}(\beta_0, s) - s^{(0)}(\beta_0, s))\lambda_0(s)ds \right\|, && (7.21)
\end{aligned}$$

where $N(s) = \sum_{i=1}^n N_i(s)$. First, (7.18) is bounded by

$$\sup_s ||V(\widehat{\beta}, s) - v(\widehat{\beta}, s))|| \left\{ \frac{1}{n} \sum_{i=1}^n N_i(\tau) \right\}$$

Conditions (C.1) and (C.2) implies $sup_s||V(\widehat{\beta}, s) - v(\widehat{\beta}, s))|| \to 0$ in probability. Also by the law of large numbers $\sum_{i=1}^n N_i(\tau) \to \mathbb{E}N_i(\tau) < 1$. Hence, (7.18) converges to 0 in probability.

Using the equicontinuity of $v(\beta, s)$ at β_0, we have $\sup_s ||v(\widehat{\beta}, s) - v(\beta_0, s))|| \to 0$ in probability as $\widehat{\beta} \to \beta_0$ in probability. As (7.19) is bounded by

$$\sup_s ||v(\widehat{\beta}, s) - v(\beta_0, s))|| \left\{ \frac{1}{n} \sum_{i=1}^n N_i(\tau) \right\},$$

it must converge to 0 in probability.

With the (i, j)-th entry, $v_{ij}(\beta_0, s)$, of $v(\beta_0, s)$, we consider the martingale process

$$\int_0^t v_{ij}(\beta_0, s) \frac{1}{n} \sum_{i=1}^n dM_i(s)$$

which has the variation process of

$$\int_0^t v_{ij}^2(\beta_0, s)\frac{1}{n^2}d\langle M_i\rangle(s) = \frac{1}{n}\int_0^t v^2(\beta_0, s)S^0(\beta_0, s)\lambda_0(s)ds.$$

With Conditions (C.1) and (C.2) and that $S_0(\beta_0, s)$ is bounded for all $s \in [0, \tau]$, we apply Lemma 7.6 and obtain that

$$n\,\mathrm{Var}\left\{\int_0^\tau v_{ij}(\beta_0, s)\frac{1}{n}\sum_{i=1}^n dM_i(s)\right\} = \mathbb{E}\left\{\int_0^\tau v^2(\beta_0, s)S^0(\beta_0, s)\lambda_0(s)ds\right\} \to \int_0^\tau v^2(\beta_0, s)s^0(\beta_0, s)\lambda_0(s)ds < \infty$$

as $n \to \infty$. Hence, $\mathrm{Var}\left\{\int_0^\tau v_{ij}(\beta_0, s)\frac{1}{n}\sum_{i=1}^n dM_i(s)\right\} \to 0$. Applying the Markov inequality, we obtain that

$$\int_0^\tau v_{ij}(\beta_0, s)\frac{1}{n}\sum_{i=1}^n dM_i(s) \to 0$$

in probability for $1 \le i, j \le p$. Hence, (7.20) converges to 0 in probability.

Finally, (7.21) is bounded by

$$\int_0^\tau ||v(\beta_0, s)|||S^{(0)}(\beta_0, s) - s^{(0)}(\beta_0, s)|\lambda_0(s)ds.$$

Because $|S^{(0)}(\beta_0, s) - s^{(0)}(\beta_0, s)| \to 0$ in probability uniformly in $s \in [0, \tau]$ under Condition (C.1), (7.21) converges to 0 in probability.

Combining the results for (7.18)-(7.21), we have

$$\left\|\frac{1}{n}\mathcal{I}(\widehat{\beta}) - \Sigma(\beta_0, \tau)\right\| \to 0$$

in probability as long as $\widehat{\beta} \to \beta_0$ in probability.

Examining the convergence result in each step, we can also conclude that

$$||n^{-1}\mathcal{I}(\widehat{\beta}(u)) - \Sigma(\beta_0, \tau)|| \to 0$$

in probability uniformly with respect to $u \in [0, 1]$.

□

Lemma 7.8. *Let $F : \mathbb{R}^p \to \mathbb{R}^p$ be continuously differentiable on an open convex set containing the points β_0 and β_1. Then*

$$F(\beta_1) - F(\beta_0) = \left(\int_0^1 J(\beta_0 + u(\beta_1 - \beta_0))du\right)(\beta_1 - \beta_0),$$

where $J(\beta)$ is the Jacobian matrix of F at β.

Proof. Define a path from β_0 to β_1 by the line segment:

$$\beta(t) = \beta_0 + t(\beta_1 - \beta_0), \quad u \in [0,1].$$

Here, the endpoints correspond to $\beta(0) = \beta_0$ and $\beta(1) = \beta_1$. The derivative of F along this path can be expressed using the chain rule:

$$\frac{d}{du}F(\beta(u)) = J(\beta(u)) \cdot \frac{d\beta}{du} = J(\beta(u)) \cdot (\beta_1 - \beta_0),$$

where $J(\beta(t))$ is the Jacobian matrix of F evaluated at $\beta(u)$, and $\frac{d\beta}{du} = \beta_1 - \beta_0$. Integrating the derivative from $u = 0$ to $u = 1$, we have:

$$F(\beta_1) - F(\beta_0) = \int_0^1 \frac{d}{du}F(\beta(u))\,du = \int_0^1 J(\beta(u)) \cdot (\beta_1 - \beta_0)\,du.$$

That is,

$$F(\beta_1) - F(\beta_0) = \left(\int_0^1 J(\beta_0 + u(\beta_1 - \beta_0))\,du\right)(\beta_1 - \beta_0).$$

□

Finally, we are ready to prove the asymptotic normality for $\widehat{\beta}$.

Proposition 7.9. *Under Conditions (C.1)-(C.4), $n^{1/2}(\widehat{\beta} - \beta_0)$ converges in distribution to a mean zero p-variate Gaussian random variable with covariance matrix $\{\Sigma(\beta_0, \tau)\}^{-1}$.*

Proof. Applying Lemma 7.8, we have

$$U(\widehat{\beta}, \tau) = U(\beta_0, \tau) - \left(\int_0^1 \mathcal{I}(\beta_0 + u(\widehat{\beta} - \beta_0))\,du\right)(\widehat{\beta} - \beta_0).$$

Recalling $U(\widehat{\beta}, \tau) = 0$, we have

$$\left(\int_0^1 \frac{1}{n}\mathcal{I}(\beta_0 + u(\widehat{\beta} - \beta_0))\,du\right)\sqrt{n}(\widehat{\beta} - \beta_0) = n^{-1/2}U(\beta_0, \tau).$$

By Proposition 7.7 (Part 1), $n^{-1/2}U(\beta_0, \tau)$ is asymptotically normal with covariance matrix $\Sigma(\beta_0, \tau)$.

Now consider

$$\left\| \int_0^1 \frac{1}{n}\mathcal{I}(\beta_0 + u(\widehat{\beta} - \beta_0))\,du - \Sigma(\beta_0, \tau) \right\| \leq \int_0^1 \left\| \frac{1}{n}\mathcal{I}(\widehat{\beta}(u)) - \Sigma(\beta_0, \tau) \right\| du$$

Since $\widehat{\beta}$ is consistent, Proposition 7.7 (Part 2) shows the uniform convergence of $\frac{1}{n}\mathcal{I}(\widehat{\beta}(u))$ to the nonsingular matrix $\Sigma(\beta_0, \tau)$ for $u \in [0, 1]$. Hence,

$$\int_0^1 \frac{1}{n}\mathcal{I}(\beta_0 + u(\widehat{\beta} - \beta_0))\, du \to \Sigma(\beta_0, \tau)$$

in probability. The result follows from Slutsky's Theorem. □

The proposition justifies the use of a confidence interval based on a multivariate normal random vector for inferring β. Also, heuristically, the variance of $\widehat{\beta}$ is approximately $n^{-1}\Sigma^{-1}(\beta_0, \tau)$, which can be estimated by $I^{-1}(\widehat{\beta})$, the inverse of the observed information. This approach has been implemented by the software.

7.5 Application to the Veterans' Administration lung cancer dataset

We applied the proportional hazards model to the Veterans' Administration Lung Cancer dataset, a publicly available dataset from the `survival` package in `R`. This dataset includes 137 male patients with advanced lung cancer who participated in a randomized trial comparing a standard treatment to an experimental treatment. Among them, 128 deaths were recorded, indicating a high mortality rate. The patients' ages ranged from 39 to 82 years, with a mean age of approximately 63 years. The dataset categorizes lung cancer into four types: squamous (reference group), small cell, adenocarcinoma, and large cell. Of the 137 patients, 31 (23%) had squamous cell carcinoma, 27 (20%) had small cell carcinoma, 14 (10%) had adenocarcinoma, and 25 (18%) had large cell carcinoma, while 40 (29%) had unspecified or missing cancer cell type data. Baseline physical function was assessed using the Karnofsky Performance Score (KPS), ranging from 10 (severe disability) to 90 (minimal disability), with an average score of 60. Most patients had KPS values below 70, indicating significant functional impairment. Our objectives were to evaluate the experimental treatment's impact on survival, analyze the effects of baseline KPS and age, and assess the influence of cancer cell types on survival outcomes. The dataset includes the following variables:

- **time**: Observed survival time in days.
- **status**: Event indicator (1 = death, 0 = censored).
- **trt**: Treatment group (1 = standard treatment, 2 = test treatment).
- **age**: Age of the patient in years.
- **celltype**: Type of cancer cell (squamous, small cell, adeno, large).
- **karno**: Karnofsky performance score (0–100, higher scores indicate better functioning).
- **diagtime**: Months from diagnosis to randomization.
- **prior**: Indicator of prior therapy (0 = no, 1 = yes).

We fitted a Cox proportional hazards model to assess the effects of treatment group, Karnofsky performance score, age, and cancer cell type (using squamous cell type as the reference group) on survival:

$$\lambda(t|Z) = \lambda_0(t)\exp(\beta_1\text{trt} + \beta_2\text{karno} + \beta_3\text{age} + \beta_4\text{small cell} + \beta_5\text{adeno} + \beta_6\text{large cell}),$$

where $\lambda_0(t)$ represents the baseline hazard function, and Z denotes the vector of covariates included in the model. The table below presents the results of the Cox model. The test treatment was associated with a hazard ratio of 0.85 (95% CI: 0.67–1.09), suggesting a 15% reduction in the hazard of death, though this result was not statistically significant ($p = 0.21$); see Figure 11. The Karnofsky performance score was a strong predictor of survival, with a hazard ratio of 0.96 (95% CI: 0.95–0.98, $p < 0.001$), indicating that better baseline functioning significantly reduces the hazard of death. Age was not significantly associated with survival ($HR = 1.01$, $p = 0.45$). Among cancer cell types, large-cell carcinoma had the worst prognosis, with a 45% higher hazard of death compared to squamous cell carcinoma ($HR = 1.45$, $p < 0.001$), followed by small-cell carcinoma ($HR = 1.25$, $p = 0.04$). Adenocarcinoma showed a modest but non-significant increase in hazard ($HR = 1.10$, $p = 0.42$).

Covariate	**Estimate (β)**	**SE**	**HR**	**95% CI**	**p-value**
Treatment (test)	-0.16	0.13	0.85	$[0.67, 1.09]$	0.21
Karnofsky score	-0.04	0.01	0.96	$[0.95, 0.98]$	< 0.001
Age	0.01	0.01	1.01	$[0.99, 1.03]$	0.45
Cell type (small cell)	0.22	0.11	1.25	$[1.01, 1.54]$	0.04
Cell type (adeno)	0.09	0.12	1.10	$[0.87, 1.39]$	0.42
Cell type (large)	0.37	0.12	1.45	$[1.15, 1.83]$	< 0.001

Table 1: Analysis of the Veterans' Administration Lung Cancer dataset using the Cox proportional hazards model.

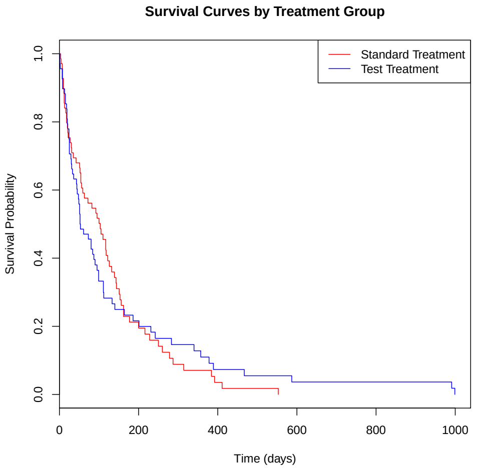


Figure 11: The survival comparison between the two treatment arms.

Below is the R code to fit the Cox model and reproduce the results:

```
# Load necessary packages
library(survival)
library(dplyr)

# Load the dataset
data(veteran)
```

```
# Fit the Cox proportional hazards model
cox_model <- coxph(Surv(time, status) ~ trt + karno + age + celltype, data =
    veteran)

# Display the results
summary(cox_model)

# You can also plot the survival curves based on the treatment groups
# Survival curves for the treatment groups
surv_fit <- survfit( Surv ( time , status ) ~ trt, data = veteran)

# Plot survival curves
plot(surv_fit, main="Survival␣Curves␣by␣Treatment␣Group", xlab="Time␣(days)", ylab
    ="Survival␣Probability", col = c("red", "blue"))
legend("topright", legend = c("Standard␣Treatment", "Test␣Treatment"), col = c("
    red", "blue"), lty = 1)
```

7.6 Cox models with time-dependent covariates

Time-dependent covariates are variables that change over time and can influence the hazard function dynamically. Unlike standard Cox proportional hazards models, which assume fixed covariates, models with time-dependent covariates allow for a more flexible and realistic representation of evolving risk factors. Time-dependent covariates can be categorized into two broad types:

- (External time-dependent covariates) These vary over time but are not influenced by the subject's survival status. Examples include: Temperature fluctuations affecting patient health; Changes in air pollution levels influencing respiratory conditions; Economic indicators affecting the risk of financial distress.
- (Internal time-dependent covariates) These depend on the subject's own history and potentially their survival status. Examples include: Blood pressure levels measured at different times in a cardiac study; Tumor size progression in an oncology study; CD4 cell count for HIV patients monitored over time.

For $i = 1, \ldots, n$, let $Z_i(t)$ be a p-dimensional time-dependent covariate vector for individual i. We also introduce

$$\widetilde{Z}_i(t) = \{Z_i(s) : 0 \le s \le t\},$$

which represents the trajectory of the covariate process from time 0 to t, i.e., the history of $Z_i(s)$ up to time t. We define the hazard function for individual i at t, given the trajectory of the covariate process up to t, as

$$\lambda_i(t) = \lim_{dt \to 0^+} \frac{1}{dt} P(t \le T_i < t + dt \mid T_i \ge t, \widetilde{Z}_i(t)).$$

If there is no ambiguity, we write

$$\lambda_i(t) = \lambda(t \mid \widetilde{Z}_i(t)),$$

emphasizing the dependence of the hazard on the covariate path. The conditional hazard function for the Cox model with time-dependent covariates is specified by:

$$\lambda(t \mid \widetilde{Z}_i(t)) = \lambda_0(t)\exp(Z_i(t)^\top \beta), \tag{7.22}$$

which states that, in this model, the hazard at time t depends only on the current covariate value $Z_i(t)$, given the past trajectory of covariates, reflecting a Markov-like assumption.

Define
$$\mathcal{F}_t = \sigma\{N_i(s), Y_i(s), Z_i(s), 1 \le i \le n, 0 \le s \le t\}.$$
It follows that $M_i(t) = N_i(t) - \int_0^t Y_i(s)\lambda(s \mid \widetilde{Z}_i(s))ds = N_i(t) - \int_0^t Y_i(s)\lambda_0(s)\exp(Z_i(s)^\top \beta_0)ds$ is a martingale with respect to $\mathcal{F}_t$. Here, we assume that (7.22) holds and β_0 is the true value of β.

As in the case of Cox models with time-independent covariates, the estimation of β_0 can be performed using partial likelihood. Suppose we observe n independent individuals, each observed with $(X_i, \Delta_i, \widetilde{Z}_(X_i))$. Define the risk set at time t is defined as: $R(t) = \{j : X_j \ge t\}$. Then similar to the derivation in Section , the partial likelihood function for time-dependent covariates is

$$L(\beta) = \prod_{i=1}^n \left[\frac{\exp(Z_i(X_i)^\top \beta)}{\sum_{j\in R(X_i)} \exp(Z_j(X_i)^\top \beta)}\right]^{\Delta_i}.$$

Taking the log of $L(\beta)$, we obtain

$$\begin{aligned}\ell(\beta) &= \sum_{i=1}^n \Delta_i \left[Z_i(X_i)^\top \beta - \log \sum_{j\in R(X_i)} \exp(Z_j(X_i)^\top \beta)\right] \\ &= \sum_{i=1}^n \int_0^\infty \left[Z_i(t)^\top \beta - \log \sum_{j=1}^n Y_j(t)\exp(Z_j(t)^\top \beta)\right] dN_i(t).\end{aligned}$$

This formulation accounts for the time-varying nature of $Z_i(\cdot)$ at the observed event times and properly handles censoring. To facilitate the theory, instead of using ∞ as the upper limit of the integral range, we consider the log partial likelihood with an upper integral limit of τ satisfying $\tau < \infty$ and $\Lambda_0(\tau) < \infty$ (normally, τ is taken to be the maximal followup time in the data):

$$\ell(\beta, \tau) = \sum_{i=1}^n \int_0^\tau \left[Z_i(t)^\top \beta - \log \sum_{j=1}^n Y_j(t)\exp(Z_j(t)^\top \beta)\right] dN_i(t). \tag{7.23}$$

Denote the maximum partial likelihood estimator as $\widehat{\beta}$. Under that $Z_i(\cdot)$ are bounded processes, we can show that $\widehat{\beta}$ is a consistent estimator of β. In addition, the MPLE satisfies:

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{d} N(0, \Sigma^{-1}(\beta_0, \tau)),$$

where $\Sigma(\beta_0, \tau)$ is as defined in (7.10) after modifying

$$S^{(k)}(\beta, t) = \frac{1}{n}\sum_{j=1}^n Y_j(t)Z_j(t)^{\otimes k} e^{Z_j(t)^\top \beta}, k = 0, 1, 2.$$

Here, $Z_j(t)^{\otimes 0} = 1, Z_j(t)^{\otimes 1} = Z_j(t), Z_j(t)^{\otimes 2}(t) = Z_j(t)Z_j(t)^\top$. The proofs of the consistency and normality results follow those of Propositions 7.3 and 7.9, respectively.

7.7 Setting up data for analysis with time-dependent covariates

Careful data preparation is crucial and we give some practical steps for structuring the dataset for proper modeling of the time-dependent nature of the covariate.. Briefly, unlike traditional Cox models, which assume fixed covariates, time-dependent covariates require a longitudinal (start-stop) format, where each individual may have multiple records corresponding to different time intervals. Each row in the dataset represents an interval $[t_{\text{start}}, t_{\text{end}}]$ during which the covariate values are constant.

Consider a simple dataset with two subjects. For Subject 1, the covariate is 2.1 from time 0 to 5 (event = 0), and it increases to 3.5 between times 5 and 8, when the subject fails at time 8 (event = 1). Subject 2 experiences the event at time 7 (event = 1), with a constant covariate value of 1.8 throughout the entire observation period. To conduct the analysis, the data must be formatted as follows:

ID	Entry Time (t_{start})	Exit Time (t_{end})	Event	Covariate ($Z(t)$)
1	0	5	0	2.1
1	5	8	1	3.5
2	0	7	1	1.8

Therefore, the first step for structuring the data is to develop a "start-stop" format. When covariates change over time, we break each individual's record into multiple rows, with each row representing a period where the covariate remains constant. The second step involves handling time-dependent covariates. It is crucial to ensure that covariate values are recorded at observed failure time points; if the data are collected intermittently, such as during medical checkups, it may be necessary to use interpolation or carry-forward methods to fill in the missing values. These methods help maintain the integrity of the data when covariate values are not observed at every time point.

The following is the R code for the example.

```r
library(survival)

# Example dataset: Start-Stop format
data <- data.frame(
  ID = c(1, 1, 2),
  start = c(0, 5, 0),
  stop = c(5, 8, 7),
  event = c(0, 1, 1),
  Z = c(2.1, 3.5, 1.8) # Time-dependent covariate
)

# Fit Cox model with time-dependent covariates
cox_model <- coxph(Surv(start, stop, event) ~ Z, data = data)
summary(cox_model)
```

The "start-stop" format implies that, at any given time t, only individuals who are still at risk of experiencing the event should be included in the analysis. The risk set at time t includes all individuals for whom the start time of the observation interval is less than or equal to t, and the end time is greater than t (i.e., $t_{\text{start}} \leq t < t_{\text{end}}$). We note that the partial likelihood based on the table or constructed by treating the three records as independent individuals remains the same, i.e.,

$$\frac{\exp(1.8\beta)}{\exp(1.8\beta) + \exp(3.5\beta)}. \tag{7.24}$$

Splitting the data does not change the risk set composition at each failure time. The start-stop format simply restructures the data to reflect periods of constant covariate values without altering the hazard function or an individual's contribution to the likelihood. Furthermore, the two sub-records for Subject 1 cannot experience an event simultaneously, so if treated as separate individuals, their corresponding counting processes would be orthogonal (see Section 8.1). As a result, these sub-records can be considered as coming from independent individuals without affecting the likelihood calculation. Therefore, even with multiple records originated from the same individuals, clustering does not need to be accounted for.

Finally, corresponding to (7.24), the maximum partial likelihood estimate of β is $-\infty$. This is due to perfect separation in the example data; larger values of the covariate Z correspond to lower hazards, meaning subjects with higher Z are less likely to fail earlier. We may add a penalty term (Ridge or Firth Correction) to the likelihood function to shrink the estimates and prevent divergence.

When incorporating time-dependent covariates in Cox models, we often consider these aspects. First, it is crucial to ensure that these covariates do not change in response to the outcome itself, as this could introduce time-dependent confounding. Such confounding arises when a covariate is associated with both the outcome and other variables in the model, potentially distorting the true relationships. Second, it is important to assess the proportional hazards assumption even when using time-dependent covariates. The Cox model assumes that the hazard ratio remains constant over time, but this assumption must be carefully validated. If the hazard ratio varies over time, alternative modeling approaches or stratification may be necessary. Finally, incorporating baseline covariates enhances the comprehensiveness of the model by accounting for factors that influence the hazard both at specific time points and over the study's duration. This improves the model's ability to capture the full complexity of survival data.

8 Competing Risks

Competing risks arise when an individual is subject to multiple potential failure types, meaning that an event can occur due to different causes. This is a common situation in medical studies, reliability engineering, and risk assessment. Unlike classical survival models, which assume a single failure type and treat all failures as the same event, competing risks explicitly account for the fact that different failure causes prevent the occurrence of others. In the following, we use T to denote the failure time (the time until the first occurrence of any failure) and J to denote the failure type, where $J \in \{1, 2, \ldots, K\}$. Thus, the observable data for each subject is a pair of (T, J), or the time of failure and its cause.

Classical survival analysis methods, particularly the Kaplan-Meier estimator, face two major limitations in the presence of competing risks. First, they typically assume non-informative censoring; however, in many clinical settings, competing events introduce dependent (informative) censoring, violating this assumption and leading to biased estimates. For example, in studies of lung cancer–specific mortality, deaths from cardiovascular disease or infection may reflect underlying health status and therefore cannot be treated as non-informative censoring. Second, these methods estimate survival under a hypothetical scenario in which all other causes of failure are removed. In hematopoietic stem cell transplantation, for instance, applying Kaplan-Meier to estimate mortality due to hematologic relapse may underestimate the true risk by ignoring transplant-related deaths, such as those due to graft-versus-host disease. To overcome these limitations, a competing risks framework should be adopted. It models each cause of failure using either cause-specific hazards or cumulative incidence functions, while properly accounting for the presence of other competing events. To illustrate, consider a study of patients undergoing heart surgery, where the outcome of interest is death. Patients may die from various causes, such as cardiac-related deaths (e.g., heart failure or stroke) or non-cardiac-related deaths (e.g., infection or cancer). Competing risks analysis allows researchers to estimate the probability of each type of event over time and address questions such as

1. What is the probability of dying from a specific cause by a given time?
   - This is captured by the cumulative incidence function (CIF) (defined later), which estimates the probability that a patient dies from a specific cause before time t, accounting for other competing causes.
   - Example: “What is the probability that a patient dies from a cardiac-related cause within five years of surgery?”
2. How do different factors influence the cause-specific failure rates?
   - Using cause-specific hazard models (defined later), we can study how risk factors (e.g., age, pre-existing conditions, lifestyle) affect the likelihood of different types of death.
   - Example: “Does smoking increase the risk of cardiac-related death more than non-cardiac-related death?”
3. What is the relative contribution of each failure type to overall mortality?
   - The CIF can be used to compare the relative proportion of deaths due to each cause over time.
   - Example: “Among patients who die within five years, what percentage die due to cardiac-related versus non-cardiac-related causes?”
4. How do treatments affect the risk of different types of death?
   - A competing risks framework helps evaluate whether a treatment reduces mortality for all causes or just for specific ones.
   - Example: “Does a new heart surgery technique reduce cardiac-related mortality without increasing the risk of non-cardiac-related death?”
5. What is the expected time to failure for each failure type?

- Instead of treating time-to-event as a single outcome, competing risks models can estimate time-to-failure distributions separately for each cause.
- Example: "On average, how long after surgery do cardiac-related versus non-cardiac-related deaths occur?"

These questions center on the following two concepts in competing risks.

- Cause-specific hazard function, which represents the instantaneous rate of failure from a cause, say cause k, as
$$\alpha_k(t) = \lim_{dt \to 0} \frac{P(T \in [t, t+dt), J = k \mid T \geq t)}{dt}. \tag{8.1}$$
Obviously, it connects with the overall hazard function via $\lambda(t) = \sum_{k=1}^{K} \alpha_k(t)$.
- Cumulative incidence function (CIF) (also referred to as the subdistribution function), which gives the probability of failing from a cause, say, cause k, by time t, as
$$F_k(t) = P(T \leq t, J = k). \tag{8.2}$$
Corresponding to (8.3), we also define the CIF density function (commonly known as subdistribution density) as
$$f_k(t) = \lim_{dt \to 0+} \frac{1}{dt} P(t \leq T_i < t + dt, J_i = k). \tag{8.3}$$

8.1 Multivariate counting processes

For each independent individual $i = 1, \ldots, n$, we use the notion of counting process to represent the number of events of type $k = 1, \ldots, K$, occurring by time t, which is a step function with at most one jump from 0 to 1. In the absence of censoring, we would define:

- $N_{k,i}(t)$ as the counting process for failure type k, where $k = 1, \ldots, K$, for subject $i = 1, \ldots, n$.
- Define $N_i(t) = (N_{1,i}(t), \ldots, N_{K,i}(t))$, a multivariate counting process for subject i with all K types of failure, that is, each component tracks the occurrence of a specific type of failure, ensuring only one can happen per subject.

Each $N_{k,i}(t)$ is a right-continuous, increasing process that takes values in $\{0, 1\}$ for each individual, where $N_{k,i}(t) = 1$ if failure type k has occurred by time t and 0 otherwise. That $N_i(t)$ is a multivariate counting process follows from Definition 2.5.1 of Fleming & Harrington (2013) as no two component processes jump at the same time.

To incorporate censoring, we introduce the censoring time C_i, and define the observed time as $X_i = \min(T_i, C_i)$ and censoring indicator $\Delta_i = I(T_i \leq C_i)$. We observe $(X_i, \Delta_i, J_i^* = J_i\Delta_i)$, where J_i^* ensures that we only observe the failure type J_i for subject i when $\Delta_i = 1$. The counting process is then modified to:
$$N_{k,i}(t) = I(X_i \leq t, \Delta_i = 1, J_i^* = k). \tag{8.4}$$
Additionally, we define the at-risk process for subject i, as $Y_i(t) = I(X_i \geq t)$, which indicates

whether an individual is still under observation at time t. With the modified definition of $N_{k,i}$, $N_i(t) = (N_{1,i}(t), \ldots, N_{K,i}(t))$ is still a multivariate counting process as each $N_{k,i}(t)$ is a counting process, with no two processes, e.g., $N_{k,i}(t)$ and $N_{k',i}(t), k \neq k'$, jump at the same time.

Proposition 8.1. *Assume T_i, J_i are independent of C_i and T_i is continuous. If $P(X_i \geq t) > 0$, then*

$$\lim_{dt \to 0^+} \frac{1}{dt} P(t \leq X_i < t + dt,\ \Delta_i = 1, J_i^* = k \mid X_i \geq t) = \alpha_k(t).$$

Proof. We consider

$$P(t \leq X_i < t + dt, \Delta_i = 1, J_i^* = k \mid X_i \geq t) = \frac{P(t \leq X_i < t + dt, \Delta_i = 1, J_i = k)}{P(X_i \geq t)},$$

while expanding the numerator,

$$\begin{aligned} P(t \leq X_i < t + dt, \Delta_i = 1, J_i = k) &= P(t \leq T_i < t + dt, J_i = k, T_i \leq C_i) \\ &= P(t \leq T_i < t + dt, J_i = k) P(T_i \leq C_i \mid T_i \in [t, t + dt], J_i = k). \end{aligned}$$

Since T_i is continuous, we have

$$\lim_{dt \to 0^+} \frac{1}{dt} \frac{P(t \leq T_i < t + dt, J_i = k)}{P(T_i \geq t)} = \alpha_k(t).$$

With $f_k(s)$ defined in (8.8), we consider

$$\begin{aligned} & P(T_i \leq C_i \mid T_i \in [t, t + dt), J_i = k) \\ = & \frac{P(T_i \leq C_i, T_i \in [t, t + dt), J_i = k)}{P(T_i \in [t, t + dt), J_i = k)} \\ = & \frac{\int_t^{t+dt} f_k(s) P(C_i \geq s) ds}{\int_t^{t+dt} f_k(s) ds}, \end{aligned}$$

where the last equality comes from the independence of C_i with T_i and J_i. Let $dt \to 0^+$ and apply L'Hôpital's rule, we have

$$\lim_{dt \to 0^+} P(T_i \leq C_i \mid T_i \in [t, t + dt), J_i = k) = \frac{f_k(t) P(C_i \geq t)}{f_k(t)} = P(C_i \geq t).$$

Putting all pieces together, we have

$$\begin{aligned} & \lim_{dt \to 0^+} \frac{1}{dt} P(t \leq X_i < t + dt,\ \Delta_i = 1 \mid X_i \geq t) \\ = & \lim_{dt \to 0+} \frac{1}{dt} \frac{P(t \leq T_i < t + dt, J_i = k, T_i \leq C_i)}{P(X_i \geq t)} \\ = & \lim_{dt \to 0+} \frac{1}{dt} \frac{P(t \leq T_i < t + dt, J_i = k)}{P(T_i \geq t)} \times \frac{P(T_i \geq t)}{P(X_i \geq t)} \lim_{dt \to 0+} P(T_i \leq C_i \mid T_i \in [t, t + dt), J_i = k) \\ = & \alpha_k(t) \times \frac{P(T_i \geq t) P(C_i \geq t)}{P(X_i \geq t)} = \alpha_k(t). \end{aligned}$$

□

The result indicates that under independent censoring, the observed data can be used to estimate the cause-specific hazard function, for example, by using the Nelson-Aalen. Moreover, it can help develop the martingale framework for competing risks, as shown below.

8.2 Martingale representation

Proposition 8.2. *Define*

$$\mathcal{F}_t = \sigma\{N_{k,i}(s), Y_i(s) : 0 \leq s \leq t, k = 1, \ldots, K, i = 1, \ldots, n\}.$$

Under the independent censoring assumption, the compensated process:

$$M_{k,i}(t) = N_{k,i}(t) - \int_0^t Y_i(s)\alpha_k(s)ds, \tag{8.5}$$

is a martingale with respect to $\mathcal{F}_t$.

Proof. To prove that $M_{k,i}(t)$ is a martingale with respect to $\mathcal{F}_t$, we need to show

(i) the process $M_{k,i}(t)$ is adapted to $\mathcal{F}_t$;

(ii)
$$\mathbb{E}[|M_{k,i}(t)|] < \infty \quad \text{for all } t;$$

(iii)
$$\mathbb{E}[dM_{k,i}(t) \mid \mathcal{F}_{t-}] = 0.$$

First, the process $M_{k,i}(t)$ involves $N_{k,i}(t)$, which is adapted to $\mathcal{F}_t$, and the integral $\int_0^t Y_i(s)\alpha_k(s)\,ds$, where $Y_i(s)$ is adapted to $\mathcal{F}_s$ and $\alpha_k(s)$ is non-random. Hence, the integral is predictable and adapted to $\mathcal{F}_t$. Therefore, $M_{k,i}(t)$ is adapted to $\mathcal{F}_t$.

Second, as $N_{k,i}(t) \leq 1$, we have that $\mathbb{E}[N_{k,i}(t)] \leq 1$. With $\int_0^t Y_i(s)\alpha_k(s)ds \leq \int_0^t Y_i(s)\lambda(s)ds$, it follows that

$$\mathbb{E}\int_0^t Y_i(s)\lambda(s)ds \leq 1$$

as shown in the proof of Proposition 2.7. Therefore, $M_{k,i}(t)$ is integrable as $\mathbb{E}[|M_{k,i}(t)|] \leq 2$.

Finally, with

$$dM_{k,i}(t) = dN_{k,i}(t) - Y_i(t)\alpha_k(t)\,dt,$$

we compute conditional expectation of $dM_{k,i}(t)$ given $\mathcal{F}_{t-}$ by considering the conditional expectation of each term. In particular, using Proposition 8.1 and following the proof of Proposition 2.7, we have

$$\mathbb{E}[dN_{k,i}(t) \mid \mathcal{F}_{t-}] = Y_i(t)\alpha_k(t)\,dt.$$

The term $Y_i(t)\alpha_k(t)\,dt$ is predictable and hence measurable with respect to $\mathcal{F}_{t^-}$, implying

$$\mathbb{E}[Y_i(t)\alpha_k(t)\,dt \mid \mathcal{F}_{t-}] = Y_i(t)\alpha_k(t)\,dt.$$

Therefore,

$$\mathbb{E}[dM_{k,i}(t) \mid \mathcal{F}_{t-}] = Y_i(t)\alpha_k(t)\,dt - Y_i(t)\alpha_k(t)\,dt = 0.$$

□

Further, define $N_k(t) = \sum_{i=1}^n N_{k,i}(t)$, $Y(t) = \sum_{i=1}^n Y_i(t)$ and $M_k(t) = \sum_{i=1}^n M_{k,i}(t)$, then $M_k(t) = N_k(t) - \int_0^t Y(s)\alpha_k(s)ds$ is a martingale with respect to $\mathcal{F}_t$. This follows because each $M_{k,i}$ is a martingale with respect to $\mathcal{F}_t$. Immediately, we have

$$\langle M_k\rangle(t) = \int_0^t Y(s)\alpha_k(s)ds$$

and

$$\langle M_k, M_{k'}\rangle(t) = 0 \tag{8.6}$$

which follow from Theorem 2.5.2 of Fleming & Harrington (2013). This means, under the competing risk framework, any two cause-specific martingale process of $M_k, M_{k'}$ are orthogonal (or uncorrelated) as the corresponding counting processes cannot jump at the same time.

8.3 The Nelson-Aalen estimator of the cause-specific hazard

Define the cumulative cause-specific hazard function as

$$A_k(t) = \int_0^t \alpha_k(s)ds.$$

It does not represent a probability, but rather a measure of the expected number of failures (per unit population) from cause k by time t in the presence of competing risks. We are interested in estimating it because CIF depends on it while accounting for the risk of failure from other causes.

In the following, we use the Nelson-Aalen estimator to nonparametrically estimate $A_k(t)$ and then discuss its property.

$$\widehat{A}_k(t) = \int_0^t \frac{dN_k(s)}{Y(s)}. \tag{8.7}$$

Proposition 8.3. *If $u \in (0, \infty]$ is such that*

$$Y(s) \to \infty \quad \textit{in probability as } n \to \infty,$$

for any $s \leq u$, then

$$\sup_{0\leq s\leq u} |\widehat{A}_k(s) - A_k(s)| \to 0 \quad \textit{as } n \to \infty.$$

Proof. Note that

$$\widehat{A}_k(s) - A_k(s) = \int_0^t \frac{dM_k(s)}{Y(s)},$$

and follow the proof of Proposition 8.3. □

8.4 The estimator of cumulative incidence function

The cumulative incidence function (CIF) as defined in (8.3) gives the probability of failing from cause k by time t and it can be shown that it is related to the cause-specific hazard via

$$F_k(t) = \int_0^t S(s^-)\alpha_k(s)ds = \int_0^t S(s^-)dA_k(s), \tag{8.8}$$

where the overall survival function $S(t) = P(T > t)$ and $S(t^-) = P(T \geq t)$. Obviously, $S(t) = S(t^-)$ when T is continuous. With (8.8), it is natural to estimate the CIF with

$$\widehat{F}_k(t) = \int_0^t \widehat{S}(s^-)d\widehat{A}_k(s),$$

where $\widehat{S}$ is the Kaplan-Meier estimate of $S(t)$ and $\widehat{A}_k(s)$ is as defined in (8.7). We now prove the uniform consistency of $\widehat{F}_k(t)$.

Proposition 8.4. *If $u \in (0, \infty]$ is such that*

$$Y(s) \to \infty \quad \textit{in probability as } n \to \infty,$$

for any $s \leq u$, then

$$\sup_{t \leq u} |\widehat{F}_k(t) - F_k(t)| \to 0$$

in probability.

Proof. We first consider

$$\begin{aligned}
&\widehat{F}_k(t) - F_k(t) \\
= &\int_0^t \widehat{S}(s^-)d(\widehat{A}_k(s) - A_k(s)) + \int_0^t (\widehat{S}(s^-) - S(s^-))dA_k(s) \\
= &\int_0^t \frac{\widehat{S}(s^-)}{Y(s)}dM_k(s) + \int_0^t (\widehat{S}(s^-) - S(s^-))dA_k(s).
\end{aligned}$$

We apply the Lenglart inequality for the first item. Specifically, we let $Z_k(t) = \int_0^t \frac{\widehat{S}(s-)}{Y(s)}dM_k(s)$. As $\widehat{S}(s-)$ and $Y(s)$ are measurable with respect to $\mathcal{F}_{t^-}$, it follows that $Z_k(t)$ is a locally square integrable martingale with respect to $\mathcal{F}_t$. In addition, its quadratic varition process is

$$\langle Z_k \rangle(t) = \int_0^t \frac{\widehat{S}^2(s-)}{Y^2(s)}d\langle M_k \rangle(s) = \int_0^t \frac{\widehat{S}^2(s-)}{Y(s)}dA_k(s).$$

So, for any $\epsilon > 0$, we are ready to use the Lenglart inequality to quantify $P\left\{\sup_{0\le s\le u} Z_k^2(s) > \varepsilon\right\}$. In fact, by the Lenglart inequality, for any $\eta > 0$, it holds that

$$P\left\{\sup_{0\le s\le u} |Z_k(s)| > \sqrt{\varepsilon}\right\} = P\left\{\sup_{0\le s\le u} Z_k^2(s) > \varepsilon\right\} < \frac{\eta}{\varepsilon} + P\left\{\int_0^u \frac{\widehat{S}^2(s-)}{Y(s)} dA_k(s) > \eta\right\}$$

$$< \frac{\eta}{\varepsilon} + P\left\{\frac{A_k(u)}{Y(u)} > \eta\right\} < \frac{\eta}{\varepsilon} + P\left\{\frac{\Lambda(u)}{Y(u)} > \eta\right\}$$

as $A_k(u) < \Lambda(u) < \infty$, the overall cumulative hazard. Since $Y(u) \to \infty$ in probability as $n \to \infty$, the second term on the right-hand side above converges to zero as $n \to \infty$ for any $\eta > 0$. Since η and ϵ are arbitrary, the uniformly convergence holds.

Finally, because

$$\sup_{t\le u}\left|\int_0^t (\widehat{S}(s^-) - S(s^-)) dA_k(s)\right| \le \sup_{t\le u} |\widehat{S}(t) - S(t)| A_k(u)$$

$$\le \sup_{t\le u} |\widehat{S}(t) - S(t)| \Lambda(u),$$

it converges to 0 in probability because $\Lambda(u) < \infty$ and by the uniform consistency of $\widehat{S}(t)$ for $t \le u$ (Proposition 4.4). □

8.5 Cause-specific proportional hazards models

To model the impact of covariates on the *cause-specific hazard function*, say, $\alpha_k(t)$ for failure type k, we may use a *Cox-type proportional hazards framework*, which evaluates how covariates influence the instantaneous risk of experiencing a specific type of failure. This modeling approach is useful in competing risks settings, where subjects are at risk of multiple mutually exclusive failure types.

Let Z_i denote the vector of covariates associated with subject i. The *cause-specific hazard* for failure type k, conditional on Z_i, is specified as:

$$\alpha_k(t \mid Z_i) = \alpha_{k,0}(t) \exp\left(\beta_k^\top Z_i\right),$$

where $\alpha_{k,0}(t)$ is the *baseline hazard function* for cause k, representing the hazard function when all covariates set to zero; and β_k is a vector of *regression coefficients* specific to failure type k, quantifying the log-relative effect of covariates on the cause-specific hazard. This formulation assumes *proportional hazards* for each cause: the hazard for failure type k is proportional across individuals with different covariate profiles, and the proportionality factor is given by $\exp(\beta_k^\top Z_i)$. The coefficients β_k describe the effect of covariates on the hazard of failing from cause k, in the presence of competing risks. A positive coefficient $\beta_{kj} > 0$ implies that the j-th covariate increases the risk of failure from cause k. In the following, we use $\beta_k^0 \in \mathbb{R}^p$ to denote the true value of β_k^0 and our main goal is to estimate β_k^0 using the observed data.

The cause-specific proportional hazards model can be estimated using the partial likelihood method from Cox regression, treating failures from other causes as censored at their failure times.

For each cause k, define the risk set $\mathcal{R}(t) = \{i : X_i \geq t\}$, and consider only the individuals who fail from cause k as events, while others (including failures from different causes and censored observations) contribute to the risk set. The partial likelihood for cause k is given by:

$$L_k(\beta_k) = \prod_{i:J_i=k,\Delta_i=1} \frac{\exp(\beta_k^\top Z_i)}{\sum_{j\in\mathcal{R}(T_i)} \exp(\beta_k^\top Z_j)}. \tag{8.9}$$

The log partial likelihood is:

$$\ell_k(\beta_k) = \sum_{i:J_i=k,\Delta_i=1} \left[\beta_k^\top Z_i - \log\left(\sum_{j\in\mathcal{R}(T_i)} \exp(\beta_k^\top Z_j) \right) \right]. \tag{8.10}$$

Maximizing $\ell_k(\beta_k)$ provides estimates of β_k for each cause. As shown later, we can fit separate Cox models for each failure type as the estimates for each failure type are asymptotically independent.

8.6 Large Sample Theory for the Cause-Specific Proportional Hazards Estimator

Suppose we observe n independent and identically distributed survival data with competing risks:

$$\{(X_i, \Delta_i, J_i, Z_i),\ i = 1, \ldots, n\},$$

where $X_i = \min(T_i, C_i)$ is the observed time, $\Delta_i = I(T_i \leq C_i)$ is the event indicator, and $J_i \in \{1, \ldots, K\}$ is the cause of failure (if $\Delta_i = 1$), $Z_i \in \mathbb{R}^p$ is the covariate vector.

Recalling $N_{k,i}(t) = I\left(X_i \leq t, \Delta i = 1, J_i^* = k\right)$ where $J_i^* = J_i\Delta_i$ and $Y_i(t) = I(X_i \geq t)$, we can rewrite (8.10) as

$$l_k(\beta_k) = \sum_{i=1}^n \int_0^\infty [\beta_k^\top Z_i - \log(\sum_{j=1}^n Y_j(s) e^{\beta_k^\top Z_j})] dN_{k,i}(s).$$

To facilitate the theory, instead of using ∞ as the upper limit of the integral range, we consider the log partial likelihood with an upper integral limit of τ satisfying $\tau < \infty$ and $\max_k \int_0^\tau \alpha_{k,0}(s)ds < \infty$. The choice of τ instead of ∞ as the upper limit of integration helps prevent divergence and ensures that the integral is restricted to a finite observation period. This avoids unrealistic assumptions about unobserved or censored times. In practice, τ is often chosen as the maximum observation period in the study.

Then the MPLE $\widehat{\beta}_k$ is obtained by maximizing

$$l_k(\beta_k, \tau) = \sum_{i=1}^n \int_0^\tau [\beta_k^\top Z_i - \log(\sum_{j=1}^n Y_j(s) e^{\beta_k^\top Z_j})] dN_{k,i}(s). \tag{8.11}$$

The added τ (or t in later development) in the likelihood emphasizes the time-dependent nature of the information, specifically the use of data available up to time τ (or time t), which will be critical in the theoretical development. We study the large sample results for $\widehat{\beta}_k, k = 1, \dots, K$.

As in Section 7.2, we first introduce notation adapted to the competing risk setting. For $k = 1, \dots, K$, we introduce

$$S_k^{(0)}(\beta, t) = \frac{1}{n}\sum_{j=1}^{n} Y_j(t) e^{Z_j^\top \beta_k}, S_k^{(1)}(\beta, t) = \frac{1}{n}\sum_{j=1}^{n} Y_j(t) Z_j e^{Z_j^\top \beta_k}, S_k^{(2)}(\beta, t) = \frac{1}{n}\sum_{j=1}^{n} Y_j(t) Z_j^{\otimes 2} e^{Z_j^\top \beta_k}.$$

Here, $Z_j^{\otimes 2} = Z_j Z_j^\top$. Then the score function associated with (8.11) is

$$U^{(k)}(\beta_k, \tau) = \sum_{i=1}^{n} \int_0^\tau \left(Z_i - \bar{Z}_k(\beta_k, t)\right) dN_{k,i}(t), \tag{8.12}$$

where

$$\bar{Z}_k(\beta_k, t) = \frac{S_k^{(1)}(\beta_k, t)}{S_k^{(0)}(\beta_k, t)}, \tag{8.13}$$

and the Hessian matrix of (8.11) is

$$H_k(\beta_k) = -\int_0^\tau \left[\frac{S_k^{(2)}(\beta_k, t)}{S_k^{(0)}(\beta, t)} - \left\{ \frac{S_k^{(1)}(\beta_k, t)}{S_k^{(0)}(\beta_k, t)} \right\}^{\otimes 2} \right] dN_k(t) \overset{def}{=} -\mathcal{I}_k(\beta_k), \tag{8.14}$$

where $N_k(t) = \sum_{i=1}^n N_{k,i}(t)$, and $\mathcal{I}_k(\beta_k)$ is the observed information matrix. Moreover, following the convention used in the univariate Cox model asymptotics, we use $\|\cdot\|$ to denote the maximum absolute value (sup-norm) of the elements of a vector or matrix, and $|\cdot|$ to indicate the Euclidean norm for vectors or the absolute value for scalars.

We then introduce the regularity conditions adapted to the competing risk setting. For $k = 1, \dots, K$, we assume

(C.1') there exists an open and convex neighborhood $\mathcal{B}_k$ of $\beta_k^0 \in \mathbb{R}^p$ and, respectively, scalar, vector, and matrix functions, $s_k^{(0)}, s_k^{(1)}, s_k^{(2)}$ such that

$$\sup_{t \in [0,\tau], \beta_k \in \mathcal{B}_k} ||S_k^{(0)}(\beta_k, t) - s_k^{(0)}(\beta_k, t)|| \to 0,$$

$$\sup_{t \in [0,\tau], \beta_k \in \mathcal{B}_k} ||S_k^{(1)}(\beta_k, t) - s_k^{(1)}(\beta_k, t)|| \to 0,$$

$$\sup_{t \in [0,\tau], \beta_k \in \mathcal{B}_k} ||S_k^{(2)}(\beta_k, t) - s_k^{(2)}(\beta_k, t)|| \to 0$$

in probability.

(C.2') In the same $\mathcal{B}_k$, it holds that, for any $\beta_k \in \mathcal{B}_k$ and $t \in [0, \tau]$,

$$s_k^{(1)}(\beta_k, t) = \frac{\partial}{\partial \beta_k} s^{(0)}(\beta_k, t),\ s_k^{(2)}(\beta_k, t) = \frac{\partial}{\partial \beta_k} s_k^{(1)}(\beta_k, t) = \frac{\partial^2}{\partial \beta_k \partial \beta_k^\top} s_k^{(0)}(\beta_k, t).$$

We also assume each element of $s_k^{(0)}(\beta_k, t), s_k^{(1)}(\beta_k, t), s_k^{(2)}(\beta_k, t)$ is bounded, and in addition, $s_k^{(0)}(\beta_k, t)$ is bounded away from 0 in $\mathcal{B}_k \times [0, \tau]$. In addition, for any $t \in [0, \tau]$, $s_k^{(0)}(\beta_k, t), s_k^{(1)}(\beta_k, t), s_k^{(2)}(\beta_k, t)$ are equicontinuous at β_k^0.

(C.3') Define

$$v_k(\beta_k, t) = \frac{s_k^{(2)}(\beta_k, t)}{s_k^{(0)}(\beta_k, t)} - \left(\frac{s_k^{(1)}(\beta_k, t)}{s_k^{(0)}(\beta_k, t)}\right)^{\otimes 2}. \tag{8.15}$$

We assume

$$\Sigma_k(\beta_k^0, \tau) \overset{def}{=} \int_0^\tau v_k(\beta_k^0, s) s_k^{(0)}(\beta_k^0, s) \alpha_{k,0}(s) ds \tag{8.16}$$

is positive definite.

(C.4') There exists a $\delta > 0$ so that

$$\sup_{1 \le i \le n, t \in [0,\tau]} n^{-1/2} |Z_i| Y_i(t) I(Z_i^\top \beta_k^0 > -\delta |Z_i|) \to 0$$

in probability.

Under these conditions, we establish the consistency and asymptotic normality of the MPLE $\widehat{\beta}_k$ as follows.

Asymptotic Consistency

Proposition 8.5. *Under Assumptions (C.1')-(C.3'), the MPLE $\widehat{\beta}_k$ where $1 \le k \le K$, is* ***consistent****, i.e.,*

$$\widehat{\beta}_k \xrightarrow{P} \beta_k^0,$$

as $n \to \infty$, where β_k^0 is the true value of the parameter vector for cause k.

Proof. Let us introduce $S_k^{(m)}(\beta_k, t) = \frac{1}{n} \sum_{j=1}^n Y_j(t) Z_j^{\otimes m} e^{\beta_k^\top Z_j}, m = 0, 1, 2$, where $S_k^{(m)}(\cdot, \cdot)$ be a functions on $\mathcal{B}_k \times [0, \tau]$, where $\mathcal{B}_k$ be the open neighborhood for β_k^0 from (C.1). Now we define the MPLE estimate $\widehat{\beta}_k$ for β_k^0 as the solution by maximizing the partial likelihood over $[0, \tau]$ which is $l_k(\beta_k, \tau)$ where $l_k(\beta_k, \tau) = \sum_{i=1}^n \int_0^\tau [\beta_k^\top Z_i - \log(\sum_{j=1}^n Y_j(s) e^{\beta_k^\top Z_j}) dN_{k,i}(s)]$. We introduce $nX_{k,n}(\beta_k, .)$ as the process which, at time t, is the difference in log partial likelihoods over $[0, t]$ evaluated at an arbitrary β_k and the true value β_k^0, i.e.,

$$X_{k,n}(\beta_k, t) = n^{-1}\{l_k(\beta_k, t) - l_k(\beta_k^0, t)\} = n^{-1} \sum_{i=1}^n \int_0^t \left[(\beta_k - \beta_k^0)^\top Z_i - \log \frac{S_k^{(0)}(\beta_k, s)}{S_k^{(0)}(\beta_k^0, s)}\right] dN_{k,i}(s).$$

Also define,

$$A_{k,n}(\beta_k, t) = n^{-1} \sum_{i=1}^n \int_0^t \Big[(\beta_k - \beta_k^0)^\top Z_i - \log \frac{S_k^{(0)}(\beta_k, s)}{S_k^{(0)}(\beta_k^0, s)}\Big] Y_i(s) e^{\beta_k^\top Z_i} \alpha_{k,0}(s) ds.$$

Next we define the right continuous squared-integrable Martingale w.r.t $\mathcal{F}_t$,

$$M_{k,i}(t) = N_{k,i}(t) - \int_0^t Y_i(s) e^{\beta_k^\top Z_i} \alpha_{k,0}(s) ds$$

with $\langle M_{k,i}\rangle(t) = \int_0^t Y_i(s) e^{\beta_k^\top Z_i} \alpha_{k,0}(s) ds$. Thus,

$$X_{k,n}(\beta_k, t) - A_{k,n}(\beta_k, t) = n^{-1} \sum_{i=1}^n \int_0^t \Big[(\beta_k - \beta_k^0)^\top Z_i - \log \frac{S_k^{(0)}(\beta_k, s)}{S_k^{(0)}(\beta_k^0, s)}\Big] dM_{k,i}(s).$$

Now, choose $\tau_{n,i,k} = n \wedge \sup\{s : \big|(\beta_k - \beta_k^0)^\top Z_i - \log \frac{S_k^{(0)}(\beta_k,s)}{S_k^{(0)}(\beta_k^0,s)}\big| \leq n\}$, then

$$\int_0^{t\wedge\tau_{n,i,k}} \Big[(\beta_k - \beta_k^0)^\top Z_i - \log \frac{S_k^{(0)}(\beta_k, s)}{S_k^{(0)}(\beta_k^0, s)}\Big] dM_{k,i}(s)$$

is a square integrable martingale by **Property 2** since the integrand is bounded over $t \wedge \tau_{n,i,k}$. Therefore, for any given $\beta_k \in \mathcal{B}_k$ (convex neighbourhood of β_k^0), the process $X_{k,n}(\beta_k, \cdot) - A_{k,n}(\beta_k, \cdot)$ is a local square integrable martingale with the predictable variation process at t,

$$\begin{aligned}
&\langle X_{k,n}(\beta_k, \cdot) - A_{k,n}(\beta_k, \cdot)\rangle(t) \\
=&n^{-2} \sum_{i=1}^n \int_0^t \big[(\beta_k - \beta_k^0)^\top Z_i - \log \frac{S_k^{(0)}(\beta_k, s)}{S_k^{(0)}(\beta_k^0, s)}\big]^2 d\langle M_{k,i}\rangle(s) \\
=&n^{-2} \sum_{i=1}^n \int_0^t \big[(\beta_k - \beta_k^0)^\top Z_i - \log \frac{S_k^{(0)}(\beta_k, s)}{S_k^{(0)}(\beta_k^0, s)}\big]^2 Y_i(s) e^{\beta_k^{0\top} Z_i} \alpha_{k,0}(s) ds \\
=&n^{-1} \int_0^t \Big[(\beta_k - \beta_k^0)^\top S_k^{(2)}(\beta_k^0, s)(\beta_k - \beta_k^0)^\top - 2(\beta_k - \beta_k^0)^\top S_k^{(1)}(\beta_k^0, s) \log \frac{S_k^{(0)}(\beta_k, s)}{S_k^{(0)}(\beta_k^0, s)} \\
&+ \big\{\log \frac{S_k^{(0)}(\beta_k, s)}{S_k^{(0)}(\beta_k^0, s)}\big\}^2 S_k^{(0)}(\beta_k^0, s)\Big] \alpha_{k,0}(s) ds \\
&\xrightarrow{p} 0.
\end{aligned}$$

The last step is due to (C.1') and (C.2') being valid for any $\beta_k \in \mathcal{B}_k, k = 1, \ldots, K$. Now choosing $H_{k,n}(\beta_k, t) = X_{k,n}(\beta_k, t) - A_{k,n}(\beta_k, t)$ for $t \in [0, \tau]$ we apply the Lenglart inequality (Lemma 7.2 for fixed k to conclude $X_{k,n}(\beta_k, t) - A_{k,n}(\beta_k, t) \xrightarrow{P} 0$ uniformly over the range of $t \in [0, \tau]$, since the RHS in the Lenglart equation (Lemma 7.2) can be made arbitrarily small for given $\epsilon > 0$. Under Condition (C.1') for all $\beta_k \in \mathcal{B}_k$, $k = 1, \ldots, K$, with $s_k^{(0)}(\cdot, \cdot), s_k^{(1)}(\cdot, \cdot)$ and $s_k^{(2)}(\cdot, \cdot)$ being respective dominant functions, we have

$$A_{k,n}(\beta_k, \tau) \to A_k(\beta_k, \tau) = \int_0^\tau \Big[(\beta_k - \beta_k^0)^\top s_k^{(1)}(\beta_k^0, s) - \log \frac{s_k^{(0)}(\beta_k, s)}{s_k^{(0)}(\beta_k^0, s)} s_k^{(0)}(\beta_k^0, s)\Big] \alpha_{k,0}(s) ds.$$

It follows that $X_{k,n}(\beta_k,\tau)$ must converge in probability to the same limit, as long as $\beta_k\in\mathcal{B}_k$. Clearly, $X_n(,\tau)$ is a concave function of $\beta_k\in\mathcal{B}_k$ with unique maxima and under conditions (C.2') and (C.3') (positive definity)(for each k) $A_k(\beta_k,\tau)$ has unique maxima at β_k^0. Thus applying lemma (7.2) of notes, we claim $\widehat{\beta}_k$ which is the solution of maximizing partial likelihood over $[0,\tau]$ converges in probability to β_k^0 i.e. $\widehat{\beta}_k$ is consistent estimator for β_k^0. □

Asymptotic Normality

Proposition 8.6. *Under Conditions (C.1')-(C.4'), $n^{1/2}(\widehat{\beta}_k-\beta_k^0)$ converges in distribution to a mean zero p-variate Gaussian random variable with covariance matrix $\{\Sigma_k(\beta_k^0,\tau)\}^{-1}$.*

Proof. Recall that $U^{(k)}(\beta_k^0,\tau)=\sum_{i=1}^n\int_0^\tau\{Z_i-\bar{Z}_k(\beta_k^0,t)\}dN_{k,i}(t)$, where $\bar{Z}_k(\beta_k^0,t)=\frac{S_k^{(1)}(\beta_k,t)}{S_k^{(0)}(\beta_k^0,t)}$. Since,

$$\sum_{i=1}^n\{Z_i-\bar{Z}_k(\beta_k^0,t)\}Y_i(t)e^{\beta_k^{0\top}Z_i}\alpha_{k,0}(t)=0,$$

we can alternatively write

$$U^{(k)}(\beta_k^0,\tau)=\sum_{i=1}^n\int_0^\tau\{Z_i-\bar{Z}_k(\beta_k^0,t)\}dM_{k,i}(t)$$

and introduce its normalized version

$$U_n^{(k)}(\beta_k^0,\tau)=n^{-1/2}U^{(k)}(\beta_k^0,\tau). \tag{8.17}$$

Applying Lemma 7.8 choosing $F(\cdot)=U^{(k)}(\cdot,\tau)$ we get

$$U^{(k)}(\widehat{\beta}_k,\tau)-U^{(k)}(\beta_k^0,\tau)=\Big\{-\int_0^1\mathcal{I}_k(\beta_k^0+s(\widehat{\beta}_k-\beta_k^0))ds\Big\}(\widehat{\beta}_k-\beta_k^0). \tag{8.18}$$

where $\mathcal{I}_k(\beta_k)$ is as defined in (8.16) and $N_k(s)=\sum_i N_{k,i}(s)$. For later developments, we recall $v_k(\beta_k,s)=\Big[\frac{S_k^{(2)}(\beta_k,s)}{S_k^{(0)}(\beta_k,s)}-\Big\{\frac{S_k^{(1)}(\beta_k,s)}{S_k^{(0)}(\beta_k,s)}\Big\}^{\otimes 2}\Big]$. Since $\widehat{\beta}_k$ satisfies $U^{(k)}(\widehat{\beta}_k,\tau)=0$ and because of (8.17), (8.18) becomes

$$U_n^{(k)}(\beta_k^0,\tau)=\Big\{\int_0^1\frac{1}{n}\mathcal{I}_k(\beta_k^0+s(\widehat{\beta}_k-\beta_k^0))ds\Big\}\sqrt{n}(\widehat{\beta}_k-\beta_k^0). \tag{8.19}$$

With all these, we will show the theorem by two steps.

Step 1: We show the asymptotic normality of $U_n^{(k)}(\beta_k^0,\tau)$. Let us introduce $U_{k,l}^n(\beta_k^0,t)$ as the l^{th} component of $U_n^{(k)}(\beta_k^0,\tau)$, or $U_{k,l}^n(\beta_k^0,t)=n^{-1/2}\sum_{i=1}^n\int_0^\tau\{Z_{i,l}-\bar{Z}_{k,l}(\beta_k^0,t)\}dM_{k,i}(t)$. Defining

$$H_{k,i,l}^n=n^{-1/2}(Z_{i,l}-\bar{Z}_l(\beta_k^0,s)),$$

we can express

$$U^n_{k,l}(\beta^0_k, t) = \sum_{i=1}^{n} \int_o^t H^n_{k,i,l} dM_{k,i}(s)$$

as the terms similar to Lemma 7.5. Since we can verify the above term is locally bounded and predictable, by **Property 4** $U^n_{k,l}(\beta^0_k, t)$ is a local square integrable martingale, and,

$$\begin{aligned}\left\langle U^n_{k,l}\left(\beta^0_k, t\right), U^n_{k,l'}\left(\beta^0_k, t\right)\right\rangle &= \sum_{i=1}^{n} \int_0^t H^n_{k,i,l}(s) H^n_{k,i,l}(s) d\left\langle M_{k,i}\right\rangle(s) \\ &= \frac{1}{n} \sum_{i=1}^{n} \int_0^t \left(Z_{i,l} - \bar{Z}_l\left(\beta^0_k, s\right)\right)\left(Z_{i,l'} - \bar{Z}_{l'}\left(\beta^0_k, s\right)\right) Y_i(s) e^{\beta^{0\top}_k Z_i} \alpha_{k,0}(s) ds.\end{aligned}$$

Now directly from (C.1') and (C.2') we claim the right hand side (RHS) converges in probability to

$$\int_0^t v_k(\beta^0_k, s)_{ll'} s^{(0)}_k\left(\beta^0_k, s\right) \alpha_{k,0}(s) ds$$

for $t \in [0, \tau]$.

We next look at the Lindeberg condition. For any $\epsilon > 0$, define, for all l and t, that

$$U^n_{k,l,\epsilon}\left(\beta^0_k, t\right) = \sum_{i=1}^{n} \int_0^t H^n_{k,i,l}(s) I\left(n^{-1/2}\left|H_{k,i,l}(s)\right| \geq \epsilon\right) dM_{k,i}(s).$$

Again by **Property 4**, the above term is a square integrable martingale with,

$$\begin{aligned}\left\langle U^n_{k,l,\epsilon}\left(\beta^0_k, t\right)\right\rangle =& \sum_{i=1}^{n} \int_0^t \left\{H^n_{k,i,l}(s)\right\}^2 I\left(\left|H^n_{k,i,l}(s)\right| \geq \epsilon\right) d\left\langle M_{k,i}\right\rangle(s) \\ =& n^{-1} \sum_{i=1}^{n} \int_0^t \left\{H_{k,i,l}(s)\right\}^2 I\left(\left|H^n_{k,i,l}(s)\right| \geq \epsilon\right) Y_i(s) e^{\beta^{0\top}_k Z_i} \alpha_{k,0}(s) ds.\end{aligned}$$

By Lemma 7.4, the RHS of the above expression is bounded by

$$\begin{aligned}&\frac{4}{n} \sum_{i=1}^{n} \int_0^t Z^2_{i,l} I\left(n^{-1/2}\left|Z_{i,l}\right| \geq \epsilon/2\right) Y_i(s) e^{\beta^{0\top}_k Z_i} \alpha_{k,0}(s) ds \\ +&\frac{4}{n} \sum_{i=1}^{n} \int_0^t \bar{Z}^2_l\left(\beta^0_k, s\right) I\left(n^{-1/2}\left|\bar{Z}_l\left(\beta^0_k, s\right)\right| \geq \epsilon/2\right) Y_i(s) e^{\beta^{0\top}_k Z_i} \alpha_{k,0}(s) ds \\ =&\mathrm{I} + \mathrm{II}.\end{aligned}$$

Now II can be expressed as

$$4 \sum_{i=1}^{n} \int_0^t \bar{Z}^2_l\left(\beta^0_k, s\right) I\left(n^{-1/2}\left|\bar{Z}_l\left(\beta^0_k, s\right)\right| \geq \epsilon/2\right) S^{(0)}_k(\beta^0_k, s) ds.$$

By Conditions (C.1') and (C.2') for $\beta_k \in \mathcal{B}_k$, we can verify $I\left(n^{-1/2}\left|\bar{Z}_l\left(\beta^0_k, s\right)\right| \geq \epsilon/2\right) = 0$ with probability going to 1 uniformly in s, resulting $\mathrm{II} \xrightarrow{P} 0$. Now looking at (C.4'), we split I as I_1 and

I$_2$, where

$$\mathrm{I}_1 = \frac{4}{n}\sum_{i=1}^{n}\int_0^t Z_{i,l}^2 I\left(n^{-1/2}\left|Z_{i,l}\right| \geq \epsilon/2, \beta_k^{0\top} Z_i > -\delta\left|Z_i\right|\right) Y_i(s) e^{\beta_k^{0\top} Z_i}\alpha_{k,0}(s)ds,$$

$$\mathrm{I}_2 = \frac{4}{n}\sum_{i=1}^{n}\int_0^t Z_{i,l}^2 I\left(n^{-1/2}\left|Z_{i,l}\right| \geq \epsilon/2, \beta_k^{0\top} Z_i \leq -\delta\left|Z_i\right|\right) Y_i(s) e^{\beta_k^{0\top} Z_i}\alpha_{k,0}(s)ds.$$

A direct consequence of (C.4') implies that there is at least one $\delta > 0$ such that, for a fixed $c' > 0$ and for large n, there again exists a set A with $P(A) > 1-\epsilon$ and on which $I\left(n^{-1/2}\left|Z_{i,l}\right| \geq \epsilon/2, \beta_k^{0\top} Z_i \leq -\delta\left|Z_i\right|\right) Y_i(s) = 0$ uniformly for all $s \in [0,\tau]$, resulting in I$_1 \xrightarrow{P} 0$. Note that,

$$Z_{i,l}^2 I\left(n^{-1/2}\left|Z_{i,l}\right| \geq \epsilon/2, \beta_k^{0\top} Z_i \leq -\delta\left|Z_i\right|\right) Y_i(s) e^{\beta_k^{0\top} Z_i} \leq I\left(n^{-1/2}\left|Z_{i,l}\right| \geq \epsilon/2\right) Z_{i,l}^2 e^{-\delta|Z_i|}.$$

Now when $n^{-1/2}\left|Z_{i,l}\right| < \epsilon/2$ the results holds trivially since LHS becomes 0. When $n^{-1/2}\left|Z_{i,l}\right| \geq \epsilon/2$, also $Z_{i,l}^2 e^{-\delta|Z_i|} \leq Z_{i,l}^2 e^{-\delta\left|Z_{i,l}\right|}$. Because $x^2 e^{-\delta x} \to 0$ when $\delta > 0$ as $x \to \infty$, for any $\eta > 0$, there exists an n_0 such that when $n > n_0$, $Z_{i,l}^2 e^{-\delta\left|Z_{i,l}\right|} < \eta$. This implies I$_2$ is bounded by $4\eta\int_0^\tau \alpha_{k,0}(s)ds$ making it arbitrarily small resulting in I$_2 \xrightarrow{P} 0$. Hence by Lemma 7.5 we claim that

$$n^{-1/2}U_k(\beta_k^0,\tau) \xrightarrow{d} N(0, \Sigma_k(\beta_k^0,\tau)).$$

Here, the $(l,l')^{th}$ entry of $\Sigma_k(\beta_k^0,\tau)$ is $\int_0^t v_k(\beta_k^0,s)_{ll'} s_k^{(0)}\left(\beta_k^0,s\right)\alpha_{k,0}(s)ds$.

Step 2: By the definition of $\|\cdot\|$, it follows that

$$\left\|\int_0^1 \frac{1}{n}\mathcal{I}_k(\beta_k^0 + s(\widehat{\beta}_k - \beta_k^0))ds - \Sigma(\beta_k^0,\tau)\right\| \leq \int_0^1 \left\|\frac{1}{n}\mathcal{I}_k(\widehat{\beta}_k(s)) - \Sigma(\beta_k^0,\tau)\right\|ds \tag{8.20}$$

where $\widehat{\beta}_k(s) = \beta_k^0 + s(\widehat{\beta}_k - \beta_k^0)$. We next consider

$$\begin{aligned}
&\left\|\frac{1}{n}\mathcal{I}(\widehat{\beta}_k) - \Sigma\left(\beta_k^0,\tau\right)\right\| \\
\leq&\left\|\int_0^\tau (V_k(\widehat{\beta}_k,s) - v_k(\widehat{\beta}_k,s))\frac{1}{n}dN_k(s)\right\| \\
+&\left\|\int_0^\tau \left(v_k(\widehat{\beta}_k,s) - v_k\left(\beta_k^0,s\right)\right)\frac{1}{n}dN_k(s)\right\| \\
+&\left\|\int_0^\tau v_k\left(\beta_k^0,s\right)\frac{1}{n}\sum_{i=1}^n dM_{k,i}(s)\right\| \\
+&\left\|\int_0^\tau v_k\left(\beta_k^0,s\right)\left(S_k^{(0)}\left(\beta_k^0,s\right) - s_k^{(0)}\left(\beta_k^0,s\right)\right)\alpha_{k,0}(s)ds\right\| \\
=&\widetilde{\mathrm{I}} + \widetilde{\mathrm{II}} + \widetilde{\mathrm{III}} + \widetilde{\mathrm{IV}},
\end{aligned}$$

where $V_k(\beta_k,t) = \frac{S_k^{(2)}(\beta_k,t)}{S_k^{(0)}(\beta_k,t)} - \left(\frac{S_k^{(1)}(\beta_k,t)}{S_k^{(0)}(\beta_k,t)}\right)^{\otimes 2}$.

Now by the law of large number, we can see $\frac{1}{n}\sum_i N_{k,i}(\tau) \to \mathbb{E}(N_{k,i}(\tau)) < 1$. Also (C.1') and (C.2') imply $\sup_s \|V_k(\widehat{\beta}_k, s) - v_k(\widehat{\beta}_k, s)\| \xrightarrow{P} 0$. Note that we can also verify, by Lenglart inequality that for $c > 0$ arbitrary $\delta > 0$,

$$\mathbb{P}(n^{-1}N_k(\tau) > c) \leq \frac{\delta}{c} + \mathbb{P}(\int_0^\tau S_k^{(0)}(\beta_k^0, s)\alpha_{k,0}(s)ds > \delta)$$

and by (C.2') we can claim for large $c > 0$ there exists n_0 such that for $n \geq n_0$ we have $\mathbb{P}(n^{-1}N_k(\tau) > c) < \delta$. This all together along with the fact $\widetilde{\mathrm{I}} \leq \sup_s \left\|V_k(\widehat{\beta}_k, s) - v_k(\widehat{\beta}_k, s))\right\| \{\frac{1}{n}\sum_{i=1}^n N_{k,i}(\tau)\}$ implies that $\widetilde{\mathrm{I}} \xrightarrow{P} 0$.

Similarly the equicontinuity through (C.2') on $v_k(\beta_k, s)$ at β_k^0 and $\widehat{\beta}_k \xrightarrow{P} \beta_k^0$ along with the result established before on $n^{-1}N_k(\tau)$ yields $\widetilde{\mathrm{II}} \xrightarrow{P} 0$. Similarly direct consequence of (C.1') results in $\widetilde{\mathrm{IV}} \xrightarrow{P} 0$.

We are only left with the third term. We can prove the result by using Lenglart inequality and thereby controlling the term by the conditions. Otherwise we can control using simple Markov inequality. We consider $(i,j)^{th}$ element, $v_{k,i,j}\left(\beta_k^0, s\right)$, of $v_k\left(\beta_k^0, s\right)$, and consider the martingale process $\int_0^t v_{k,i,j}\left(\beta_k^0, s\right)\frac{1}{n}\sum_{i=1}^n dM_{k,i}(s)$ with variation

$$\int_0^t v_{k,i,j}^2\left(\beta_k^0, s\right)\frac{1}{n^2}d\left\langle M_{k,i}\right\rangle(s) = \frac{1}{n}\int_0^t v_k^2\left(\beta_k^0, s\right)S_k^{(0)}\left(\beta_k^0, s\right)\alpha_{k,0}(s)ds.$$

With Conditions (C.1') and (C.2') , boundedness of $S_k^{(0)}\left(\beta_k^0, s\right)$ is bounded for $s \in [0,\tau]$ and weak DCT, we get,

$$\begin{aligned} n\mathrm{Var}\left\{\int_0^\tau v_{k,i,j}\left(\beta_k^0, s\right)\frac{1}{n}\sum_{i=1}^n dM_{k,i}(s)\right\} =& \mathbb{E}\left\{\int_0^\tau v^2\left(\beta_k^0, s\right)S_k^0\left(\beta_k^0, s\right)\alpha_{k,0}(s)ds\right\} \\ \to& \int_0^\tau v_k^2\left(\beta_k^0, s\right)s_k^0\left(\beta_k^0, s\right)\alpha_{k,0}(s)ds \\ <&\infty. \end{aligned}$$

By simple application of Markov inequality we proved $\widetilde{\mathrm{III}} \xrightarrow{P} 0$. These all together shows,

$$\left\|\frac{1}{n}\mathcal{I}_k(\widehat{\beta}_k) - \Sigma_k\left(\beta_k^0, \tau\right)\right\| \xrightarrow{P} 0.$$

And thereafter examining the convergence result in each step, we can also conclude that

$$\left\|n^{-1}\mathcal{I}_k(\widehat{\beta}_k(s)) - \Sigma_k(\beta_k^0, \tau)\right\| \to 0$$

in probability uniformly with respect to $s \in [0,1]$. With (8.20), this implies

$$\left\|\int_0^1 \frac{1}{n}\mathcal{I}_k(\beta_k^0 + s(\widehat{\beta}_k - \beta_k^0))ds - \Sigma_k(\beta_k^0, \tau)\right\| \to 0$$

in probability. Combining **Step 1**, **Step 2**, and (8.19) and applying Slutsky's theorem, we claim that

$$\sqrt{n}(\widehat{\beta}_k - \beta_k^0) \xrightarrow{d} N(0, \Sigma_k^{-1}(\beta_k^0, \tau)).$$

□

The proposition establishes that inference for β_k^0 can be conducted using an approximate multivariate normal distribution. Moreover, based on the proof in "Step 2," the variance of $\widehat{\beta}_k$ is approximately $n^{-1}\Sigma^{-1}(\beta_k^0, \tau)$, which can be consistently estimated by $\mathcal{I}_k^{-1}(\widehat{\beta}_k)$, the inverse of the observed information matrix evaluated at $\widehat{\beta}_k$. Confidence intervals for the components of β_k are typically constructed using the normal approximation. For the jth component, the confidence interval is given by

$$\widehat{\beta}_{kj} \pm z_{1-\alpha/2} \cdot \mathrm{SE}(\widehat{\beta}_{kj}),$$

where $\mathrm{SE}(\widehat{\beta}_{kj})$ denotes the estimated standard error of $\widehat{\beta}_{kj}$. Additionally, hypothesis testing for the regression coefficients, such as using the Wald test, score test, or likelihood ratio test, is based on the same asymptotic theory. These tests assess whether specific covariate effects are statistically significant within the cause-specific hazard framework.

We next consider the joint distribution of $(\widehat{\beta}_1, \ldots, \widehat{\beta}_K)$ to understand their collective behavior across failure types. The results may inform potential dependence, and support valid simultaneous inference and multivariate testing. We have the following results.

Proposition 8.7. *Denote by* $\widehat{\beta} = (\widehat{\beta}_1^\top, \ldots, \widehat{\beta}_K^\top)^\top$ *and* $\beta^0 = ((\beta_1^0)^\top, \ldots, (\beta_K^0)^\top)^\top$. *Under (C.1')–(C.4'),*

$$\sqrt{n}(\widehat{\beta} - \beta^0) \xrightarrow{d} N\big(0, \mathit{diag}\big\{\Sigma_1^{-1}(\beta_1^0, \tau), \Sigma_2^{-1}(\beta_2^0, \tau) \ldots \Sigma_K^{-1}(\beta_K^0, \tau)\big\}\big).$$

Proof. We stack the the normalized score vectors, established in Proposition 8.6, for each cause as

$$U_n\left(\beta^0, t\right) = \left(U_n^{(1)}\left(\beta_1^0, t\right)^\top, U_n^{(2)}\left(\beta_2^0, t\right)^\top \ldots, U_n^{(K)}\left(\beta_K^0, t\right)^\top\right)^\top,$$

where, for each cause $l = 1, \ldots, K$,

$$U_n^{(l)}\left(\beta^0, t\right) = \left(U_{1,l}^n\left(\beta_1^0, t\right), U_{2,l}^n\left(\beta_2^0, t\right) \ldots, U_{K,l}^n\left(\beta_K^0, t\right)\right)^\top.$$

We apply Lemma 7.5. Clearly,

$$\left\langle U_{k,l}^n\left(\beta_k^0, t\right), U_{k',l'}^n\left(\beta_{k'}^0, t\right)\right\rangle = \sum_{i=1}^n \int_0^t H_{k,i,l}^n(s) H_{k',i,l'}^n(s) d\left\langle M_{k,i}, M_{k',i}\right\rangle(s).$$

Therefore,

$$\left\langle U_{k,l}^n\left(\beta_k^0, t\right), U_{k',l'}^n\left(\beta_{k'}^0, t\right)\right\rangle = \begin{cases} \frac{1}{n}\sum_{i=1}^n \int_0^t \left(Z_{i,l} - \bar{Z}_l(\beta_k^0, s)\right)\left(Z_{i,l'} - \bar{Z}_{l'}(\beta_k^0, s)\right) Y_i(s) e^{\beta_k^{0\top} Z_i} \alpha_{k,0}(s) ds & k = k' \\ 0 & k \neq k' \end{cases}$$

because $\langle M_{k,i}\rangle(t) = \int_0^t Y_i(s) e^{\beta_k^{0\top} Z_i} \alpha_{k,0}(s) ds$ as well as due to the orthogonality under the competing risk setting, i.e., $\left\langle M_{k,i}, M_{k',i}\right\rangle(s) = 0$ for $k \neq k'$; see (8.6).

Let $B_k = \{(k-1)p+1, \ldots, kp\}$, for $k = 1, 2 \ldots, K$, be the k^{th} block. Then,

$$\left\langle U_n^{(l)}\left(\beta^0, t\right), U_n^{(l')}\left(\beta^0, t\right)\right\rangle = \begin{cases} \left\langle U_{1,l}^n\left(\beta_1^0, t\right), U_{1,l'}^n\left(\beta_1^0, t\right)\right\rangle & (l, l') \in B_1 \\ \left\langle U_{2,l}^n\left(\beta_2^0, t\right), U_{2,l'}^n\left(\beta_2^0, t\right)\right\rangle & (l, l') \in B_2 \\ \vdots & \\ \left\langle U_{K,l}^n\left(\beta_K^0, t\right), U_{K,l'}^n\left(\beta_K^0, t\right)\right\rangle & (l, l') \in B_K \\ 0 & \text{otherwise} \end{cases}$$

which converges in probability to $\int_0^t v_k(\beta_k^0, s)_{ll'} s_k^{(0)}\left(\beta_k^0, s\right)\alpha_{k,0}(s)ds$ for $(l, l') \in B_k$ for each k and $t \in [0, \tau]$, which is nothing but $(l, l')^{th}$ element of $\Sigma_k(\beta_k^0, \tau)$ for each k. We can verify the Lindeberg condition as done in the proof of Proposition 8.6 and conclude

$$\sqrt{n}(\widehat{\beta} - \beta^0) \xrightarrow{d} N\big(0, \operatorname{diag}\big\{\Sigma_1^{-1}(\beta_1^0, \tau), \Sigma_2^{-1}(\beta_2^0, \tau) \ldots \Sigma_K^{-1}(\beta_K^0, \tau)\big\}\big).$$

□

The results show that the estimators for each failure type are asymptotically independent. This property justifies analyzing cause-specific Cox models separately for each event type. In medical studies involving competing risks, such as cardiovascular vs. non-cardiovascular death, various forms of cancer recurrence, or progression to organ failure vs. mortality, this separation is especially useful, because covariate effects can be interpreted without needing to adjust for correlations between outcomes, and inference procedures remain valid without incorporating cross-covariance. For instance, one might find that a treatment lowers the risk of cardiovascular death while raising the risk of non-cardiovascular death, or that a biomarker predicts distant but not local recurrence, each conclusion supported by its own confidence interval and p-value. More broadly, this independence is of practical value in fields like oncology, organ transplantation, and chronic disease management, where understanding risk factors for distinct failure modes is crucial for clinical decision-making.

References

Aalen, O. (1978), ‘Nonparametric inference for a family of counting processes’, *The Annals of Statistics* **6**(4), 701–726.
URL: *https://projecteuclid.org/journals/annals-of-statistics/volume-6/issue-4/Nonparametric-Inference-for-a-Family-of-Counting-Processes/10.1214/aos/1176344247.full*

Anderson, P. K. & Gill, R. D. (1982), ‘Cox’s regression model for counting processes: A large sample study’, *The Annals of Statistics* **10**(4), 1100–1120.

Berg, H. C. (2023), ‘Random walks in biology: Brownian motion and cellular processes’, *Biophysical Journal* **124**(3), 567–580.

Billingsley, P. (2013), *Convergence of probability measures*, John Wiley & Sons.

Black, F. & Scholes, M. (1973), 'The pricing of options and corporate liabilities', *Journal of Political Economy* **81**(3), 637–654. Pioneering work in financial modeling using Brownian motion.
URL: *https://www.jstor.org/stable/1831029*

Brown, B. M. (1971), 'Martingale central limit theorems', *The Annals of Mathematical Statistics* **42**(1), 59–66.
URL: *https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-42/issue-1/Martingale-Central-Limit-Theorems/10.1214/aoms/1177693494.full*

Brown, R. (1828), 'A brief account of microscopical observations made in the months of june, july, and august 1827 on the particles contained in the pollen of plants, and on the general existence of active molecules in organic and inorganic bodies', *Philosophical Magazine* **4**, 161–173. First observation of the phenomenon now known as Brownian motion.
URL: *https://www.jstor.org/stable/615607*

Carter, M., Van Brunt, B., Carter, M. & Van Brunt, B. (2000), *The lebesgue-stieltjes integral*, Springer.

Chung, K. L. (1974), *A Course in Probability Theory*, 2nd edn, Academic Press.

Cox, D. R. (1972), 'Regression models and life-tables', *Journal of the Royal Statistical Society: Series B (Methodological)* **34**(2), 187–220.
URL: *https://www.jstor.org/stable/2985181*

Fleming, T. R. & Harrington, D. P. (2013), *Counting processes and survival analysis*, Vol. 625, John Wiley & Sons.

Franklin, G. F., Powell, J. D. & Emami-Naeini, A. (2023), 'Applications of brownian motion in signal processing, noise modeling, and control systems', *IEEE Transactions on Automatic Control* **68**(4), 1123–1137.

Harrington, D. P. & Fleming, T. R. (1982), 'A class of rank test procedures for censored survival data', *Biometrika* **69**(3), 553–566.

Mantel, N. (1966), 'Evaluation of survival data and two new rank order statistics arising in its consideration', *Cancer Chemotherapy Reports* **50**(3), 163–170.

Meyer, P.-A. (1963), 'Decomposition of supermartingales: the uniqueness theorem', *Illinois Journal of Mathematics* **7**(1), 1–17.

Mörters, P. & Peres, Y. (2010), *Brownian Motion*, Vol. 30 of *Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge University Press.
URL: *https://www.cambridge.org/core/books/brownian-motion/F639B9A8403BD465F896F3E18A9C3382*

Peto, R. & Peto, J. (1972), 'Asymptotically efficient rank invariant test procedures', *Journal of the Royal Statistical Society. Series A (General)* **135**(2), 185–207.

Shorack, G. R. & Wellner, J. A. (1986), *Empirical Processes with Applications to Statistics*, Wiley Series in Probability and Statistics, Wiley.

Tsiatis, A. A. (2006), *Semiparametric Theory and Missing Data*, Springer, New York.

van der Vaart, A. W. & Wellner, J. A. (1996), *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer Series in Statistics, Springer.
URL: *https://link.springer.com/book/10.1007/978-1-4757-2545-2*

Wand, M. P. & Jones, M. C. (1995), *Kernel Smoothing*, Vol. 60 of *Monographs on Statistics and Applied Probability*, Chapman and Hall/CRC.
URL: *https://link.springer.com/book/10.1007/978-1-4899-4493-1*

Øksendal, B. (2003), *Stochastic Differential Equations: An Introduction with Applications*, 6th edn, Springer, Berlin, Heidelberg.