NMST531 Censored Data Analysis

Extended Course Notes

Michal Kulich

Last modified on January 5, 2021.

Department of Probability and Mathematical Statistics
Faculty of Mathematics and Physics, Charles University

These course notes contain the whole content of the lectures for the course "NMST531 Censored Data Analysis", which is a part of the curriculum of the Master's program "Probability, Mathematical Statistics and Econometrics".

This material is under development. The author will appreciate notifications by the reader of potential typos or misprints.

Michal Kulich
kulich@karlin.mff.cuni.cz

In Karlín on January 5, 2021

Contents

1. Introduction

1.1. Time-to-event data

Consider a non-negative random variable $T \geq 0$, which can be interpreted as the moment when a certain event occurs, or a waiting time for the event. There are numerous examples of such random variables in various applications in different fields:

- The event is the **death** of a person; T is the time of death or the **survival time**.
- The event is an **occurrence of a disease** in a previously healthy person; T is the **time of diagnosis**.
- The event is the **failure** of a machine or device; T is **time to failure**.
- The event is the **repair** of a broken machine; T is **duration of maintenance**.
- The event is the **default of a debtor**; T is **time to default**.
- The event is an **occurrence of an insurance claim**; T is the **time when the claim is made**.
- The event terminates the **payment of a pension** (e.g. the death of the pensioner); T is the **total amount paid on the pension**. (In this example, *time is money*.)

In all examples, there is some time scale involved, and time 0 must be appropriately defined.

The theory of **time-to-event analysis** bears various names depending on the field of application: it is called *survival analysis*[*] in biomedical applications, *life tables*[†] in demographics and life insurance, *reliability theory*[‡] in technical applications, *credit risk*[§] and *insurance risk*[¶] in financial applications and non-life insurance. We usually use the term *"censored data analysis"*[‖] because it is application-neutral. The term "censoring" is introduced in the next section.

[*] Česky *analýza přežití* [†] Česky *úmrtnostní tabulky* [‡] Česky *teorie spolehlivosti* [§] Česky *kreditní riziko* [¶] Česky *pojistné riziko* [‖] Česky *analýza censorovaných dat*

1.2. Censoring

Standard statistical methods for estimation and testing could be used for the analysis of time-to-event data if event times could be observed on all participating subjects. However, this is usually not the case. Events such as default on a debt or diagnosis of a specific disease may not occur at all for some subjects. Even if the event does eventually occur, it may take so long to develop that it would not be observed. Thus, time-to-event data are usually incompletely observed and require specialized analysis methods.

In practice, event times are recorded through an *observation process*. Time 0 is the time when the observation starts. We wait for the event to occur starting at time 0. If the observation ends before the event occurs, the event is never observed. This feature can be formally handled by the probabilistic model by considering two latent random variables for each subject: $T \geq 0$ is the time to event* (*failure time, survival time*) and $C \geq 0$ is the *censoring time*†, which expresses the duration of observation of the subject. If the duration of observation is shorter than the failure time, the failure time is not available. Thus, the pair (T, C) generates the following two cases:

- If $T \leq C$, T is observed and C is not observed.
- If $C < T$, C is observed and T is not observed.

The possibility that the event may not occur at all can be expressed by setting T to a very large value or even taking $T = \infty$.

Notation. Let A be a random event, define $\mathbb{1}(A) = 1$ if A occurs and $\mathbb{1}(A) = 0$ if A does not occur. The random variable $\mathbb{1}(A)$ is called *the indicator of* A.

Notation. Let s, t be real numbers. Denote $s \wedge t = \min(s, t)$.

Let $X = T \wedge C$ be called *the censored failure time*‡ and let $\delta = \mathbb{1}(T \leq C)$ be called *the failure indicator*§. These two random variables capture what can be actually observed in censored data problems. The analysis methods must be modified to work with the partial information contained in (X, δ).

Consider latent failure and censoring times $(T_1, C_1), \dots, (T_n, C_n)$ generated for n independent subjects. If we could observe $T_1, \dots, T_n$, we could perform standard statistical procedures on this random sample. However, we only observe censored failure times and failure indicators $(X_1, \delta_1), \dots, (X_n, \delta_n)$. Such censored data requires specialized methods of statistical analysis.

In general, the censoring variables $C_1, \dots, C_n$ are considered random variables with certain distributions (in principle, each C_i may have a different distribution function). This is called the *random censorship model*. There are two special case of this

* Česky *doba do události* † Česky *čas censorování* ‡ Česky *censorovaná doba do události* § Česky *indikátor události*

general model that acquired their own names:

Type I censoring All censoring variables are equal to a pre-specified constant τ that expresses the common maximal duration of observation, i.e., $C_i = \tau$ for all i almost surely[*].

Type II censoring All remaining observations are censored when the k-th failure occurs (where $k \in \{1, \dots, n\}$ is pre-specified), i.e., $C_i = T_{(k)}$ for all i, $T_{(k)}$ is the k-th order statistic of the random sample $T_1, \dots, T_n$[†].

These two censoring schemes are used primarily in technical applications. They are unrealistic for most biomedical and financial applications.

We will only deal with Type I and Type II censoring in the first part of the course. The rest of the course will consider the general random censorship model.

Think:

A standard task in classical statistics is the estimation of the expectation from the random sample. If we observe i.i.d. random variables $T_1, \dots, T_n$, we can estimate the expectation $\mathsf{E}\, T_i$ by the arithmetic mean $\overline{T}_n = \frac{1}{n}\sum_{i=1}^{n} T_i$. By the weak law of large numbers, this estimator is consistent as long as the expectation exists and is finite.

With censored data, we could only use the observations (X_i, δ_i) for the estimation. Is there any obvious and easy way to modify the arithmetic mean to this data so that we end up with a consistent estimator of the expectation?

1.3. Survival function, hazard function

The distribution of a non-negative random variable T is usually described by its distribution function $F(t) = \mathsf{P}[T \leq t]$ or a density $f(t)$ with respect to a σ-finite measure μ, a function such that $F(t) = \int_0^t f(s)\, d\mu(s)$.

When working with censored failure time data, it is of advantage to work with survival functions instead of distribution functions.

Definition 1.1. The function $S(t) = 1 - F(t) = \mathsf{P}[T > t]$ is called the *survival function* of a random variable T with distribution function $F(t)$. ∇

Notation. Let f be a right-continuous function. Define $f(t-) = \lim_{h \searrow 0} f(t-h)$ (if the limit exists). This is a left-continuous function.

Note. Let $S(t)$ be the survival function of a non-negative random variable T with distribution function $F(t)$. Then it is known that:

[*] Česky *censorování typu I., censorování časem* [†] Česky *censorování typu II., censorování poruchou*

- $S(t)$ is non-increasing right-continuous function.
- $S(0) = 1 - \mathrm{P}[T = 0]$, $\lim_{t\to\infty} S(t) = 0$.
- If T is continuous with density $f(t)$ w.r.t. the Lebesgue measure then $f(t) = -S'(t)$ and $S(t) = \int_t^\infty f(s)\,ds$.
- If T is discrete with values $t_1, t_2, \ldots$ and $p_i = \mathrm{P}[T = t_i]$ then $p_i = S(t-) - S(t)$ and $S(t) = \sum_{\{i: t_i > t\}} p_i$.

Frequently, we assume that $\mathrm{P}[T = 0] = 0$ (the failure cannot occur at the time 0). Then $S(0) = 1$.

Relationship between survival function and expectation

The expectation of a non-negative random variable is defined as $\mathrm{E}\, T = \int_0^\infty t\, dF(t)$. The following lemma shows that the expectation can be obtained by integrating the survival function.

Lemma 1.1. *Let $T \geq 0$ a.s. and $\mathrm{E}\, T < \infty$. Then*

$$\mathrm{E}\, T = \int_0^\infty S(t)\, dt.$$

$\diamond$

The end of lecture 1 (Oct. 5)

Proof. The proof will use integration by parts for Lebesgue-Stieltjes integral, see Theorem A.1 in the Appendix. We take $G(s) = s$, $F(s) = \mathrm{P}[T \leq s]$. The Theorem gives us the equality

$$F(s)s - 0 = \int_0^s F(t-)dt + \int_0^s t\, dF(t).$$

Hence

$$\int_0^s t\, dF(t) = F(s)s - \int_0^s F(t-)dt.$$

We subtract s from the right-hand side and add it back realizing that s can be written as $\int_0^s 1\, dt$. We get

$$\begin{aligned}\int_0^s t\, dF(t) &= [F(s) - 1]s + \int_0^s [1 - F(t-)]dt \\ &= [F(s) - 1]s + \int_0^s S(t)\, dt.\end{aligned}$$

Now take the limit as $s \to \infty$ on both sides. The left-hand side converges to $\mathrm{E}\, T$ and the right hand side to $\int_0^\infty S(t)\, dt$, as long as $\lim_{s\to\infty} s[1 - F(s)] = 0$.

It remains to be shown that the limit of the superfluous term is really zero. We assume that $\mathsf{E}\,T$ is finite. Thus, for any $0 \le s < \infty$,

$$\begin{aligned}\mathsf{E}\,T = \int_0^\infty t\,dF(t) &= \int_0^s t\,dF(t) + \int_s^\infty t\,dF(t)\\ &\ge \int_0^s t\,dF(t) + s\int_s^\infty dF(t) = \int_0^s t\,dF(t) + s[1-F(s)].\end{aligned}$$

Hence

$$0 \le s[1-F(s)] \le \mathsf{E}\,T - \int_0^s t\,dF(t) \to 0 \quad \text{as } s \to \infty,$$

because $\mathsf{E}\,T = \lim_{s\to\infty} \int_0^s t\,dF(t)$. The proof is finished. □

This result can be easily generalized to random variables that can attain negative values and to higher moments.

Corollary. Let X be a random variable such that $\mathsf{E}\,|X|^\alpha < \infty$. Then

$$\mathsf{E}\,|X|^\alpha = \alpha \int_0^\infty t^{\alpha-1}\mathrm{P}\,[|X| > t]\;dt.$$

Homework 1:

Prove the corollary.

Lemma 1.1 is useful for **estimation of the expectation from censored data**. It is unclear how to generalize the arithmetic mean to censored data. However, if there is an estimator $\widehat{S}(t)$ of the survival function, which is consistent over the whole interval $\langle 0, \infty)$, the expectation could be consistently estimated by $\int_0^\infty \widehat{S}(t)\,dt$.

Hazard function

It is known that the distribution of a random variable can be described by density, distribution function, survival function, quantile function or characteristic function. However, there is another way to describe the distribution, called *the hazard function**; it is especially useful for time-to-event data.

Definition 1.2. Let T be a continuous non-negative random variable. Then the *hazard function* $\lambda(t)$ of T is defined as

$$\lambda(t) = \lim_{h\searrow 0} \frac{1}{h}\,\mathrm{P}\big[t \le T < t+h \,\big|\, T \ge t\big].$$

* Česky *risiková funkce*

Let T be discrete with values $0 \le t_1 < t_2 < \cdots$. Then the *hazard function* $\lambda(t)$ of T is defined at $t_1, t_2, \ldots$ by

$$\lambda(t_i) \equiv \lambda_i = \mathrm{P}\big[T = t_i \,\big|\, T \ge t_i\big]. \qquad \nabla$$

Loosely speaking, the hazard function measures the probability of having the event at the time t (or shortly thereafter) given that the event had not occurred earlier. Thus, it expresses the risk of having the event at t. The hazard function may have different names in different application areas: in reliability theory, where the event of interest is a failure of a machine, it is called *failure rate*[*] (or failure intensity); in epidemiology, where the event of interest is occurrence of disease, it is called *incidence rate*[†] (or incidence function); in demography or insurance, where the event of interest is death, it is called *mortality rate*[‡].

Note. Realize that the density $f(t)$ of a continuous random variable can be written as

$$f(t) = \lim_{h \searrow 0} \frac{1}{h} \mathrm{P}\left[t \le T < t + h\right].$$

So, the hazard function differs by adding the condition that $T \ge t$. The same is true for discrete random variables.

Notation. The function $\Lambda(t)$ defined as

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

for continuous t, and

$$\Lambda(t) = \sum_{\{i : t_i \le t\}} \lambda(t_i)$$

for discrete T, is called the *cumulative hazard function*[§].

The next theorem shows that the hazard function indeed characterizes the whole distribution. It also reveals the relationship between the hazard function and the density/the survival function.

Theorem 1.2. *Let T be a non-negative random variable with hazard function λ, density f, distribution function F and survival function $S = 1 - F$. Then*

(i)

$$\lambda(t) = \frac{f(t)}{S(t-)}.$$

(ii)

$$\Lambda(t) = \int_0^t \frac{dF(s)}{S(s-)}. \tag{1.1}$$

[*] Česky *intensita poruch* [†] Česky *incidence choroby* [‡] Česky *úmrtnost* [§] Česky *kumulativní riziko*

(iii) For continuous T*,*

$$S(t) = \mathrm{e}^{-\Lambda(t)}. \tag{1.2}$$

For discrete T *with values* $0 \le t_1 < t_2 < \cdots$*,*

$$S(t) = \prod_{\{i: t_i \le t\}} (1 - \lambda_i). \tag{1.3}$$

◊

Part (i) of the theorem shows how to calculate the hazard function from the density and the survival function. Part (iii) reverts the relationship and proves that the survival function can be calculated from the hazard function. Hence, the hazard function fully specifies the distribution of T.

Corollary. For continuous T,

$$f(t) = \lambda(t)\mathrm{e}^{-\Lambda(t)},$$

for discrete T with values $0 \le t_1 < t_2 < \cdots$,

$$\mathrm{P}[T = t_i] = \lambda_i \prod_{\{j: t_j < t_i\}} (1 - \lambda_j).$$

Notice that $\mathrm{P}[T = t_i] = (1-\lambda_1)(1-\lambda_2)\cdots(1-\lambda_{i-1})\lambda_i$. The probability of death at t_i is the product of conditional probabilities of surviving all the previous death opportunities times the conditional probability of dying at t_i (given that the subject survived till then).

Proof (of Theorem 1.2).

(i) Suppose that T is continuous. Then

$$\lambda(t) = \lim_{h \searrow 0} \frac{1}{h} \mathrm{P}\left[t \le T < t+h \,\middle|\, T \ge t\right] = \frac{\lim_{h\searrow 0} h^{-1}\mathrm{P}[t \le T < t+h]}{\mathrm{P}[T \ge t]} = \frac{f(t)}{S(t-)}.$$

Now take a discrete T with values $0 \le t_1 < t_2 < \cdots$. Obviously,

$$\lambda(t_i) = \mathrm{P}\left[T = t_i \,\middle|\, T \ge t_i\right] = \frac{\mathrm{P}[T = t_i]}{S(t_{i-1})} = \frac{f(t_i)}{S(t_i-)}. \tag{*}$$

(ii) Denote μ the measure with respect to which f is a density. Then

$$\Lambda(t) = \int_0^t \frac{f(s)}{S(s-)}\, d\mu(s) = \int_0^t \frac{dF(s)}{1 - F(s-)}.$$

(iii) Consider a continuous T first. We have

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

Hence

$$\Lambda(t) = -\log S(t) + C.$$

Because $\Lambda(0) = 0$, C must be 0. Therefore

$$S(t) = \mathrm{e}^{-\Lambda(t)}.$$

Now take a discrete T. Denote $p_i \equiv \mathrm{P}[T = t_i]$ for ordered values $0 \le t_1 < t_2 < \cdots$, and, according to (*),

$$\lambda_i \equiv \lambda(t_i) = \frac{p_i}{\sum_{j\ge i} p_j} \qquad i = 1, 2, , \ldots \tag{**}$$

Note that $\lambda_1 = p_1$. From this,

$$S(t_i) = \mathrm{P}[T > t_i] = \sum_{j>i} p_j \overset{(**)}{=} \frac{p_{i+1}}{\lambda_{i+1}}. \tag{†}$$

Also from (**), we get:

$$\begin{aligned}
\lambda_{i+1}\Bigl(\sum_{j\ge i} p_j - p_i\Bigr) &= p_{i+1},\\
\lambda_{i+1}\Bigl(\frac{\sum_{j\ge i} p_j}{p_i} - 1\Bigr) &= \frac{p_{i+1}}{p_i},\\
\lambda_{i+1}\Bigl(\frac{1}{\lambda_i} - 1\Bigr) &= \frac{p_{i+1}}{p_i},\\
\frac{p_i}{\lambda_i}(1-\lambda_i) &= \frac{p_{i+1}}{\lambda_{i+1}}.
\end{aligned}$$

From this, from (†), and from the fact that $p_1/\lambda_1 = 1$, we get the desired result

$$S(t_i) = \frac{p_{i+1}}{\lambda_{i+1}} = \prod_{j\le i}(1-\lambda_j),$$

which is equivalent to (1.3). This completes the proof. □

There is yet another way to characterize a failure time distribution, which is useful especially in engineering, demographics and life insurance: mean residual lifetime.

Definition 1.3. Let $T \ge 0$ a.s. The function $r(t) = \mathsf{E}\bigl[T - t \,\big|\, T \ge t\bigr]$ is called *the mean residual lifetime*[*]. ∇

Clearly, $r(0) = \mathsf{E}\,T$.

[*] Česky *střední zbytková doba života*

Theorem 1.3. *Let T be a non-negative random variable with survival function S and mean residual lifetime r. Then*

(i) *The conditional survival function of T given $T \geq t$ is $\mathrm{P}\left[T > s \,\middle|\, T \geq t\right] = S(s)/S(t-)$ for $s \geq t$.*

(ii) *The mean residual lifetime of T can be expressed as*

$$r(t) = \frac{\int_t^\infty S(s)\,ds}{S(t-)}.$$

(iii) *For continuous T and any $t \geq 0$,*

$$S(t) = \frac{\mathsf{E}\,T}{r(t)} \exp\left\{-\int_0^t \frac{ds}{r(s)}\right\}. \qquad \diamond$$

The last point of this theorem proves that the mean residual lifetime as a function defined on $\langle 0, \infty)$ completely specifies the failure time distributioon.

Independent study task:

Go through Appendix A.1. Notice the following facts:

- *Properties of exponential distribution*
- *Relationship of exponential distribution to Weibull distribution*
- *Relationship of exponential and Weibull distribution to Gumbel distribution*
- *Appreciate the memoryless property of exponential and geometric distributions and the relationship between those two distributions*

The end of lecture 2 (Oct. 5)

1.4. Independent censoring

Suppose the failure time T is censored – instead of T, we observe the pair (X, δ), where $X = \min(T, C) = T \wedge C$, $\delta = \mathbb{1}(T \leq C)$ and C is the censoring variable. We are interested in the distribution of T, which can be described, e.g., by the survival function $S(t) = \mathrm{P}[T > t]$ or by the hazard function $\lambda(t) = f(t)/S(t-)$. It is clear that the task requires imposing certain conditions on the censoring variable C.

Suppose now that the random variables T and C are *independent*. We can only observe $X = T \wedge C$. Consider the survival function of X:

$$S_X(t) = \mathrm{P}[X > t] = \mathrm{P}[T > t, C > t] = S(t)\mathrm{P}[C > t] \leq S(t).$$

Clearly, it is hard to relate the survival function of X to the survival function of T, unless we know the distribution of C.

Next, consider the survival function of X when X is uncensored:

$$S_X^*(t) = \mathrm{P}\big[T > t \,\big|\, T \le C\big] = \frac{\mathrm{P}\,[t < T \le C]}{\mathrm{P}\,[T \le C]} \ne S(t).$$

This is even less useful then $S_X(t)$.

Now suppose that T has a continuous distribution and consider its hazard function $\lambda(t)$

$$\begin{aligned}\lambda(t) &= \lim_{h \searrow 0} \frac{1}{h} \mathrm{P}\big[t \le T < t + h \,\big|\, T \ge t\big] \overset{(*)}{=} \lim_{h \searrow 0} \frac{1}{h} \mathrm{P}\big[t \le T < t + h \,\big|\, T \ge t, C \ge t\big] \\ &= \lim_{h \searrow 0} \frac{1}{h} \mathrm{P}\big[t \le T < t + h \,\big|\, X \ge t\big],\end{aligned} \tag{1.4}$$

where the equation $(*)$ holds because of independence. Thus, the hazard function of T can be recovered from censored data under certain conditions if we look at the occurrence of death among subjects who are alive and still uncensored at the particular time of interest. This is the reason why the hazard function is so convenient tool for the analysis of censored data.

Stochastic independence between T and C is a sufficient but not a necessary condition for equation (1.4). Therefore we take that equation and make it a definition of *independent censoring*.

Definition 1.4. Let T be continuous and let

$$\lambda(t) = \lim_{h \searrow 0} \frac{1}{h} \mathrm{P}\big[t \le T < t + h \,\big|\, T \ge t\big]$$

be its true hazard function (called *the net hazard* in this context). Let

$$\lambda^{\#}(t) = \lim_{h \searrow 0} \frac{1}{h} \mathrm{P}\big[t \le T < t + h \,\big|\, X \ge t\big]$$

be the hazard function of T in the presence of censoring (called *the crude hazard*).

The censoring variable C satisfies *the independent censoring condition*[*] if and only if $\lambda(t) = \lambda^{\#}(t)$ a.e., that is, when the net and crude hazards are equal. ∇

A generalization of the independent censoring condition to arbitrary failure time distributions (which need not be continuous) will be considered in Chapter 3.

We will always assume that independent censoring condition holds. Below is a rather trivial example where T and C are clearly not independent but the independent censoring condition is still satisfied.

Example (Type II censoring). Consider independent latent failure times $T_1, \dots, T_n$ and define $C_i = T_{(k)}$ for all i, $1 \le k \le n$. Then C_i is not independent of T_i but the independent censoring condition (1.4) holds for each i. △

[*] Česky *nezávislé censorování*

Homework 2:

Assume that the failure time has a continuous distribution. Prove that Type II censoring satisfies the independent censoring condition (1.4).

2. Parametric Models

In this chapter, we briefly discuss basic parametric models for censored data. The most important result is the formation of the likelihood function for parametric models when T and C are independent in Section 2.1. Next, we develop some exact inference methods for the exponential model with Type II censoring. In the final section, we present an easy method for analyzing regression models with censored exponentially distributed response.

2.1. Parametric likelihood for arbitrary random censoring

Let $(T_1, C_1), \ldots, (T_n, C_n)$ be independent, let $X_i = \min(T_i, C_i) = T_i \wedge C_i$ be the censored failure times and $\delta_i = \mathbb{1}(T_i \le C_i)$ the failure indicators. The data consist of independent pairs $(X_1, \delta_1), \ldots, (X_n, \delta_n)$.

Let $T_1, \ldots, T_n$ be identically distributed with survival function $S(x; \boldsymbol{\theta})$, density $f(x; \boldsymbol{\theta})$, and hazard function $\lambda(x; \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta$ is a p-dimensional parameter vector. Suppose the family of densities $f(x; \boldsymbol{\theta})$ satisfies the regularity assumptions of the maximum likelihood theory.

Denote by $G_i(x)$ the survival function of the censoring variable C_i and by $g_i(x)$ its density. We are not assuming that the censoring times are equally distributed; arbitrary distributions are allowed for each of them. However, we will assume throughout this section *that T_i and C_i are stochastically independent.*

The likelihood function is based on the product (over i) of joint densities of the observations (X_i, δ_i).

Lemma 2.1. *Let T_i and C_i be stochastically independent. Then the joint density of (X_i, δ_i) is*

$$q_i(x, \delta) = \left[f(x; \boldsymbol{\theta}) G_i(x-)\right]^{\delta} \left[g_i(x) S(x; \boldsymbol{\theta})\right]^{1-\delta}. \qquad \diamond$$

We include two proofs of Lemma 2.1. The first version assumes that the distribution of X_i is continuous and it is relatively easy and straightforward. However, continuity of X_i requires all C_i's to have continuous distributions. This is not true in many real applications because censoring times often follow mixtures of discrete and continuous distributions. Therefore we also present a proof of the most general case, when both T_i and C_i have discrete and continuous components.

It is enough to understand the simpler proof with continuous X_i (the first version).

Proof (Version 1 – continuous case). Assume that T_i and C_i have continuous distributions (in the proof, we ignore the parameter θ). Hence, X_i also has a continuous distribution. The joint density of (X_i, δ_i) at the point (x, δ_0) can be calculated as

$$-\frac{\partial}{\partial x}\mathrm{P}\left[X > x, \delta = \delta_0\right].$$

First, let $\delta_0 = 1$. Then

$$\mathrm{P}\left[X > x, \delta = 1\right] = \mathrm{P}\left[x < T < C\right] = \int_x^\infty \underbrace{\left[\int_t^\infty h(t,s)\,ds\right]}_{\psi(t)} dt,$$

where $h(t,s)$ is the joint density of (T, C). Due to independence, $h(t,s) = f(t)g_i(s)$. Next,

$$-\frac{\partial}{\partial x}\mathrm{P}\left[X > x, \delta = 1\right] = -\frac{\partial}{\partial x}\int_x^\infty \psi(t)\,dt = \psi(x) = \int_x^\infty f(x)g_i(s)\,ds = f(x)G_i(x).$$

For $\delta_0 = 0$, the proceeds in the same way and we obtain

$$\int_x^\infty f(s)g_i(x)\,ds = g_i(x)S(x).$$

Note: *The fact that G_i needs to be made left-continuous and S right-continuous follows from the second version of the proof. Here, both are continuous.* □

Proof (Version 2 – general case). Let $S_1 = \{x \in \mathbb{R} : \mathrm{P}\left[T_i = x\right] > 0\}$ and $S_2 = \{x \in \mathbb{R} : \mathrm{P}\left[C_i = x\right] > 0 \text{ for some } i\}$ be countable sets that include the possible discrete values of failure times and censoring times, respectively. Suppose the sets have at most finitely many points within any bounded subset of $\mathbb{R}$. Suppose the distributions of T_i and C_i are all absolutely continuous with respect to the measure $\lambda + \mu_{S_1 \cup S_2}$, where λ is the Lebesgue measure and $\mu_{S_1 \cup S_2}$ is the counting measure on the set $S_1 \cup S_2$. Then there exists a density $f(t)$ of T_i such that $S(t) = \int_{(t,\infty)} f(s)\,d\mu(s)$, which can be written down as $f(t) = f^*(t) + \Delta S(t)$, where

$$f^*(t) = \lim_{h \searrow 0} \frac{S(t) - S(t+h)}{h}$$
$$\text{and}\quad \Delta S(t) = S(t-) - S(t) = \mathrm{P}\left[T_i = t\right].$$

Similarly, there exist densities $g_i(t)$ of C_i such that $G_i(t) = \int_{(t,\infty)} g_i(s)\,d\mu(s)$, which can be written down as $g_i(t) = g_i^*(t) + \Delta G_i(t)$, where

$$g_i^*(t) = \lim_{h \searrow 0} \frac{G_i(t) - G_i(t+h)}{h}$$
$$\text{and}\quad \Delta G_i(t) = G_i(t-) - G_i(t) = \mathrm{P}\left[C_i = t\right].$$

Because T_i and C_i are independent, they have a joint density $h_i(t,s)$ with respect to the product measure $\mu \otimes \mu$ and $h_i(t,s) = f(t)g_i(s)$.

Since $X_i = T_i \wedge C_i$, the observation (X_i, δ_i) has a joint density $q_i(x,\delta)$ with respect to the product measure $\mu \otimes \mu_{\{0,1\}}$. First, evaluate it at $\delta = 1$. It is a sum of a continuous and a discrete component that can be obtained as

$$\lim_{h \searrow 0} \frac{\mathrm{P}[X_i > x, \delta_i = 1] - \mathrm{P}[X_i > x+h, \delta_i = 1]}{h} + \mathrm{P}[X_i = x, \delta_i = 1]$$

The continuous component can be also written as $-\frac{d^+}{dx}\mathrm{P}[X_i > x, \delta_i = 1]$, where the derivative is taken from the right. Calculate

$$\begin{aligned}
\mathrm{P}[X_i > x, \delta_i = 1] = \mathrm{P}[x < T_i \leq C_i] &= \int_{(x,\infty)} \left[\int_{\langle t,\infty)} h_i(t,s)\, d\mu(s) \right] d\mu(t) \\
&= \int_{(x,\infty)} f(t) \int_{\langle t,\infty)} g_i(s)\, d\mu(s)\, d\mu(t) = \int_{(x,\infty)} f(t) G_i(t-)\, d\mu(t).
\end{aligned}$$

The right derivative of this expression with respect to x is $f^*(x)G_i(x-)$, because $\Delta S(t) = 0$ on $\langle x, x+h)$ for h sufficiently small.

Next, calculate

$$\mathrm{P}[X_i = x, \delta_i = 1] = \mathrm{P}[T_i = x, C_i \geq x] = \mathrm{P}[T_i = x]\,\mathrm{P}[C_i \geq x] = \Delta S(x) G_i(x-).$$

Summing the two results, we get

$$q_i(x,1) = [f^*(x) + \Delta S(x)]G_i(x-) = f(x)G_i(x-).$$

For $\delta = 0$, we show by the same technique that

$$\mathrm{P}[X_i > x, \delta_i = 0] = \mathrm{P}[x < C_i < T_i] = \int_{(x,\infty)} g_i(t) S(t)\, d\mu(t)$$

and $\mathrm{P}[X_i = x, \delta_i = 0] = \Delta G_i(x) S(x)$. This leads to the desired result. $\square$

How to remember Lemma 2.1:

- *Let $\delta = 0$. We observe a censoring time x. We know that $C = x$ and $T > x$. Therefore the density is $g(x)S(x)$.*
- *Let $\delta = 1$. We observe a failure time x. We know that $T = x$ and $C \geq x$. Therefore the density is $f(x)G(x-)$.*

Now we can use Lemma 2.1 to construct the likelihood for parametric models with censored data. We recall that the distribution of T depends on a finite-dimensional parameter $\boldsymbol{\theta}$ to be estimated.

Theorem 2.2. *Let $T_1, \ldots, T_n$ be identically distributed with survival function $S(x;\boldsymbol{\theta})$, density $f(x;\boldsymbol{\theta})$, and hazard function $\lambda(x;\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$. Let C_i be independent of each other and independent of T_i, with an arbitrary survival function $G_i(x)$ and density $g_i(x)$.*

Then the likelihood function for $\boldsymbol{\theta}$ has the form

$$L(\boldsymbol{\theta}) = C \prod_{i=1}^{n} \left[\lambda(X_i;\boldsymbol{\theta}) \frac{S(X_i-;\boldsymbol{\theta})}{S(X_i;\boldsymbol{\theta})} \right]^{\delta_i} S(X_i;\boldsymbol{\theta}).$$

When the distribution of T_i is continuous,

$$L(\boldsymbol{\theta}) = C \prod_{i=1}^{n} \big[\lambda(X_i;\boldsymbol{\theta})\big]^{\delta_i} S(X_i;\boldsymbol{\theta}), \tag{2.1}$$

and the log-likelihood can be written as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left[\delta_i \log \lambda(X_i;\boldsymbol{\theta}) - \int_0^{X_i} \lambda(t;\boldsymbol{\theta})\, dt \right] + c.$$ ◊

Thus, the likelihood does not depend on the censoring distributions — as long as the censoring distributions do not involve the parameter $\boldsymbol{\theta}$. This requirement is called *uninformative censoring*.

Proof. Using Lemma 2.1, we get:

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^{n} \big[f(X_i;\boldsymbol{\theta})G_i(X_i-)\big]^{\delta_i} \big[g_i(X_i)S(X_i;\boldsymbol{\theta})\big]^{1-\delta_i} \\ &= \prod_{i=1}^{n} f(X_i;\boldsymbol{\theta})^{\delta_i} S(X_i;\boldsymbol{\theta})^{1-\delta_i} \underbrace{\left[\prod_{i=1}^{n} G_i(X_i-)^{\delta_i} g_i(X_i)^{1-\delta_i} \right]}_{=C} \\ &= C \prod_{i=1}^{n} \left[\frac{f(X_i;\boldsymbol{\theta})}{S(X_i;\boldsymbol{\theta})} \right]^{\delta_i} S(X_i;\boldsymbol{\theta}) = C \prod_{i=1}^{n} \left[\lambda(X_i;\boldsymbol{\theta}) \frac{S(X_i-;\boldsymbol{\theta})}{S(X_i;\boldsymbol{\theta})} \right]^{\delta_i} S(X_i;\boldsymbol{\theta}). \end{aligned}$$

For continuous T_i, we get $S(X_i-;\boldsymbol{\theta}) = S(X_i;\boldsymbol{\theta})$ and

$$S(X_i;\boldsymbol{\theta}) = \exp\{-\Lambda(X_i;\boldsymbol{\theta})\} = \exp\left\{ -\int_0^{X_i} \lambda(t;\boldsymbol{\theta})\, dt \right\}.$$ □

Standard results of the maximum likelihood theory can be applied to obtain the maximum likelihood estimator of $\boldsymbol{\theta}$ and its asymptotic distribution. For most failure time distributions, however, the score function and the information matrix are not easy to calculate. In the next two sections, we consider two special cases involving the exponential failure time distribution, which is the easiest to handle.

The end of lecture 3 (Oct. 12)

2.2. Exponential distribution with Type II censoring

Let $T_1, \dots, T_n$ be identically distributed with exponential distribution $\mathsf{Exp}(\lambda)$. The density and survival function of T_i is

$$f(t;\lambda) = \lambda \mathrm{e}^{-\lambda t} \quad \text{and} \quad S(t;\lambda) = \mathrm{e}^{-\lambda t},$$

respectively. The hazard function is $\lambda(t) = \lambda$.

In this section, we derive some useful results about the estimation and testing of the parameter λ under Type II censoring. We have a fixed $k \in \{1, \dots, n\}$ and set $C_i = T_{(k)}$ for all i, where $T_{(k)}$ is the k-th order statistic of the random sample $T_1, \dots, T_n$. The independent censoring condition introduced in Definition 1.4 is fulfilled but the conditions of Theorem 2.2 are not – C_i is not independent of T_i.

The statement of Theorem 2.2 is still true but we have to prove it separately, taking into account Type II censoring. This is done in the following Lemma 2.3 and Theorem 2.4.

The observed data $(X_1, \delta_1), \dots, (X_n, \delta_n)$ are determined by the values of the first k order statistics, so the likelihood can be obtained from the joint density of $(T_{(1)}, \dots, T_{(k)})$.

Lemma 2.3. *The joint density of $(T_{(1)}, \dots, T_{(k)})$ is*

$$h(t_1, \dots, t_k) = \frac{n!}{(n-k)!} \lambda^k \mathrm{e}^{-\lambda \left[\sum_{i=1}^k t_i + (n-k) t_k\right]} \quad \textit{when } 0 < t_1 < t_2 < \dots < t_k,$$

and $h(t_1, \dots, t_k) = 0$ otherwise. ◊

Proof. The joint density of the first k order statistics can be obtained by discretizing the continuous distribution of the vector and taking a limit. It proceeds as follows: take $0 = t_0 < t_1 < t_2 < \dots < t_k$ fixed and a small $h > 0$ (smaller than the smallest difference between t_j and t_{j-1}). Divide the positive half-line into $2k+1$ intervals

$$\langle t_0, t_1), \langle t_1, t_1 + h), \langle t_1 + h, t_2), \langle t_2, t_2 + h), \dots, \langle t_k, t_k + h), \langle t_k + h, \infty). \tag{*}$$

The probability

$$\mathrm{P}\left[T_{(1)} \in \langle t_1, t_1 + h), T_{(2)} \in \langle t_2, t_2 + h), \dots, T_{(k)} \in \langle t_k, t_k + h)\right]$$

is equal to the probability

$$\mathrm{P}\left[\boldsymbol{N} = (0, 1, 0, 1, \dots, 0, 1, n-k)\right]$$

where $\boldsymbol{N}$ is a random vector containing the numbers of observations $T_1, \dots, T_n$ that fell into the $2k+1$ successive intervals (*). The distribution of $\boldsymbol{N}$ is multinomial, in particular, $\mathsf{Mult}_{2k+1}(n, \boldsymbol{p})$, where $\boldsymbol{p}$ are the probabilities of the $2k+1$ intervals for a single exponential observation.

The survival function of the exponential distribution is $S(t) = \mathrm{P}[T_i > t] = \mathrm{e}^{-\lambda t}$. From this, the vector of probabilities $\boldsymbol{p}$ is equal to

$$\boldsymbol{p} = (1-\mathrm{e}^{-\lambda t_1}, \mathrm{e}^{-\lambda t_1}-\mathrm{e}^{-\lambda(t_1+h)}, \mathrm{e}^{-\lambda(t_1+h)}-\mathrm{e}^{-\lambda t_2}, \mathrm{e}^{-\lambda t_2}-\mathrm{e}^{-\lambda(t_2+h)}, \dots, \mathrm{e}^{-\lambda t_k}-\mathrm{e}^{-\lambda(t_k+h)}, \mathrm{e}^{-\lambda(t_k+h)}).$$

From the form of the multinomial density, we get

$$\mathrm{P}\left[T_{(1)} \in \langle t_1, t_1+h), T_{(2)} \in \langle t_2, t_2+h), T_{(k)} \in \langle t_k, t_k+h)\right] = \\ = \frac{n!}{(n-k)!}\left[\prod_{j=1}^{k} \mathrm{e}^{-\lambda t_j}\left(1-\mathrm{e}^{-\lambda h}\right)\right] \cdot \left[\mathrm{e}^{-\lambda(t_k+h)}\right]^{n-k}.$$

To find the desired joint density $h(t_1, \dots, t_k)$, divide this by h^k and take a limit as $h \to 0$ from above. This gives us

$$h(t_1, \dots, t_k) = \\ = \frac{n!}{(n-k)!}\exp\left\{-\lambda\left[\sum_{j=1}^{k} t_j + (n-k)t_k\right]\right\}\left[\prod_{j=1}^{k} \lim_{h\searrow 0}\frac{1-\mathrm{e}^{-\lambda h}}{h}\right]\lim_{h\searrow 0}\mathrm{e}^{-\lambda h(n-k)}.$$

Because $\lim_{h\searrow 0}\frac{1-\mathrm{e}^{-\lambda h}}{h} = \lambda$ and $\lim_{h\searrow 0}\mathrm{e}^{-\lambda h(n-k)} = 1$, we finally get

$$h(t_1, \dots, t_k) = \frac{n!}{(n-k)!}\lambda^k \exp\left\{-\lambda\left[\sum_{j=1}^{k} t_j + (n-k)t_k\right]\right\},$$

which was to be proven. □

The joint density is transformed into the likelihood by evaluating it at the observed data and considering it a function of the unknown parameter.

Theorem 2.4. *The likelihood function for exponential data with Type II censoring is*

$$L(\lambda \mid T_{(1)}, \dots, T_{(k)}) = \frac{n!}{(n-k)!}\lambda^k \mathrm{e}^{-\lambda S_{k,n}},$$

where

$$S_{k,n} = \sum_{i=1}^{k} T_{(i)} + (n-k)T_{(k)}$$

is the sufficient statistic. ◊

Theorem 2.4 follows directly from Lemma 2.3.

Comparing the result of Theorem 2.4 to the likelihood (2.1) from Theorem 2.2, we can see that they are the same, except the irrelevant multiplicative constant.

Maximizing the likelihood, we get the likelihood equation $k/\widehat{\lambda} - S_{k,n} = 0$, leading to the MLE $\widehat{\lambda} = k/S_{k,n}$. Notice that with our general notation, the MLE of the constant hazard rate λ can be written as

$$\widehat{\lambda} = \frac{\sum_{i=1}^n \delta_i}{\sum_{i=1}^n X_i},$$

that is, the number of observed failures divided by the total observation time.

The MLE of the expected failure time is

$$\widehat{\mu} = \frac{1}{\widehat{\lambda}} = \frac{1}{k}\sum_{i=1}^k T_{(i)} + \frac{n-k}{n} T_{(k)}.$$

The sufficient statistic $S_{k,n}$ is not a sum of independent and identically distributed random variables. However, it can be written as such after a simple transformation. This idea is the key to the proof of the following Theorem.

Theorem 2.5. *Let $T_1, \dots, T_n$ be independent and identically distributed with distribution* $\mathsf{Exp}(\lambda)$*, let $S_{k,n} = \sum_{i=1}^k T_{(i)} + (n-k)T_{(k)}$. Then*

$$2\lambda S_{k,n} \sim \chi^2_{2k}. \qquad \diamond$$

Proof. Consider the linear transformation $U_i = (n-i+1)(T_{(i)} - T_{(i-1)})$ for $i = 1, \dots, k$ with $T_{(0)} \equiv 0$. Notice that $\sum_{i=1}^k U_i = S_{k,n}$. Let us calculate the joint density of $(U_1, \dots, U_k)$ using the transformation theorem. The inverse transformation is also linear:

$$T_{(i)} = \sum_{j=1}^i \frac{U_j}{n-j+1}.$$

The derivative matrix of the inverse transformation is a triangular matrix and the Jacobian can be written as $(n-k)!/n!$.

From Lemma 2.3 and the transformation theorem, we get the density of $(U_1, \dots, U_k)$ as follows:

$$\lambda^k \mathrm{e}^{-\lambda \sum_{i=1}^k u_i}.$$

But this is the joint density of k iid random variables with the distribution $\mathsf{Exp}(\lambda)$, which is also a gamma distribution $\Gamma(\lambda, 1)$.

Next, $2\lambda U_i$ are iid random variables with $\Gamma(1/2, 1)$ distribution and hence

$$2\lambda S_{k,n} = 2\lambda \sum_{i=1}^k U_i \sim \Gamma(1/2, k) \equiv \chi^2_{2k}.$$

This completes the proof. $\square$

This theorem provides the basis for conducting exact parametric inference for exponential distribution with Type II censoring. Consider testing

$$H_0 : \lambda = \lambda_0 \quad \text{and} \quad H_1 : \lambda \neq \lambda_0.$$

Take $2\lambda_0 S_{k,n}$ as the test statistic, which has χ^2_{2k} distribution under the null hypothesis. The test that rejects H_0 when the test statistic is either too small or too large:

$$2\lambda_0 S_{k,n} < \chi^2_{2k}(\alpha/2) \quad \text{or} \quad 2\lambda_0 S_{k,n} > \chi^2_{2k}(1-\alpha/2),$$

where $\chi^2_f(\alpha)$ is the α-quantile of the χ^2_f distribution. The level of this test is exactly α and it is the most powerful test of H_0 in the current model. Similarly, an exact confidence interval for λ with coverage $1-\alpha$ is

$$\left(\frac{\chi^2_{2k}(\alpha/2)}{2S_{k,n}}, \frac{\chi^2_{2k}(1-\alpha/2)}{2S_{k,n}} \right).$$

This is the only case when exact inference is possible with censored data.

The end of lecture 4 (Oct. 12)

2.3. Exponential regression with arbitrary random censoring

In this section, we use the results of Section 2.1 to investigate a regression model for censored exponentially distributed response. Let $T_1, \ldots, T_n$ be distributed according to $\mathsf{Exp}(\lambda_i)$ and let $C_1, \ldots, C_n$ be independent of each other and independent of $T_1, \ldots, T_n$, with arbitrary distributions. We observe independent triplets

$$(X_1, \delta_1, \boldsymbol{Z}_1), \ldots, (X_n, \delta_n, \boldsymbol{Z}_n),$$

where $X_i = T_i \wedge C_i$, $\delta_i = \mathbb{1}(T_i \leq C_i)$, and $\boldsymbol{Z}_i$ are random covariate vectors of dimension p, typically with the first component equal to one.

Suppose there exists a p-vector $\boldsymbol{\beta}$ of regression parameters such that

$$\lambda_i = \mathrm{e}^{\boldsymbol{\beta}^\top \boldsymbol{Z}_i}.$$

We would like to estimate the vector of regression coefficients $\boldsymbol{\beta}$ by maximum likelihood methods. According to Theorem 2.2, the likelihood function for $\boldsymbol{\beta}$ has the form

$$L(\boldsymbol{\beta}) = C \prod_{i=1}^{n} \lambda_i^{\delta_i} \mathrm{e}^{-\lambda_i X_i},$$

and the log-likelihood is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\delta_i \boldsymbol{\beta}^\top \boldsymbol{Z}_i - \mathrm{e}^{\boldsymbol{\beta}^\top \boldsymbol{Z}_i} X_i \right) + c.$$

Differentiating this with respect to the vector $\boldsymbol{\beta}$, we obtain the score statistic

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\delta_i - \mathrm{e}^{\log X_i + \boldsymbol{\beta}^\top \boldsymbol{Z}_i} \right) \boldsymbol{Z}_i.$$

This is equivalent to the score statistic of a Poisson loglinear model with δ_i as the response and $\log X_i$ as the offset (see the notes for Advanced Regression Models).

Algorithms for finding the maximum likelihood estimator of $\boldsymbol{\beta}$, calculating the observed information matrix, approximating the distribution of the estimated $\boldsymbol{\beta}$, performing tests about $\boldsymbol{\beta}$ and building confidence intervals are all the same as in the Poisson loglinear model.

In particular, the expected information matrix is

$$I(\boldsymbol{\beta}) = \mathsf{E}\, \boldsymbol{Z}_i \boldsymbol{Z}_i^\top \mathrm{e}^{\boldsymbol{Z}_i^\top \boldsymbol{\beta}} X_i.$$

It can be consistently estimated by

$$\widehat{I}_n = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{Z}_i \boldsymbol{Z}_i^\top \mathrm{e}^{\boldsymbol{Z}_i^\top \widehat{\boldsymbol{\beta}}} X_i$$

and

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathsf{D}} \mathsf{N}_p(\mathbf{0}, I^{-1}(\boldsymbol{\beta})).$$

Any software that can fit loglinear models can be used to perform exponential regression with arbitrary independent censoring.

An important special case arises when $\boldsymbol{Z}_i = 1$ for all i. Then $\boldsymbol{Z}_i^\top \boldsymbol{\beta} = \beta$ and the failure times are i.i.d. with the distribution $\mathsf{Exp}(\lambda)$ where $\lambda = \mathrm{e}^{\beta}$. The score statistic is

$$U(\beta) = \sum_{i=1}^{n} \left(\delta_i - \mathrm{e}^{\log X_i + \beta} \right) = \sum_{i=1}^{n} (\delta_i - \lambda X_i).$$

and the MLE of the parameter λ has an explicit form

$$\widehat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} X_i},$$

that is, the number of observed failures divided by the total follow-up time. This is the same estimator as for the case of Type II censoring.

3. Counting Processes and Martingales

Starting with this chapter, we turn our attention to nonparametric methods for censored data. It turns out that it helps to view censored data as stochastic processes evolving over time and to use theoretical results that are available for stochastic processes to develop nonparametric estimators and tests and to establish their properties. We will see that the theory of martingales and martingale integrals will become particularly useful to our purposes. In this chapter, we introduce notation and summarize the most important results from martingale theory that will be used in subsequent chapters.

We start by introducing a new notation that translates the censored observation (X, δ), where $X = T \wedge C$ and $\delta = \mathbb{1}(T \leq C)$, into a pair of stochastic processes. At the origin, time $t = 0$, we start following the subject and wait for the failure. Once we observe one, we mark the subject as having failed at that moment. If the subject is censored before a failure occurs, the follow-up is terminated and the subject is no longer at risk for failure.

This consideration motivates the following formulation of the problem. Let $N(t)$ be a stochastic process defined as

$$N(t) = \mathbb{1}(T \leq t, \delta = 1).$$

It is a process that counts the number of failures that were observed prior to (and including) t. It is an example of *a counting process*[*] specified by Definition A.5 in the Appendix; its initial value $N(0)$ is zero, it is finite, it has piecewise constant right-continuous paths and jumps of size 1. This particular counting process jumps to 1 at the failure time, and stays at 1 thereafter. If the subject is censored, $N(t)$ never jumps and stays at 0 indefinitely.

In this chapter, we summarize some useful results about general counting processes and apply them to the censored data problem. When we need to distinguish the specific counting process $N(t) = \mathbb{1}(T \leq t, \delta = 1)$ from the general counting process $N(t)$, we will call it *the censored data counting process*. Ideally, we would prefer to work with the "uncensored" counting process $\mathbb{1}(T \leq t)$ but we cannot because its paths are not fully observed.

Let $Y(t)$ be another process such that

$$Y(t) = \mathbb{1}(X \geq t).$$

[*] Česky *čítací proces*

This is the *at-risk process*[*]. It indicates whether or not the subject is under observation at the time t. It starts at 1 at $t = 0$ and drops to 0 as soon as the failure occurs or the subject is censored. The at-risk process $Y(t)$ is left-continuous.

Obviously, the original censored data setup (X, δ) is equivalent to providing the observed paths of $N(t)$ and $Y(t)$, $t > 0$.

The counting process notation is useful in several ways. First, by emphasizing the development of the censored observation over time, it facilitates the utilization of the martingale theory based on conditioning upon the past. This simplifies the problem in many important ways. Second, it can be easily generalized to allow for late entry into the observation period, repeated change of the at-risk status, presence of repeated failures for the same subject, and direct modeling of various time-varying features (such as covariates that change over time). However, most of these topics will not be adressed in the present course.

Before we start exploring the properties of censored data via counting processes, we generalize the independent censoring condition first formulated for continuous data in Definition 1.4 to arbitrary failure time distributions.

Definition 3.1. The censoring variable C satisfies *the independent censoring condition* for the failure time T with cumulative hazard Λ if and only if

$$\Lambda(t) = -\int_0^t \frac{d\mathrm{P}\left[T \geq s, C \geq T\right]}{\mathrm{P}\left[T \geq s, C \geq s\right]} \quad \forall t \text{ such that } \mathrm{P}\left[T \geq t, C \geq t\right] > 0. \tag{3.1}$$

∇

Note. When the distribution of T is continuous, condition (3.1) is equivalent to equality

$$\begin{aligned}\lambda(t) &= \frac{-\frac{\partial}{\partial s}\mathrm{P}\left[T \geq s, C \geq t\right]\Big|_{s=t}}{\mathrm{P}\left[T \geq t, C \geq t\right]} \\ &= \lim_{h \searrow 0} \frac{1}{h}\mathrm{P}\left[t \leq T < t + h \,\middle|\, T \geq t, C \geq t\right] \quad \forall t \geq 0.\end{aligned}$$

The net hazard on the left should be equal to the crude hazard on the right, as required by Definition 1.4. Definition 3.1 is written in a less intuitive way but applies to distributions with discrete components as well.

We will always assume that the independent censoring condition holds.

Note: Many theorems and other statements in this chapter are left without proof; however, many of the proofs have been covered in the course NMTP436 *Continuous martingales and counting processes.*

[*] Česky *pozorovací proces*

3.1. Doob-Meyer decomposition

Our most important tool will be the Doob-Meyer decomposition of a submartingale.

Theorem 3.1 (Doob-Meyer). *Let $X(t)$ be a right-continuous non-negative $\mathcal{F}_t$-submartingale. Then there exists a unique (up to sets of measure zero) right-continuous martingale $M(t)$ and a non-decreasing right-continuous $\mathcal{F}_t$-predictable process $A(t)$ such that $A(0) = 0$, $\mathsf{E}\, A(t) < \infty$, and*

$$X(t) = M(t) + A(t) \quad \textit{almost surely}$$

for any $t \geq 0$. In addition, if $X(t)$ is bounded then $M(t)$ is uniformly integrable and $A(t)$ is integrable. ◊

Note.

- The process $A(t)$ is called the *compensator** for the submartingale $X(t)$. In general, it depends on the filtration $\mathcal{F}_t$.
- Suppose $X(0) = 0$. Then $M(0) = 0$, $\mathsf{E}\, M(t) = 0$, and the martingale $M(t)$ represents the "random noise" part of $X(t)$ while the compensator $A(t)$ can be regarded as the "systematic" part of X.
- Left-continuous adapted processes are always predictable. The compensator $A(t)$ from the Doob-Meyer theorem is right-continuous and still predictable.
- Any counting process satisfies the conditions of the Doob-Meyer theorem.

Recall that a single censored observation can be described as the pair of stochastic processes $N(t) = \mathbb{1}(T \leq t, \delta = 1)$ and $Y(t) = \mathbb{1}(X \geq t)$ or, equivalently, as the pair of counting processes $N(t) = \mathbb{1}(T \leq t, \delta = 1)$ and $N^U(t) = \mathbb{1}(C \leq t, \delta = 0)$. Introduce the natural filtration summarizing the history of observed failure and censoring times up to time t:

$$\mathcal{F}_t = \sigma\{N(s), Y(s),\ 0 \leq s \leq t\} = \sigma\{N(s), N^U(s),\ 0 \leq s \leq t\}. \tag{3.2}$$

Then $N(t)$ is a counting process with respect to this filtration in the sense of Definition A.5. It is also a right-continuous non-negative $\mathcal{F}_t$-submartingale. Thus, according to the Doob-Meyer Theorem, there exists a non-decreasing right-continuous $\mathcal{F}_t$-predictable compensator $A(t)$ such that $M(t) = N(t) - A(t)$ is an $\mathcal{F}_t$-martingale. The next theorem shows that under independent censoring condition we know the form of this compensator.

Theorem 3.2. *Let*

$$A(t) = \int_0^t Y(s)\, d\Lambda(s) = \Lambda(t \wedge X). \tag{3.3}$$

* Česky *kompenzátor*

This is a right-continuous $\mathcal{F}_t$-predictable process. The process

$$M(t) = N(t) - A(t)$$

is an $\mathcal{F}_t$-martingale if and only if the independent censoring condition (3.1) *holds.* ◊

Note. Because $M(0) = 0$ a.s., we have $\mathsf{E}\, M(t) = 0$ and hence $\mathsf{E}\, N(t) = \mathsf{E}\, \Lambda(t \wedge X)$ for all $t > 0$.

Note. The claim of Theorem 3.2 can be extended to the case of multiple independent censored observations. Let $(T_1, C_1), \ldots, (T_n, C_n)$ be independent, $X_i = T_i \wedge C_i$ and $\delta_i = \mathbb{1}(T_i \le C_i)$. We observe independent pairs $(X_1, \delta_1), \ldots, (X_n, \delta_n)$. Let $\Lambda_i(t)$ be the cumulative hazard of T_i. Let the independent censoring condition (3.1) hold for each pair (T_i, C_i). Define $N_i(t) = \mathbb{1}(T_i \le t, \delta_i = 1)$, $Y_i(t) = \mathbb{1}(X_i \ge t)$, and $N_i^U(t) = \mathbb{1}(C_i \le t, \delta_i = 0)$. Define the extended filtration summarizing the history of observed failure and censoring times for all subjects up to time t:

$$\mathcal{F}_t = \sigma\{N_i(s), Y_i(s),\ 0 \le s \le t, i = 1, \ldots, n\} = \sigma\{N_i(s), N_i^U(s),\ 0 \le s \le t, i = 1, \ldots, n\}. \tag{3.4}$$

Let

$$A_i(t) = \int_0^t Y_i(s)\, d\Lambda_i(s) = \Lambda_i(t \wedge X_i). \tag{3.5}$$

Then $M_i(t) = N_i(t) - A_i(t)$ is a martingale with respect to the extended filtration (3.4).

It is easy to see that for a right-continuous martingale $M(t)$ such that $\mathsf{E}\, M^2(t) < \infty$, the process $M^2(t)$ is a right-continuous submartingale. The Doob-Meyer decomposition can be applied to M^2, justifying the following corollary.

Corollary. For each right-continuous $\mathcal{F}_t$-martingale $M(t)$ with $\mathsf{E}\, M^2(t) < \infty$ for all $t > 0$, there exists a non-decreasing right-continuous $\mathcal{F}_t$-predictable process $\langle M, M\rangle(t)$ with $\langle M, M\rangle(0) = 0$ and finite expectation such that

$$M^2(t) - \langle M, M\rangle(t) \quad \text{is an } \mathcal{F}_t\text{-martingale.}$$

The process $\langle M, M\rangle(t)$ is uniquely determined (up to sets of measure zero).

The process $\langle M, M\rangle$ introduced in the corollary is called *the predictable variation process*[*] of the martingale $M(t)$.

Note. If $M(0) = 0$ a.s. and $\mathsf{E}\, M^2(t) < \infty$ then $\mathsf{var}\, M(t) = \mathsf{E}\, M^2(t) = \mathsf{E}\, \langle M, M\rangle(t)$.

The product of two martingales is not in general a submartingale, however, the Doob-Meyer theorem can be extended to guarantee the existence of a "pseudo-compensator" for martingale products.

[*] Česky *prediktabilní variančnı́ proces*

Theorem 3.3. *Let $M_1(t)$, $M_2(t)$ be right-continuous $\mathcal{F}_t$-martingales with* $\mathsf{E}\, M_j^2(t) < \infty$ *for all* $t > 0$, $j = 1, 2$. *Then there exists a process* $\langle M_1, M_2\rangle(t)$ *with the following properties:*

(i) $\langle M_1, M_2\rangle(t)$ *is right-continuous,* $\mathcal{F}_t$*-predictable,* $\langle M_1, M_2\rangle(0) = 0$ *a.s., and its expectation is finite* $\forall t \geq 0$*;*

(ii) $\langle M_1, M_2\rangle(t)$ *is a difference of two non-decreasing right-continuous* $\mathcal{F}_t$*-predictable processes;*

(iii)

$$M_1(t)M_2(t) - \langle M_1, M_2\rangle(t) \quad \text{is an } \mathcal{F}_t\text{-martingale.} \qquad \diamond$$

The process $\langle M_1, M_2\rangle$ of the preceding theorem is called *the predictable covariation process*[*] of the martingales $M_1(t)$ and $M_2(t)$.

Note.

- If $M_1(0)$ and $M_2(0)$ are uncorrelated then $\operatorname{cov}(M_1(t), M_2(t)) = \mathsf{E}\,\langle M_1, M_2\rangle(t)$.
- M_1M_2 is a martingale if and only if $\langle M_1, M_2\rangle(t) = 0$ at all $t \geq 0$. If this is the case, the martingales M_1 and M_2 are called *orthogonal*.

In the next section we will show that it is possible to derive explicit forms of predictable variation and covariation processes for counting process martingales.

Homework 3:

Prove Theorem 3.3 by application of the Doob-Meyer Theorem to the submartingale $(M_1(t) + M_2(t))^2$.

The end of lecture 5 (Oct. 19)

3.2. Martingale integrals

In this section, we consider processes of the type

$$L(t) = \int_0^t H(s)\, dM(s),$$

where H is a bounded $\mathcal{F}_t$-predictable process and M is an $\mathcal{F}_t$-martingale having paths with total variation bounded by a constant almost surely. For any right-continuous function F with left-hand limits, the notation $\Delta F(x) = F(x) - F(x-)$ means the jump of F at x.

The results of this section are not formulated in their most general versions. They can be extended in several ways. First, to processes H that are only locally bounded:

[*] Česky *prediktabilní kovarianční proces*

the details can be found in Fleming and Harrington (1991) and Andersen et al. (1993). Second, to martingales M that do not have paths of bounded variation such as the Brownian motion. This extension leads to Itô-type integrals.

Theorem 3.4. *Let N be a general counting process and let A be its compensator according to Theorem 3.1 such that $M = N - A$ is an $\mathcal{F}_t$-martingale. Let $\Delta M(0) = 0$ a.s. Let H be a bounded $\mathcal{F}_t$-predictable process. Then*

$$L(t) = \int_0^t H(s)\, dM(s)$$

is an $\mathcal{F}_t$-martingale. ◊

Note.

- Since $L(0) = 0$ a.s., it follows that $\mathsf{E} \int_0^t H(s)\, dM(s) = 0$ for all $t \geq 0$.
- Consider processes N_i, A_i, H_i for $i = 1, \ldots, n$. Let $M_i = N_i - A_i$. Suppose that the conditions of Theorem 3.4 are satisfied for each i with a common filtration $\mathcal{F}_t$. Then

$$L(t) = \sum_{i=1}^{n} \int_0^t H_i(s)\, dM_i(s) \tag{3.6}$$

  is an $\mathcal{F}_t$-martingale.

Predictable covariation processes for martingale integrals

Now we consider predictable covariation processes for martingale integrals. When we write expressions such as $\int Z\, dX$ without limits and dummy arguments, they are to be interpreted as $\int_0^t Z(s)\, dX(s)$.

Theorem 3.5. *Let the conditions of Theorem 3.4 hold for N_j, A_j, H_j, $j = 1, 2$, take $M_j = N_j - A_j$ and assume $\mathsf{E}\, M_j^2(t) < \infty$. Denote $L_j(t) = \int H_j\, dM_j$. Then there exists a predictable covariation process $\langle L_1, L_2 \rangle$ and*

$$\langle L_1, L_2 \rangle = \int H_1 H_2\, d\langle M_1, M_2 \rangle.$$

In particular,

$$\int H_1\, dM_1 \int H_2\, dM_2 - \int H_1 H_2\, d\langle M_1, M_2 \rangle$$

is an $\mathcal{F}_t$-martingale. ◊

Corollary.

- $\mathrm{cov}\left(\int H_1\, dM_1, \int H_2\, dM_2 \right) = \mathsf{E} \int H_1 H_2\, d\langle M_1, M_2 \rangle.$

- If M_1 and M_2 are orthogonal then $\operatorname{cov}\left(\int H_1\,dM_1, \int H_2\,dM_2\right) = 0$ for any bounded predictable H_1 and H_2.
- $\operatorname{var} \int H\,dM = \mathsf{E} \int H^2\,d\langle M, M\rangle$.

Note. Let $U_1 = \sum_{i=1}^n \int H_i\,dM_i$ and $U_2 = \sum_{i=1}^n \int H_i^*\,dM_i$ with all H_i and H_i^* bounded and $\mathcal{F}_t$-predictable. Then $\mathsf{E}\,U_1 = \mathsf{E}\,U_2 = 0$,

$$\begin{aligned}
\operatorname{var} U_1 &= \mathsf{E} \sum_{i=1}^n \sum_{j=1}^n \int H_i H_j\,d\langle M_i, M_j\rangle, \\
\operatorname{var} U_2 &= \mathsf{E} \sum_{i=1}^n \sum_{j=1}^n \int H_i^* H_j^*\,d\langle M_i, M_j\rangle, \quad \text{and} \\
\operatorname{cov}(U_1, U_2) &= \mathsf{E} \sum_{i=1}^n \sum_{j=1}^n \int H_i H_j^*\,d\langle M_i, M_j\rangle.
\end{aligned}$$

When M_i and M_j are orthogonal martingales for all $i \neq j$, then

$$\begin{aligned}
\operatorname{var} U_1 &= \mathsf{E} \sum_{i=1}^n \int H_i^2\,d\langle M_i, M_i\rangle, \\
\operatorname{var} U_2 &= \mathsf{E} \sum_{i=1}^n \int (H_i^*)^2\,d\langle M_i, M_i\rangle, \quad \text{and} \\
\operatorname{cov}(U_1, U_2) &= \mathsf{E} \sum_{i=1}^n \int H_i H_i^*\,d\langle M_i, M_i\rangle.
\end{aligned}$$

These results will become useful when the we learn how to calculate predictable variation and covariation processes.

Theorem 3.6. *Let $A(t)$ be the compensator for a general counting process $N(t)$ and denote $M(t) = N(t) - A(t)$ the associated martingale. Then*

$$\langle M, M\rangle(t) = \int_0^t [1 - \Delta A(s)]\,dA(s).$$

If the compensator $A(t)$ is continuous then $\langle M, M\rangle(t) = A(t)$. ◇

Corollary. The martingale $M(t)$ is square integrable:

$$\operatorname{var} M(t) = \mathsf{E}\,M^2(t) = \mathsf{E} \int_0^t (1 - \Delta A)\,dA \leq \mathsf{E}\,A(t) < \infty.$$

Proof (of Theorem 3.6). We apply integration by parts for Lebesgue-Stieltjes integral (Theorem A.1 in the Appendix). Take $F(t) = G(t) = M(t)$ and write

$$M^2(t) = 2\int_0^t M(s-)\,dM(s) + \int_0^t \Delta M(s)\,dM(s). \tag{*}$$

In the first integral, the integrand $M(s-)$ is bounded on $(0,t)$, left continuous and therefore predictable. The first integral is a martingale. In the second integral, the integrand $\Delta M(s)$ is not predictable.

In the sequel, we drop the time arguments and integral bounds whenever possible. Take $\int \Delta M\, dM$ and write $M = N - A$. Then

$$\begin{aligned}\int \Delta M\, dM &= \int (\Delta N - \Delta A)\, d(N-A) \\ &= \sum_{s\le t} \Delta N(\Delta N - \Delta A) - \int \Delta A\, dM \\ &= \sum_{s\le t} \Delta N - \int \Delta A\, dN - \int \Delta A\, d(N-A) \\ &= N - A + \int 1\, dA - \int \Delta A\, dA - 2\int \Delta A\, d(N-A) \\ &= M + \int (1-\Delta A)\, dA - 2\int \Delta A\, dM.\end{aligned}$$

Now, the compensator A is bounded and predictable, hence $\int \Delta A\, dM$ is a martingale. We can write

$$\int \Delta M\, dM - \int (1-\Delta A)\, dA = M - 2\int \Delta A\, dM.$$

The right-hand side is a martingale. Hence, $\int (1-\Delta A)\, dA$ is a compensator to $\int \Delta M\, dM$, and, because of (*), also to M^2. This completes the proof. □

Let us return to the special case of censored data counting process $N(t) = \mathbb{1}(T \le t, \delta = 1)$ accompanied by the at-risk process $Y(t) = \mathbb{1}(X \ge t)$. Let the filtration be defined by (3.2). According to Theorem 3.2, when the independent censoring condition holds the compensator for $N(t)$ is $A(t) = \int_0^t Y(s)\, d\Lambda(s)$ and $M(t) = N(t) - A(t)$ is an $\mathcal{F}_t$-martingale.

Theorem 3.6 tells us that the predictable variation process for $M(t)$ is $\langle M, M\rangle(t) = \int_0^t [1-\Delta\Lambda(s)]Y(s)\, d\Lambda(s)$. For continuous failure times, we have $\Delta\Lambda(s) = 0$ and thus the same process $A(t) = \int_0^t Y(s)\, d\Lambda(s)$ compensates both $N(t)$ and $M^2(t)$.

The end of lecture 6 (Oct. 19)

Multivariate counting process

Definition 3.2. Let $N_i(t)$, $i = 1, \dots, n$, be general counting processes adapted to a common filtration $\mathcal{F}_t$. The collection $\{N_1(t), \dots, N_n(t)\}$ is called a *multivariate counting process*[*] if and only if $\mathrm{P}\left[\Delta N_i(t) = 1, \Delta N_j(t) = 1\right] = 0$ for all $i \ne j$ and all $t \ge 0$. ∇

[*] Česky *mnohorozměrný čítací proces*

Note. Individual counting processes included in a multivariate counting process cannot jump at the same time. In our censored data special case, if failure times $T_1, \dots, T_n$ are independent with continuous distributions, their counting processes $N_i(t) = \mathbb{1}(T_i \leq t, \delta_i = 1)$ form a multivariate counting process.

In the next part we calculate predictable covariation processes for martingales associated with a multivariate counting process. The following theorem was proven in the course "Continuous Martingales and Counting Processes".

Theorem 3.7. *Let $\{N_1(t), \dots, N_n(t)\}$ be a multivariate counting process. Let A_i, a compensator for N_i, be* ***continuous*** *for each $i = 1, \dots, n$, let $M_i = N_i - A_i$. Then $\langle M_i, M_j \rangle = 0$ a.s. for all $i \neq j$.*

Note. In the censored data setting, it follows that if the underlying failure time variables are continuous, their martingales are orthogonal and the processes $M_i M_j$ are martingales for any $i \neq j$.

The previous theorem can be extended to any multivariate counting process.

Theorem 3.8. *Let $\{N_1(t), \dots, N_n(t)\}$ be a multivariate counting process. Let A_i be a compensator for N_i with respect to the common filtration $\mathcal{F}_t$. Let $M_i = N_i - A_i$, $i = 1, \dots, n$. Then*

$$\langle M_i, M_j \rangle = -\int \Delta A_i \, dA_j \quad \text{a.s. for all } i \neq j.$$

◊

Note. In the censored data setting, if the underlying failure time variables have discrete components so that their compensators have jumps, their martingales are negatively correlated. This agrees with the definition of the multivariate counting process, where jumps are prohibited for all other processes at the time when one of them jumps.

Proof. Consider the process $N_i + N_j$ for $i \neq j$. Because $\{N_1(t), \dots, N_n(t)\}$ is a multivariate counting process, the process $N_i + N_j$ has jumps of size at most one (almost surely) and hence it is a counting process.

The compensator for $N_i + N_j$ is $A_i + A_j$. $M_i + M_j$ is a martingale that satisfies the conditions of Theorem 3.6. According to that theorem,

$$\begin{aligned}\langle M_i + M_j, M_i + M_j \rangle &= \int (1 - \Delta A_i - \Delta A_j) \, d(A_i + A_j) \\ &= A_i + A_j - \int (\Delta A_i + \Delta A_j) \, d(A_i + A_j).\end{aligned}$$

Use the equality

$$\langle M_i, M_j\rangle = \frac{1}{2}\big[\langle M_i + M_j, M_i + M_j\rangle - \langle M_i, M_i\rangle - \langle M_j, M_j\rangle\big]$$

and the known forms of the predictable covariance processes on the right-hand side to get

$$\begin{aligned}\langle M_i, M_j\rangle &= \frac{1}{2}\Big[A_i + A_j - \int (\Delta A_i + \Delta A_j)\, d(A_i + A_j)\\ &\qquad - \int (1-\Delta A_i)\, dA_i - \int (1-\Delta A_j)\, dA_j\Big]\\ &= \frac{1}{2}\Big(-\int \Delta A_i\, dA_j - \int \Delta A_j\, dA_i\Big)\\ &= -\int \Delta A_i\, dA_j.\end{aligned}$$

This completes the proof. □

The final theorem does not require a multivariate counting process but makes a conditional independence assumption.

Theorem 3.9. *Let $\Delta N_1(t), \ldots, \Delta N_n(t)$ be independent given $\mathcal{F}_{t-}$. Then $\langle M_i, M_j\rangle(t) = 0$ almost surely for all $i \neq j$ and all $t \geq 0$.*

Proof. Use integration by parts on $M_i M_j$:

$$\begin{aligned}M_i(t)M_j(t) - \underbrace{M_i(0)M_j(0)}_{=0} &= \int_0^t M_i(s-)\, dM_j(s) + \int_0^t M_j(s-)\, dM_i(s)\\ &\qquad + \sum_{s\leq t} \Delta M_i(s)\Delta M_j(s).\end{aligned}$$

The first two terms are martingales. We need to show that $\sum_{s\leq t} \Delta M_i(s)\Delta M_j(s)$ is an $\mathcal{F}_t$-martingale, too. It will suffice to show that the following conditional expectation is zero for any $u < s$.

$$\mathsf{E}\Big[\sum_{u<s\leq t} \Delta M_i(s)\Delta M_j(s)\,\big|\mathcal{F}_u\Big] = \sum_{u<s\leq t} \mathsf{E}\big[\mathsf{E}\big[\Delta M_i(s)\Delta M_j(s)\,\big|\mathcal{F}_{s-}\big]\,\big|\mathcal{F}_u\big].$$

Now decompose

$$\mathsf{E}\big[\Delta M_i(s)\Delta M_j(s)\,\big|\mathcal{F}_{s-}\big] = \mathsf{E}\big[M_i(s)M_j(s) - M_i(s-)M_j(s) - M_i(s)M_j(s-) + M_i(s-)M_j(s-)\,\big|\mathcal{F}_{s-}\big].$$

Because M_i and M_j are martingales,

$$\mathsf{E}\big[M_i(s)\,\big|\mathcal{F}_{s-}\big] = M_i(s-) \quad \text{and} \quad \mathsf{E}\big[M_j(s)\,\big|\mathcal{F}_{s-}\big] = M_j(s-).$$

It remains to show that $\mathsf{E}\big[M_i(s)M_j(s)\,\big|\mathcal{F}_{s-}\big] = M_i(s-)M_j(s-)$ as well. Decompose both martingales as the counting process minus the compensator.

$$\mathsf{E}\big[M_i(s)M_j(s)\,\big|\mathcal{F}_{s-}\big] = \mathsf{E}\big[N_i(s)N_j(s) - A_i(s)N_j(s) - N_i(s)A_j(s) + A_i(s)A_j(s)\,\big|\mathcal{F}_{s-}\big].$$

The compensators A_i, A_j are predictable and hence $\mathcal{F}_{s-}$-measurable. Now use the assumption of independence of the jumps in $N_i(s)$ and $N_j(s)$ given $\mathcal{F}_{s-}$ to get

$$\begin{aligned}\mathsf{E}\big[M_i(s)M_j(s)\,\big|\mathcal{F}_{s-}\big] &= \mathsf{E}\big[N_i(s) - A_i(s)\,\big|\mathcal{F}_{s-}\big]\mathsf{E}\big[N_j(s) - A_j(s)\,\big|\mathcal{F}_{s-}\big] \\ &= \mathsf{E}\big[M_i(s)\,\big|\mathcal{F}_{s-}\big]\mathsf{E}\big[M_j(s)\,\big|\mathcal{F}_{s-}\big] = M_i(s-)M_j(s-).\end{aligned}$$

This completes the proof. □

Summary of main results

Let us summarize the important properties of martingale integral sums of the form (3.6) that we explained in this section. Consider censored data counting processes $N_i(t)$ and at-risk processes $Y_i(t)$, $i = 1, \dots, n$, that describe n independent observations of censored failure times with cumulative hazard functions Λ_i. Let $M_i(t) = N_i(t) - A_i(t)$, where $A_i(t) = \int_0^t Y_i(t)\,d\Lambda_i(t)$ is the compensator for $N_i(t)$ under independent censoring and a common filtration $\mathcal{F}_t$. By Theorem 3.6,

$$\langle M_i, M_i\rangle(t) = \int_0^t [1 - \Delta A_i(s)]\,dA_i(s) = \int_0^t [1 - \Delta\Lambda_i(s)]Y_i(s)\,d\Lambda_i(s).$$

For continuous failure times with hazard functions λ_i, we get $\langle M_i, M_i\rangle(t) = \int_0^t Y_i(s)\lambda_i(s)\,ds$. Also, $\langle M_i, M_j\rangle(t) = 0$ for all $i \neq j$ by Theorem 3.9 (because of independence).

Take $H_{ki}(t)$ bounded, $\mathcal{F}_t$-predictable processes $k = 1, 2$, $i = 1, \dots, n$. Consider the sums

$$U_k(t) = \sum_{i=1}^n \int_0^t H_{ki}(s)\,dM_i(s), \qquad k = 1, 2.$$

We have established the following facts about these processes:

- $U_k(t)$ are $\mathcal{F}_t$-martingales by Theorem 3.4.
- $\mathsf{E}\,U_k(t) = 0$.
- $\operatorname{var} U_k(t) = \sum_{i=1}^n \int_0^t \mathsf{E}\left[H_{ki}^2(s)Y_i(s)\right][1 - \Delta\Lambda_i(s)]\,d\Lambda_i(s)$ by Theorems 3.5, 3.6, and 3.9.
- $\operatorname{cov}(U_1(t), U_2(t)) = \sum_{i=1}^n \int_0^t \mathsf{E}\left[H_{1i}(s)H_{2i}(s)Y_i(s)\right][1 - \Delta\Lambda_i(s)]\,d\Lambda_i(s)$ by Theorems 3.5, 3.6, and 3.9.

In the next section, we provide central limit theorems for such sums as $n \to \infty$.

3.3. Central limit theorems for sums of martingale integrals

We consider two central limit theorems for two somewhat different cases. Both assume continuous compensators, though they could be extended to other cases as well.

Central limit theorem, case 1

We will be working under the following conditions:

- Let $\{N_{ki}^{(n)} : k = 1, \dots, r,\ i = 1, \dots, n\}$ be a multivariate counting process with respect to the stochastic basis $(\Omega, \mathcal{A}, \{\mathcal{F}_t\}_{t\geq 0}, P)$.
- Let the compensator $A_{ki}^{(n)}$ for $N_{ki}^{(n)}$ be continuous.
- Let $H_{ki}^{(n)}$, $k = 1, \dots, r$, $i = 1, \dots, n$, be bounded* $\mathcal{F}_t$-predictable processes on the interval $\langle 0, \tau\rangle$.

Let $M_{ki}^{(n)} = N_{ki}^{(n)} - A_{ki}^{(n)}$ be the $\mathcal{F}_t$-martingale for $N_{ki}^{(n)}$. Denote

$$U_{ki}^{(n)}(t) = \int_0^t H_{ki}^{(n)}(s)\, dM_{ki}^{(n)}(s) \quad \text{and} \quad U_k^{(n)}(t) = \sum_{i=1}^n U_{ki}^{(n)}(t).$$

Take any $\varepsilon > 0$ and denote

$$U_{ki,\varepsilon}^{(n)}(t) = \int_0^t H_{ki}^{(n)}(s)\mathbb{1}(|H_{ki}^{(n)}(s)| > \varepsilon)\, dM_{ki}^{(n)}(s) \quad \text{and} \quad U_{k,\varepsilon}^{(n)}(t) = \sum_{i=1}^n U_{ki,\varepsilon}^{(n)}(t).$$

All of these processes are square integrable martingales and, by Theorems 3.5, 3.6, and 3.7,

$$\langle U_k^{(n)}, U_k^{(n)}\rangle(t) = \sum_{i=1}^n \int_0^t \left[H_{ki}^{(n)}(s)\right]^2 dA_{ki}^{(n)}(s)$$

and

$$\langle U_{k,\varepsilon}^{(n)}, U_{k,\varepsilon}^{(n)}\rangle(t) = \sum_{i=1}^n \int_0^t \left[H_{ki}^{(n)}(s)\right]^2 \mathbb{1}(|H_{ki}^{(n)}(s)| > \varepsilon)\, dA_{ki}^{(n)}(s).$$

Theorem 3.10 (Central limit theorem I.). *Let for all $t \in \langle 0, \tau\rangle$ and all $k = 1, \dots, r$*

$$\langle U_k^{(n)}, U_k^{(n)}\rangle(t) \xrightarrow{\mathsf{P}} \int_0^t f_k^2(s)\, ds < \infty$$

as $n \to \infty$, where f_k are non-negative measurable functions, and, for all $\varepsilon > 0$,

$$\langle U_{k,\varepsilon}^{(n)}, U_{k,\varepsilon}^{(n)}\rangle(t) \xrightarrow{\mathsf{P}} 0 \tag{3.7}$$

* Boundedness is not a necessary condition, it can be relaxed to *local boundedness*.

as $n \to \infty$. Then

$$\left(U_1^{(n)}, U_2^{(n)}, \dots, U_r^{(n)}\right) \Longrightarrow \left(\int f_1\, dW_1, \int f_2\, dW_2, \dots, \int f_r\, dW_r\right) \text{ on } D^r\langle 0,\tau\rangle,$$

where $W_1, W_2, \dots, W_r$ are independent Brownian motions. ◇

Note.

- The processes $\int f_k\, dW_k$, $k = 1, \dots, r$, are independent time-transformed Brownian motions. See Appendix A.3.
- The symbol "$\Longrightarrow$" means weak convergence of a multivariate stochastic process in the space $D^r\langle 0,\tau\rangle$ of left-continuous functions with right-hand limits defined on the r-dimensional Carthesian product of $\langle 0,\tau\rangle$. See Appendix A.4.
- The condition (3.7) is analogous to the Feller-Lindeberg condition for sums of random variables. It can be shown that it is automatically satisfied when both sequences $N_{k1}^{(n)}, \dots, N_{kn}^{(n)}$ and $A_{k1}^{(n)}, \dots, A_{kn}^{(n)}$ are identically distributed for each k.

The most important consequence of Theorem 3.10 is that the random vector of values $(U_1^{(n)}, U_2^{(n)}, \dots, U_r^{(n)})$ evaluated at a single fixed time $t \in \langle 0,\tau\rangle$ converges in distribution to an r-dimensional normal random vector with zero mean, independent components and variances $\int_0^t f_k^2(s)\, ds$.

Central limit theorem, case 2

Now take a single set of counting processes with multiple integrands.

- Let $\{N_i^{(n)} : i = 1, \dots, n\}$ be a multivariate counting process with respect to the stochastic basis $(\Omega, \mathcal{A}, \{\mathcal{F}_t\}_{t\geq 0}, P)$.
- Let the compensator $A_i^{(n)}$ for $N_i^{(n)}$ be continuous.
- Let $H_{ki}^{(n)}$, $k = 1, \dots, r$, $i = 1, \dots, n$, be bounded* $\mathcal{F}_t$-predictable processes on the interval $\langle 0,\tau\rangle$.

Let $M_i^{(n)} = N_i^{(n)} - A_i^{(n)}$ be the $\mathcal{F}_t$-martingale for $N_i^{(n)}$. Denote

$$U_{ki}^{(n)}(t) = \int_0^t H_{ki}^{(n)}(s)\, dM_i^{(n)}(s) \quad \text{and} \quad U_k^{(n)}(t) = \sum_{i=1}^n U_{ki}^{(n)}(t).$$

Take any $\varepsilon > 0$ and denote

$$U_{ki,\varepsilon}^{(n)}(t) = \int_0^t H_{ki}^{(n)}(s)\mathbb{1}(|H_{ki}^{(n)}(s)| > \varepsilon)\, dM_i^{(n)}(s) \quad \text{and} \quad U_{k,\varepsilon}^{(n)}(t) = \sum_{i=1}^n U_{ki,\varepsilon}^{(n)}(t).$$

* Again, boundedness can be relaxed.

All of these processes are square integrable martingales and, by Theorems 3.5, 3.6, and 3.7,

$$\langle U_k^{(n)}, U_l^{(n)} \rangle(t) = \sum_{i=1}^{n} \int_0^t H_{ki}^{(n)}(s) H_{li}^{(n)}(s)\, dA_i^{(n)}(s)$$

and

$$\langle U_{k,\varepsilon}^{(n)}, U_{l,\varepsilon}^{(n)} \rangle(t) = \sum_{i=1}^{n} \int_0^t H_{ki}^{(n)}(s) H_{li}^{(n)}(s) \mathbb{1}(|H_{ki}^{(n)}(s)| > \varepsilon) \mathbb{1}(|H_{li}^{(n)}(s)| > \varepsilon)\, dA_i^{(n)}(s).$$

Theorem 3.11 (Central limit theorem II.). *Let for all $t \in \langle 0, \tau \rangle$ and all $k, l = 1, \ldots, r$*

$$\langle U_k^{(n)}, U_l^{(n)} \rangle(t) \xrightarrow{\mathsf{P}} c_{kl}(t) < \infty$$

as $n \to \infty$, where c_{kl} are continuous functions, and, for all $\varepsilon > 0$ and all $k = 1, \ldots, r$,

$$\langle U_{k,\varepsilon}^{(n)}, U_{k,\varepsilon}^{(n)} \rangle(t) \xrightarrow{\mathsf{P}} 0$$

as $n \to \infty$. Then

$$\left(U_1^{(n)}, U_2^{(n)}, \ldots, U_r^{(n)}\right) \Longrightarrow \left(W_1^*, \ldots, W_r^*\right) \text{ on } D^r \langle 0, \tau \rangle,$$

where $W_1^*, \ldots, W_r^*$ are dependent zero-mean Gaussian processes with independent increments, a.s. continuous sample paths, and covariance functions $\operatorname{cov}(W_k^*(s), W_l^*(t)) = c_{kl}(s)$ for all k, l and all $0 \le s \le t \le \tau$. ◊

By this theorem, the random vector $\left(U_1^{(n)}, U_2^{(n)}, \ldots, U_r^{(n)}\right)$ evaluated at a single fixed time $t \in \langle 0, \tau \rangle$ converges in distribution to an r-dimensional normal random vector with zero mean and covariance matrix $c_{kl}(t)$, $k, l \in 1, \ldots, r$.

The end of lecture 7 (Oct. 26)

4. Nonparametric Estimation of Failure Time Distribution

4.1. Estimating cumulative hazard function and survival function

Let $(T_1, C_1), \ldots, (T_n, C_n)$ be independent, let $T_1, \ldots, T_n$ be identically distributed with survival function S and cumulative hazard function Λ.

Let $X_i = T_i \wedge C_i$ be censored failure times and $\delta_i = \mathbb{1}(T_i \le C_i)$ failure indicators. We would like to estimate the survival function S and the cumulative hazard function Λ from the independent observations $(X_1, \delta_1), \ldots, (X_n, \delta_n)$ without making any assumptions on the distribution of T_i.

If the data were not censored, the survival function S could be estimated by $\widehat{S} = 1 - \widehat{F}$, where $\widehat{F}(t) = n^{-1} \sum_{i=1}^n \mathbb{1}(T_i \le t)$ is the empirical distribution function. So our task can be viewed as extending the empirical distribution function to censored data.

Consider the counting processes $N_i(t) = \mathbb{1}(T_i \le t, \delta_i = 1)$ and at-risk processes $Y_i(t) = \mathbb{1}(X_i \ge t)$, $i = 1, \ldots, n$. Take the extended filtration

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

and the compensator $A_i(t) = \int_0^t Y_i(u)\, d\Lambda(u)$. If the independent censoring condition (3.1) holds (we always assume this) for each pair (T_i, C_i) the process $M_i(t) = N_i(t) - A_i(t)$ is an $\mathcal{F}_t$-martingale (see Section 3.1 on p. 28).

Let $\overline{N}(t) = \sum_{i=1}^n N_i(t)$ and $\overline{Y}(t) = \sum_{i=1}^n Y_i(t)$. It follows that $\overline{M}(t) = \sum_{i=1}^n M_i(t) = \overline{N}(t) - \int_0^t \overline{Y}(u)\, d\Lambda(u)$ is an $\mathcal{F}_t$-martingale.

Denote by T_* the time when the data run out, i.e., $T_* = \inf\{s : \overline{Y}(s) = 0\}$. Take the bounded $\mathcal{F}_t$-predictable process

$$H(u) = \frac{\mathbb{1}(\overline{Y}(u) > 0)}{\overline{Y}(u)}.$$

For $u \ge T_*$, the numerator is 0 and the whole process is defined as 0. By Theorem 3.4, $\int H\, d\overline{M}$ is a martingale and its expectation is zero. Write

$$\int_0^t H(u)\, d\overline{M}(u) = \int_0^t \frac{\mathbb{1}(\overline{Y}(u) > 0)}{\overline{Y}(u)}\, d\overline{N}(u) - \int_0^t \mathbb{1}(\overline{Y}(u) > 0)\, d\Lambda(u) = \int_0^t \frac{d\overline{N}(u)}{\overline{Y}(u)} - \Lambda(t \wedge T_*).$$

The left-hand side has zero expectation. The random part of the right-hand side appears to be a good candidate for an unbiased estimator of $\Lambda(t)$ at times t when data are still observed.

Definition 4.1. The function

$$\widehat{\Lambda}(t) = \int_0^t \frac{d\overline{N}(u)}{\overline{Y}(u)}$$

is called the *Nelson-Aalen estimator* of the cumulative hazard function. ∇

Note.

- This estimator was proposed by Nelson (1969). Its consistency and weak convergence were first proven by Breslow and Crowley (1974) using standard methods and then by Aalen (1978) using martingale theory.
- The Nelson-Aalen estimator is constant for $t \geq T_*$. There is no information in the data about the hazard after the last observation fails or is censored.
- Denote by $t_1 < \cdots < t_d$ the ordered distinct failure times observed in the data. Then

$$\widehat{\Lambda}(t) = \sum_{\{j:t_j \leq t\}} \frac{\Delta\overline{N}(t_j)}{\overline{Y}(t_j)} = \sum_{\{j:t_j \leq t\}} \widehat{\lambda}_j.$$

  This is how the estimator is calculated. The contribution $\widehat{\lambda}_j$ is an empirical estimate of the discrete hazard at t_j: the ratio of the number of subjects who failed at t_j divided by the number of subjects who could have failed at t_j.

Having an estimator $\widehat{\Lambda}$ for Λ, we can use it to obtain an estimator for the survival function S. By equation (1.2), we have $S(t) = \mathrm{e}^{-\Lambda(t)}$ for continuous failure time distributions. So we could take

$$\widehat{S}(t) = \mathrm{e}^{-\widehat{\Lambda}(t)}.$$

This is called the *Fleming-Harrington estimator* of survival function. However, (1.2) only holds for continuous failure time distributions, which allow no ties among failure times. So let us use equality (1.1) instead, which is more universal.

We have

$$\Lambda(t) = \int_0^t \frac{dF(u)}{S(u-)}.$$

Hence, for any measurable function $g(t)$,

$$\int_0^t g(u)\, d\Lambda(u) = \int_0^t \frac{g(u)}{S(u-)}\, dF(u).$$

Take $g(u) = 1 - F(u-)$ to get

$$\int_0^t [1 - F(u-)]\, d\Lambda(u) = F(t) = 1 - S(t)$$

and finally,

$$S(t) = 1 - \int_0^t S(u-)\, d\Lambda(u).$$

Plug in the Nelson-Aalen estimator of Λ and obtain an estimator of the survival function $\widehat{S}$ that satisfies the equation

$$\widehat{S}(t) = 1 - \int_0^t \widehat{S}(u-)\, d\widehat{\Lambda}(u). \tag{4.1}$$

We can solve this equation recursively as follows:

$$\widehat{S}(t) = 1 - \int_0^t \widehat{S}(u-)\frac{d\overline{N}(u)}{\overline{Y}(u)},$$
$$\widehat{S}(t) = 1 - \sum_{u \le t} \widehat{S}(u-)\frac{\Delta\overline{N}(u)}{\overline{Y}(u)}.$$

We can see that $\widehat{S}(t)$ is a step function because it can only change at the observed failure times. Let us calculate the size of the jump of $\widehat{S}(t)$ at any t.

$$\widehat{S}(t-) - \widehat{S}(t) = -\Delta\widehat{S}(t) = \widehat{S}(t-)\frac{\Delta\overline{N}(t)}{\overline{Y}(t)}.$$

Hence,

$$\widehat{S}(t) = \widehat{S}(t-)\left[1 - \frac{\Delta\overline{N}(t)}{\overline{Y}(t)}\right].$$

This is the recursive equation that allows us to subsequently calculate all values of $\widehat{S}(t)$. It is easy to see that it can be solved to get the following definition.

Definition 4.2. The function

$$\widehat{S}(t) = \prod_{u \le t}\left[1 - \frac{\Delta\overline{N}(u)}{\overline{Y}(u)}\right]$$

is called the *Kaplan-Meier estimator* of survival function. ∇

Note.

- This estimator was first proposed by Kaplan and Meier (1958).
- With $t_1 < \cdots < t_d$ the ordered distinct failure times,

$$\widehat{S}(t) = \prod_{\{j:t_j \le t\}}\left[1 - \frac{\Delta\overline{N}(t_j)}{\overline{Y}(t_j)}\right] = \prod_{\{j:t_j \le t\}}(1 - \widehat{\lambda}_j).$$

  The last expression agrees with equation (1.3) for discrete hazard functions.

- The Kaplan-Meier estimator is a right-continuous piecewise constant function. When the data are not censored, $1-\widehat{S}$ equals the empirical distribution function.
- The Kaplan-Meier estimator is constant for $t \geq T_*$. It does not drop to zero at the last observed failure time t_d unless all the remaining subjects fail at that time.

Homework 4:

Prove that, with no censoring, the Kaplan-Meier estimator coincides with $1-\widehat{F}$, where $\widehat{F}$ is the empirical distribution function.

The end of lecture 8 (Oct. 26)

4.2. Properties of the Nelson-Aalen estimator

Select $\tau > 0$ a fixed time such that $\mathrm{P}[Y_i(\tau) = 1] > \delta > 0$ for all $i = 1, \dots, n$ and $\Lambda(\tau) < \infty$. The properties of nonparametric estimators will be investigated on the fixed interval $\langle 0, \tau\rangle$ because our theoretical tools (in particular the central limit theorem) require that.

In practice, we perform the analysis on the random interval $\langle 0, T_*\rangle$, where $T_* = \inf\{s : \overline{Y}(s) = 0\}$ is the time when the last observation fails or is censored. The proofs could be extended to this more general case but they would become much more complicated.

Assumptions. There exists a deterministic function $\pi : \langle 0, \tau\rangle \longrightarrow (0, 1\rangle$ such that

$$\sup_{t \in \langle 0, \tau\rangle} \left| \frac{1}{n} \overline{Y}(t) - \pi(t) \right| \xrightarrow{\mathrm{P}} 0. \tag{4.2}$$

If the censoring times $C_1, \dots, C_n$ are identically distributed (hence $N_i(t), Y_i(t)$ are), then condition (4.2) is satisfied with $\pi(t) = \mathrm{P}[Y_i(t) = 1]$, which is positive on $\langle 0, \tau\rangle$.

Lemma 4.1. *If the data are independent and identically distributed, then condition (4.2) is satisfied.* ◊

Proof. By the weak law of large numbers for iid random variables, $\overline{Y}(t)/n \xrightarrow{\mathrm{P}} \pi(t) \equiv \mathrm{P}[Y_i(t) = 1]$ at every $t \in \langle 0, \tau\rangle$. Uniformity follows from the Glivenko-Cantelli Theorem on uniform convergence of the empirical distribution function. □

As shown previously, the Nelson-Aalen estimator $\widehat{\Lambda}(t) = \int_0^t \frac{d\overline{N}(u)}{\overline{Y}(u)}$ can be written as

$$\widehat{\Lambda}(t) = \Lambda^*(t) + \int_0^t H(u)\, d\overline{M}(u),$$

where $H(u) = \mathbb{1}(\overline{Y}(u) > 0)/\overline{Y}(u)$ is predictable and $\Lambda^*(t) = \Lambda(t \wedge T_*)$. This martingale representation together with the results of Chapter 3 allows us to prove the following theorem that summarizes the important properties of the Nelson-Aalen estimator.

Theorem 4.2.

(i) For any $t \in \langle 0, \tau \rangle$, $\mathsf{E}\left[\widehat{\Lambda}(t) - \Lambda^*(t)\right] = 0$.

(ii) For any $t \in \langle 0, \tau \rangle$, $0 \geq \mathsf{E}\left[\widehat{\Lambda}(t) - \Lambda(t)\right] = -\int_0^t \mathsf{P}\left[\overline{Y}(u) = 0\right] d\Lambda(u)$. For identically distributed data, $\left|\mathsf{E}\,[\widehat{\Lambda}(t) - \Lambda(t)]\right| \leq [1 - \pi(t)]^n \Lambda(t) \to 0$ as $n \to \infty$.

(iii) If Λ is continuous with hazard function λ,

$$\sqrt{n}\left[\widehat{\Lambda}(t) - \Lambda(t)\right] \Longrightarrow \int_0^t \sqrt{\frac{\lambda(u)}{\pi(u)}}\, dW(u) \quad \text{on } D\langle 0, \tau \rangle.$$

$\diamond$

Proof. (i)

$$\widehat{\Lambda}(t) - \Lambda^*(t) = \int_0^t H(u)\, d\overline{M}(u),$$

where $H(u) = \mathbb{1}(\overline{Y}(u) > 0)/\overline{Y}(u)$ is bounded and predictable. By Theorem 3.4, $\widehat{\Lambda}(t) - \Lambda^*(t)$ is a martingale a thus it has zero expectation at all $t \in \langle 0, \tau \rangle$.

(ii)

Homework 5:

Prove this part of Theorem 4.2.

Solution:

$$\begin{aligned}\mathsf{E}\,[\widehat{\Lambda}(t) - \Lambda(t)] = \mathsf{E}\,[\Lambda^*(t) - \Lambda(t)] = -\mathsf{E} \int_0^t \mathbb{1}(\overline{Y}(u) = 0)\, d\Lambda(u) \\ = -\int_0^t \mathsf{P}\left[\overline{Y}(u) = 0\right] d\Lambda(u).\end{aligned}$$

Thus, there is a negative bias in $\widehat{\Lambda}(t)$ that starts to appear after the data have run out. When the data are iid, $\mathsf{P}\left[\overline{Y}(u) = 0\right] = \mathsf{P}\,[Y_1(u) = Y_2(u) = \cdots = Y_n(u) = 0] = [1 - \pi(u)]^n$. Hence,

$$\left|\mathsf{E}\,[\widehat{\Lambda}(t) - \Lambda(t)]\right| = \int_0^t [1 - \pi(u)]^n\, d\Lambda(u) \leq [1 - \pi(t)]^n \Lambda(t)$$

and the right-hand side converges to 0 as $n \to \infty$.

(iii) Now we assume that T has a continuous failure time distribution with hazard function $\lambda(t)$. The proof of weak convergence uses Theorem 3.10, the first central limit theorem for martingale integrals, with $r = 1$ group (the subscript k is dropped). We take

$$U_i^{(n)}(t) = \int_0^t \sqrt{n} H(u)\, dM_i(u)$$

with $H(u) = \mathbb{1}(\overline{Y}(u) > 0)/\overline{Y}(u)$ bounded and predictable and

$$U^{(n)}(t) = \sum_{i=1}^n U_i^{(n)}(t) = \int_0^t \sqrt{n} H(u)\, d\overline{M}(u) = \sqrt{n}\big[\widehat{\Lambda}(t) - \Lambda^*(t)\big].$$

Let us verify the conditions of Theorem 3.10. First, by Theorems 3.5, 3.6, and 3.7 or 3.9,

$$\begin{aligned}\langle U^{(n)}, U^{(n)}\rangle(t) &= \int_0^t nH^2(u)\overline{Y}(u)\, d\Lambda(u) = \int_0^t \frac{\mathbb{1}(\overline{Y}(u) > 0)}{\overline{Y}(u)/n}\, d\Lambda(u) \\ &= \int_0^t \frac{1}{\pi(u)}\, d\Lambda(u) + \int_0^t \left[\frac{\mathbb{1}(\overline{Y}(u) > 0)}{\overline{Y}(u)/n} - \frac{1}{\pi(u)}\right] d\Lambda(u).\end{aligned}$$

Denote the second summand A_n. If we show that $A_n \xrightarrow{\mathrm{P}} 0$ we will have proven the first confition,

$$\langle U^{(n)}, U^{(n)}\rangle(t) \xrightarrow{\mathrm{P}} \int_0^t f^2(u)\, du$$

with $f(u) = \sqrt{\lambda(u)/\pi(u)}$. By the Cauchy-Schwartz inequality,

$$|A_n| \le \int_0^t \left|\frac{n}{\overline{Y}(u)} - \frac{1}{\pi(u)}\right| \mathbb{1}(\overline{Y}(u) > 0)\, d\Lambda(u) + \int_0^t \mathbb{1}(\overline{Y}(u) = 0)\frac{1}{\pi(u)}\, d\Lambda(u).$$

The first term can be bounded above by

$$\sup_{0<u<t} \left|\frac{n}{\overline{Y}(u)} - \frac{1}{\pi(u)}\right| \Lambda(t)$$

and the supremum converges to zero by uniform convergence of $\overline{Y}(u)/n$ to $\pi(u)$ and the fact that $\pi(u)$ is bounded away from zero. The second term is bounded above by

$$\mathbb{1}(\overline{Y}(\tau) = 0) \int_0^t \frac{1}{\pi(u)}\, d\Lambda(u).$$

Let us show that $\mathbb{1}(\overline{Y}(\tau) = 0) \xrightarrow{\mathrm{P}} 0$. Take any $\varepsilon > 0$.

$$\mathrm{P}\left[\mathbb{1}(\overline{Y}(\tau) = 0) > \varepsilon\right] = \mathrm{P}\left[\overline{Y}(\tau) = 0\right] \to 0$$

because $\mathrm{P}[Y_i(\tau) = 1] > \delta > 0$ for all i. This shows that the first condition of Theorem 3.10 is satisfied.

Turn our attention to the second condition (Feller-Lindebergh-type). Choose any $\varepsilon > 0$ and consider the process

$$U_\varepsilon^{(n)}(t) = \int_0^t \sqrt{n}H(u)\mathbb{1}(\sqrt{n}H(u) > \varepsilon)\, d\overline{M}(u).$$

Its predictable variation process is

$$\langle U_\varepsilon^{(n)}, U_\varepsilon^{(n)}\rangle(t) = \int_0^t nH^2(u)\overline{Y}(u)\mathbb{1}(\sqrt{n}H(u) > \varepsilon)\, d\Lambda(u).$$

The integrand is zero whenever $\overline{Y}(u) = 0$. Otherwise, the process is

$$\int_0^t \frac{n}{\overline{Y}(u)}\mathbb{1}(\overline{Y}(u) < \sqrt{n}/\varepsilon)\, d\Lambda(u) = \int_0^t \frac{1}{\pi(u)}\mathbb{1}\left(\frac{\overline{Y}(u)}{n} < \frac{1}{\sqrt{n}\varepsilon}\right) d\Lambda(u) + o_P(1),$$

where the replacement of $n/\overline{Y}$ by $1/\pi$ is justified by the same argument as earlier and $o_P(1)$ is a term converging in probability to zero as $n \to \infty$. The integral is bounded above by

$$\frac{1}{\pi(\tau)}\Lambda(\tau)\mathbb{1}\left(\frac{\overline{Y}(\tau)}{n} < \frac{1}{\sqrt{n}\varepsilon}\right)$$

and this converges to 0 in probability because $\frac{\overline{Y}(\tau)}{n} \xrightarrow{\mathrm{P}} \pi(\tau) > 0$ as $n \to \infty$.

The conditions of Theorem 3.10 have been verified. Hence

$$\sqrt{n}\big[\widehat{\Lambda}(t) - \Lambda^*(t)\big] \Longrightarrow \int f(u)\, dW(u)$$

with $f(u) = \sqrt{\lambda(u)/\pi(u)}$.

Last, we show that $\sqrt{n}\big[\Lambda^*(t) - \Lambda(t)\big] \xrightarrow{\mathrm{P}} 0$. This is true because

$$\sqrt{n}\big[\Lambda^*(t) - \Lambda(t)\big] = \sqrt{n}\big[\Lambda(T_*) - \Lambda(t)\big]\mathbb{1}(T_* \le t)$$

and

$$\mathrm{P}\left[\sqrt{n}\mathbb{1}(T_* \le t) > \varepsilon\right] = \mathrm{P}\left[\sqrt{n}\mathbb{1}(\overline{Y}(t) = 0) > \varepsilon\right] = \mathrm{P}\left[\overline{Y}(t) = 0\right] \to 0$$

for any $t \in \langle 0, \tau\rangle$ and any $\varepsilon > 0$. This completes the proof. □

Note.

- Part (iii) of Theorem 4.2 implies that for any fixed $t \in \langle 0, \tau\rangle$, $\sqrt{n}\Big[\widehat{\Lambda}(t) - \Lambda(t)\Big]$ has asymptotically normal distribution with zero mean and variance

$$\int_0^t \frac{\lambda(u)}{\pi(u)}\, du.$$

- Part (iii) of Theorem 4.2 implies uniform consistency, i.e.

$$\sup_{t\in\langle 0,\tau\rangle}\left|\widehat{\Lambda}(t)-\Lambda(t)\right| \xrightarrow{\mathsf{P}} 0 \quad \text{as } n\to\infty.$$

- Asymptotic normality and consistency of $\widehat{\Lambda}$ also hold for discrete failure time distributions. We presented a proof for continuous distributions, which is actually the more difficult case than with discrete distributions.

The next theorem introduces a variance estimator for $\widehat{\Lambda}(t)$ and establishes its consistency and unbiasedness.

Theorem 4.3.

(i) For any finite n,

$$\sigma^2_{\Lambda}(t) \equiv \mathsf{var}\left[\widehat{\Lambda}(t)-\Lambda^*(t)\right] = \int_0^t \mathsf{E}\,H(u)\,[1-\Delta\Lambda(u)]\,d\Lambda(u).$$

As $n\to\infty$, $n\sigma^2_{\Lambda}(t)\xrightarrow{\mathsf{P}}\int_0^t \frac{1-\Delta\Lambda(u)}{\pi(u)}\,d\Lambda(u)$.

(ii) Define

$$S^2_{\Lambda}(t) = \int_0^t \frac{1}{\overline{Y}^2(u)}\left[1-\frac{\Delta\overline{N}(u)-1}{\overline{Y}(u)-1}\right]d\overline{N}(u).$$

Then

$$nS^2_{\Lambda}(t)\xrightarrow{\mathsf{P}}\int_0^t \frac{1}{\pi(u)}[1-\Delta\Lambda(u)]\,d\Lambda(u).$$

(iii)

$$\mathsf{E}\left[S^2_{\Lambda}(t)-\sigma^2_{\Lambda}(t)\right] = \int_0^t \mathsf{P}\left[\overline{Y}(u)=1\right]\Delta\Lambda(u)\,d\Lambda(u),$$

which is zero if the failure time distribution is continuous. ◊

Note. For continuous distributions, $\overline{N}$ cannot jump by more than one and the estimator can be simplified to:

$$S^2_{\Lambda}(t) = \int_0^t \frac{1}{\overline{Y}^2(u)}\,d\overline{N}(u).$$

Proof.

(i) By Theorems 3.5, 3.6, and 3.9,

$$\begin{aligned}\mathsf{var}\int_0^t H(u)\,d\overline{M}(u) &= \sum_{i=1}^n \mathsf{E}\left[H^2(u)Y_i(u)\right][1-\Delta\Lambda(u)]\,d\Lambda(u)\\ &= \int_0^t \mathsf{E}\left[H^2(u)\overline{Y}(u)\right][1-\Delta\Lambda(u)]\,d\Lambda(u) = \int_0^t \mathsf{E}\,H(u)[1-\Delta\Lambda(u)]\,d\Lambda(u).\end{aligned}$$

(ii) We only present the proof for continuous failure time distributions.

$$nS_\Lambda^2(t) = \int_0^t \frac{n}{\overline{Y}^2(u)}\,d\overline{N}(u) = \int_0^t \frac{n}{\overline{Y}(u)}\,d\widehat{\Lambda}(u) \xrightarrow{\mathsf{P}} \int_0^t \frac{1}{\pi(u)}\,d\Lambda(u)$$

because $n/\overline{Y}(u) \xrightarrow{\mathsf{P}} 1/\pi(u)$ and $\widehat{\Lambda}(u) \xrightarrow{\mathsf{P}} \Lambda(u)$, both uniformly in time.

(iii) This part is also done only for continuous failure time distributions.

$$\begin{aligned}\mathsf{E}\left[S_\Lambda^2(t) - \sigma_\Lambda^2(t)\right] &= \mathsf{E}\int_0^t \frac{H(u)}{\overline{Y}(u)}\,d\overline{N}(u) - \mathsf{E}\int_0^t \frac{H(u)}{\overline{Y}(u)}\overline{Y}(u)\,d\Lambda(u)\\ &= \mathsf{E}\int_0^t \frac{H(u)}{\overline{Y}(u)}\,d\overline{M}(u) = 0.\end{aligned}$$

□

The end of lecture 9 (Nov. 2)

4.3. Properties of the Kaplan-Meier estimator

Let us return to the Kaplan-Meier estimator of survival function

$$\widehat{S}(t) = \prod_{u\le t}\left[1 - \frac{\Delta\overline{N}(u)}{\overline{Y}(u)}\right].$$

Because the Kaplan-Meier estimator has the form of a product rather than a sum, its properties need to be investigated as ratios rather than differences. The key statements are formulated by the following two lemmas.

Lemma 4.4. *At all $t \ge 0$ such that $S(t) > 0$, it holds*

$$\frac{\widehat{S}(t)}{S(t)} = 1 - \int_0^t \frac{\widehat{S}(u-)}{S(u)}\,d(\widehat{\Lambda} - \Lambda)(u).$$

◊

Proof. We start with integration by parts on the product $\widehat{S}$ and $1/S$.

$$\frac{\widehat{S}(t)}{S(t)} - \frac{\widehat{S}(0)}{S(0)} = \int_0^t \widehat{S}(u-)\,d\Big(\frac{1}{S}\Big)(u) + \int_0^t \frac{1}{S(u)}\,d\widehat{S}(u). \qquad (*)$$

We know that $\widehat{S}(0) = S(0) = 1$. This step does not seem to help because we do not know how to calculate $\int \cdot\, d(1/S)$ but this is solved in the next step by another application of integration by parts. This time we apply it on the product $S \cdot 1/S$ over the interval $(v, t\rangle$ for $0 \le v < t$. We get

$$0 = \frac{S(t)}{S(t)} - \frac{S(v)}{S(v)} = \int_v^t S(u-)\,d\Big(\frac{1}{S}\Big)(u) + \int_v^t \frac{1}{S(u)}\,dS(u). \qquad (**)$$

We use this equality to argue that, for any measurable function h,

$$\int_0^t h(u)S(u-)\,d\Big(\frac{1}{S}\Big)(u) = -\int_0^t \frac{h(u)}{S(u)}\,dS(u). \tag{†}$$

This can be shown by a technique known from measure theory. First, show the validity of (†) for $h(u) = \mathbb{1}(u \in (v, t\rangle)$. But this is already done by (**). Next, show that it still holds for h being a simple function of the form $h(u) = \sum_{l=1}^{L} \mathbb{1}(u \in (v_l, t_l\rangle)$ for disjoint intervals $(v_l, t_l\rangle \subset (0, t\rangle$. This is trivial. Last, we take a sequence of simple functions h_L such that $h_L \nearrow h$ for $L \to \infty$ and use the monotone convergence theorem to justify the equality (†) in the limit.

Now we use (†) with $h(u) = \widehat{S}(u-)/S(u-)$, which is measurable for almost any trajectory of $\widehat{S}$. We get

$$\begin{aligned}\int_0^t \frac{\widehat{S}(u-)}{S(u-)}S(u-)\,d\Big(\frac{1}{S}\Big)(u) &= -\int_0^t \frac{\widehat{S}(u-)}{S(u-)}\frac{1}{S(u)}\,dS(u)\\ &= \int_0^t \frac{\widehat{S}(u-)}{S(u)}\frac{d(1-S(u))}{S(u-)} = \int_0^t \frac{\widehat{S}(u-)}{S(u)}d\Lambda(u).\end{aligned} \tag{4.3}$$

Now we are done with the first term on the right-hand side of (*). The second term is easier. We use equation (4.1)

$$\widehat{S}(t) = 1 - \int_0^t \widehat{S}(u-)\,d\widehat{\Lambda}(u)$$

to get

$$\int_0^t \frac{1}{S(u)}\,d\widehat{S}(u) = -\int_0^t \frac{\widehat{S}(u-)}{S(u)}\,d\widehat{\Lambda}(u).$$

Putting the last two results together and plugging them into (*), we get

$$\frac{\widehat{S}(t)}{S(t)} = 1 + \int_0^t \frac{\widehat{S}(u-)}{S(u)}d\Lambda(u) - \int_0^t \frac{\widehat{S}(u-)}{S(u)}\,d\widehat{\Lambda}(u) = 1 - \int_0^t \frac{\widehat{S}(u-)}{S(u)}\,d\big[\widehat{\Lambda}(u) - \Lambda(u)\big]. \qquad \square$$

Lemma 4.5. *At all $t \geq 0$ such that $S(t) > 0$, it holds*

$$\frac{\widehat{S}(t) - S(t)}{S(t)} = -\int_0^t H(u)\,d\overline{M}(u) + B(t),$$

where

$$H(u) = \frac{\widehat{S}(u-)}{S(u)}\frac{\mathbb{1}(\overline{Y}(u) > 0)}{\overline{Y}(u)}$$

is a predictable process and

$$B(t) = \frac{\widehat{S}(T_*)}{S(T_*)}\frac{S(T_*) - S(t)}{S(t)}\mathbb{1}(T_* < t). \qquad \diamond$$

Proof. By Lemma 4.4,

$$\begin{aligned}\frac{\widehat{S}(t)-S(t)}{S(t)} &= -\int_0^t \frac{\widehat{S}(u-)}{S(u)}\left\{\frac{d\overline{N}(u)}{\overline{Y}(u)} - \mathbb{1}(\overline{Y}(u)>0)\frac{\overline{Y}(u)}{\overline{Y}(u)}\,d\Lambda(u) - \mathbb{1}(\overline{Y}(u)=0)\,d\Lambda(u)\right\}\\ &= -\int_0^t \frac{\widehat{S}(u-)}{S(u)}\frac{\mathbb{1}(\overline{Y}(u)>0)}{\overline{Y}(u)}\,d\overline{M}(u) + \int_0^t \frac{\widehat{S}(u-)}{S(u)}\mathbb{1}(\overline{Y}(u)=0)\,d\Lambda(u)\\ &= -\int_0^t H(u)\,d\overline{M}(u) + B(t).\end{aligned}$$

It remains to rewrite the bias term $B(t)$. Notice that $B(t)=0$ as long as $\overline{Y}(t)>0$ or $t \le T_*$. We use equation (4.3) derived during the proof of Lemma 4.4

$$\int_0^t \widehat{S}(u-)\,d\Big(\frac{1}{S}\Big)(u) = \int_0^t \frac{\widehat{S}(u-)}{S(u)}d\Lambda(u)$$

together with the fact that $\widehat{S}(u-)=\widehat{S}(T_*)$ for $u>T_*$ to get

$$\begin{aligned}B(t) &= \int_{T_*}^t \frac{\widehat{S}(u-)}{S(u)}\,d\Lambda(u) = \int_{T_*}^t \widehat{S}(u-)\,d\Big(\frac{1}{S}\Big)(u)\\ &= \mathbb{1}(T_*<t)\widehat{S}(T_*)\int_{T_*}^t d\Big(\frac{1}{S}\Big)(u)\\ &= \mathbb{1}(T_*<t)\widehat{S}(T_*)\left[\frac{1}{S(t)}-\frac{1}{S(T_*)}\right] = \frac{\widehat{S}(T_*)}{S(T_*)}\frac{S(T_*)-S(t)}{S(t)}\mathbb{1}(T_*<t).\end{aligned}$$ □

Lemma 4.5 states that the relative error in the Kaplan-Meier estimator can be expressed as a martingale integral plus a bias term. The bias term $B(t)$ converges to zero in probability for $t \le \tau$ because $\mathrm{P}[T_*<t]\to 0$.

The next theorem specifies the first two moments of the Kaplan-Meier estimator.

Theorem 4.6. *At all $t \ge 0$ such that $S(t)>0$, it holds*

(i)

$$\mathsf{E}\,\widehat{S}(t) = S(t) + \mathsf{E}\,\mathbb{1}(T_*<t)\frac{\widehat{S}(T_*)}{S(T_*)}\big[S(T_*)-S(t)\big] \ge S(t)$$

(ii)

$$\mathsf{E}\,\widehat{S}(t) - S(t) \le [1-S(t)]\,\mathrm{P}[T_*<t] \to 0 \quad \textit{as } n\to\infty \textit{ for } t\le\tau.$$

(iii)

$$\begin{aligned}\mathsf{var}\big[\widehat{S}(t)-S(t)B(t)\big] &= S^2(t)\int_0^t \mathsf{E}\,\frac{\widehat{S}^2(u-)}{S^2(u)}\frac{\mathbb{1}(\overline{Y}(u)>0)}{\overline{Y}(u)}\big[1-\Delta\Lambda(u)\big]\,d\Lambda(u)\\ &= \mathsf{var}\,\widehat{S}(t) + o(1).\end{aligned}$$ ◊

Proof.

(i) By Lemma 4.5,

$$\frac{\mathrm{E}\,\widehat{S}(t)}{S(t)} - 1 = -\mathrm{E}\int_0^t H(u)\,d\overline{M}(u) + \mathrm{E}\,B(t) = \mathrm{E}\,B(t).$$

Hence

$$\mathrm{E}\,\widehat{S}(t) = S(t) + S(t)\mathrm{E}\,B(t) \geq S(t).$$

(ii)

$$\mathrm{E}\,\widehat{S}(t) - S(t) = S(t)\mathrm{E}\,B(t) = \mathrm{E}\,\widehat{S}(T_*)\left[1 - \frac{S(t)}{S(T_*)}\right]\mathbb{1}(T_* < t).$$

Because $\widehat{S}(T_*) \leq 1$ and $S(T_*) \leq 1$,

$$\mathrm{E}\,\widehat{S}(t) - S(t) \leq [1 - S(t)]\mathrm{E}\,\mathbb{1}(T_* < t) = [1 - S(t)]\mathrm{P}\,[T_* < t]\,.$$

(iii) By Lemma 4.5,

$$\begin{aligned}
\mathrm{var}\,[\widehat{S}(t) - B(t)S(t)] &= S^2(t)\,\mathrm{var}\int_0^t H(u)\,d\overline{M}(u) \\
&= S^2(t)\,\mathrm{E}\int_0^t H^2(u)\overline{Y}(u)[1 - \Delta\Lambda(u)]\,d\Lambda(u) \\
&= S^2(t)\int_0^t \mathrm{E}\,\frac{\widehat{S}^2(u-)}{S^2(u)}\frac{\mathbb{1}(\overline{Y}(u) > 0)}{\overline{Y}(u)}\big[1 - \Delta\Lambda(u)\big]\,d\Lambda(u).
\end{aligned}$$

For $t \leq \tau$, the term $B(t) \xrightarrow{\mathrm{P}} 0$ as $n \to \infty$, hence its variance is asymptotically negligible and $\mathrm{var}\,\widehat{S}(t) \approx \mathrm{var}\,[\widehat{S}(t) - B(t)S(t)]$. □

Note. An estimator $\widehat{V}(t)$ for $\mathrm{var}\,\sqrt{n}[\widehat{S}(t) - S(t)]$ can be obtained from item (iii) of the previous theorem by replacing S with $\widehat{S}$ and Λ with $\widehat{\Lambda}$.

Definition 4.3. The estimator

$$\widehat{V}(t) = n\widehat{S}^2(t)\int_0^t \frac{d\overline{N}(u)}{[\overline{Y}(u) - \Delta\overline{N}(u)]\overline{Y}(u)} \equiv \widehat{S}^2(t)\widehat{\sigma}(t). \tag{4.4}$$

for $\mathrm{var}\,\sqrt{n}[\widehat{S}(t) - S(t)]$ is called *the Greenwood formula.* ∇

Homework 6:

Derive the Greenwood formula.

The following proposition states the uniform consistency of the Kaplan-Meier estimator for continuous failure time distributions. Its proof is relatively complicated and so is omitted.

Proposition 4.7. *Let the observations be independent and identically distributed and Λ be continuous. Then*

$$\sup_{0 \le t \le \tau} \left| \widehat{S}(t) - S(t) \right| \xrightarrow{\mathsf{P}} 0.$$

◇

We proceed to claim weak convergence of the normalized Kaplan-Meier estimator to a zero-mean Gaussian process. Again, we only state and proof this result for continuous failure time distributions.

Theorem 4.8. *Let the observations be independent and identically distributed and Λ be continuous and differentiable almost everywhere with hazard function $\Lambda' = \lambda$. Denote $\sigma(t) = \int_0^t \pi^{-1}(u)\lambda(u)\,du$. Then*

(i)

$$\sqrt{n}\big[\widehat{S}(t) - S(t)\big] \Longrightarrow S(t)W(\sigma(t)) \quad \text{on } D\langle 0, \tau\rangle,$$

where $W(\sigma(t)) = \int_0^t \sqrt{\frac{\lambda(u)}{\pi(u)}}\,dW(u)$.

(ii)

$$\sqrt{n}\,\frac{\widehat{S}(t) - S(t)}{\widehat{S}(t)} \Longrightarrow W(\sigma(t)) \quad \text{on } D\langle 0, \tau\rangle.$$

◇

This theorem was first proven by Breslow and Crowley (1974) under somewhat stronger conditions.

Corollary. For a fixed time t in $\langle 0, \tau\rangle$, $\sqrt{n}[\widehat{S}(t) - S(t)] \xrightarrow{\mathsf{D}} \mathsf{N}(0, V(t))$, where

$$V(t) = S^2(t)\sigma(t) = S^2(t)\int_0^t \pi^{-1}(u)\lambda(u)\,du.$$

Note. The limiting variance $\sigma(t)$ in part (ii) is the same as that for the Nelson-Aalen estimator.

Note. The Greenwood formula $\widehat{V}(t)$ introduced in (4.4) is a uniformly consistent estimator for $V(t) = S^2(t)\sigma(t)$ on $\langle 0, \tau\rangle$.

Proof (of Theorem 4.8).

(i) By Lemma 4.5,

$$\sqrt{n}[\widehat{S}(t) - S(t)] = -S(t)\int_0^t \sqrt{n}H(u)\,d\overline{M}(u) + \sqrt{n}S(t)B(t),$$

where $\sup_{t\in\langle 0,\tau\rangle}\sqrt{n}S(t)B(t)\xrightarrow{\mathsf{P}}0$ because, by Theorem 4.6, part (ii), $0<\sqrt{n}B(t)\le\sqrt{n}[1-S(t)]\mathbb{1}(T_*<t)$ and $\sup_{0<t\le\tau}\sqrt{n}\mathbb{1}(T_*<t)\xrightarrow{\mathsf{P}}0$.

It suffices to investigate the weak convergence of $U^{(n)}(t)=\int_0^t\sqrt{n}H(u)\,d\overline{M}(u)$ with

$$H(u)=\frac{\widehat{S}(u-)}{S(u)}\,\frac{\mathbb{1}(\overline{Y}(u)>0)}{\overline{Y}(u)}$$

bounded and predictable on $\langle 0,\tau\rangle$.

As in the proof of weak convergence of the Nelson-Aalen estimator, we will use Theorem 3.10 with $r=1$ group. Let us verify the conditions of the theorem. First, by Theorems 3.5, 3.6, and 3.7 or 3.9,

$$\begin{aligned}\langle U^{(n)},U^{(n)}\rangle(t)&=\int_0^t nH^2(u)\overline{Y}(u)\,d\Lambda(u)\\&=\int_0^t\frac{\widehat{S}^2(u-)}{S^2(u)}\,\frac{\mathbb{1}(\overline{Y}(u)>0)}{\overline{Y}(u)/n}\,d\Lambda(u)\\&\xrightarrow{\mathsf{P}}\int_0^t\frac{\lambda(u)}{\pi(u)}\,du\equiv\sigma(t).\end{aligned}$$

because S is continuous, $\widehat{S}\xrightarrow{\mathsf{P}}S$ uniformly in $\langle 0,\tau\rangle$ according to Proposition 4.7, and $\overline{Y}(u)/n\xrightarrow{\mathsf{P}}\pi(u)$ also uniformly, by assumption (4.2).

Next,

$$\langle U_\varepsilon^{(n)},U_\varepsilon^{(n)}\rangle(t)=\int_0^t nH^2(u)\overline{Y}(u)\mathbb{1}(\sqrt{n}H(u)>\varepsilon)\,d\Lambda(u)\xrightarrow{\mathsf{P}}0,$$

because $\sup_{t\in\langle 0,\tau\rangle}\sqrt{n}H(t)\xrightarrow{\mathsf{P}}0$ (details are omitted). By Theorem 3.10,

$$\sqrt{n}\big[\widehat{S}(t)-S(t)\big]\Longrightarrow S(t)W(\sigma(t))\quad\text{on }D\langle 0,\tau\rangle,$$

where $W(\sigma(t))=\int_0^t\sqrt{\frac{\lambda(u)}{\pi(u)}}\,dW(u)$.

The negative sign that we encountered at the beginning of this proof does not matter because the limiting process has the same distribution as its negative.

(ii) Since $S(t)/\widehat{S}(t)\xrightarrow{\mathsf{P}}1$ uniformly on $\langle 0,\tau\rangle$ (Proposition 4.7),

$$\sqrt{n}\frac{\widehat{S}(t)-S(t)}{\widehat{S}(t)}\Longrightarrow W(\sigma(t))\quad\text{on }D\langle 0,\tau\rangle.$$

□

The end of lecture 10 (Nov. 9)

4.4. Confidence bounds for the survival function

It is easy to contruct pointwise confidence intervals for $S(t)$ at a fixed $t \in \langle 0, \tau \rangle$. Based on corollary to Theorem 4.8 and using the Greenwood formula, we get

$$\mathrm{P}\left[\widehat{S}(t) - u_{1-\alpha/2}\sqrt{\frac{\widehat{V}(t)}{n}} < S(t) < \widehat{S}(t) + u_{1-\alpha/2}\sqrt{\frac{\widehat{V}(t)}{n}}\right] \to 1 - \alpha.$$

The lower and upper bounds of a confidence interval for $S(t)$ with asymptotic coverage probability $1 - \alpha$ are

$$\widehat{S}(t)\left(1 - u_{1-\alpha/2}\sqrt{\frac{\widehat{\sigma}(t)}{n}}\right) \quad \text{and} \quad \widehat{S}(t)\left(1 + u_{1-\alpha/2}\sqrt{\frac{\widehat{\sigma}(t)}{n}}\right),$$

respectively.

Let us turn our attention to confidence bounds that cover the whole curve with the desired probability, not just at one point. We are looking for random functions $C_L(t)$ and $C_U(t)$ calculated from the data such that

$$\mathrm{P}\left[C_L(t) < S(t) < C_U(t) \text{ for all } t \in \langle 0, \tau \rangle\right] \to 1 - \alpha.$$

The following lemma is based on Theorem 4.8, point (ii), and the continuous mapping theorem for weak convergence.

Lemma 4.9. *Under the conditions of Theorem 4.8,*

$$\sqrt{\frac{n}{\widehat{\sigma}(\tau)}} \sup_{t \in \langle 0, \tau \rangle} \frac{1}{\widehat{S}(t)} \left|\widehat{S}(t) - S(t)\right| \xrightarrow{\mathrm{D}} \sup_{0 \le u \le 1} |W(u)|. \qquad \diamond$$

Proof. According to Theorem 4.8, part (ii),

$$\sqrt{n}\,\frac{\widehat{S}(t) - S(t)}{\widehat{S}(t)} \Longrightarrow W(\sigma(t)) \quad \text{on } D\langle 0, \tau \rangle.$$

Hence, the process

$$\sqrt{\frac{n}{\sigma(\tau)}}\,\frac{\widehat{S}(t) - S(t)}{\widehat{S}(t)}$$

converges weakly to the process $Q(t) = \frac{1}{\sqrt{\sigma(\tau)}} W(\sigma(t))$. Because the mapping $Q \to \sup_{0 \le t \le \tau} |Q(t)|$ is continuous with respect to the Skorokhod metric (see the note on p. 97 in the Appendix), we have

$$\sup_{0 \le t \le \tau} \left|\sqrt{\frac{n}{\sigma(\tau)}}\,\frac{\widehat{S}(t) - S(t)}{\widehat{S}(t)}\right| \xrightarrow{\mathrm{D}} \sup_{0 \le t \le \tau} |Q(t)|. \qquad (*)$$

The process $Q(t)$ is zero-mean Gaussian process with independent increments and continuous sample paths and its variance function is

$$\operatorname{var} Q(t) = \frac{1}{\sigma(\tau)} \operatorname{var} W(\sigma(t)) = \frac{\sigma(t)}{\sigma(\tau)}.$$

Now define $u = \frac{\sigma(t)}{\sigma(\tau)} \in \langle 0, 1\rangle$ and define a process $Q^*(u)$ by making a change of variables:

$$Q^*(u) = \frac{1}{\sqrt{\sigma(\tau)}} W(u\sigma(\tau)), \quad u \in \langle 0, 1\rangle.$$

Clearly,

$$\sup_{0\le t\le\tau} |Q(t)| = \sup_{0\le u\le 1} |Q^*(u)|.$$

But $Q^*(u)$ is the standard Brownian motion because

$$\operatorname{var} Q^*(u) = \frac{1}{\sigma(\tau)} \operatorname{var} W(u\sigma(\tau)) = \frac{u\sigma(\tau)}{\sigma(\tau)} = u.$$

Hence, the limiting distribution on the right-hand side of (*) is the same as the distribution of $\sup_{0\le u\le 1} |W(u)|$.

Replacing $\sigma(\tau)$ by $\widehat{\sigma}(\tau)$ does not change the limiting distribution because $\widehat{\sigma}(\tau)$ is a consistent estimator of $\sigma(\tau)$. □

The distribution function of the limiting random variable can be expressed as

$$\mathrm{P}\left[\sup_{0\le u\le 1} |W(u)| \le y\right] = \frac{4}{\pi} \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} \exp\left\{-\frac{\pi^2(2k+1)^2}{8y^2}\right\}$$

for any $y > 0$ (Billingsley 1999). Denote by c_α the α-quantile of this distribution.

Based on this result, Gill (1980) proposed asymptotic confidence bounds for the whole survival curve on the interval $\langle 0, \tau\rangle$ with lower and upper boundaries

$$\widehat{S}(t)\left(1 - c_{1-\alpha}\sqrt{\frac{\widehat{\sigma}(\tau)}{n}}\right) \quad \text{and} \quad \widehat{S}(t)\left(1 + c_{1-\alpha}\sqrt{\frac{\widehat{\sigma}(\tau)}{n}}\right).$$

Notice that the Gill bounds differ from the pointwise confidence intervals not only by the quantile but also by using $\widehat{\sigma}(\tau)$ in place of $\widehat{\sigma}(t)$.

The Gill bounds have the largest width at t close to zero when $\widehat{S}(t) \approx 1$. To overcome this weakness, alternative bounds were proposed by Hall and Wellner (1980). They are based on the following extension of Theorem 4.8.

Theorem 4.10. *Let* $K(t) = \frac{\sigma(t)}{1+\sigma(t)}$ *and* $\widehat{K}(t) = \frac{\widehat{\sigma}(t)}{1+\widehat{\sigma}(t)}$. *Under the conditions of Theorem 4.8,*

$$\sqrt{n}\,\frac{1 - \widehat{K}(t)}{\widehat{S}(t)} [\widehat{S}(t) - S(t)] \Longrightarrow B(K(t)) \quad \text{on } D\langle 0, \tau\rangle.$$

◊

Here, the process $B(t)$ is Brownian bridge discussed in Appendix A.3.3 on p. 95. It is a Gaussian process defined on the interval $\langle 0,\ 1\rangle$, with zero mean, variance function $\operatorname{var} B(t) = t(1-t)$, and covariance function at $s \le t$ given by $\operatorname{cov}(B(s), B(t)) = s(1-t)$. Notice that $K(t) \in \langle 0,\ 1)$ and $\widehat{K}(t) \in \langle 0,\ 1)$. The limiting process is a time-transformed Brownian bridge, with $K(t)$ playing the role of a non-decreasing time transformation from $\langle 0, \tau\rangle$ to $\langle 0,\ 1)$.

Proof (of Theorem 4.10). By Theorem 4.8, part (ii),

$$\sqrt{n}(\widehat{S} - S) \Longrightarrow SW(\sigma).$$

Since $K(s) = \sigma(s)/(1+\sigma(s))$, we have $\sigma(s) = K(s)/(1-K(s))$. Let us calculate the covariance function of the limiting process $SW(\sigma)$ at $s < t$.

$$\begin{aligned}\operatorname{cov}\left[S(s)W(\sigma(s)), S(t)W(\sigma(t))\right] &= S(s)S(t)\sigma(s) = S(s)S(t)\frac{K(s)}{1-K(s)}\\ &= \frac{S(s)}{1-K(s)}\,\frac{S(t)}{1-K(t)}K(s)[1-K(t)].\end{aligned}$$

So, the limiting process of

$$\frac{1-K(t)}{S(t)}\sqrt{n}[\widehat{S}(t) - S(t)]$$

has the covariance function $K(s)[1-K(t)]$ at $s < t$ and this is exactly the covariance function of the Brownian bridge B evaluated at $K(t)$.

Finally, one can replace S by $\widehat{S}$ and K by $\widehat{K}$ because they are uniformly consistent estimators. □

It follows from Theorem 4.10 and the continuous mapping theorem A.2 that

$$\mathrm{P}\left[\sup_{0\le t\le\tau} \sqrt{n}\,\frac{1-\widehat{K}(t)}{\widehat{S}(t)}\left|\widehat{S}(t) - S(t)\right| \ge y\right] \to \mathrm{P}\left[\sup_{0\le t\le\tau} |B(K(t))| \ge y\right]$$

We have

$$\mathrm{P}\left[\sup_{0\le t\le\tau} |B(K(t))| \ge y\right] = \mathrm{P}\left[\sup_{0\le u\le K(\tau)} |B(u)| \ge y\right]$$

Denote the α-quantile of the distribution of $\sup_{0\le u\le K(\tau)}|B(u)|$ by $k_\alpha(\tau)$. This can be calculated numerically.

The Hall-Wellner confidence bounds for survival function have lower and upper boundaries

$$\widehat{S}(t)\left(1 - k_{1-\alpha}(\tau)\frac{1}{\sqrt{n}[1-\widehat{K}(t)]}\right) \quad\text{and}\quad \widehat{S}(t)\left(1 + k_{1-\alpha}(\tau)\frac{1}{\sqrt{n}[1-\widehat{K}(t)]}\right).$$

Using the relationship $\frac{1}{1-\widehat{K}(t)} = 1 + \widehat{\sigma}(t)$, we can rewrite the Hall-Wellner bounds as

$$\widehat{S}(t)\left(1 - k_{1-\alpha}(\tau)\frac{1+\widehat{\sigma}(t)}{\sqrt{n}}\right) \quad \text{and} \quad \widehat{S}(t)\left(1 + k_{1-\alpha}(\tau)\frac{1+\widehat{\sigma}(t)}{\sqrt{n}}\right).$$

We can get conservative Hall-Wellner bounds that do not require the calculation of $k_{1-\alpha}(\tau)$ for a specific τ as follows: Since

$$\mathrm{P}\left[\sup_{0\le u\le K(\tau)} |B(u)| \ge y\right] \le \mathrm{P}\left[\sup_{0\le u\le 1} |B(u)| \ge y\right] = 2\sum_{j=1}^{\infty}(-1)^{j+1}\mathrm{e}^{-2j^2y^2},$$

where the distribution on the right-hand side can be calculated (Billingsley 1999) and is the same as the asymptotic distribution of the Kolmogorov-Smirnov test statistic, we can replace $k_{1-\alpha}(\tau)$ by the critical value of the Kolmogorov-Smirnov test $k_{1-\alpha}$ to obtain confidence bounds with asymptotic coverage $\ge 1-\alpha$.

The end of lecture 11 (Nov. 16)

5. Two-Sample Tests for Censored Data

5.1. Notation

Consider two independent samples of censored data obtained from two groups of subjects. We assume that (T_{ki}, C_{ki}), $i = 1, \dots, n_k$, $k = 1, 2$, are independent random vectors. Let $T_{k1}, \dots, T_{kn_k}$ be identically distributed with survival function S_k and cumulative hazard function Λ_k, $k = 1, 2$.

The goal is to test whether the failure time distributions in the two groups are the same. In particular,

$$H_0 : S_1(t) = S_2(t) \text{ for all } t \geq 0 \quad \text{against} \quad H_1 : \text{There exists } t \geq 0 \text{ s.t. } S_1(t) \neq S_2(t).$$

Of course, the hypothesis can be equivalently formulated as equality of cumulative hazard functions.

> *Notice that we do not assume that the censoring mechanisms in the two groups are the same. The censoring variables C_{ki} may have arbitrary distinct distributions.*

Denote $X_{ki} = T_{ki} \wedge C_{ki}$ censored failure times and $\delta_{ki} = \mathbb{1}(T_{ki} \leq C_{ki})$ failure indicators. The observed data are (X_{ki}, δ_{ki}), $i = 1, \dots, n_k$, $k = 1, 2$. The observed data can be also expressed in terms of counting processes $N_{ki}(t) = \mathbb{1}(T_{ki} \leq t, \delta_{ki} = 1)$ and at-risk processes $Y_{ki}(t) = \mathbb{1}(X_{ki} \geq t)$, $i = 1, \dots, n_k$, $k = 1, 2$.

We will work with the filtration

$$\mathcal{F}_t = \sigma\{N_{ki}(u), Y_{ki}(u),\ 0 \leq u \leq t, i = 1, \dots, n_k, k = 1, 2\}.$$

Take the compensator $A_{ki}(t) = \int_0^t Y_{ki}(u)\, d\Lambda_k(u)$. Under the independent censoring condition, $M_{ki}(t) = N_{ki}(t) - A_{ki}(t)$ are all $\mathcal{F}_t$-martingales. Define

$$\overline{N}_k(t) = \sum_{i=1}^{n_k} N_{ki}(t) \quad \text{and} \quad \overline{Y}_k(t) = \sum_{i=1}^{n_k} Y_{ki}(t).$$

Then

$$\overline{M}_k(t) = \sum_{i=1}^{n_k} M_{ki}(t) = \overline{N}_k(t) - \int_0^t \overline{Y}_k(u)\, d\Lambda_k(u)$$

are $\mathcal{F}_t$-martingales, $k = 1, 2$. Also define $\overline{N}(t) = \overline{N}_1(t) + \overline{N}_2(t)$ and $\overline{Y}(t) = \overline{Y}_1(t) + \overline{Y}_2(t)$.

Table 5.1.: Contingency table of failing and non-failing subjects in the two groups at the j-th ordered failure time t_j.

Failure	Group 1	Group 2	Total
Yes	D_{1j}	D_{2j}	D_j
No	$R_{1j} - D_{1j}$	$R_{2j} - D_{2j}$	$R_j - D_j$
Total	R_{1j}	R_{2j}	R_j

5.2. A heuristic derivation of the logrank test

Take all distinct failure times observed in both groups and order them. Denote the ordered failure times $t_1 < t_2 < \cdots < t_d$. Denote the number of observed failures in the k-th group at the time t_j by $D_{kj} = \Delta\overline{N}_k(t_j)$ and the number of subjects at risk in the k-th group at the time t_j by $R_{kj} = \overline{Y}_k(t_j)$. Let $D_j = D_{1j} + D_{2j}$ and $R_j = R_{1j} + R_{2j}$ be the number of failures and the risk set size in the combined sample. With this notation, the data observed at the time t_j can be summarized in the form of a two-way contingency table, see Table 5.1.

If H_0 is true then the two discrete variables (failure status and group membership) that formed the classification given in Table 5.1 are independent. It can be shown that* conditionally on the marginals D_j, R_{1j}, and R_{2j}, the number D_{1j} of failures in the first group has a hypergeometric distribution under H_0. Thus, conditionally on D_j, R_{1j}, and R_{2j},

$$\operatorname{E} D_{1j} = D_j \frac{R_{1j}}{R_j} \equiv E_j$$

$$\text{and} \quad \operatorname{var} D_{1j} = D_j \frac{R_{1j} R_{2j}}{R_j^2} \frac{R_j - D_j}{R_j - 1} \equiv V_j,$$

if H_0 holds. The test statistic we are going to consider compares the number of failures observed in the first group with the conditional expectation under H_0 at each failure time, and accumulates these contributions. Thus,

$$W = \sum_{j=1}^{d} (D_{1j} - E_j) = \sum_{j=1}^{d} \left(D_{1j} - D_j \frac{R_{1j}}{R_j} \right). \tag{5.1}$$

To standardize the statistic, we divide W by $\sqrt{\widehat{\operatorname{var}}\, W}$, the estimated standard deviation of W. If W were the sum of independent terms, we could take $\widehat{\operatorname{var}}\, W = \sum_{j=1}^{d} V_j$. However, $D_{1j} - E_j$ are clearly not independent. Nevertheless, it can be shown that

* See the development of the Fisher exact test of independence for 2×2 contingency tables with small cell counts.

the naive variance estimator that ignores the lack of independence is asymptotically correct and so

$$\frac{\sum_{j=1}^{d}(D_{1j}-E_j)}{\sqrt{\sum_{j=1}^{d}V_j}} \xrightarrow{\mathsf{D}} \mathsf{N}(0,1) \tag{5.2}$$

under H_0. Thus, we reject H_0 when

$$\frac{\left|\sum_{j=1}^{d}(D_{1j}-E_j)\right|}{\sqrt{\sum_{j=1}^{d}V_j}} \geq u_{1-\alpha/2} \quad \text{or} \quad \frac{\left[\sum_{j=1}^{d}(D_{1j}-E_j)\right]^2}{\sum_{j=1}^{d}V_j} \geq \chi_1^2(1-\alpha).$$

This test is called the two-sample *logrank test*. It was proposed (without any proof of its properties) by Mantel (1966).

We will prove (5.2) using martingale theory. In order to do that, we need to write the numerator of the logrank test statistic as a difference of stochastic integrals. Recall that $D_{kj} = \Delta\overline{N}_k(t_j)$ and $R_{kj} = \overline{Y}_k(t_j)$ and express (5.1) as follows:

$$\begin{aligned}
W &= \int_0^\infty 1\, d\overline{N}_1(s) - \int_0^\infty \frac{\overline{Y}_1(s)}{\overline{Y}(s)} d(\overline{N}_1 + \overline{N}_2)(s) \\
&= \int_0^\infty \left(1 - \frac{\overline{Y}_1(s)}{\overline{Y}(s)}\right) \overline{Y}_1(s) \frac{d\overline{N}_1(s)}{\overline{Y}_1(s)} - \int_0^\infty \frac{\overline{Y}_1(s)}{\overline{Y}(s)} \overline{Y}_2(s) \frac{d\overline{N}_2(s)}{\overline{Y}_2(s)} \\
&= \int_0^\infty \frac{\overline{Y}_1(s)\overline{Y}_2(s)}{\overline{Y}(s)}\, d(\widehat{\Lambda}_1 - \widehat{\Lambda}_2)(s),
\end{aligned}$$

where $\widehat{\Lambda}_k(t) = \int_0^t \frac{d\overline{N}_k(s)}{\overline{Y}_k(s)}$ is the Nelson-Aalen estimator of the cumulative hazard for the k-th group.

This also shows that W is the integrated weighted difference between the Nelson-Aalen estimators of cumulative hazards in the two groups. The weight $\overline{Y}_1(s)\overline{Y}_2(s)/\overline{Y}(s)$ takes into account the number of subjects that are observed at both groups at the time s. When either of the groups runs out of observations ($\overline{Y}_k = 0$), the weight is zero.

5.3. Linear rank statistics for censored data, weighted logrank tests

Definition, connections to rank tests

We will consider a class of test statistics of the form

$$W_K(t) = \int_0^t K(s) d(\widehat{\Lambda}_1 - \widehat{\Lambda}_2)(s),$$

with $W_K \equiv W_K(\infty)$. The process $K(s)$ is a bounded non-negative predictable process such that $K(s) = 0$ whenever $\overline{Y}_1(s) = 0$ or $\overline{Y}_2(s) = 0$. Every process $K(s)$ with these properties can be written in the form

$$K(s) = \sqrt{\frac{n_1 + n_2}{n_1 n_2}}\, W(s)\, \frac{\overline{Y}_1(s)\overline{Y}_2(s)}{\overline{Y}(s)},$$

where $W(s)$ is a bounded non-negative predictable process. The logrank test is obtained by setting $W(s) \equiv 1$. The statistics W_K are called *weighted logrank statistics*. In the notation of equation (5.1), a weighted logrank statistic can be expressed as

$$W_K = \sum_{j=1}^{d} W_j \left(D_{1j} - D_j \frac{R_{1j}}{R_j} \right),$$

where $W_j = W(t_j)$ is a weight for the j-th observed failure time.

We require the process W to be predictable, which implies that $W(s)$ must not depend on the data observed after s. If we choose $W(s)$ so that it depends only on the observed numbers of failures before s and numbers of subjects who were at risk when those failures occurred (but not on failure and censoring times directly), the statistic W_K becomes invariant with respect to strictly increasing transformations of time (they do not change the order of the observed failure or censoring times). Because of this, W_K represents a class of *linear rank statistics* for censored data.

With non-censored data, nonparametric *two-sample linear rank statistics* are defined as $\sum_{i=1}^{n_1} \varphi\left(\frac{R_i}{n+1}\right)$, where the nondecreasing function φ defined on $(0, 1)$ is called *the score*, R_i are ranks of the first sample among all observations from both samples, and $n = n_1 + n_2$ is the total sample size (Lehmann 1975). These test statistics are also invariant with respect to any strictly increasing transformations of data because such transformations do not change the ranks. We are going to note that some of these non-censored linear rank statistics are special cases of weighted logrank statistics.

*It is difficult to generalize the term "**rank**" to censored data because censoring makes ordering of failure times unclear. The class W_K provides a generalization of linear rank statistics to censored data through its invariance property even though it avoids any direct reference to the ranks.*

Examples of weighted logrank tests

1. For $W(s) = 1$, we get the *logrank test* (Mantel 1966).

   In non-censored data, the logrank test is equivalent to the *Savage exponential* *scores test* (Savage 1956) with scores

$$\varphi\left(\frac{R_i}{n+1}\right) = \sum_{j=1}^{R_i} \frac{1}{n-j+1}.$$

These scores are expressions for $\mathsf{E}\, X_{(R_i)}$, expected values of order statistics for a random sample of size n from the exponential distribution with parameter 1. Savage test is the most powerful test against changes in scale between two exponentially distributed samples or against shifts in location between two samples with Gumbel distributions.

2. For $W(s) = \frac{\overline{Y}(s)}{n+1}$, we get the *Gehan-Wilcoxon test* (Gehan 1965).

   In uncensored data, the Gehan-Wilcoxon test is equivalent to the *Wilcoxon rank-sum test* with scores
$$\varphi\left(\frac{R_i}{n+1}\right) = \frac{R_i}{n+1}.$$
   Wilcoxon test is the most powerful test against shifts in location between two samples with logistic distributions.

   This test puts more weight on early differences in hazard functions than on differences that occur later.

> ***Homework 7:***
>
> *Prove that, with uncensored data, the Gehan-Wilcoxon test is equivalent to the Wilcoxon two sample test.*

3. For $W(s) = \widehat{S}(s-)$, we get the *Prentice-Wilcoxon test* (Prentice 1978). This is another generalization of the *Wilcoxon rank-sum test* to censored data. It uses the Kaplan-Meier estimator as the weight (left-continuous version is used to assure predictability).

> *The Prentice test differs from the Gehan test by using the Kaplan-Meier estimator $\widehat{S}$ in place of the empirical distribution of the censored failure time. If the data are uncensored, $\frac{\overline{Y}}{n+1}$ and $\widehat{S}$ are both estimators of the survival function. However, in censored data $\frac{\overline{Y}(s)}{n+1}$ estimates the probability of being at risk, which is affected by the censoring distribution, unlike the Kaplan-Meier estimator, which estimates the survival function. This is why the Prentice-Wilcoxon test is the preferred variant.*

4. For $W(s) = \widehat{S}(s-)^{\rho}[1-\widehat{S}(s-)]^{\gamma}$, where $\rho, \gamma \geq 0$ are selected constants, we get the *Fleming-Harrington* $G(\rho,\gamma)$ *class of test statistics* (Fleming and Harrington 1981; Harrington and Fleming 1982). This class includes increasing, decreasing, and non-monotone weights depending on the choice of ρ and γ. The logrank test

is a special case for $\rho = \gamma = 0$, the Prentice-Wilcoxon test is a special case for $\rho = 1, \gamma = 0$.

Moments of weighted logrank statistics

Lemma 5.1. *The weighted logrank statistic*

$$W_K(t) = \int_0^t K(s) d(\widehat{\Lambda}_1 - \widehat{\Lambda}_2)(s),$$

can be written as

$$W_K(t) = \int_0^t \frac{K(s)}{\overline{Y}_1(s)} d\overline{M}_1(s) - \int_0^t \frac{K(s)}{\overline{Y}_2(s)} d\overline{M}_2(s) + \int_0^t K(s)[d\Lambda_1(s) - d\Lambda_2(s)]. \tag{5.3}$$

◊

The first two terms are martingale integrals because they have bounded and predictable integrands; the third term vanishes when the null hypothesis holds. This representation is the key to the theoretical investigation of weighted logrank statistics.

Proof. Consider the decomposition $\overline{N}_k = \overline{M}_k + \int \overline{Y}_k \, d\Lambda_k$ for $k = 1, 2$. We have

$$\int_0^t K(s) d\widehat{\Lambda}_k(s) = \int_0^t \frac{K(s)}{\overline{Y}_k(s)} \, d\overline{N}_k(s) = \int_0^t \frac{K(s)}{\overline{Y}_k(s)} \, d\overline{M}_k(s) + \int_0^t \frac{K(s)}{\overline{Y}_k(s)} \overline{Y}_k(s) \, d\Lambda_k(s).$$

Since $K(s) = 0$ whenever $\overline{Y}_k(s) = 0$, this directly leads to 5.3. □

Theorem 5.2.

(i) $\mathsf{E}\, W_K = \int_0^\infty \mathsf{E}\, K(s) \, d[\Lambda_1(s) - \Lambda_2(s)]$. *Under* $H_0 : \Lambda_1 = \Lambda_2$, $\mathsf{E}\, W_K = 0$.

(ii) Under $H_0 : \Lambda_1 = \Lambda_2 \equiv \Lambda$,

$$\sigma_K^2 \equiv \mathsf{var}\, W_K = \int_0^\infty \mathsf{E}\left\{\frac{\overline{Y}(s)}{\overline{Y}_1(s)\overline{Y}_2(s)} K^2(s)\right\}[1 - \Delta\Lambda(s)] \, d\Lambda(s).$$

◊

Proof. By Lemma 5.1,

$$W_K(t) = \widetilde{M}_1(t) - \widetilde{M}_2(t) + \int_0^t K(s) \, d(\Lambda_1 - \Lambda_2)(s),$$

where $\widetilde{M}_k = \int K/\overline{Y}_k \, d\overline{M}_k$ are martingales. Hence,

(i)

$$\mathsf{E}\, W_K = \mathsf{E}\, \widetilde{M}_1(\infty) - \mathsf{E}\, \widetilde{M}_2(\infty) + \int_0^\infty \mathsf{E}\, K(s) \, d(\Lambda_1 - \Lambda_2)(s)$$

and the expectations of the two martingales are zero. Under H_0, $\Lambda_1(s) = \Lambda_2(s)$ at all $s \geq 0$ and hence $EW_K = 0$.

(ii) Under H_0, $W_K(t) = \widetilde{M}_1(t) - \widetilde{M}_2(t)$. By Theorems 3.6 and 3.9, we have

$$\operatorname{var} \widetilde{M}_k(t) = \mathsf{E} \int_0^t \frac{K^2(s)}{\overline{Y}_k^2(s)} \overline{Y}_k(s)[1 - \Delta\Lambda(s)]\, d\Lambda(s)$$

and (the two samples are independent),

$$\operatorname{cov}(\widetilde{M}_1(t), \widetilde{M}_2(t)) = \mathsf{E} \int_0^t \frac{K^2(s)}{\overline{Y}_1(s)\overline{Y}_2(s)}\, d\langle \widetilde{M}_1, \widetilde{M}_2\rangle(s) = 0.$$

Thus,

$$\operatorname{var} W_K(t) = \operatorname{var} \widetilde{M}_1(t) + \operatorname{var} \widetilde{M}_2(t) = \int_0^t \mathsf{E}\, K^2(s)\left(\frac{1}{\overline{Y}_1(s)} + \frac{1}{\overline{Y}_2(s)}\right)[1 - \Delta\Lambda(s)]\, d\Lambda(s)$$

and this leads to the desired result. □

The next theorem introduces an unbiased estimator of σ_K^2.

Theorem 5.3. *Let the null hypothesis be true. Define*

$$\begin{aligned}
\widehat{\sigma}_K^2(t) &= \int_0^t K^2(s)\left(\frac{1}{\overline{Y}_1(s)} + \frac{1}{\overline{Y}_2(s)}\right)\left(1 - \frac{\Delta\overline{N}(s) - 1}{\overline{Y}(s) - 1}\right) d\widehat{\Lambda}(s) \\
&= \int_0^t \frac{K^2(s)}{\overline{Y}_1(s)\overline{Y}_2(s)}\left(1 - \frac{\Delta\overline{N}(s) - 1}{\overline{Y}(s) - 1}\right) d\overline{N}(s),
\end{aligned}$$

where $\widehat{\Lambda}(t) = \int_0^t d\overline{N}(s)/\overline{Y}(s)$ is the Nelson-Aalen estimator of the common cumulative hazard calculated from both samples. Then $\mathsf{E}\,\widehat{\sigma}_K^2(\infty) = \sigma_K^2$. ◇

It is not difficult to verify that for the logrank test, $\widehat{\sigma}_K^2$ is equal to the variance estimator $\sum V_j$ proposed in the previous section by considering hypergeometric distribution and ignoring non-independence of the terms included in the statistic.

Proof. Calculate

$$\begin{aligned}
E(\widehat{\sigma}_K^2(\infty) - \sigma_K^2) &= \mathsf{E}\left\{\int \frac{K^2}{\overline{Y}_1\overline{Y}_2}\left(1 - \frac{\Delta\overline{N} - 1}{\overline{Y} - 1}\right) d\overline{N} - \int \frac{K^2\overline{Y}}{\overline{Y}_1\overline{Y}_2}(1 - \Delta\Lambda)\, d\Lambda\right\} \\
&= \mathsf{E}\int \frac{K^2}{\overline{Y}_1\overline{Y}_2}\, d\overline{M} - \mathsf{E}\int \frac{K^2}{\overline{Y}_1\overline{Y}_2(\overline{Y} - 1)}\left[(\Delta\overline{N} - 1)\, d\overline{N} - \overline{Y}(\overline{Y} - 1)\Delta\Lambda\, d\Lambda\right].
\end{aligned}$$

The first term is a martingale integral with zero expectation. In the second term, the square bracket is zero for continuous failure times and is a martingale for discrete failure times (the proof of this is omitted). □

The end of lecture 12 (Nov. 23)

5.4. Asymptotic results for weighted logrank statistics

Take $\tau > 0$ such that $\mathrm{P}[Y_{ki}(\tau) = 1] > \delta > 0$ for $k = 1, 2$ and all $i = 1, \dots, n_k$. Assume that $\Lambda_k(\tau) < \infty$ for $k = 1, 2$. If (T_{ki}, C_{ki}), $i = 1, \dots, n_k$ are identically distributed within group k then by Lemma 4.1 there exist deterministic non-increasing functions $\pi_k(t) = \mathrm{P}[Y_{ki}(t) = 1]$ such that

$$\sup_{t \in \langle 0, \tau \rangle} \left| \frac{1}{n_k} \overline{Y}_k(t) - \pi_k(t) \right| \xrightarrow{\mathsf{P}} 0 \quad \text{and} \quad \pi_k(t) > \delta > 0 \text{ for } t \in \langle 0, \tau \rangle. \tag{5.4}$$

If the data (i.e., censoring times) are not identically distributed, the existence of functions π_1, π_2 satisfying (5.4) is taken as an assumption. Denote $n = n_1 + n_2$ and assume that $n_k / n \to a_k > 0$ as $n \to \infty$, $k = 1, 2$. It follows that $n^{-1} \overline{Y}(s)$ converges in probability to the function $\pi(s) = a_1 \pi_1(s) + a_2 \pi_2(s)$, uniformly in time.

Note. Under the null hypothesis, the distribution of T_{ki} is the same in both groups but the censoring distributions may not be the same, so in general $\pi_1(t) \neq \pi_2(t)$ even when H_0 holds.

We will formulate a result on the weak convergence of the weighted logrank statistic under the null hypothesis. The statistic is viewed as a process developing over time, i.e.,

$$W_K(t) = \int_0^t K(s) d(\widehat{\Lambda}_1 - \widehat{\Lambda}_2)(s),$$

with

$$K(s) = \sqrt{\frac{n}{n_1 n_2}}\, W(s) \frac{\overline{Y}_1(s) \overline{Y}_2(s)}{\overline{Y}(s)} = \sqrt{\frac{n_1 n_2}{n}}\, W(s) \frac{\overline{Y}_1(s)}{n_1} \frac{\overline{Y}_2(s)}{n_2} \frac{n}{\overline{Y}(s)}.$$

Theorem 5.4. *Let $W_K(t)$ be a weighted logrank statistic with the weight $W(s)$ of the form $W(s) = g(\widehat{S}(s-))$, where g is a bounded nonnegative continuous function with bounded variation on $\langle 0, 1 \rangle$ and $\widehat{S}(s)$ is the pooled Kaplan-Meier estimator at s. Suppose that the failure times in the two groups have the same distribution with cumulative hazard Λ. Let*

$$\sigma^2(t) = \int_0^t (h_1(s) + h_2(s)) (1 - \Delta\Lambda(s))\, d\Lambda(s) < \infty$$

for all $t \leq \tau$, where $h_k(s)$ is the limit in probability of $K^2(s) / \overline{Y}_k(s)$. Denote

$$\widehat{\sigma}^2(t) = \int_0^t \frac{K^2(s)}{\overline{Y}_1(s) \overline{Y}_2(s)} \left[1 - \frac{\Delta \overline{N}(s) - 1}{\overline{Y}(s) - 1} \right] d\overline{N}(s).$$

Then $W_K(t)$ (taken as a process over time) converges weakly to a time-transformed Brownian motion $W(\sigma^2(t))$ on $D\langle 0, \tau \rangle$ and $\widehat{\sigma}^2(t) \xrightarrow{\mathsf{P}} \sigma^2(t)$ as $n \to \infty$ uniformly over $t \in \langle 0, \tau \rangle$. In particular,

$$\frac{W_K(\tau)}{\sqrt{\widehat{\sigma}^2(\tau)}} \xrightarrow{\mathsf{D}} \mathsf{N}(0, 1).$$

◊

Note.

- We present the proof with the additional condition that the distribution of T_{ki} is continuous, however, the theorem also holds for distributions that are not continuous.
- The theorem also holds for $W(s) = g(\widehat{\pi}(s))$, where $\widehat{\pi}(s) = \overline{Y}(s)/n$ (Gehan-Wilcoxon test statistic).
- Asymptotic normality of W_K also holds when the statistic is calculated over the whole range of the data, i.e., when τ is replaced by $\inf\{t : \overline{Y}_1(t) = 0 \text{ or } \overline{Y}_2(t) = 0\}$. However, the conditions must be formulated a little bit more carefully and the proof needs additional work at some places.
- Weighted logrank statistics can be extended to test the equality of failure time distributions in several groups.

The hypothesis $H_0 : S_1(t) = S_2(t)$ is rejected when

$$\frac{|W_K(\tau)|}{\sqrt{\widehat{\sigma}^2(\tau)}} \geq u_{1-\alpha/2} \quad \text{or, equivalently,} \quad \frac{W_K^2(\tau)}{\widehat{\sigma}^2(\tau)} \geq \chi_1^2(1-\alpha).$$

Theorem 5.4 assures that the level of this test converges to α as $n \to \infty$.

Proof. Assume that the failure time distribution is continuous with a common hazard function λ and cumulative hazard Λ. According to Lemma 5.1,

$$W_K(t) = \int_0^t \frac{K(s)}{\overline{Y}_1(s)} d\overline{M}_1(s) - \int_0^t \frac{K(s)}{\overline{Y}_2(s)} d\overline{M}_2(s) + \int_0^t K(s)[d\Lambda_1(s) - d\Lambda_2(s)].$$

Since $K(s)$ is a bounded predictable process, $W_K(t)$ is a difference between two martingale integrals under the null hypothesis ($\Lambda_1 = \Lambda_2 \equiv \Lambda$). We will prove the joint weak convergence of the two martingale integrals.

Take $U_1 = \int \frac{K}{\overline{Y}_1} d\overline{M}_1$ and $U_2 = \int \frac{K}{\overline{Y}_2} d\overline{M}_2$ and apply Theorem 3.10. To verify the conditions, we need to show that $\langle U_k, U_k \rangle$ converges in probability to a deterministic function. We have

$$\langle U_k, U_k \rangle(t) = \int_0^t \frac{K^2(s)}{\overline{Y}_k^2(s)} \overline{Y}_k(s)\, d\Lambda(s) = \int_0^t \frac{K^2(s)}{\overline{Y}_k(s)}\, d\Lambda(s).$$

Now,

$$\frac{K^2(s)}{\overline{Y}_k(s)} = g^2(\widehat{S}(s-)) \frac{n_{3-k}}{n_1 + n_2} \frac{n_k}{\overline{Y}_k(s)} \left(\frac{\overline{Y}_1(s)}{n_1}\right)^2 \left(\frac{\overline{Y}_2(s)}{n_2}\right)^2 \left(\frac{n}{\overline{Y}(s)}\right)^2$$

and we know that $\overline{Y}_k(s)/n_k$ converges in probability to $\pi_k(s)$ uniformly in time, $\overline{Y}(s)/n$ converges in probability to $\pi(s)$ uniformly in time, and, because g is continuous, $g^2(\widehat{S}(s-))$ converges in probability to $g^2(S(s))$ uniformly in time. Thus,

$$\langle U_k, U_k \rangle(t) = \int_0^t \frac{K^2(s)}{\overline{Y}_k(s)}\, d\Lambda(s) \xrightarrow{\mathsf{P}} \int_0^t h_k(s)\, d\Lambda(s) = \int_0^t h_k(s)\lambda(s)\, ds,$$

where

$$h_k(s) = g^2(S(s))\frac{a_{3-k}}{\pi_k(s)}\frac{\pi_1^2(s)\pi_2^2(s)}{\pi^2(s)}.$$

The condition $\langle U_{k,\varepsilon}, U_{k,\varepsilon}\rangle \xrightarrow{\mathsf{P}} 0$ can be shown to hold by similar technique as in the proof of Theorem 4.2, part (iii).

So, by Theorem 3.10,

$$\begin{pmatrix} U_1 \\ U_2 \end{pmatrix} \Longrightarrow \begin{pmatrix} \int \sqrt{h_1\lambda}\, dW_1 \\ \int \sqrt{h_2\lambda}\, dW_2 \end{pmatrix} \quad \text{on } D^2\langle 0,\tau\rangle,$$

where W_1 and W_2 are two independent Brownian motions.

When evaluated at $t = \tau$, we get convergence in distribution of $(U_1(\tau), U_2(\tau))$ to a bivariate normal distribution with zero mean and diagonal covariance matrix with elements $\int_0^\tau h_k(s)\, d\Lambda(s)$. Thus, by the Cramér-Wold theorem,

$$W_K(\tau) = U_1(\tau) - U_2(\tau) \xrightarrow{\mathsf{D}} \mathsf{N}\Big(0, \int_0^\tau [h_1(s) + h_2(s)]\, d\Lambda(s)\Big).$$

The integrand in the asymptotic variance is

$$h_1(s) + h_2(s) = g^2(S(s))\frac{\pi_1^2(s)\pi_2^2(s)}{\pi^2(s)}\left[\frac{a_2}{\pi_1(s)} + \frac{a_1}{\pi_2(s)}\right] = g^2(S(s))\frac{\pi_1(s)\pi_2(s)}{\pi(s)}.$$

It remains to show that $\widehat{\sigma}^2(\tau)$ is a consistent estimator of this asymptotic variance. Indeed, with continuous failure times,

$$\begin{aligned} \widehat{\sigma}^2(\tau) &= \int_0^\tau \frac{K^2(s)}{\overline{Y}_1(s)\overline{Y}_2(s)}\, d\overline{N}(s) = \int_0^\tau \frac{K^2(s)}{\overline{Y}_1(s)\overline{Y}_2(s)}[\overline{Y}_1(s) + \overline{Y}_2(s)]\frac{d\overline{N}(s)}{\overline{Y}(s)} \\ &= \int_0^\tau \left[\frac{K^2(s)}{\overline{Y}_1(s)} + \frac{K^2(s)}{\overline{Y}_2(s)}\right] d\widehat{\Lambda}(s) \xrightarrow{\mathsf{P}} \int_0^\tau [h_1(s) + h_2(s)]\, d\Lambda(s) \end{aligned}$$

because of uniform convergences of the functions in the integrand as well as of the Nelson-Aalen estimator. □

5.5. Behavior of weighted logrank tests under the alternative

Consistency

First, let us investigate consistency of weighted logrank tests.

Definition 5.1. Let W_n be a sequence of test statistics with α-level rejection regions R_n, $n = 1, 2, \ldots$. The sequence W_n is consistent against the alternative H_A if

$$\lim_{n\to\infty} \mathrm{P}\big[W_n \in R_n \,\big|H_A\big] = 1. \qquad \nabla$$

Let $S_k(t)$ be the survival function of T in group k and let $\lambda_k(t)$ be the associated hazard function. We will be interested in two special alternatives. The alternative $H_1 : \lambda_1(t) \geq \lambda_2(t)$ (with strict inequality at some t) is called the *ordered hazards alternative*. The alternative $H_2 : S_2(t) \geq S_1(t)$ (with strict inequality at some t) is called the *alternative of stochastic ordering*. Clearly, H_1 implies H_2.

Let $\lambda_1(t) \geq \lambda_2(t)$ on $\langle 0, \tau\rangle$ and $\Lambda_1(\tau) > \Lambda_2(\tau)$. Consider weighted logrank statistics with $W(t) = g(\widehat{S}(t-))$ or $W(t) = g(\widehat{\pi}(t))$. We have

$$K(s) = \sqrt{\frac{n_1 n_2}{n}} W(s) \frac{\widehat{\pi}_1(s)\widehat{\pi}_2(s)}{\widehat{\pi}(s)}.$$

Since $\widehat{\pi}_k(s) \xrightarrow{\mathsf{P}} \pi_k(s)$ and $W(s) \xrightarrow{\mathsf{P}} w(s)$, a left continuous function such that $w(s) > 0$ on $\langle 0, \tau\rangle$, $K(s)$ converges to ∞ on a non-null set. Under the ordered hazards alternative, according to Theorem 5.2(i), the mean of $W_K(\tau)$ converges to infinity; since its variance estimator is bounded in probability, it follows that W_K is consistent against ordered hazards.

Consistency against stochastic ordering does not hold in general. It can be shown that W_K is consistent against H_2 if

$$\int_0^\tau w(s) \frac{\pi_1(s)\pi_2(s)}{\pi(s)} [d\Lambda_1(s) - d\Lambda_2(s)] > 0.$$

After performing integration by parts, this condition can be expressed as

$$\int_0^\tau [\Lambda_1(s) - \Lambda_2(s)]\, d\left[w(s)\frac{\pi_1(s)\pi_2(s)}{\pi(s)}\right] < 0.$$

The integrand $\Lambda_1(s) - \Lambda_2(s)$ is positive under H_2. The whole integral on the left-hand side is negative if and only if $w(s)\frac{\pi_1(s)\pi_2(s)}{\pi(s)}$ is a decreasing function of s. Since $\pi_1(s)\pi_2(s)/\pi(s)$ is decreasing, a sufficient condition is that $w(s)$ is non-increasing in s, in other words that the function g that defines the weight is non-decreasing. Then W_K is consistent against stochastic ordering. However, when g decreases consistency need not hold. Thus, $G(\rho, 0)$ statistics, including the logrank and Prentice-Wilcoxon, are always consistent against stochastic ordering. On the other hand, $G(\rho, \gamma)$ statistics with $\gamma > 0$ may not be.

Power

The power of weighted logrank tests is investigated in the local asymptotic sense. For a given $n = n_1 + n_2$, let the survival functions in the two groups be specified as $S_k^{(n)}(t) = S(h(t) + \theta_k^{(n)})$ where S is a known continuous survival function defined on $\mathbb{R}$ with a differentiable density, $h(t)$ is some differentiable increasing function from $\langle 0, \infty)$ to $\mathbb{R}$, $\theta_1^{(n)} = \theta_0 + c/\sqrt{n}$, and $\theta_2^{(n)} = \theta_0 - c/\sqrt{n}$, where c is a positive constant. This setup specifies so called *time-transformed shift alternatives*. Because the size of the shift is of the order $1/\sqrt{n}$, the distributions in the two groups converge to a single common distribution as $n \to \infty$.

It can be shown that, under these conditions, the test statistic is asymptotically normal with a finite non-zero mean so that the asymptotic power (probability of rejection) lies within the interval $(0, 1)$. The weight that maximizes this asymptotic power is

$$W(t) = g(\widehat{S}(t-)) = \psi'(S^{-1}(\widehat{S}(t-))),$$

where $\psi = \log(-S'/S)$ is the logarithm of the hazard for the distribution S and $\widehat{S}$ is the pooled Kaplan-Meier estimator. The test that maximizes power in this sense is called *locally asymptotically efficient*.

For example, if the data arise from a time-transformed shift in an extreme-value distribution with survival function $S(t) = \exp(-\mathrm{e}^t)$, we get $\psi(t) = \log(e^t) = t$ and $\psi' = 1$. Hence, the statistic with $W(t) = 1$, i.e., the logrank, *is locally efficient against shift alternatives in the extreme value distribution.*

Next, take the logistic distribution with $S(t) = 1 - (1 + \exp(-t))^{-1}$. Then $\psi(t) = -\log(1+\exp(-t))$, $\psi'(t) = \exp(-t)(1+\exp(-t))^{-1} = S(t)$ and $\psi'(S^{-1}(\widehat{S}(t-))) = S(S^{-1}(\widehat{S}(t-))) = \widehat{S}(t-)$. Hence, the Prentice-Wilcoxon statistic *is locally efficient against shift alternatives in the logistic distribution.*

These results can be extended by taking advantage of the generality of the time transformation h. The logrank can be shown to be efficient not only against shifts in the extreme value distribution, but against any proportional hazards alternatives, that is, alternatives $\lambda_2(t) = \theta\lambda_1(t)$ for $0 < \theta \neq 1$ independent of time and any hazard function λ_1.

The Gehan-Wilcoxon statistic uses a weight that is not a function of $\widehat{S}(t-)$; therefore it cannot be efficient against any location-shift alternative.

See Section 7.4 of Fleming and Harrington (1991) for more detailed discussion of local asymptotic efficiency of weighted logrank tests.

- *The (unweighted) logrank test has the best power against alternatives with constant hazard ratios.*
- *The Prentice-Wilcoxon test is a good choice for alternatives with decreasing hazard ratios (an early effect on the hazard that dissipates over time).*
- *Tests with increasing weights, such as $G(0, 1)$, are suitable for detecting increasing hazard ratios (a delayed effect on the hazard).*
- *For crossing-hazard alternatives, the weighted logrank tests may be inconsistent.*

The end of lecture 13 (Nov. 30)

6. Cox Proportional Hazards Model

6.1. Definition and interpretation

Consider n independent observations of the triplet $(X_i, \delta_i, \boldsymbol{Z}_i)$, $i = 1, \ldots, n$, where $X_i = \min(T_i, C_i)$ is a censored failure time, δ_i is the failure indicator, and $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ip})^\mathsf{T}$ is a p-vector of covariates. We would like to express the potential influence of the covariate vector $\boldsymbol{Z}_i$ on the distribution of T_i (which we assume to be *continuous*) through some regression model. Our ultimate goal will be to estimate the effect of $\boldsymbol{Z}_i$ on T_i and to test whether the components of $\boldsymbol{Z}_i$ affect T_i or not.

As usual, we view censored failure time data for each subject as a pair of processes: the counting process $N_i(t) = \mathbb{1}(T_i \leq t, \delta_i = 1)$ and the at-risk process $Y_i(t) = \mathbb{1}(X_i \geq t)$. In this context, we can allow the covariate vector to vary with time as well. So, let the covariates $\boldsymbol{Z}_i(t)$ be considered as vectors of p stochastic processes. Of course, this concept allows some (or all) components of $\boldsymbol{Z}_i(t)$ to be constant in time.

Note. In practice, fixed (time-independent) covariates represent characteristics of the subjects that cannot change (or are not allowed to change by the design of the study), such as gender or genotype. Time-varying covariates describe factors that change values during the follow-up of the subject, such as blood pressure, cholesterol concentration in blood, or cumulative amount of alcohol consumption during lifetime.

The independent censoring condition needs to take into account that hazard can be affected by the covariates. It will be expressed in terms of conditional hazard given the covariates, as follows:

$$\begin{aligned}\lambda(t \mid \boldsymbol{Z}) \equiv \lim_{h \searrow 0} \frac{1}{h} \mathrm{P}\big[t \leq T < t+h \,\big|\, T \geq t, \boldsymbol{Z}(t)\big] = \\ \lim_{h \searrow 0} \frac{1}{h} \mathrm{P}\big[t \leq T < t+h \,\big|\, T \geq t, C \geq t, \boldsymbol{Z}(t)\big]\end{aligned} \tag{6.1}$$

This condition is weaker than the original independent censoring condition (1.4). A sufficient condition for independent censoring is that T and C are conditionally independent given the covariates. It allows censoring times to depend on the covariates (e.g., men can have a different censoring distribution than women as long as gender is included in the model as a covariate).

Because we work with censored data, it is awkward to specify the regression model by expressing the influence of the covariates on the expected failure time. Instead, we will specify a model for the conditional hazard function defined in the top row of (6.1). The proportional hazards model proposed by Cox (1972) assumes a specific form for the effect of the covariate on the hazard function.

Definition 6.1. The observations $(X_i, \delta_i, \boldsymbol{Z}_i(t))$, $i = 1, \ldots, n$, satisfy the *Cox proportional hazards model* if the following two conditions hold:

(i) they are independent across different subjects;
(ii) the conditional hazard function given the covariate process has the form

$$\lambda(t \mid \boldsymbol{Z}) = \lambda_0(t) \exp\{\boldsymbol{\beta}_0^\top \boldsymbol{Z}(t)\}, \tag{6.2}$$

where $\lambda_0(t)$ is some unknown unspecified hazard function and $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is an unknown vector of regression coefficients. ∇

Note.

- The function $\lambda_0(t)$ is called *the baseline hazard*. It is the hazard of a subject with all covariate components equal to zero.
- The model does not include any intercept term (the role of the intercept is played by the baseline hazard).
- If $\lambda_0(t)$ were specified up to a finite-dimensional parameter vector the model would be fully parametric and the maximum likelihood theory could be used to estimate the parameters $\boldsymbol{\beta}_0$. E.g., if λ_0 were assumed to be constant over time, we would obtain the parametric exponential regression model discussed in Section 2.3.
- The Cox model makes assumptions on the form of the association between the covariate and the hazard but does not put any conditions on the shape of the hazard function. This type of statistical model is called a *semiparametric model*.
- If the covariates are time-varying, the hazard at t is only allowed to depend on the covariate value at the same time. However, the covariate may be transformed before inclusion into the model so that the value at t summarizes the past covariate history in some sense. The covariate cannot depend on anything measured after t (that would violate predictability).

Definition 6.2. The function

$$\Lambda_0(t) = \int_0^t \lambda_0(u)\, du$$

∇

is called the cumulative baseline hazard.

Suppose the covariates are constant over time, i.e., $\boldsymbol{Z}(t) \equiv \boldsymbol{Z}$. Then it follows from (6.2) that for any covariate values $\boldsymbol{Z}$ and $\boldsymbol{Z}^*$,

$$\frac{\lambda(t \mid \boldsymbol{Z}^*)}{\lambda(t \mid \boldsymbol{Z})} = \exp\{\boldsymbol{\beta}_0^\top (\boldsymbol{Z}^* - \boldsymbol{Z})\},$$

that is, the hazard ratio (*relative risk*) for any two subjects does not change over time. This is called the *proportional hazards assumption*.

Taking $\boldsymbol{Z}^* = \boldsymbol{Z} + \boldsymbol{e}_j$, where $\boldsymbol{e}_j$ is a p-vector with the j-th component equal to 1 and all other components zero, we get

$$\exp\{\beta_j\} = \frac{\lambda(t \mid \boldsymbol{Z} + \boldsymbol{e}_j)}{\lambda(t \mid \boldsymbol{Z})}$$

for any $\boldsymbol{Z}$ and t. This equation gives a meaning to the regression parameters: when exponentiated, they express relative risk for the event due to a unit increase in the associated covariate, while keeping all other covariates unchanged.

If the covariates are constant, the Cox model can be also expressed in terms of survival functions. Denote $S_0(t) = \exp\{-\Lambda_0(t)\}$, the baseline survival function. Then the conditional survival function for a subject with covariates $\boldsymbol{Z}$ is

$$S(t \mid \boldsymbol{Z}) = \exp\left\{-\int_0^t \lambda(s \mid \boldsymbol{Z})\, ds\right\} = [S_0(t)]^{\exp\{\boldsymbol{\beta}_0^\top \boldsymbol{Z}\}}.$$

With time-varying covariates, the conditional hazard function cannot be integrated easily and the survival function cannot be expressed in this way.

Note. In practice, time-varying covariates arise in two different ways.

1. They represent observations of some random process developing along with the follow-up of the subject. The observations are usually taken at discrete occassions. Such a covariate usually has a left-continuous piecewise-constant trajectory determined by the last observation of the random process.
2. They are created during the analysis as interactions of a time-invariant covariate Z with some transformation of time $g(t)$. Such interactions allow to circumvent the proportional hazards assumption by explicit modeling of the change in the covariate effect over time.

6.2. Parameter estimation via partial likelihood

Parameter estimation in the Cox proportional hazards model cannot be done by maximum likelihood methods because the model is not parametric. The problem is that the baseline hazard λ_0 is an unknown and arbitrary function. Sir David Cox (1972) proposed a modification of the likelihood function so that it does not depend on $\lambda_0(t)$. He called the modified likelihood *the partial likelihood*.

Let us describe here one of the possible approaches to derive the partial likelihood. Denote $t_1 < t_2 < \cdots < t_d$ the ordered distinct failure times. Because the failure time distribution is continuous there is exactly one failure at each t_i. Overall, d subjects failed and $n - d$ subjects were censored. Denote $t_0 = 0$ and $t_{d+1} = \infty$.

Denote by D_i the index of the subject that failed at t_i, and let B_i store all the other information that was accrued in the data during the interval $\langle t_i, t_{i+1})$, in particular:

- the time t_{i+1} of the next failure;
- indices of subjects who were censored in $\langle t_i, t_{i+1})$;
- censoring times in $\langle t_i, t_{i+1})$;
- covariate values of all subjects in $\langle t_i, t_{i+1})$.

All the information contained in the original data $(X_i, \delta_i, \boldsymbol{Z}_i(\cdot))$, $i = 1, \ldots, n$ is also contained in the sequence $(B_0, D_1, B_1, D_2, \ldots, D_d, B_d)$. The likelihood, that is the joint density of all the data, can be written as

$$\begin{aligned}
f(B_0, D_1,&B_1, D_2, \ldots, D_d, B_d) = \\
&= f(B_0)f(D_1 \mid B_0)f(B_1 \mid B_0, D_1) \times \cdots \times f(D_d \mid B_0, \ldots, B_{d-1}, D_1, \ldots, D_{d-1}) \\
&\quad \times f(B_d \mid B_0, \ldots, B_{d-1}, D_1, \ldots, D_d) \\
&= \underbrace{\prod_{i=1}^{d} f(D_i \mid B_0, \ldots, B_{i-1}, D_1, \ldots, D_{i-1})}_{\equiv L(\boldsymbol{\beta})} \prod_{i=1}^{d} f(B_i \mid B_0, \ldots, B_{i-1}, D_1, \ldots, D_i).
\end{aligned}$$

The first part contains most of the information about the effect of covariates on the hazard of failure. It is denoted by $L(\boldsymbol{\beta})$ and called *the partial likelihood*. The second part is ignored. To evaluate the partial likelihood, we need to find an expression for $f(D_i \mid B_0, \ldots, B_{i-1}, D_1, \ldots, D_{i-1})$. Because D_i contains one of the values $1, \ldots, n$, this is interpreted as the conditional probability $\mathrm{P}\left[D_i = l \,\middle|\, B_0, \ldots, B_{i-1}, D_1, \ldots, D_{i-1}\right]$ that the subject $l \in \{1, \ldots, n\}$ fails at the time t_i, knowing that exactly one subject failed at t_i, and knowing which subjects failed or were censored before t_i and what were the covariates of the subjects who were at risk for failure at t_i. For subjects that are not at risk at t_i the conditional probability is zero. For subjects that are still at risk, the failure probability at t_i is proportional to their hazard at this time, which is expressed via the Cox model specification (6.2). Since the failure probabilities must sum into one across all subjects who are at risk at t_i, we get

$$\begin{aligned}
\mathrm{P}&\left[D_i = l \,\middle|\, B_0, \ldots, B_{i-1}, D_1, \ldots, D_{i-1}\right] = \\
&= \frac{\lambda_0(t_i)\exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_l(t_i)\}}{\sum_{j=1}^{n} Y_j(t_i)\lambda_0(t_i)\exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_j(t_i)\}} = \frac{\exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_l(t_i)\}}{\sum_{j=1}^{n} Y_j(t_i)\exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_j(t_i)\}}.
\end{aligned}$$

Note that this does not depend on the baseline hazard. Taking these terms for all failure times as likelihood contributions to be multiplied, we get the partial likelihood in an explicit form

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{d} \frac{\exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_{l(i)}(t_i)\}}{\sum_{j=1}^{n} Y_j(t_i)\exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_j(t_i)\}},$$

where $l(i)$ denotes the index of the subject that failed at t_i. After a simple manipulation, we get the definition below.

Definition 6.3. The function

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{s>0} \left[\frac{Y_i(s) \exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_i(s)\}}{\sum_{j=1}^{n} Y_j(s) \exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_j(s)\}} \right]^{\Delta N_i(s)}$$

is called the partial likelihood [PL] function for parameters $\boldsymbol{\beta}$ in the Cox proportional hazards model. The value

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta} \in \mathbb{R}^p} L(\boldsymbol{\beta})$$

is called the maximum partial likelihood estimator [MPLE] of Cox model parameters (Cox 1972). ∇

Notation. Let

$$S_n^{(k)}(\boldsymbol{\beta}, t) = \frac{1}{n} \sum_{i=1}^{n} Y_i(t) \boldsymbol{Z}_i^{\otimes k}(t) \mathrm{e}^{\boldsymbol{\beta}^\top \boldsymbol{Z}_i(t)},$$

where $\boldsymbol{z}^{\otimes 0} = 1$, $\boldsymbol{z}^{\otimes 1} = \boldsymbol{z}$, and $\boldsymbol{z}^{\otimes 2} = \boldsymbol{z}\boldsymbol{z}^\top$, for any vector $\boldsymbol{z}$. Let

$$\overline{\boldsymbol{Z}}_n(\boldsymbol{\beta}, t) = \frac{S_n^{(1)}(\boldsymbol{\beta}, t)}{S_n^{(0)}(\boldsymbol{\beta}, t)}.$$

Notice that $S_n^{(0)}$ is a random variable, $S_n^{(1)}$ is a random p-vector, and $S_n^{(2)}$ is a random $p \times p$ matrix. The denominator of each term in the partial likelihood is equal to $nS_n^{(0)}$. Differentiating $S_n^{(0)}$ once and twice with respect to $\boldsymbol{\beta}$, we get

$$\frac{\partial S_n^{(0)}(\boldsymbol{\beta}, t)}{\partial \boldsymbol{\beta}} = S_n^{(1)}(\boldsymbol{\beta}, t) \quad \text{and} \quad \frac{\partial S_n^{(1)}(\boldsymbol{\beta}, t)}{\partial \boldsymbol{\beta}^\top} = S_n^{(2)}(\boldsymbol{\beta}, t). \tag{6.3}$$

Also,

$$\frac{\partial \log S_n^{(0)}(\boldsymbol{\beta}, t)}{\partial \boldsymbol{\beta}} = \frac{S_n^{(1)}(\boldsymbol{\beta}, t)}{S_n^{(0)}(\boldsymbol{\beta}, t)} = \overline{\boldsymbol{Z}}_n(\boldsymbol{\beta}, t). \tag{6.4}$$

Because

$$\overline{\boldsymbol{Z}}_n(\boldsymbol{\beta}, t) = \sum_{i=1}^{n} w_i(\boldsymbol{\beta}, t) \boldsymbol{Z}_i(t), \quad \text{where} \quad w_i(\boldsymbol{\beta}, t) = \frac{Y_i(t) \exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_i(t)\}}{\sum_{j=1}^{n} Y_j(t) \exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_j(t)\}} \tag{6.5}$$

are weights summing up into one, $\overline{\boldsymbol{Z}}_n(\boldsymbol{\beta}, t)$ can be viewed as a weighted average of the covariates of subjects who are at risk at the time t. The weights are equal to the conditional probabilities that the i-th subject fails at t given that a failure occurred at t (these weights are equal to the partial likelihood contributions).

The maximum partial likelihood estimator (MPLE) is obtained by maximizing log partial likelihood

$$\ell(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^\infty \left[\boldsymbol{\beta}^\top \boldsymbol{Z}_i(s) - \log n S_n^{(0)}(\boldsymbol{\beta}, s) \right] dN_i(s).$$

Differentiating this expression with respect to $\boldsymbol{\beta}$ and using (6.3) and (6.4), we obtain the score statistic

$$\boldsymbol{U}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\infty} \left[\boldsymbol{Z}_i(t) - \overline{\boldsymbol{Z}}_n(\boldsymbol{\beta}, t)\right] dN_i(t).$$

The maximum partial likelihood estimator $\widehat{\boldsymbol{\beta}}$ solves the system of equations

$$\boldsymbol{U}_n(\widehat{\boldsymbol{\beta}}) = \mathbf{0}.$$

Take the filtration

$$\mathcal{F}_t = \sigma\{N_i(u), Y_i(u+), \boldsymbol{Z}_i(u),\ 0 \le u \le t, i = 1, \ldots, n\}.$$

Let $\boldsymbol{Z}_i(t)$ be $\mathcal{F}_t$-predictable. Denote

$$A_i(t) = \int_0^t Y_i(u) \exp\{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i(u)\}\, d\Lambda_0(u).$$

If the independent censoring conditon holds then $M_i(t) = N_i(t) - A_i(t)$ is an $\mathcal{F}_t$-martingale according to Theorem 3.2.

In the sequel, we will follow the development of the score statistic as a process evolving over time. We denote

$$\boldsymbol{U}_n(\boldsymbol{\beta}, t) \equiv \sum_{i=1}^{n} \int_0^{t} \left[\boldsymbol{Z}_i(s) - \overline{\boldsymbol{Z}}_n(\boldsymbol{\beta}, s)\right] dN_i(s)$$

so that $\boldsymbol{U}_n(\boldsymbol{\beta}) = \boldsymbol{U}_n(\boldsymbol{\beta}, \infty)$.

The following lemma is the key for investigating the properties of the MPLE.

Lemma 6.1. *At the true parameter $\boldsymbol{\beta}_0$ and at any $t \in \langle 0, \infty)$ we have*

$$\boldsymbol{U}_n(\boldsymbol{\beta}_0, t) = \sum_{i=1}^{n} \int_0^{t} \left[\boldsymbol{Z}_i(s) - \overline{\boldsymbol{Z}}_n(\boldsymbol{\beta}_0, s)\right] dM_i(s)$$

where the integrand is a predictable process. Thus, $\boldsymbol{U}_n(\boldsymbol{\beta}_0, t)$ is an $\mathcal{F}_t$-martingale. ◊

The end of lecture 14 (Dec. 7)

Proof. In the proof, we leave out most of the arguments to make the expressions easier to read.

The difference between $\sum \int (\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}_n)\, dN_i$ and $\sum \int (\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}_n)\, dM_i$ is

$$\sum \int (\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}_n)\, dA_i = \sum \int (\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}_n) Y_i e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i}\, d\Lambda_0 = \int \left(\sum \boldsymbol{Z}_i Y_i e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i} - \overline{\boldsymbol{Z}}_n \sum Y_i e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i}\right) d\Lambda_0$$

and the parenthesis is

$$\sum \boldsymbol{Z}_i Y_i e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i} - \overline{\boldsymbol{Z}}_n \sum Y_i e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i} = n S_n^{(1)} - \frac{S_n^{(1)}}{S_n^{(0)}} n S_n^{(0)} = n(S_n^{(1)} - S_n^{(1)}) = \mathbf{0}.$$

Hence the two expressions for the score at $\boldsymbol{\beta}_0$ are the same. □

Lemma 6.1 would not be true if the covariates were not $\mathcal{F}_t$-predictable processes or if the score was evaluated at a parameter value other than $\boldsymbol{\beta}_0$.

Note. (about the Cox model score statistic)

- The score statistic can be written in a form suitable for calculation as follows:

$$\boldsymbol{U}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[\boldsymbol{Z}_i(T_i) - \overline{\boldsymbol{Z}}_n(\boldsymbol{\beta}, T_i)\right].$$

- The score statistic includes a term for each of the failures. A subject who was censored does not contribute a term to the score but appears in $\overline{\boldsymbol{Z}}_n(\boldsymbol{\beta}, T_i)$ (as long as the censoring occurred after T_i). Thus, the score statistic is not a sum of independent terms.
- The likelihood equations can be written as

$$\sum_{i=1}^{n} \delta_i \boldsymbol{Z}_i(T_i) = \sum_{i=1}^{n} \delta_i \overline{\boldsymbol{Z}}_n(\widehat{\boldsymbol{\beta}}, T_i).$$

Notation. Let

$$\mathcal{I}_n(\boldsymbol{\beta}, t) \equiv -\frac{1}{n} \frac{\partial \boldsymbol{U}_n(\boldsymbol{\beta}, t)}{\partial \boldsymbol{\beta}^\top} = \frac{1}{n} \sum_{i=1}^{n} \int_0^t \left[\frac{S_n^{(2)}(\boldsymbol{\beta}, s)}{S_n^{(0)}(\boldsymbol{\beta}, s)} - \overline{\boldsymbol{Z}}_n^{\otimes 2}(\boldsymbol{\beta}, s)\right] dN_i(s). \tag{6.6}$$

This matrix is a counterpart of the observed information matrix in ordinary likelihood theory.

First, let us show that the observed information can be expressed as shown on the right-hand side of (6.6). We have

$$-\frac{\partial \boldsymbol{U}_n(\boldsymbol{\beta}, t)}{\partial \boldsymbol{\beta}^\top} = \sum_{i=1}^{n} \int_0^t \frac{\partial \overline{\boldsymbol{Z}}_n(\boldsymbol{\beta}, s)}{\partial \boldsymbol{\beta}^\top} dN_i(s)$$

and

$$\frac{\partial \overline{\boldsymbol{Z}}_n}{\partial \boldsymbol{\beta}^\top} = \frac{\partial \left(S_n^{(1)}/S_n^{(0)}\right)}{\partial \boldsymbol{\beta}^\top} = \frac{\partial S_n^{(1)}}{\partial \boldsymbol{\beta}^\top} \frac{1}{S_n^{(0)}} - S_n^{(1)} \frac{{S_n^{(1)}}^\top}{\left(S_n^{(0)}\right)^2} = \frac{S_n^{(2)}}{S_n^{(0)}} - \overline{\boldsymbol{Z}}_n \overline{\boldsymbol{Z}}_n^\top.$$

Now let us investigate the existence and uniqueness of the solution to the system of equations $\boldsymbol{U}_n(\widehat{\boldsymbol{\beta}}) = \boldsymbol{0}$. By (6.5), we have

$$\frac{S_n^{(2)}}{S_n^{(0)}} - \overline{\boldsymbol{Z}}_n^{\otimes 2} = \sum_{i=1}^{n} w_i \boldsymbol{Z}_i^{\otimes 2} - \left(\sum_{i=1}^{n} w_i \boldsymbol{Z}_i\right)^{\otimes 2} = \sum_{i=1}^{n} w_i \left[\boldsymbol{Z}_i - \left(\sum_{j=1}^{n} w_j \boldsymbol{Z}_j\right)\right]^{\otimes 2} \geq 0.$$

This matrix is in fact a weighted covariance matrix of the covariates. It is positive definite as long as it is non-singular. Singularity can only occur if there exists a linear combination of the covariates with zero variance, i.e., if the covariates of the subjects who are at risk are linearly dependent (at all times!). Thus, if we assume that $\mathcal{I}_n(\boldsymbol{\beta}, t)$ is non-sigular, it must be positive definite at all t and for all $\boldsymbol{\beta}$. This proves the following lemma.

Lemma 6.2. *Let $\mathcal{I}_n(\boldsymbol{\beta}, \infty)$ be non-singular. Then $\ell(\boldsymbol{\beta})$ is strictly concave at all $\boldsymbol{\beta} \in \mathbb{R}^p$, has a unique maximum $\widehat{\boldsymbol{\beta}}$, and the maximum is the unique solution to the system of equations $\boldsymbol{U}_n(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$.* ◊

The likelihood equations are solved numerically by the Newton-Raphson algorithm. Choose an initial value $\widehat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$ and iterate

$$\widehat{\boldsymbol{\beta}}^{(r+1)} = \widehat{\boldsymbol{\beta}}^{(r)} + \left[n\mathcal{I}_n(\widehat{\boldsymbol{\beta}}^{(r)}, \infty) \right]^{-1} \boldsymbol{U}_n(\widehat{\boldsymbol{\beta}}^{(r)})$$

until convergence.

6.3. Properties of the maximum PL estimator

We work with the filtration

$$\mathcal{F}_t = \sigma\{N_i(u), Y_i(u+), \boldsymbol{Z}_i(u),\ 0 \le u \le t, i = 1, \dots, n\},$$

we assume that $\boldsymbol{Z}_i(t)$ are $\mathcal{F}_t$ predictable and that the independent censoring condition (6.1) is fulfilled.

According to Lemma 6.1, $\boldsymbol{U}_n(\boldsymbol{\beta}_0, t) = \sum_{i=1}^n \int_0^t H_i(s)\, dM_i(s)$, where $H_i(s) = \boldsymbol{Z}_i(s) - \overline{\boldsymbol{Z}}_n(\boldsymbol{\beta}_0, s)$ is predictable. In the subsequent proofs, we will also assume that all the components of $\boldsymbol{Z}_i(t)$ are bounded, which implies that the process $H_i(s)$ is bounded. The boundedness condition could be relaxed, however – only the proofs would become a bit more complicated. The distribution of the failure time T_i is assumed to be continuous throughout the whole chapter.

The following theorem shows that, under the given conditions, the partial likelihood score statistic has the same moment properties as the ordinary likelihood score statistic.

Theorem 6.3. *At any $t \ge 0$,*

(i) $\mathsf{E}\, \boldsymbol{U}_n(\boldsymbol{\beta}_0, t) = \mathbf{0}$

(ii) $\operatorname{var} \boldsymbol{U}_n(\boldsymbol{\beta}_0, t) = \mathsf{E} \displaystyle\int_0^t \left[\frac{S_n^{(2)}(\boldsymbol{\beta}_0, s)}{S_n^{(0)}(\boldsymbol{\beta}_0, s)} - \overline{\boldsymbol{Z}}_n^{\otimes 2}(\boldsymbol{\beta}_0, s) \right] n S_n^{(0)}(\boldsymbol{\beta}_0, s)\, d\Lambda_0(s) = -\mathsf{E}\, \frac{\partial \boldsymbol{U}_n(\boldsymbol{\beta}_0, t)}{\partial \boldsymbol{\beta}^\mathsf{T}}.$ ◊

Proof.

(i) By Theorem 3.4, $\boldsymbol{U}_n(\boldsymbol{\beta}_0, t)$ is a martingale integral, hence it has zero expectation at all t.

(ii) By Theorems 3.5, 3.6, and 3.9 (see the bottom of page 35)

$$
\begin{aligned}
\operatorname{var} \boldsymbol{U}_n(\boldsymbol{\beta}_0, t) &= \mathsf{E} \sum_{i=1}^{n} \int_0^t H_i^{\otimes 2}(s) Y_i(s)\, d\Lambda_i(s) \\
&= \mathsf{E} \sum_{i=1}^{n} \int_0^t \left[\boldsymbol{Z}_i(s) - \overline{\boldsymbol{Z}}_n(\boldsymbol{\beta}_0, s)\right]^{\otimes 2} Y_i(s) \mathrm{e}^{\boldsymbol{\beta}_0^\mathsf{T} \boldsymbol{Z}_i(s)}\, d\Lambda_0(s) \\
&= \mathsf{E} \int_0^t \sum_{i=1}^{n} \left[\boldsymbol{Z}_i^{\otimes 2} - \boldsymbol{Z}_i \overline{\boldsymbol{Z}}_n^\mathsf{T} - \overline{\boldsymbol{Z}}_n \boldsymbol{Z}_i^\mathsf{T} + \overline{\boldsymbol{Z}}_n^{\otimes 2}\right] Y_i(s) \mathrm{e}^{\boldsymbol{\beta}_0^\mathsf{T} \boldsymbol{Z}_i(s)}\, d\Lambda_0(s) \\
&= \mathsf{E}\, n \int_0^t \left[S_n^{(2)} - S_n^{(1)} \overline{\boldsymbol{Z}}_n^\mathsf{T} - \overline{\boldsymbol{Z}}_n S_n^{(1)\mathsf{T}} + \overline{\boldsymbol{Z}}_n^{\otimes 2} S_n^{(0)}\right] d\Lambda_0(s) \\
&= \mathsf{E}\, n \int_0^t \left[S_n^{(2)} - \overline{\boldsymbol{Z}}_n^{\otimes 2} S_n^{(0)}\right] d\Lambda_0(s) \\
&= \mathsf{E} \int_0^t \left[\frac{S_n^{(2)}(\boldsymbol{\beta}_0, s)}{S_n^{(0)}(\boldsymbol{\beta}_0, s)} - \overline{\boldsymbol{Z}}_n^{\otimes 2}(\boldsymbol{\beta}_0, s)\right] n S_n^{(0)}(\boldsymbol{\beta}_0, s)\, d\Lambda_0(s)
\end{aligned}
$$

Next, from (6.6),

$$
\begin{aligned}
-\mathsf{E} \frac{\partial \boldsymbol{U}_n(\boldsymbol{\beta}_0, t)}{\partial \boldsymbol{\beta}^\mathsf{T}} &= \mathsf{E} \sum_{i=1}^{n} \int_0^t \left[\frac{S_n^{(2)}(\boldsymbol{\beta}_0, s)}{S_n^{(0)}(\boldsymbol{\beta}_0, s)} - \overline{\boldsymbol{Z}}_n^{\otimes 2}(\boldsymbol{\beta}_0, s)\right] dN_i(s) \\
&= \mathsf{E} \sum_{i=1}^{n} \int_0^t \left[\frac{S_n^{(2)}(\boldsymbol{\beta}_0, s)}{S_n^{(0)}(\boldsymbol{\beta}_0, s)} - \overline{\boldsymbol{Z}}_n^{\otimes 2}(\boldsymbol{\beta}_0, s)\right] Y_i(s) \mathrm{e}^{\boldsymbol{\beta}_0^\mathsf{T} \boldsymbol{Z}_i(s)}\, d\Lambda_0(s) \\
&= \operatorname{var} \boldsymbol{U}_n(\boldsymbol{\beta}_0, t) \qquad \square
\end{aligned}
$$

To prove asymptotic properties of the partial likelihood estimator, a set of additional regularity conditions is needed.

Assumptions.

A.1 The data are observed on an interval $\langle 0, \tau \rangle$, such that $\tau > 0$ is fixed and the probability of being observed is $\mathrm{P}\left[Y_i(\tau) = 1\right] > \delta$ for all i and some $\delta > 0$. Let $\Lambda_0(\tau) < \infty$.

A.2 There exists a neighborhood $\mathcal{B}$ of $\boldsymbol{\beta}_0$ and functions $s^{(0)}$, $s^{(1)}$, and $s^{(2)}$ defined on $\mathcal{B} \times \langle 0, \tau \rangle$ such that

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}, t \in \langle 0, \tau \rangle} \left\| S_n^{(j)}(\boldsymbol{\beta}, t) - s^{(j)}(\boldsymbol{\beta}, t) \right\| \xrightarrow{\mathsf{P}} 0,$$

for each $j = 0, 1, 2$, where $\|\boldsymbol{a}\| \equiv \max |a_k|$.

A.3 The functions $s^{(j)}$ are bounded on $\mathcal{B} \times \langle 0, \tau \rangle$, $s^{(0)}$ is bounded away from 0 on $\mathcal{B} \times \langle 0, \tau \rangle$. The family $\{s^{(j)}(\boldsymbol{\beta}, t) : t \in \langle 0, \tau \rangle\}$ is equicontinuous at $\boldsymbol{\beta}_0$, i.e.

$$\forall \varepsilon > 0\ \exists \delta > 0\ \forall \boldsymbol{\beta} \in \mathcal{B} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| < \delta \Rightarrow \|s^{(j)}(\boldsymbol{\beta}, t) - s^{(j)}(\boldsymbol{\beta}_0, t)\| < \varepsilon\ \forall t \in \langle 0, \tau \rangle.$$

A.4 $\forall \boldsymbol{\beta} \in \mathcal{B}, \forall t \in \langle 0, \tau \rangle$

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

A.5 Let $\boldsymbol{e}(\boldsymbol{\beta}, t) \equiv \frac{s^{(1)}(\boldsymbol{\beta},t)}{s^{(0)}(\boldsymbol{\beta},t)}$. The matrix

$$I(\boldsymbol{\beta}_0, t) \equiv \int_0^t \left[\frac{s^{(2)}(\boldsymbol{\beta}_0, s)}{s^{(0)}(\boldsymbol{\beta}_0, s)} - \boldsymbol{e}^{\otimes 2}(\boldsymbol{\beta}_0, s) \right] s^{(0)}(\boldsymbol{\beta}_0, s)\, d\Lambda_0(s)$$

is positive definite at $t = \tau$.

The last assumption defines the information matrix and assures its regularity. If the data are independent and identically distributed, assumptions A.2 – A.4 can be replaced by the single condition

$$\mathsf{E} \sup_{\substack{\boldsymbol{\beta} \in \mathcal{B} \\ t \in \langle 0, \tau \rangle}} Y_i(t) \|\boldsymbol{Z}_i(t)\|^2 \mathrm{e}^{\boldsymbol{\beta}^\top \boldsymbol{Z}_i(t)} < \infty.$$

If all covariate components have bounded support, this condition is automatically fulfilled.

Now we can state and prove weak convergence of the partial likelihood score statistic.

Theorem 6.4. *Let assumptions A.1 – A.5 hold. Then*

$$\frac{1}{\sqrt{n}} \boldsymbol{U}_n(\boldsymbol{\beta}_0, t) \Longrightarrow \boldsymbol{W}(t) \quad \text{on } D^p \langle 0, \tau \rangle,$$

where $\boldsymbol{W}(t)$ is a p-variate zero-mean Gaussian process with continuous sample paths, independent increments and variance function $\operatorname{var} \boldsymbol{W}(t) = I(\boldsymbol{\beta}_0, t)$. ◊

Corollary. Under conditions A.1 – A.5,

$$\frac{1}{\sqrt{n}} \boldsymbol{U}_n(\boldsymbol{\beta}_0, \tau) \xrightarrow{\mathsf{D}} \mathsf{N}_p(\boldsymbol{0}, I(\boldsymbol{\beta}_0, \tau)).$$

Proof. We will use Theorem 3.11 to show joint weak convergence of the components of the score statistic. Denote

$$\boldsymbol{U}^{(n)}(t) \equiv \frac{1}{\sqrt{n}} \boldsymbol{U}_n(\boldsymbol{\beta}_0, t) = \sum_{i=1}^n \int_0^t H_i^{(n)}(s)\, dM_i(s)$$

where

$$H_i^{(n)}(s) = \frac{1}{\sqrt{n}} \left[\boldsymbol{Z}_i(s) - \overline{\boldsymbol{Z}}_n(\boldsymbol{\beta}_0, s) \right]$$

is a bounded predictable process. The predictable covariance process for the k-th and l-th component of $\boldsymbol{U}^{(n)}(t)$ is

$$\langle U_k^{(n)}, U_l^{(n)} \rangle(t) = \sum_{i=1}^{n} \frac{1}{n} \int_0^t [Z_{ik}(s) - \overline{Z}_k(\boldsymbol{\beta}_0, s)][Z_{il}(s) - \overline{Z}_l(\boldsymbol{\beta}_0, s)] Y_i(s) \mathrm{e}^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i(s)} \, d\Lambda_0(s)$$

Writing this in a matrix form and following the same steps as the proof of Theorem 6.3, we can express the whole matrix of predictable covariance processes as

$$\int_0^t \left[S_n^{(2)}(\boldsymbol{\beta}_0, s) - \overline{\boldsymbol{Z}}_n^{\otimes 2}(\boldsymbol{\beta}_0, s) S_n^{(0)}(\boldsymbol{\beta}_0, s) \right] d\Lambda_0(s).$$

By Condition A.3, the integrand converges in probability to

$$s^{(2)}(\boldsymbol{\beta}_0, s) - \boldsymbol{e}^{\otimes 2}(\boldsymbol{\beta}_0, s) s^{(0)}(\boldsymbol{\beta}_0, s),$$

uniformly in $s \in \langle 0, \tau \rangle$. The matrix $c_{kl}(t)$of deterministic limiting functions in Theorem 3.11 has the form

$$\int_0^t \left[s^{(2)}(\boldsymbol{\beta}_0, s) - \boldsymbol{e}^{\otimes 2}(\boldsymbol{\beta}_0, s) s^{(0)}(\boldsymbol{\beta}_0, s) \right] d\Lambda_0(s) = I(\boldsymbol{\beta}_0, t).$$

This is also the variance matrix of the limiting Gaussian process.

The proof of the remaining condition

$$\langle U_{k,\varepsilon}^{(n)}, U_{k,\varepsilon}^{(n)} \rangle(t) \xrightarrow{\mathsf{P}} 0$$

is omitted. □

The next theorem shows that the observed information matrix is a uniformly consistent estimator of the theoretical information matrix.

Theorem 6.5. *Let assumptions A.1–A.5 hold. Let $\widehat{\boldsymbol{\beta}}$ be any consistent estimator of $\boldsymbol{\beta}_0$. Then*

$$\sup_{t \in \langle 0, \tau \rangle} \| \mathcal{I}_n(\widehat{\boldsymbol{\beta}}, t) - I(\boldsymbol{\beta}_0, t) \| \xrightarrow{\mathsf{P}} 0.$$

◇

Proof. We only present a partial proof, with observed information evaluated at the true parameter.

We have

$$\begin{aligned}
\mathcal{I}_n(\beta_0,t) - I(\boldsymbol{\beta}_0,t) &= \frac{1}{n}\int_0^t \left[\frac{S_n^{(2)}(\boldsymbol{\beta},s)}{S_n^{(0)}(\boldsymbol{\beta},s)} - \overline{\boldsymbol{Z}}_n^{\otimes 2}(\boldsymbol{\beta},s)\right] d\overline{N}(s)\\
&\quad - \frac{1}{n}\int_0^t \left[\frac{S_n^{(2)}(\boldsymbol{\beta},s)}{S_n^{(0)}(\boldsymbol{\beta},s)} - \overline{\boldsymbol{Z}}_n^{\otimes 2}(\boldsymbol{\beta},s)\right] d\overline{A}(s)\\
&\quad + \int_0^t \left[\frac{S_n^{(2)}(\boldsymbol{\beta},s)}{S_n^{(0)}(\boldsymbol{\beta},s)} - \overline{\boldsymbol{Z}}_n^{\otimes 2}(\boldsymbol{\beta},s)\right] S_n^{(0)}(\boldsymbol{\beta},s)\,d\Lambda_0(s)\\
&\quad - \int_0^t \left[\frac{s^{(2)}(\boldsymbol{\beta}_0,s)}{s^{(0)}(\boldsymbol{\beta}_0,s)} - \boldsymbol{e}^{\otimes 2}(\boldsymbol{\beta}_0,s)\right] s^{(0)}(\boldsymbol{\beta}_0,s)\,d\Lambda_0(s)
\end{aligned}$$

The difference in the first two terms gives

$$\int_0^t \frac{1}{n}\left[\frac{S_n^{(2)}(\boldsymbol{\beta},s)}{S_n^{(0)}(\boldsymbol{\beta},s)} - \overline{\boldsymbol{Z}}_n^{\otimes 2}(\boldsymbol{\beta},s)\right] d\overline{M}(s) \equiv \int_0^t H^{(n)}(s)\,d\overline{M}(s),$$

which is a martingale integral with zero mean and variance converging to zero. Also, $\int_0^t \sqrt{n}H^{(n)}(s)\,d\overline{M}(s)$ converges weakly to a zero-mean Gaussian process. It follows that $\int_0^t H^{(n)}(s)\,d\overline{M}(s)$ converges to zero in probability uniformly in time.

The difference in the second two terms converges to zero uniformly because the integrand of the first term converges to the integrand of the second term uniformly in time. □

The end of lecture 15 (Dec. 14)

The next theorem states the weak consistency of $\widehat{\boldsymbol{\beta}}$.

Theorem 6.6. *Under conditions A.1–A.5,*

$$\widehat{\boldsymbol{\beta}} \xrightarrow{\mathsf{P}} \boldsymbol{\beta}_0. \qquad \diamond$$

Proof. Consider the log partial likelihood as a random process

$$\ell(\boldsymbol{\beta},t) = \sum_{i=1}^n \int_0^t \left[\boldsymbol{\beta}^\top \boldsymbol{Z}_i(s) - \log nS_n^{(0)}(\boldsymbol{\beta},s)\right] dN_i(s)$$

and define

$$\begin{aligned}
X_n(\boldsymbol{\beta},t) &= \frac{1}{n}\left[\ell(\boldsymbol{\beta},t) - \ell(\boldsymbol{\beta}_0,t)\right]\\
&= \frac{1}{n}\left[\sum_{i=1}^n \int_0^t (\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top \boldsymbol{Z}_i(s)\,dN_i(s) - \int_0^t \log\frac{S_n^{(0)}(\boldsymbol{\beta},s)}{S_n^{(0)}(\boldsymbol{\beta}_0,s)}\,d\overline{N}(s)\right]
\end{aligned}$$

and

$$A_n(\boldsymbol{\beta}, t) = \int_0^t (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top S_n^{(1)}(\boldsymbol{\beta}_0, s)\, d\Lambda_0(s) - \int_0^t \log \frac{S_n^{(0)}(\boldsymbol{\beta}, s)}{S_n^{(0)}(\boldsymbol{\beta}_0, s)} S_n^{(0)}(\boldsymbol{\beta}_0, s)\, d\Lambda_0(s).$$

Then

$$X_n(\boldsymbol{\beta}, t) - A_n(\boldsymbol{\beta}, t) = \int_0^t H(s)\, d\overline{M}(s),$$

where

$$H(s) = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \boldsymbol{Z}_i(s) - \frac{1}{n} \log \frac{S_n^{(0)}(\boldsymbol{\beta}, s)}{S_n^{(0)}(\boldsymbol{\beta}_0, s)}$$

is a bounded and predictable process. We have

$$\mathsf{E}\left[X_n(\boldsymbol{\beta}, t) - A_n(\boldsymbol{\beta}, t)\right] = 0$$
$$\mathsf{var}\left[X_n(\boldsymbol{\beta}, t) - A_n(\boldsymbol{\beta}, t)\right] = \mathsf{E} \int_0^t H^2(s)\, d\overline{A}(s)$$

Since $\mathsf{E} \int_0^\tau H^2(s)\, d\overline{A}(s) \to 0$, we get

$$X_n(\boldsymbol{\beta}, \tau) - A_n(\boldsymbol{\beta}, \tau) \xrightarrow{\mathsf{P}} 0.$$

Also,

$$A_n(\boldsymbol{\beta}, \tau) \xrightarrow{\mathsf{P}} A(\boldsymbol{\beta}, \tau) \equiv \int_0^\tau (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top s^{(1)}(\boldsymbol{\beta}_0, s)\, d\Lambda_0(s) - \int_0^\tau \log \frac{s^{(0)}(\boldsymbol{\beta}, s)}{s^{(0)}(\boldsymbol{\beta}_0, s)} s^{(0)}(\boldsymbol{\beta}_0, s)\, d\Lambda_0(s).$$

It follows that $X_n(\boldsymbol{\beta}, \tau) \xrightarrow{\mathsf{P}} A(\boldsymbol{\beta}, \tau)$. Since X_n is a concave function with a unique maximum at $\widehat{\boldsymbol{\beta}}$ and A is a concave function with a unique maximum at $\boldsymbol{\beta}_0$, we get $\widehat{\boldsymbol{\beta}} \xrightarrow{\mathsf{P}} \boldsymbol{\beta}_0$. □

Now we are ready to state and prove the asymptotic normality of $\widehat{\boldsymbol{\beta}}$.

Theorem 6.7. *Under conditions A.1–A.5,*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{\mathsf{D}} \mathsf{N}_p(\mathbf{0}, I^{-1}(\boldsymbol{\beta}_0, \tau)).$$ ◇

Proof. By the Taylor expansion of $\boldsymbol{U}_n(\widehat{\boldsymbol{\beta}}, \tau)$ around $\boldsymbol{\beta}_0$,

$$\mathbf{0} = \boldsymbol{U}_n(\widehat{\boldsymbol{\beta}}, \tau) = \boldsymbol{U}_n(\boldsymbol{\beta}_0, \tau) + \frac{\partial \boldsymbol{U}_n(\boldsymbol{\beta}^*, \tau)}{\partial \beta^\top}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0).$$

Now,

$$\frac{\partial \boldsymbol{U}_n(\boldsymbol{\beta}^*, \tau)}{\partial \beta^\top} = -n\mathcal{I}_n(\boldsymbol{\beta}^*, \tau),$$

where $\boldsymbol{\beta}^*$ lies on the line segment between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$, and hence $\boldsymbol{\beta}^* \xrightarrow{\mathrm{P}} \boldsymbol{\beta}_0$. We get

$$\frac{1}{\sqrt{n}} \boldsymbol{U}_n(\boldsymbol{\beta}_0, \tau) = \mathcal{I}_n(\boldsymbol{\beta}^*, \tau)\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$

and

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \mathcal{I}_n^{-1}(\boldsymbol{\beta}^*, \tau)\frac{1}{\sqrt{n}} \boldsymbol{U}_n(\boldsymbol{\beta}_0, \tau).$$

Because $\boldsymbol{\beta}^* \xrightarrow{\mathrm{P}} \boldsymbol{\beta}_0$, we get from Theorem 6.5 and from the continuity of matrix inverse that $\mathcal{I}_n^{-1}(\boldsymbol{\beta}^*, \tau) \xrightarrow{\mathrm{P}} I^{-1}(\boldsymbol{\beta}_0, \tau)$. From Theorem 6.4 and from Slutski's Theorem, we get the desired result. □

The following result is an easy corollary to theorems 6.5 and 6.7.

Theorem 6.8. *Let $\boldsymbol{c}$ be any non-zero p-vector of constants. Under conditions A.1–A.5,*

$$\frac{\sqrt{n}(\boldsymbol{c}^\top \widehat{\boldsymbol{\beta}} - \boldsymbol{c}^\top \boldsymbol{\beta}_0)}{\sqrt{\boldsymbol{c}^\top \mathcal{I}_n^{-1}(\widehat{\boldsymbol{\beta}}, \tau)\boldsymbol{c}}} \xrightarrow{\mathrm{D}} \mathrm{N}(0, 1).$$

◇

Consider the hypothesis $H_0 : \boldsymbol{c}^\top \boldsymbol{\beta}_0 = \gamma_0$ tested against the two-sided alternative $H_1 : \boldsymbol{c}^\top \boldsymbol{\beta}_0 \neq \gamma_0$. We reject the hypothesis if

$$\frac{\sqrt{n}\left|\boldsymbol{c}^\top \widehat{\boldsymbol{\beta}} - \gamma_0\right|}{\sqrt{\boldsymbol{c}^\top \mathcal{I}_n^{-1}(\widehat{\boldsymbol{\beta}}, \tau)\boldsymbol{c}}} \geq u_{1-\alpha/2}.$$

According to Theorem 6.8, this test has a level converging to α as $n \to \infty$. In particular, with $\boldsymbol{c}$ having a single component equal to one and all other components equal to zero, and taking $\gamma_0 = 0$, we get Wald-type tests of the individual regression coefficients.

Similarly, we can use Theorem 6.8 to calculate Wald-type confidence intervals for $\boldsymbol{c}^\top \boldsymbol{\beta}_0$ with asymptotic coverage $1 - \alpha$. These intervals have boundary points

$$\boldsymbol{c}^\top \widehat{\boldsymbol{\beta}} \mp \sqrt{\frac{\boldsymbol{c}^\top \mathcal{I}_n^{-1}(\widehat{\boldsymbol{\beta}}, \tau)\boldsymbol{c}}{n}} u_{1-\alpha/2}.$$

Theorem 6.9. *Under conditions A.1–A.5,*

$$\frac{1}{n} \boldsymbol{U}_n(\boldsymbol{\beta}_0, \tau)^\top \mathcal{I}_n^{-1}(\boldsymbol{\beta}_0, \tau)\boldsymbol{U}_n(\boldsymbol{\beta}_0, \tau) \xrightarrow{\mathrm{D}} \chi^2_p.$$

◇

This theorem is an immediate consequence of Theorem 6.4 and Theorem 6.5. It can be used to construct score tests of simple and composite hypotheses about the components of $\boldsymbol{\beta}_0$. For example, to test the hypothesis $H_0 : \boldsymbol{\beta}_0 = \boldsymbol{b}$ against $H_1 : \boldsymbol{\beta}_0 \neq \boldsymbol{b}$ we take the test statistic

$$U = \frac{1}{n}\boldsymbol{U}_n(\boldsymbol{b},\tau)^\top \mathcal{I}_n^{-1}(\boldsymbol{b},\tau)\boldsymbol{U}_n(\boldsymbol{b},\tau)$$

and reject H_0 if $U \geq \chi^2_p(1-\alpha)$.

An important special case arises when the Cox model is set up to compare hazards in two groups of subjects. Take $Z_i = 1$ when the i-th subject belongs to the second group and $Z_i = 0$ when the subject belongs to the first group. The conditional hazard in the Cox model has the form

$$\lambda(t \mid \text{group}) = \lambda_0(t)\mathrm{e}^{\beta_0 Z_i},$$

where λ_0 is the hazard function of the first group and e^{β_0} is the time-invariant hazard ratio between the second and the first group. The hypothesis of interest, $H_0 : \beta_0 = 0$, can be tested by the score test statistic U with critical value $\chi^2_1(1-\alpha)$. The score test statistic can be shown to be the square of the unweighted logrank test statistic discussed in Section 5.2. Thus, the Cox model provides another derivation of the logrank test, as a score test in a proportional hazards model. Within the regression framework, the logrank test can be easily generalized to comparing survival distributions in $K > 2$ groups by performing the score test in a model with $K-1$ dummy regressors factorizing the groups.

The next theorem shows that even likelihood ratio tests work with partial likelihood.

Theorem 6.10. *Suppose conditions A.1–A.5 are fulfilled. Let $\ell_M(\widehat{\boldsymbol{\beta}})$ be the maximized partial log-likelihood in a larger model M and let $\ell_S(\widetilde{\boldsymbol{\beta}})$ be the maximized partial log-likelihood in a submodel S. If the submodel holds then*

$$2[\ell_M(\widehat{\boldsymbol{\beta}}) - \ell_S(\widetilde{\boldsymbol{\beta}})] \xrightarrow{\mathsf{D}} \chi^2_m,$$

where m is the difference in the number of parameters in the larger model and the submodel. ◊

This theorem (which will be left without proof) is used to perform submodel testing when building the regression model. The submodel S is rejected in favor of the larger model M when $2[\ell_M(\widehat{\boldsymbol{\beta}})-\ell_S(\widetilde{\boldsymbol{\beta}})] \geq \chi^2_m(1-\alpha)$. This is a couterpart of the deviance test in generalized linear models.

6.4. Estimation of the baseline hazard and conditional survival

The partial likelihood eliminates the baseline hazard λ_0 and thus carries no information about it. An estimator for the baseline hazard must be developed by other means. It is needed for two reasons. First, to estimate the survival function for a particular subject; second, it appears at several other important quantities that need to be estimated.

An estimator for the cumulative baseline hazard $\Lambda_0(t)$ can be derived from moment considerations. Take the martingale $\overline{M} = \sum M_i = \overline{N} - \int nS_n^{(0)} d\Lambda_0$, where $\overline{N} = \sum N_i$. For any bounded predictable function H, $\sum \int H \, d\overline{M}$ is a martingale and hence

$$0 = \mathsf{E} \int H(d\overline{N} - nS_n^{(0)} d\Lambda_0) = \mathsf{E} \int nS_n^{(0)} H\left(\frac{d\overline{N}}{nS_n^{(0)}} - d\Lambda_0\right).$$

If we take $H = \left(nS_n^{(0)}\right)^{-1}$, we get

$$\Lambda_0(t) = \mathsf{E} \int_0^t \frac{d\overline{N}(s)}{nS_n^{(0)}(\boldsymbol{\beta}_0, s)}.$$

So define

$$\widehat{\Lambda}_0(t) = \int_0^t \frac{d\overline{N}(s)}{\sum_{i=1}^n Y_i(s) \exp\{\widehat{\boldsymbol{\beta}}^\mathsf{T} \boldsymbol{Z}_i(s)\}}.$$

This is called the Breslow estimator of the cumulative baseline hazard (Breslow 1972).

Note. Compare the Breslow estimator to the Nelson-Aalen estimator and note the similarities and differences.

Theorem 6.11. *Let assumptions A.1–A.5 hold. Then*

$$\sqrt{n}[\widehat{\Lambda}_0(t) - \Lambda_0(t)] \Longrightarrow W(\sigma(t)) \quad \text{on } D^p\langle 0, \tau\rangle.$$

The variance of the limiting process is

$$\sigma^2(t) = \int_0^t \frac{d\Lambda_0(s)}{s^{(0)}(\boldsymbol{\beta}_0, s)} + \boldsymbol{Q}(\boldsymbol{\beta}_0, t)^\mathsf{T} I(\boldsymbol{\beta}_0, t) \boldsymbol{Q}(\boldsymbol{\beta}_0, t),$$

where $\boldsymbol{Q}(\boldsymbol{\beta}_0, t) = \int_0^t \boldsymbol{e}(\boldsymbol{\beta}_0, s) \, d\Lambda_0(s)$. ◇

This theorem implies the uniform consistency of the Breslow estimator on $\langle 0, \tau\rangle$ and it allows the construction of confidence intervals for $\Lambda_0(t)$ at fixed t as well as confidence bounds covering the baseline hazard on the whole interval $\langle 0, \tau\rangle$. The limiting variance $\sigma^2(t)$ can be consistently estimated by replacing all the unknown quantities by their consistent estimators.

Suppose that all covariates are time-invariant and consider a subject with an observed covariate vector $\boldsymbol{z}$. The conditional cumulative hazard function for this specific subject is

$$\Lambda(t \mid \boldsymbol{z}) = \Lambda_0(t) \exp\{\boldsymbol{\beta}_0^\top \boldsymbol{z}\},$$

which can be estimated by

$$\widehat{\Lambda}(t \mid \boldsymbol{z}) = \widehat{\Lambda}_0(t) \exp\{\widehat{\boldsymbol{\beta}}^\top \boldsymbol{z}\},$$

The conditional survival function for this subject is

$$S(t \mid \boldsymbol{z}) = \exp\{-\Lambda(t \mid \boldsymbol{z})\} = \exp\left\{-\Lambda_0(t) \exp\{\boldsymbol{\beta}_0^\top \boldsymbol{z}\}\right\},$$

which can be estimated by

$$\widehat{S}(t \mid \boldsymbol{z}) = \exp\left\{-\widehat{\Lambda}_0(t) \exp\{\widehat{\boldsymbol{\beta}}^\top \boldsymbol{z}\}\right\}.$$

Confidence intervals and confidence bounds for the conditional survival can be obtained from Theorem 6.11 by the same approach we used for the Kaplan-Meier estimator.

The end of lecture 16 (Dec. 21)

6.5. Cox model with non-proportional hazards

There are two ways how to incorporate covariates that do not satisfy the proportional hazards assumption: stratification and time-dependent effects. The latter approach also allows to test the validity of the proportional hazards assumption against certain alternatives.

Stratified Cox model

Stratified Cox model consists in fitting different hazard functions within strata defined by the levels of a categorical variable.

Consider a categorical variable V with values $1, \ldots, q$ that affects the hazard function in a non-proportional way. Instead of including such variable in the linear predictor of the standard Cox model, we modify the model formula as follows:

$$\lambda(t \mid \boldsymbol{Z}, V = j) = \lambda_{0j}(t) \exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}(t)\}.$$

Thus, the influence of V on the hazard is expressed by introducing separate baseline hazard functions $\lambda_{0j}(t)$, $j = 1, \ldots, q$, depending on the value of V ("strata") . This is called a *stratified Cox model.* Here, different levels of V have totally unrestricted hazards. The effect of the other covariates is still modeled under the proportional hazards assumption.

Denote the observed data be $(N_{ji}(t), Y_{ji}(t), \boldsymbol{Z}_{ji}(t))$, $j = 1, \ldots, q$, $i = 1, \ldots, n_j$. The index j indicates the stratum (level of V), i indicates subjects within strata. The partial likelihood is taken as a product of standard partial likelihood functions for the individual strata, that is

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{q} L_j(\boldsymbol{\beta}) = \prod_{j=1}^{q} \prod_{i=1}^{n_j} \prod_{s>0} \left[\frac{Y_{ji}(s) \exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_{ji}(s)\}}{\sum_{k=1}^{n_j} Y_{jk}(s) \exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_{jk}(s)\}} \right]^{\Delta N_{ji}(s)}.$$

The score statistic has the form

$$\boldsymbol{U}_n(\boldsymbol{\beta}) = \sum_{j=1}^{q} \sum_{i=1}^{n_j} \int_0^\infty \left[\boldsymbol{Z}_{ji}(t) - \overline{\boldsymbol{Z}}_j(\boldsymbol{\beta}, t) \right] dN_{ji}(t),$$

where

$$\overline{\boldsymbol{Z}}_j(\boldsymbol{\beta}, t) = \frac{\sum_{i=1}^{n_j} Y_{ji}(t) \boldsymbol{Z}_{ji}(t) \mathrm{e}^{\boldsymbol{\beta}^\top \boldsymbol{Z}_{ji}(t)}}{\sum_{i=1}^{n_j} Y_{ji}(t) \mathrm{e}^{\boldsymbol{\beta}^\top \boldsymbol{Z}_{ji}(t)}}.$$

In the score statistic, the covariates of each failing subject are compared only to the covariates of the subjects from the same stratum. The stratified estimator $\widehat{\boldsymbol{\beta}}$ is asymptotically normal as in Theorem 6.7, but the information matrix becomes a linear combination of stratum-specific information matrices (with weights equal to the probabilities of the individual strata).

The stratum-specific cumulative baseline hazard may estimated by an obvious extension of Breslow estimator:

$$\widehat{\Lambda}_{0j}(t) = \int_0^t \frac{d\overline{N}_j(s)}{\sum_{i=1}^{n_j} Y_{ji}(s) \exp\{\widehat{\boldsymbol{\beta}}^\top \boldsymbol{Z}_{ji}(s)\}}.$$

Stratification represents a reasonable strategy to incorporate non-proportionality when the non-proportional variable is discrete with just a few levels and sufficient representation of each level in the data set. Also, we must keep in mind that the effect of the stratification variable cannot be tested or expressed by a finite number of parameters. Sometimes, stratification is used with continuous covariates, which must be dicretized into a relatively small number of levels. Clearly, this approach entails a serious loss of precision.

Modeling non-proportionality by interactions with time

Consider a covariate V that has a non-proportional effect on the hazard. Such an effect can be directly modelled by the Cox model by including interactions of V with time and treating them as time-varying covariates.

Select a set of linearly independent basis functions $g_1(t), \dots, g_r(t)$ and specify the model as

$$\lambda(t \mid \boldsymbol{Z}, V) = \lambda_0(t) \exp\Big\{\boldsymbol{\beta}^\top \boldsymbol{Z}(t) + \beta_V V + \sum_{l=1}^{r} \gamma_l g_l(t) V\Big\}.$$

Here, $\boldsymbol{Z}$ are the other covariates in the model and β_V is the main effect of V. The rest of the linear predictor specifies interactions of V with time transformed by the basis functions g_l. Then, the relative risk (hazard ratio) associated with a unit increase in V is

$$RR_V(t) \equiv \frac{\lambda(t \mid \boldsymbol{Z}, V = v + 1)}{\lambda(t \mid \boldsymbol{Z}, V = v)} = \exp\Big\{\beta_V + \sum_{l=1}^{r} \gamma_l g_l(t)\Big\}.$$

We can test the hypothesis of the proportional hazards assumption against alternatives of non-proportionality expressed as linear combinations of the basis functions $g_l(t)$ by testing $H_0 : \gamma_1 = \cdots = \gamma_r = 0$, for example by a likelihood ratio test. If the hypothesis is rejected, we can keep these interactions terms in the model and describe the time-varying effect of V by the parameters $\gamma_1, \dots, \gamma_r$.

The following special cases might be interesting in practical applications:

- $g(t) = t$
  This adds an interaction of V with linear time. The time-varying relative risk is
  $$RR_V(t) = \mathrm{e}^{\beta_V + \gamma t}.$$
  The relative risk at the time $t = 0$ is e^{β_V}. From that point, the RR increases (for $\gamma > 0$) or decreases (for $\gamma < 0$) exponentially. The test of the hypothesis $H_0 : \gamma = 0$ tests the proportional hazards assumption against the alternative of exponential change in the relative risk.
- $g(t) = \log(t + 1)$
  This adds an interaction of V with logarithmic time. The time-varying relative risk is
  $$RR_V(t) = \mathrm{e}^{\beta_V + \gamma \log(t+1)} = \mathrm{e}^{\beta_V}(t + 1)^\gamma.$$
  Again, the relative risk at the time $t = 0$ is e^{β_V} but the RR then changes as a power function, which is slower than an exponential change. The test of the hypothesis $H_0 : \gamma = 0$ tests the proportional hazards assumption against the alternative of power function change in the relative risk.
- $g_1(t) = \mathbb{1}(s_1 \le t < s_2)$, $g_2(t) = \mathbb{1}(s_2 \le t)$
  Now we factorize time into three intervals: $\langle 0, s_1)$, $\langle s_1, s_2)$, $\langle s_2, \tau)$ and let that interact with V. The relative risk is
  $$\begin{aligned} RR_V(t) &= \mathrm{e}^{\beta_V} && \text{for } t \in \langle 0, s_1), \\ RR_V(t) &= \mathrm{e}^{\beta_V + \gamma_1} && \text{for } t \in \langle s_1, s_2), \\ RR_V(t) &= \mathrm{e}^{\beta_V + \gamma_2} && \text{for } t \in \langle s_2, \tau). \end{aligned}$$

The test of the hypothesis $H_0 : \gamma_1 = \gamma_2 = 0$ tests the proportional hazards assumption against the alternative of piecewise constant relative risk function with breaks at s_1 and s_2.

Of course, we can extend this idea to an arbitrary number of time intervals but we need to have enough failures in each of them to estimate the relative risks separately.

6.6. Generalizations of the Cox model

In this section, we give brief suggestions about other possible generalizations of the Cox model.

Proportional intensity model

Suppose we observe data in the form of independent processes $(N_i(t), Y_i(t), \boldsymbol{Z}_i(t))$, $i = 1, \ldots, n$. Hower, we do not assume that the processes $N_i(t)$ and $Y_i(t)$ arised from a random censorship model. Instead, we allow $N_i(t)$ to be any counting process, giving the number of observed events for subject i until time t. The events can be recurrent and the process $N_i(t)$ may have multiple jumps. The process $Y_i(t)$ is binary and indicates whether an event occurring at t can be observed or not. It can jump repeatedly between 1 and 0 and it does not have to stay at zero after the first observed event. The times between successive jumps in $N_i(t)$ are assumed to have continuous distributions.

Take the right-continuous filtration

$$\mathcal{F}_t = \sigma\{N_i(u), Y_i(u+), \boldsymbol{Z}_i(u+),\ 0 \le u \le t, i = 1, \ldots, n\}$$

and assume that $\boldsymbol{Z}_i(t)$ and $Y_i(t)$ are $\mathcal{F}_t$ predictable. Define the process

$$A_i(t) = \int_0^t Y_i(s) \mathrm{e}^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i(s)} \lambda_0(s)\, ds,$$

the same process that plays the role of a compensator in the Cox model. We say that the data satisfy *the proportional intensity model* (Aalen 1978) if $M_i = N_i - A_i$ is an $\mathcal{F}_t$-martingale, that is, A_i is the right compensator for N_i even under our extended conditions.

It can be shown that, under the proportional intensity model,

$$\lim_{h \searrow 0} \frac{1}{h} \mathrm{P}\big[N_i(t+h) - N_i(t) = 1 \,\big|\mathcal{F}_t\big] = Y_i(s) \mathrm{e}^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i(s)} \lambda_0(s),$$

that is, the parameters $\boldsymbol{\beta}_0$ still express the influence of the covariates on the rate of occurrence of events (even though we can no longer call it the hazard rate in the context of recurring events).

Most of the results of Section 6.3 still hold under this more general model and most of their proofs come through without great changes.

The proportional intensity model can be also used to model left truncation – a case when an event cannot be observed if it occurs before a random entry time of the subject into the study. This is achieved by setting the process Y_i to zero prior to the entry time.

Generalized proportional hazards models

Another possible generalization of the Cox model is achieved by considering a general link function g for the relationship between the linear predictor $\boldsymbol{\beta}_0^\top \boldsymbol{Z}$ and the hazard $\lambda(t \mid \boldsymbol{Z})$. The model can be written as

$$\lambda(t \mid \boldsymbol{Z}(t)) = \lambda_0(t) g(\boldsymbol{\beta}^\top \boldsymbol{Z}(t)),$$

where $g(\cdot)$ is increasing, twice differentiable, and satisfies $g(0) = 1$. This is still a proportional hazards model because the hazard ratio is independent of time. The link function $g(y) = 1 + y$ generates so called *additive relative risk model* $\lambda(t \mid \boldsymbol{Z}(t)) = \lambda_0(t)(1 + \boldsymbol{\beta}^\top \boldsymbol{Z}(t))$. This model is used, e.g., in radiation epidemiology to model the effect of a radiation exposure on cancer occurrence.

The end of lecture 17 (Jan. 4)

A. Appendix

A.1. Useful failure time distributions

Unless stated otherwise, the argument t of densities, distribution functions, survival functions and hazard functions always takes values in the interval $\langle 0, \infty)$.

A.1.1. Exponential distribution

$T \sim \mathsf{Exp}(\lambda)$, $\lambda > 0$

Density: $f(t) = \lambda \mathrm{e}^{-\lambda t}$

Distribution function: $F(t) = 1 - \mathrm{e}^{-\lambda t}$

Survival function: $S(t) = \mathrm{e}^{-\lambda t}$

Hazard function: $\lambda(t) = \lambda$

Expectation: $\mathsf{E}\, T = 1/\lambda$

Mean residual lifetime: $r(t) = 1/\lambda$

Exponential distribution is the only continuous distribution that possesses so called *memoryless property*:

$$\forall s > 0, \forall t > 0: \quad \mathrm{P}\big[T > t + s \,\big|\, T > s\big] = \mathrm{P}\left[T > t\right] = \mathrm{e}^{-\lambda t}.$$

Relationship to Gumbel distribution

Take $U \sim \mathsf{Exp}(1)$ and consider the random variable $W = \log U$, which can take on any real value. The distribution function of W is

$$\mathrm{P}\left[W \le t\right] = \mathrm{P}\left[U \le e^t\right] = 1 - e^{-e^t}, \quad t \in \mathbb{R}.$$

The density of W is

$$f_W(t) = e^{t - e^t}, \quad t \in \mathbb{R}.$$

This distribution is called *the extreme value (Gumbel) distribution.*

Take $T \sim \mathsf{Exp}(\lambda)$. Then $\lambda T \sim \mathsf{Exp}(1)$, $\log \lambda T = W$ and $\log T = -\log \lambda + W$, where W is a Gumbel random variable. Consider the loglinear model $\lambda = \mathrm{e}^{\boldsymbol{\beta}^\top \boldsymbol{Z}}$. Then $\log T$ satisfies the linear model

$$\log T = -\boldsymbol{\beta}^\top \boldsymbol{Z} + W,$$

where W is a random error term distributed according to Gumbel distribution.

A.1.2. Weibull distribution

$T \sim \mathsf{W}(\lambda, \alpha)$, $\lambda > 0$, $\alpha > 0$

Density: $f(t) = \alpha \lambda^\alpha t^{\alpha-1} \mathrm{e}^{-(\lambda t)^\alpha}$

Survival function: $S(t) = \mathrm{e}^{-(\lambda t)^\alpha}$

Hazard function: $\lambda(t) = \alpha \lambda^\alpha t^{\alpha-1}$

Expectation: $\mathsf{E}\, T = \Gamma(1 + \alpha^{-1})/\lambda$

Relationship to exponential distribution

- Let $T \sim \mathsf{W}(\lambda, 1)$. Then $T \sim \mathsf{Exp}(\lambda)$.
- Let $U \sim \mathsf{Exp}(1)$. Define $T = \frac{1}{\lambda} U^{1/\alpha}$. Then $T \sim \mathsf{W}(\lambda, \alpha)$.
- Let $T \sim \mathsf{W}(\lambda, \alpha)$. Then $U = (\lambda T)^\alpha \sim \mathsf{Exp}(1)$.

Relationship to Gumbel distribution

Take $T \sim \mathsf{W}(\lambda, \alpha)$. Then $(\lambda T)^\alpha \sim \mathsf{Exp}(1)$, $\log(\lambda T)^\alpha = W$, and $\log T = -\log \lambda + \alpha^{-1} W$, where W is a Gumbel random variable. Thus, $\log T$ satisfies a location-scale model where $-\log \lambda$ represents the location parameter and $1/\alpha$ represents the scale parameter.

Consider the loglinear model $\lambda = \mathrm{e}^{\boldsymbol{\beta}^\top \boldsymbol{Z}}$. Then $\log T$ satisfies the linear model

$$\log T = -\boldsymbol{\beta}^\top \boldsymbol{Z} + \alpha^{-1} W,$$

where W is a random error term distributed according to Gumbel distribution and α^{-1} controls the variability of the error term.

A.1.3. Gamma distribution

$T \sim \Gamma(a, p)$, $a > 0$, $p > 0$

Density: $f(t) = \dfrac{a^p}{\Gamma(p)} t^{p-1} \mathrm{e}^{-at}$

Expectation: $\operatorname{E} T = \dfrac{p}{a}$

Survival function: $S(t) = 1 - \mathrm{IG}(p, at)$, where $\mathrm{IG}(p,t) = \dfrac{1}{\Gamma(p)} \int_0^t x^{p-1} \mathrm{e}^{-x}\, dx$ is the incomplete Gamma function.

Hazard function: Does not have a tractable form. When $p > 1$ then $\lambda(0) = 0$, $\lambda(t)$ is increasing, and $\lim_{t\to\infty} \lambda(t) = a$. When $p < 1$ then $\lambda(0) = \infty$, $\lambda(t)$ is decreasing, and $\lim_{t\to\infty} \lambda(t) = a$.

Relationship to exponential distribution

- Let $T \sim \Gamma(a, 1)$. Then $T \sim \mathrm{Exp}(a)$.

A.1.4. Raleigh distribution

Density: $f(t) = (\lambda_0 + \lambda_1 t)\, \mathrm{e}^{-\left(\lambda_0 t + \frac{1}{2}\lambda_1 t^2\right)}, \qquad \lambda_0 > 0, \lambda_1 > 0$

Survival function: $S(t) = \mathrm{e}^{-\left(\lambda_0 t + \frac{1}{2}\lambda_1 t^2\right)}$

Hazard function: $\lambda(t) = \lambda_0 + \lambda_1 t$

A.1.5. Gompertz distribution

Density: $f(t) = \lambda_1 \exp\left\{-\frac{\lambda_1}{\lambda_2}\left(\mathrm{e}^{\lambda_2 t} - 1\right) + \lambda_2 t\right\}, \qquad \lambda_1 > 0, \lambda_2 > 0$

Survival function: $S(t) = \exp\left\{-\frac{\lambda_1}{\lambda_2}\left(\mathrm{e}^{\lambda_2 t} - 1\right)\right\}$

Hazard function: $\lambda(t) = \lambda_1 \mathrm{e}^{\lambda_2 t}$

A.1.6. Log-logistic distribution

Density: $f(t) = \kappa\varrho \dfrac{(\varrho t)^{\kappa-1}}{\left[1 + (\varrho t)^\kappa\right]^2}, \qquad \varrho > 0, \kappa > 0$

Survival function: $S(t) = \dfrac{1}{1 + (\varrho t)^\kappa}$

Hazard function: $\lambda(t) = \kappa\varrho^\kappa \dfrac{t^{\kappa-1}}{1 + (\varrho t)^\kappa}$

A.1.7. Geometric distribution

$T \sim \mathsf{Geo}(p),\ p \in (0,1)$

This is a discrete distribution with values $0, 1, 2, \ldots$

Density: $\mathsf{P}[T = t] = p(1-p)^t, \quad t = 0, 1, 2, \ldots$

Expectation: $\mathsf{E}\,T = \dfrac{1-p}{p}$

Survival function: $S(t) = (1-p)^{[t]+1}$, $t > 0$, where $[t] = \max\{j \in \mathbb{Z} : j \leq t\}$ is the lower whole part of the real argument t

Hazard function: $\lambda(t) = p$, $t = 0, 1, 2, \ldots$

Relationship to exponential distribution

- Let $U \sim \mathsf{Exp}(\lambda)$. Then $T = [U] \sim \mathsf{Geo}(p)$, where $p = 1 - e^{-\lambda}$.

Geometric distribution is the only discrete distribution that possesses the *memoryless property*:

$$\forall s > 0, \forall t > 0: \quad \mathsf{P}\big[T > t + s \,\big|\, T > s\big] = \mathsf{P}[T > t] = (1-p)^{[t]+1}.$$

A.2. Results from mathematical analysis and martingale theory

A.2.1. Integration by parts for Lebesgue-Stieltjes integral

Theorem A.1. *(Fleming & Harrington, Theorem A.1.2) Let $F : \langle 0, \infty) \to \mathbb{R}$ and $G : \langle 0, \infty) \to \mathbb{R}$ be right-continuous functions of bounded variation on any finite interval. Let $\Delta F(x) = F(x) - F(x-)$, $\Delta G(x) = G(x) - G(x-)$. Then*

$$\begin{aligned} F(t)G(t) - F(0)G(0) &= \int_0^t F(x-)dG(x) + \int_0^t G(x)\,dF(x) \\ &= \int_0^t F(x-)dG(x) + \int_0^t G(x-)\,dF(x) + \sum_{0<x\leq t} \Delta F(x)\Delta G(x). \end{aligned}$$ ◇

Note.

$$\int_0^t F(x)dG(x) = \int_0^t F(x-)dG(x) + \sum_{0<x\leq t} \Delta F(x)\Delta G(x).$$

A.2.2. Random processes and martingales

Consider a probability space $(\Omega, \mathcal{F}, P)$.

Definition A.1. A family $\{\mathcal{F}_t : t \geq 0\}$ of sub-σ-algebras of a σ-algebra $\mathcal{F}$ is called a *filtration* if, for all $s \leq t$, $\mathcal{F}_s \subset \mathcal{F}_t$. ∇

Definition A.2. Let $\{\mathcal{F}_t : t \geq 0\}$ be a filtration. A random process $X(t)$, $t \geq 0$, is called *adapted to the filtration* $\mathcal{F}_t$ if $X(t)$ is $\mathcal{F}_t$-measurable for any $t \geq 0$. ∇

Notation.

- $X(t-) = \lim_{h \searrow 0} X(t-h)$
- $\mathcal{F}_{t-} = \sigma\left\{\bigcup_{h>0} \mathcal{F}_{t-h}\right\}$

Definition A.3. Let $X(t)$, $t \geq 0$, be a right-continuous process with left-hand limits and let $\{\mathcal{F}_t : t \geq 0\}$ be a filtration. Let $X(t)$ be adapted to $\mathcal{F}_t$ and $\mathsf{E}\,|X(t)| < \infty$ for all $t < \infty$.

(i) X is called a *martingale with respect to the filtration* $\mathcal{F}_t$ if $\mathsf{E}\big[X(t+s)\big|\mathcal{F}_t\big] = X(t)$ almost surely for all $s \geq 0$, $t \geq 0$.
(ii) X is called a *submartingale with respect to the filtration* $\mathcal{F}_t$ if $\mathsf{E}\big[X(t+s)\big|\mathcal{F}_t\big] \geq X(t)$ almost surely for all $s \geq 0$, $t \geq 0$. ∇

Note.

- Let $X(t)$ be an $\mathcal{F}_t$-martingale with $X(0) = 0$ a.s. Then $\mathsf{E}\,X(t) = 0$ for all $t \geq 0$.
- Let $X(t)$ be an $\mathcal{F}_t$-martingale. Then $\mathsf{E}\big[X(t)\big|\mathcal{F}_{t-}\big] = X(t-)$ a.s.

Definition A.4. A process $X(t)$ is called *predictable with respect to the filtration* $\mathcal{F}_t$ if it is measurable with respect to the smallest σ-algebra on $\mathbb{R}_0^+ \times \Omega$ generated by left continuous $\mathcal{F}_t$-measurable processes. ∇

Note. An equivalent definition of predictability is this: $X(t,\omega)$ is $\mathcal{F}_t$-predictable if and only if it is a mapping $\langle 0, \infty) \times \Omega \to \mathbb{R}$, which is measurable with respect to the *predictable σ-algebra*

$$\sigma\big\{\{0\} \times A : A \in \mathcal{F}_0,\ (t,s\rangle \times A : t < s \in \mathbb{R}_0^+, A \in \mathcal{F}_t\big\}.$$

Note. A left continuous $\mathcal{F}_t$-measurable process $A(t)$ is predictable with respect to $\mathcal{F}_t$.

Definition A.5. An $\mathcal{F}_t$-measurable process $\{N(t) : t \geq 0\}$ is a *counting process* if $N(0) = 0$, $N(t) < \infty$ a.s., and almost all its paths are right-continuous and piecewise constant with jumps of size 1. ∇

Note. A counting process is a submartingale.

A.3. Brownian motion

A.3.1. Standard Brownian motion

The Brownian motion (also called the Wiener process) is a random process $W(t)$, $t \in \langle 0, \infty)$, that satisfies the following requirements:

(i) $W(0) = 0$ almost surely;
(ii) almost all paths of $W(t)$ are continuous;
(iii) for any $n > 1$ and $0 \le t_1 < t_2 < \cdots < t_n$, $W(t_1)$, $W(t_2) - W(t_1)$,…,$W(t_n) - W(t_{n-1})$ are independent (independent increments);
(iv) for any $0 \le t < s$, $W(s) - W(t) \sim \mathsf{N}(0, s - t)$.

The Brownian motion has a number of additional interesting properties:

- At all $t \ge 0$, $\mathsf{E}\, W(t) = 0$, $\mathsf{var}\, W(t) = t$.
- At all $s, t \ge 0$, $\mathsf{cov}\,(W(t), W(s)) = s \wedge t$.
- If $W(t)$ is a Brownian motion then $\sigma^{-1/2} W(\sigma t)$ is also a Brownian motion.
- $W(t)$ is a martingale with respect to its history; its predictable variation process is $\langle W, W\rangle(t) = t$.
- The sample paths of $W(t)$ are not differentiable at any t a.s.
- The sample paths of $W(t)$ do not have bounded variation on any interval.

A.3.2. Time-transformed Brownian motion

The process $V = \int f\, dW$ is called *time-transformed Brownian motion.* It has all the properties of a Brownian motion except variance function. Its variance function is $\mathsf{var}\, V(t) \equiv h(t) = \int_0^t f^2(s)\, ds$. When $f(s) \equiv 1$, the time-transformed Brownian motion is a standard Brownian motion.

The variance function $h(t)$ can be viewed as a non-decreasing time transformation. We can obtain the time-transformed Brownian motion as $V(t) = W(h(t))$, where W is a standard Brownian motion.

A.3.3. Brownian bridge

Brownian bridge $B(t)$ is a stochastic process defined on the interval $\langle 0,\ 1\rangle$, with values $B(0) = B(1) = 0$. It can be obtained from the standard Brownian motion by the transformation

$$B(t) = W(t) - tW(1), \quad t \in \langle 0,\ 1\rangle.$$

Brownian bridge is a Gaussian process with zero mean and variance function $\mathsf{var}\, B(t) = t(1 - t)$. The covariance function for $s < t$ is $\mathsf{cov}\,(B(s), B(t)) = s(1 - t)$.

A.4. Weak convergence of stochastic processes

In this part we review main features of weak convergence of stochastic processes, in particular convergence of processes with right-continuous sample paths with left-hand limits defined on the interval $\langle 0, \tau \rangle$. The space of such functions is denoted $D\langle 0, \tau \rangle$.

Take a metric space $\mathcal{X}$ and the smallest σ-algebra $\mathcal{B}$ that includes all the open sets contained in $\mathcal{X}$. A stochastic process with sample paths belonging to $\mathcal{X}$ is a measurable mapping $(\Omega, \mathcal{A}) \to (\mathcal{X}, \mathcal{B})$.

The metric that defines open sets on $D\langle 0, \tau \rangle$ is called Skorokhod metric. Let Φ be the set of all strictly increasing continuous functions f mapping $\langle 0, \tau \rangle$ onto $\langle 0, \tau \rangle$, so that $f(0) = 0$ and $f(\tau) = \tau$.

Definition A.6. For any $g, h \in D\langle 0, \tau \rangle$ define

$$d(g,h) = \inf\Big\{\varepsilon > 0 : \exists f \in \Phi \text{ s.t. } \sup_{t \in \langle 0,\tau\rangle} |f(t) - t| \le \varepsilon \text{ and } \sup_{t \in \langle 0,\tau\rangle} |g(t) - h(f(t))| \le \varepsilon\Big\}.$$

The distance d is called *Skorokhod distance.* ∇

This is almost the supremal distance except that the two functions are evaluated at slightly different arguments. Skorokhod distance defines a topology of open sets on $D\langle 0, \tau \rangle$; let $\mathcal{B}^*$ be the smallest σ-algebra containing all such open sets. The Skorokhod topology can be metrized by another metric, which makes the space $(D\langle 0, \tau \rangle, \mathcal{B}^*)$ complete and separable. A stochastic process with sample paths contained in $D\langle 0, \tau \rangle$ is a measurable mapping $(\Omega, \mathcal{A}) \to (D\langle 0, \tau \rangle, \mathcal{B}^*)$.

Definition A.7. Let P_n and P be probability measures on $(\mathcal{X}, \mathcal{B})$. We say that P_n *converges weakly* to P as $n \to \infty$, (denoted $P_n \Longrightarrow P$), if and only if $P_n(A) \to P(A)$ for any $A \in \mathcal{B}$ such that $P(\partial A) = 0$, where ∂A is the boundary of the set A. ∇

If the sample space $\mathcal{X}$ is $\mathbb{R}^d$, weak convergence coincides with convergence in distribution of a random vector $\boldsymbol{X}_n$ to a multivariate distribution P.

Theorem A.2 (Continuous mapping theorem). *Let h be a continuous mapping from a metric space $(\mathcal{X}, \mathcal{B})$ to another metric space $(\mathcal{X}', \mathcal{B}')$, let $P_n \Longrightarrow P$ on $(\mathcal{X}, \mathcal{B})$. Then*

$$P_n h^{-1} \Longrightarrow P h^{-1}$$

on $(\mathcal{X}', \mathcal{B}')$. $\diamond$

Let $X_1, X_2, \dots$ be a sequence of stochastic processes on $(D\langle 0, \tau \rangle, \mathcal{B}^*)$, let X be a stochastic process on $(D\langle 0, \tau \rangle, \mathcal{B}^*)$ such that $X_n \Longrightarrow X$. Take any $k \ge 1$ and select time points $t_1, \dots, t_k \in \langle 0, \tau \rangle$. The mapping that assigns to any function $f \in D\langle 0, \tau \rangle$ the k-vector of values $(f(t_1), \dots, f(t_k))$ is continuous with respect to the Skorokhod metric. It follows from the continuous mapping theorem that the random vector $(X_n(t_1), \dots, X_n(t_k))^\mathsf{T}$ converges in distribution to $(X(t_1), \dots, X(t_k))^\mathsf{T}$. This is called the convergence of finite-dimensional distributions. It is a necessary but not sufficient condition for weak convergence of stochastic processes.

Note. It can be shown that, for $X \in D\langle 0, \tau \rangle$, the mapping $X \to \sup_{t \in \langle 0, \tau \rangle} |X(t)|$ is continuous with respect to the Skorokhod metric. It follows from the continuous mapping theorem that if $X_n \Longrightarrow X$ then

$$\sup_{t \in \langle 0, \tau \rangle} |X_n(t)| \xrightarrow{\mathsf{D}} \sup_{t \in \langle 0, \tau \rangle} |X(t)| \,.$$

Definition A.8. A collection P_n of probability measures on a metric space $(\mathcal{X}, \mathcal{B})$ is called *tight* if for any $\varepsilon > 0$ there exists a compact set $K \subset \mathcal{X}$ such that $P_n(K) > 1 - \varepsilon$ for all n. ∇

Theorem A.3. *Let $(\mathcal{X}, \mathcal{B})$ be a complete and separable metric space. Let P_n and P be probability measures on $(\mathcal{X}, \mathcal{B})$. Then*

$$P_n \Longrightarrow P$$

if and only if both of the following conditions hold:

1. *All finite-dimensional distributions of P_n converge to the respective finite-dimensional distributions of P.*
2. *The collection P_n is tight.* $\diamond$

For stochastic processes in $D\langle 0, \tau \rangle$, there is a sufficient condition for tightness, which goes as follows.

Theorem A.4. *The sequence of stochastic processes X_n with sample paths in $D\langle 0, \tau \rangle$ satisfies the tightness condition if for any $\varepsilon > 0$*

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \mathrm{P} \left[\sup_{|s-t| < \delta} |X_n(s) - X_n(t)| > \varepsilon \right] = 0.$$

$\diamond$

Bibliography

Aalen, O. (1978). Nonparametric inference for a family of counting processes, *Annals of Statistics* **6**: 701–726.

Andersen, P., Borgan, O., Gill, R. and Keiding, N. (1993). *Statistical models based on counting processes*, Springer Verlag, New York.

Billingsley, P. (1999). *Convergence of Probability Measures*, 2nd edition edn, John Wiley & Sons, Inc., New York.

Breslow, N. (1972). Contribution to the discussion on the paper by D. R. Cox, Regression and life tables, *Journal of the Royal Statistical Society, Series B* **34**: 216–217.

Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship, *The Annals of Statistics* **2**(3): 437–453.

Cox, D. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**: 187–220.

Fleming, T. and Harrington, D. (1981). A class of hypothesis tests for one and two samples of censored survival data, *Communications in Statistics* **10**: 763–794.

Fleming, T. and Harrington, D. (1991). *Counting Processes and Survival Analysis*, John Wiley & Sons, Inc., New York.

Gehan, E. (1965). A generalized Wilcoxon test for comparing arbitrarily single-censored samples, *Biometrika* **52**: 203–223.

Gill, R. D. (1980). *Censoring and stochastic integrals*, number 124 in *Mathematical Centre Tracts*, Mathematisch Centrum, Amsterdam.

Hall, W. J. and Wellner, J. A. (1980). Confidence bands for a survival curve from censored data, *Biometrika* **67**(1): 133–143.

Harrington, D. and Fleming, T. (1982). A class of rank test procedures for censored survival data, *Biometrika* **69**: 133–143.

Kaplan, E. and Meier, P. (1958). Nonparametric estimator from incomplete observations, *Journal of the American Statistical Association* **53**: 457–481.

Lehmann, E. (1975). *Nonparametrics. Statistical methods based on ranks*, Holden-Day, San Francisco.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration, *Cancer Chemotherapy Review* **50**: 163–170.

Nelson, W. (1969). Hazard plotting for incomplete failure data, *J. Qual. Technol.* **1**: 27–52.

Prentice, R. L. (1978). Linear rank tests with right censored data, *Biometrika* **65**: 167–179.

Savage, I. (1956). Contributions to the theory of rank order statistics—the two-sample case, *Annals of Mathematical Statistics* **27**: 590–615.

Index