Notes on Counting Process in Survival Analysis

Mai Zhou

Distribution of some statistics, especially rank based, with censored data are hard to get. Even if they can be worked out for small samples, they are likely to be prohibitively expensive for larger sample sizes. Therefore we need to work on the asymptotic distribution - which often works well for moderate and large sample sizes.

This notes is intended to be read along with Fleming and Harrington's book *Counting Processes and Survival Analysis*. It is not intended as a rigorous treatment of the subject of counting process martingale. The aim is to (1) present intuitions to help visualize the counting process and (2) supply simplified proofs (in special cases, or with more assumptions, perhaps), make the book easier to read and (3) select the materials that can be realistically covered in one semester or even one quarter. The students need only to have had background in Poisson process and basic probability/statistics. Some prior knowledge of survival analysis will help understand the example/application of counting process in the survival analysis.

There is an equally good book on the counting processes: by Andersen, Borgan, Keiding, Gill *Statistical Models Based on Counting Processes*, Springer 1993 This book contains many more examples of application. The book also has more materials then can be covered in one semester. At a math level similar to FH book. Chapter 5 of Kalbflisch and Prentice, second edition is similar to this note.

Review of Master level Survival Analysis. Review of central limit theorem (Linderberg condition) and Poisson process (outlined below).

1 Poisson Process and Its Properties

Review of stochastic processes.

Notation: We shall denote a Poisson process as $N^p(t)$, and a general counting process as $N(t)$.

For fixed t, $N^p(t)$ is a Poisson random variable with parameter λt. For t in $[0, T)$, it is a process with jump size = 1. Sample path (that is, for fixed ω, and changing t) is increasing, piecewise constant, with jumps of size 1 and Right Continuous with Left Limits. For every time-interval $[s, s+d)$, the increment of $N^P(t)$ on this interval, $N^P(s+d) - N^P(s)$, is a Poisson random variable with parameter λd and $N^P(s + d) - N^P(s)$ is independent of $N^P(s)$.

(This is *Stationary* and *independent increments.*)

The inter-arrival time for Poisson process are independent exponential λ r.v.s. (how to derive this from above definition?)

Filtration $\mathcal{F}_t$, is a family of increasing (wrt t) σ-algebras. Used to represent historical information (up to time t) and compute conditional expectations.

Example: $\mathcal{F}_t = \sigma\{X_1, \cdots X_k\}$ if $N(t) = k$

Example: $\mathcal{F}_t = \sigma[X(s),\ 0 \le s \le t]$ the history of the process $X(s)$ up to time t. (where it jumps, by how much, in the past).

Homework: Verify that $N^p(t) - \lambda t = M(t)$ is a martingale. Be sure to define the filtration $(\mathcal{F}_t)$.

$N^p(2t)$ is also a Poisson process. (accelerate time by 2).

$N_1^p(t) + N_2^p(t)$ is also a Poisson process, if both of them are and are independent of each other.

1.1 The intensity of a Poisson process

We call λ the intensity, λt the cumulative intensity. Intuition. It is the (conditional) mean number of hits in 1 unit of time. Unconditionally, we have $EN^p(t) = \lambda t$.

Later, intensity can change over time, we must talk about the intensity AT TIME s.

For an interval $(s, s + d]$,

$$E[N^p(s + d) - N^p(s)|N^p(s) \text{ or } \mathcal{F}_s] = E[N^p(s + d) - N^p(s)] = \lambda \times d .$$

It do not matter if we condition or not here (because of independent increments), but we will need condition for more general processes. Also it do not matter which s we use or what value d we use here, but in the counting process case (next section), the resulting λ will be depend on s and only for small d (as $d \to 0$).

The following process

$$N^p(t) - \lambda t = M(t) = N^p(t) - \int_0^t \lambda ds$$

always has mean zero. In fact, it is a martingale in t.

The waiting times between jumps are iid exponential (λ) random variables. Average (expected) weighting time is $1/\lambda$.

Example 1 Actually in WWII the number of hits of V2 bombs in the City of London, within a block of street is like a Poisson process in time. Or consider area in 2d, a Poisson field.

Example 2 World Cup soccer FIFA 2002 in Korea-Japan. Number of goals in each game (90min regular time) is approximately a Poisson random variable. How about number of goals up to time t $(t < 90)$?

1.2 How to estimate the intensity λ?

If you are given n iid Poisson processes: $N_i^p(t)$, $0 \le t \le T$, how should we estimate the unknown λ?

One possibility: since $N_i^p(1)$ are iid Poisson λ r.v.s (with mean λ) we can estimate λ by the mean of $N_i^p(1)$.

Or by the mean of $N_i^p(d)$ divide by d, or by the mean of $N_i^p(s + d) - N_i^p(s)$ divided by d.

Or by the conditional mean of $N_i^p(s+d) - N_i(s)$ divided by d, conditioning on $\mathcal{F}_s$. (independent increment property makes the conditioning irrelevant, but we need conditioning in more general counting processes).

To get a better estimator here, you would like to use the mean of $N_i^p(T)$ divided by T, (for smaller variance). (You can do that since Poisson always has λ intensity) but in the case when the intensity may change over time, then we have to use just the small interval around time t to estimate the intensity at time t, $\lambda(t) = g'(t)$.

2 Generalizing Poisson Process: Counting Processes

Suppose $N^p(t)$ is a Poisson process, we are going to look at generalzations of the form $\int_0^t f(s) dN^p(g(s))$. Notice if $f(s) = 1$ and $g(s) = s$ then we get back the Poisson process after integration: $N^p(t) = \int_0^t 1 dN^p(s)$.

The function $g(t)$ is the time accelaration/deccelaration which we study in detail in this section. The integration wrt $f(t)$ is changing the jump size, and will be studied in next section.

Counting process generalizes the Poisson process by abandon the assumption of exponential interarrival time. Therefore the jumps of counting process are still of size one, only the waiting times are no longer iid exponential r.v.s.

We shall in next section, generalize it to the case where the size of the jump can be different from one – integration.

There are at least two natural ways to generalize the Poisson process to counting process.

The first is to relax the distributional assumption of the inter-arrival times. Generalization I: inter-arrival times, $X_1, X_2, \cdots$, they are just positive random variables, instead of iid exponential r.v.s. (if they are iid with some distribution, then $N(t)$ is called a renewal process).

We may allow the distribution of X_k to depend on $X_1, X_2, \cdots, X_{k-1}$. The distribution of X_k may also depend of external covariates, or random variables.

The second approach to generalize a Poisson process to counting process is time acceleration/deceleration or crazy clock change, which we will be discussing in more detail below.

These two generalizations are equivalent and we shall adopt the view of time accelaration/deccelaration.

2.1 Accelerated/decelerated Poisson Process

Imaging you are in a hale storm, the number of hits your car gets can be modeled as a Poisson process with λ. Without loss of generalarity assume $\lambda = 1$.

Suppose someone had pre-recorded the process on a vedio tape or on a DVD disk. And you are watching it on the TV as it plays back.

One way to get more hits (at least on average) is to push the fast forward button on the VCR/DVD player. — that is time acceleration! Or time change. (Suppose the entire episode is pre recorded). If you slow it down you get fewer hits in a given time period. You fast forward it, you get (potentially) more hits.

Here we are counting the number of hits as a function of the REAL time (or calendar time), not the tape time or DVD time, (the number of hits at a given tape time never changes).

Imaging a girl/boy sitting in front of the VCR/DVD player pushing the buttons sometimes 2X, some times 4X and sometimes in slow motion. Actually this can be continuous (stepless change). Call the time acceleration dial the $g'(t)$, the possible values are $[0, \infty)$ The only requirement is that he/she has no insight into future. (for example he/she cannot FF, knowing that the event is not coming in the next 5 min. or pause the VCR/DVD just before a hit is coming, i.e. he/she did not watch the Tape/DVD before hand)

How much total expected number of hits you will get in 10min (say) if we know when he pushed what button?

Notice the reference time here is always the calendar time. The time changed Poisson process is $N(g(t))$ where $g(t)$ is a monotone increasing function representing the time flow.

Since $N^p(t)$ (the actual hale storm) is assumed to be a standard Poisson process, and what we watched, with time change, is

$$N(t) = N^p(g(t)).$$

The compensator of the above process is clearly $g(t)$, since the cumulative time passed is $g(t)$. The intensity function of this 'extended Poisson' or time-changed Poisson process is $g'(t)$. The function $g(t)$ may be called cumulative intensity. [this can be proved mathematically but it should be clear from the time accumulation point of view]

Cumulative intensity is $g(t) = \int_0^t g'(s)ds = \int_0^t \lambda(s)ds$.

$\lambda(t)$ is called the intensity function of the counting process. It is the speed of VCR at time t.

Theorem 1 *If you play the dial $g'(t)$ according to the hazard function of a CDF $F(t)$ then the inter-arrival time will be distributed according to F.*

PROOF. We compute the distribution of the first waiting time X_1. $P(X_1 > t)$ is equal to $P(N^p(g(t)) = 0) = e^{-g(t)}$ (due to the Poisson distribution). Since $g'(t) = h(t)$, we have $g(t) = H(t) = -\log(1 - F(t))$. Therefore $P(N^p(g(t)) = 0) = 1 - F(t)$. QED

Remark: For the waiting time of the first event, you turn the dial like $h(t)$ starting from $t = 0$ until the first event happen For the waiting time of the 2nd (and subsequent) event, you need to turn the dial according to $h_2(t)$ starting immediately after the arrival of the first event. (i.e. the starting time is random).

In general, if you turn the dial like $h_i(t)$ in the ith waiting time interval, the ith waiting time will be distributed like $F_i(t)$.

Censoring – as thinning

Let $X > 0$ be the failure time and $C > 0$ be the follow-up time. Assume they are independent. Let $Z = \min(X, C)$.

The counting process based on the Z is $N(t) = I[Z \leq t]$. The intensity for this (one jump) counting process is $h_z(t)I[Z \geq t]$.

Censor is to split the event of jump into two types: a death or a censor, indicated by $\delta = I[X \leq C] = I[Z = X]$.

After the thinning, we have two counting processes. $N^x(t) = I[Z \leq t, \delta = 1]$ and $N^c(t) = I[Z \leq t, \delta = 0]$. They have intensity $h_x(t)I[Z \geq t]$ and $h_c(t)I[Z \geq t]$ respectively.

Notice we have $N^x(t) + N^c(t) = N(t)$, and $h_x(t)I[Z \geq t] + h_c(t)I[Z \geq t] = h_z(t)I[Z \geq t]$.

2.2 Intensity Function [formula]

When $g(t)$ is non-random, the increment of the counting process $N(t) = N^p(g(t))$ over a small interval $N(s+d) - N(s) = N^p(g(s+d)) - N^p(g(s))$ is a Poisson r.v. with parameter $g(s+d) - g(s)$; which is $\approx g'(s)d$ for small d. The value $g'(s)$ determine (in distribution) the number of hits in this interval, and is called the intensity function (at time s).

What about a random (but predictable) $g'(t)$? When you are in a small interval, and conditional on the history, the (predictable) random $g'(t)$ is like non-random. [a function $f(t)$ is $\mathcal{F}_t$ predictable if $f(t + dt)$ is $\mathcal{F}_t$ measurable. All left continuous functions are predictable if they are adapted to $\mathcal{F}_t$].

So we have the same result except we have to conditioning on history:

$$P([N(g(s + d)) - N(g(s))] = 0|N(g(s)) or \mathcal{F}_s) = e^{-[g(s+d)-g(s)]} = e^{-g'(s)d} \approx 1 - g'(s)d$$

or

$$P(N(g(s + d)) - N(g(s)) = 1|N(g(s)) or \mathcal{F}_s) = [g(s+d) - g(s)]e^{-[g(s+d)-g(s)]} \approx g'(s)d$$

for small d. Therefore,

$$P([N^p(g(s + d)) - N^p(g(s))] = 1|\mathcal{F}_s) \approx g'(s)d$$

The probability in the above can be replaced by expectation (since Poisson distribution):

$$E\{N^p(g(s + d)) - N^p(g(s))|\mathcal{F}_s) \approx g'(s)d \ .$$

Let us put this as a theorem.

Theorem 2 For random but predictable (wrt $\mathcal{F}_t$) intensity function $g'(t)$, we have, for small d:

$$E\{N(s+d) - N(s)|\mathcal{F}_s\} = E\{N^p(g(s+d)) - N^p(g(s))|\mathcal{F}_s\} = g(s+d) - g(s) \approx g'(s)d .$$

As d shrinks down to zero this becomes an equality. Or

$$\lim_{d \searrow 0} E\{N(s+d) - N(s)|\mathcal{F}_s\}/d = g'(s) = \lambda(s) .$$

Take one more expectation and sum over small d, we have $EN(t) = Eg(t)$.

Remark: about predictability (no insight into future). Why we are not allowed to look into future? This will destroy the property in Theorem 2. Consider the following example.

Example: If a boy can look into future, then he may leave the intensity dial at $\lambda > 0$ until 0.5 second BEFORE the first event arrives, and then set the intensity to zero. Or set the intensity dial to zero at $t = X/2$, if at the time he can see what $X =$ waiting time of first event, is. This will make the $N(t) = 0$ always (since the first event will never arrive. But $Eg(t) > 0$. In the second case, clearly $Eg(t) = 1/(2\lambda) > 0$.

1. Define a history filtration F_t.

2. condition on $\mathcal{F}_{t-}$, if the value of the function $g'(t)$ is fixed, then $g'(t)$ is a predictable random function (wrt $\mathcal{F}_t$).

2' (Discrete version) condition on $\mathcal{F}_{(s-d)}$, if the value of the function $g'(s)$ is fixed and no longer random.

Then the function is called predictable wrt history $\mathcal{F}_t$.

In the continuous case, predictable means $g(t)|\mathcal{F}_{t-} =$ constant or $g'(t)$ is $\mathcal{F}_{t-}$ measurable. Or $g'(t+)$ is $\mathcal{F}_t$ measurable.

(Regular) Poisson process has constant intensity, it is related to exponential random variable (the inter-arrival time). When the inter-arrival time is not exponential, the intensity changes. (In what pattern?) For renewal processes this changes in a cyclic pattern, with the arrival of next hale as the end of a cycle, and also the beginning of next cycle.

The conditional mean of $N(t)$ minus the cumulative intensity is still zero for predictable intensity! We can conditioning one small interval at a time.

$E([N(s+d) - N(s)]|or\mathcal{F}_s N(s))$

Summation of several independent Poisson processes is again a Poisson process.

3 One jump counting process

Define $N(t) = I_{[X \leq t]}$ where X is a positive random variable, then it is a one jump process.

It is an extended (or time changed) Poisson process. What is the intensity of this extended Poisson process? There is just one cycle. The intensity will be zero after the first and only hit, since there can be no more hit other than the first one. And the waiting time for the hit is X. Using the result of previous two sections, the intensity of $N(t)$ is seen to be $h(t)I_{[X \geq t]}$.

Claim: The intensity function of the above counting process $N(t) = I_{[X \leq t]}$ is

$$h(t)I_{[X \geq t]} ;$$

and therefore

$$N(t) - \int_0^t h(s)I_{[X \geq s]}ds$$

is a martingale wrt filteration $\mathcal{F}_t$.

The rest of this section is an alternative, direct proof of the above fact and some further examples. An easier fact to verify is

$$EN(t) = F_X(t) = E \int_0^t h(s)I_{[X \geq s]}ds \ .$$

What if we force the intensity down to zero at $t = 4$, regardless of if the event has happened or not. $h(t)I_{[X \geq t]}I_{[t \leq 4]}$. (censoring at 4)

An alternative direct proof: For $N(t) = I_{[X \leq t]}$, if $N(s) = 1$ then the probability of more jump in $(s, s+d]$ is zero. If $N(s) = 0$ (not yet jumped) then

$$P(N(s+d) - N(s) = 1|N(s) = 0) = P(s < X \leq s + d|X > s) = \frac{F(s+d) - F(s)}{1 - F(s)}$$

for small d, this is

$$\frac{f(s)}{1 - F(s)}d = h(s)d$$

Combine the two cases, we see that

$$P(N(s+d) - N(s) = 1|N(s)) = I_{[X \geq s]}h(s)d \ .$$

Suppose $X_1, \ldots, X_n$ are independent, positive r.v.s the counting process

$$N(t) = \sum_i I_{[X_i \leq t]}$$

has intensity function

$$\sum_i h_i(s)I_{[X_i \geq s]} \ .$$

If there are censoring times C_i then the counting process

$$N(t) = \sum_i I_{[X_i \leq \min(t, C_i)]} = \sum_i I_{[\min(X_i, C_i) \leq t, \delta_i = 1]}$$

has intensity function

$$\sum_i h_i(s)I_{[X_i \geq s]}I_{[C_i \geq s]} = \sum_i h_i(s)I_{\min(X_i, C_i) \geq s} \ .$$

The history σ-algebra should be defined accordingly.

For left truncated data, or "late-entry" data, (or intermitent entered data). The intensity process is

$$\sum_i h_i(s)R_i(s) = \sum_i h_i(s)I_{[Y_i < s]}I_{[X_i \geq s]}I_{[C_i \geq s]}$$

where $Y_i, (Y_i < X_i)$ is the entry time or left truncation time.

3.1 Counting Processes with Discrete Cumulative Intensities

We know that $N(t) = I - H(t)$ is a martingale when $H(t)$ is continuous (or $h(t)$ exist)

How do you get the counting process when the cumulative hazard has discontinuous points?

In the VCR/DVD machine setting we can not have $g'(t)$ any more. But we have a discontinuous $g(t)$. Or in the notation of FH book $g(t) = A(t)$.

If N(t) is standard Poisson process N(g(t)) - g(t) is still a martingale.

However, the variance process is different to the continuous case.

$$\int_0^t (1 - \Delta A(t)) dA(t)$$

For binomial r.v.s $Var(Y) = p(1 - p)$

$$d\Lambda(t) = \frac{dF(t)}{1 - F(t-)} = \frac{F(t+) - F(t-)}{P(X \geq t)}$$

$$= P(X \in [t, t+dt] | X \geq t)$$

$dN(g(t))$

4 Convergence of Stochastic Processes

We shall mostly focus on the convergence of martingales to the Brownian motion.

Concepts.

Similar to the convergence in distribution of rvs. CLT of processes also exist. But what more does it gives? Compare to convergence in distribution of rvs.

Example: Poisson process when $\lambda \to \infty$.

Example: empirical process as an extended Poisson process.

$$\sqrt{n} \frac{F_n(t) - F(t)}{1 - F(t)} \to ?$$

As counting process

$$\sum I_{[X_i \leq t]}$$

has (cumulative) intensity

$$\sum H(\min(t, X_i)) = \sum \int_{-\infty}^{\infty} I_{[s \leq t]} dH(s)$$

4.1 Brownian Motion. Brownian Bridge

A BM is to stochastic process as the Normal distribution to the random variable.

What the limit process must look like: for fixed t it is just a normal rv. with mean 0 and variance t. See Applet of Brownian Motions at my web page.

The sample path of a BM is continuous in t.

Suppose $N_i^p(t)$ are iid Poisson processes. The limit (as $n \to \infty$) of

$$\lim \frac{\sum_i [N_i^p(t) - \lambda t]}{\sqrt{n\lambda}}$$

is a BM.

Given n iid uniform $[0, 1]$ random variables, we denote the empirical distribution based these observations as $\hat{F}_n(t)$. The limit (as $n \to \infty$) of

$$\sqrt{n}[\hat{F}_n(t) - F(t)]$$

is a BB. (here $0 \le t \le 1$). The meaning of the "limit" above is in distribution (for stochastic processes), or weak convergence.

If $B(t)$ is a Brownian motion, then

$$B(t) - tB(1)$$

for $0 \le t \le 1$, is a BB.

Both BM and BB are special cases of Gaussian Processes.

BM is a martingale: it has mean zero, with independent increments (inherited from Poisson process). Let $\mathcal{F}_t = \sigma\{BM(s), 0 \le s \le t\}$. For $t > s$, we have

$$E[BM(t)|\mathcal{F}_s] = E[BM(t) - BM(s) + BM(s)|\mathcal{F}_s] = E[BM(t) - BM(s)|\mathcal{F}_s] + BM(s) = BM(s)$$

since $E[BM(t) - BM(s)|\mathcal{F}_s] = E[BM(t) - BM(s)] = 0$ by independent increment.

Integration of Browning motion.

Two kinds of integration. $\int_0^t f(s)dBM(s)$ and $\int_0^t BM(s)dg(s)$.

Mean and Variance.

$$\int f(t)dBM(t) \approx \sum f(t_j)[BM(t_{j+1}) - BM(t_j)]$$

If $f()$ is non-random that is equivalent to a time changed BM. But we shall let $f()$ to be predictable.

$BM^2(t) - t$ is also a martingale.

Poisson process converge to BM, when intensity goes to infinity. (properly normalized).

5 Martingale associated with the Poisson process

See section 1.

6 Integration With Respect to a Counting Process Martingale

Counting process generalizes the Poisson process in the sense that the inter-arrival time is no longer iid exponential. But the size of the jumps are still always one.

The integration, $\int_0^t f(s)dN^p(s)$, allows us to have a process that have variable jump sizes. [The integration is Stieljes integral for a fixed ω.]

In words, the process $\int_0^t f(s)dN^p(s)$ is a pure jump process with the same locations of jump as $N^p(s)$, only the jump sizes are now $f(X_i)$ instead of size one. (where X_i is the time of the ith jump.)

The process $M^*(t) = \int_0^t f(s)dN^p(s) - \int_0^t f(s)d(\lambda s)$ is also a martingale, provided $f(s)$ is $\mathcal{F}_t$ predictable. The latter integration, $\int_0^t f(s)d(\lambda s)$ can be interpreted as in usual calculus.

Now putting the two generalizations, acceleration with jump size change, together. If $f(t)$ is $\mathcal{F}_t$ predictable, and finite, then

$$M(t) = \int_0^t f(s)d\left[N^*(s) - \int_0^s g'(u)du\right]$$

is also an $\mathcal{F}_t$ martingale, assuming $N^*(t) - \int_0^t g'(s)ds$ is an $\mathcal{F}_t$ martingale (counting process martingale) to begin with. The later is true if $g'(t)$ is $\mathcal{F}_t$ predictable.

Example: As an example we confirm now that the predictable integration wrt a Poisson process martingale is also a martingale.

Let

$$L(t) = \int_0^t f(s)dN^p(s) - \int_0^t f(s)d\lambda s = \int_0^t f(s)dM(s) \ .$$

To show $E[L(t+s)|\mathcal{F}_t] = L(t)$ for $s > 0$, we only need to go a small step at a time

$$E[L(t)|\mathcal{F}_{t-dt}] = L(t - dt) \ .$$

Observe

$$L(t) - L(t - dt) = f(t)[N^p(t) - N^p(t - dt)] - f(t)\lambda dt$$

Since $f(t)$ is $\mathcal{F}_{t-dt}$ measurable, (i.e. predictable) we have

$$E[f(t)(N^p(t) - N^p(t - dt))|\mathcal{F}_{t-dt}] = f(t)E[N^p(t) - N^p(t - dt)|\mathcal{F}_{t-dt}] = f(t)d\lambda t$$

(we also used the independent increment of Poisson process.) $\diamondsuit$

The proof (that the predictable integration wrt) a general counting process martingale (is again a martingale) is similar. We left that as an excecise.

7   Predictable Variation Process of $M(t)$

Our purpose of study the predictable variation process is to be able to compute the variance of the counting process martingale $M(t) = N(t) - \int_0^t g'(s)ds$ and the variance of the integration wrt $M(t)$. Also, the martingale CLT conditions (Lindeberg Condition) formulated in terms of predictable variation process, so we need it.

7.1   Definition

Suppose $M(t)$ is a (square integrable) martingale. If a process $V(t)$ is such that
    (1) it is predictable (i.e. $V(t)$ is $\mathcal{F}_{t-}$ measurable);
    (2) $M^2(t) - V(t)$ is again a martingale wrt $\mathcal{F}_t$;
    then $V(t)$ is called the predictable variation process of $M(t)$ and is sometimes denoted as $\langle M(t) \rangle$.

Since for all $t > 0$,
$$E\{M^2(t) - V(t)\} = 0$$
we have $Var M(t) = EM^2(t) = EV(t)$.

In words, $V(t)$ is the conditional, cumulative variance of $M(t)$.

7.2   How to compute/identify the predictable variation processes

The increment of $V(t) = \langle M(t) \rangle$ is in fact the conditional variance:

$$V(t) - V(t - dt) = E[(M(t) - M(t - dt))^2|\mathcal{F}_{t-dt}] = E[M^2(t) - M^2(t - dt)|\mathcal{F}_{t-dt}] \qquad (1)$$

(Prove the second equation as execise)

This offers a clue to identify the predictable variation process when given $M(t)$.

It is clear from the definition that for a Poisson process martingale, $N^p(t) - \lambda t$, the predictable variation process is (again) λt since λt is predictable (it is non-random, of course $\mathcal{F}_t$ predictable); and we can verify that

$$[N^p(t) - \lambda t]^2 - \lambda t$$

is again an $\mathcal{F}_t$ martingale. (the variance of a Poisson r.v. is same as the mean).

Theorem 4 For the one jump counting process martingales (with continuous hazard function) the predictable variation process is

$$\langle M(t) \rangle = \langle I_{[X \leq t]} - \int_0^t h(s) I_{[X \geq s]} ds \rangle = \int_0^t h(s) I_{[X \geq s]} ds . \tag{2}$$

Proof: Compute the increment of the predictable variation process over a small interval $(t - dt, t]$ and notice that we are dealing with the (conditional) variance of a Bernoulli r.v. The variance of Bernoulli is $p(1-p) = p - p^2$, but since here p is so small (same order as dt) we can ignor the p^2 part. (even after integration it is zero). $\diamond$

Theorem 5 The predictable variation process of the integration wrt the martingale $M(t), \mathcal{F}_t$ (asuming $f(t)$ is predictable) is given by

$$\langle \int_0^t f(s) dM(s) \rangle = \int_0^t f^2(s) d\langle M(s) \rangle . \tag{3}$$

In other words, the integrand $f(s)$ is like a constant after conditioning. This is due to the predictability of $f(t)$.

8   Law of Large Numbers for Martingales

Lenglart's Inequality. Theorem 3.4.1 and Corollary 3.4.1 of Fleming and Harrington p113.

Uniform consistency of empirical processes. [see my note on weighted empirical CDF]

Uniform strong consistency of Nelson-Aalen and Kaplan-Meier estimator. (Generalization of Gelivanko-Cantelli theorem to censored data case. More results are available. Hungarians group.)

9   CLT for Counting Process Martingales

Here is a version of the martingale CLT suitable for our use, where the § symbol indicate new concept/definition has been discussed.

(1) Suppose for $i = 1, 2, \ldots$, $M_i(t), 0 \le t \le T$ are independent L^2 martingales with respect to a common filtration $\mathcal{F}_t$. (§1)

(2) Suppose for each n, $f_{ni}(t)$ for $i = 1, 2, \cdots, n$ are $\mathcal{F}_t$ predictable functions. (jump size functions) (§2)

Then the following summation (clearly it is also a martingale process in t §3)

$$U_n(t) = \sum_{i=1}^{n} \int_0^t f_{ni}(s)dM_i(s) \qquad §4$$

will converge in distribution (§5) to a (time changed) Brownian motion $BM(V(t))$ (§6) **if** the following two conditions hold.

For every $t < T$, as $n \to \infty$ we have

$$§7 \qquad \langle U_n(t) \rangle = \sum_{i=1}^{n} \int_0^t f_{ni}^2(s)d\langle M_i(s) \rangle \to^P V(t) \tag{4}$$

where $V(t)$ is a non-random function. The second condition (Lindeberg condition) is

$$\forall \epsilon > 0, \qquad \sum_{i=1}^{n} \int_0^T f_{ni}^2(s)I_{[|f_{ni}(s)|>\epsilon]}d\langle M_i(s) \rangle \to^P 0 . \tag{5}$$

The convergence in distribution some times is also called weak convergence. The limit $BM(V(t))$ may also be written as $\int f(s)dBM(s)$ where the function $f(t)$ is such that $\int_0^t f^2(s)ds = V(t)$.

We are going to use the CLT with $M_i(t)$ constructed from the counting processes; i.e. one of those three martingales of last section.

Let $N_i(t)$ $i = 1, \ldots, n$ be independent counting processes with intensity function $g_i(t)$ such that $M_i(t) = N_i(t) - \int_0^t g_i(s)ds$ are martingales (§6) with respect to a common filtration $\mathcal{F}_t$. The predictable variation processes for these martingales are $\langle M_i(t) \rangle = \int_0^t g_i(s)ds$ (§7). The above 2 conditions in the CLT can be re-written as

$$\forall t < T, \qquad \sum_{i=1}^{n} \int_0^t f_{ni}^2(s)g_i(s)ds \to^P V(t) \tag{6}$$

and

$$\forall \epsilon > 0, \qquad \sum_{i=1}^{n} \int_0^T f_{ni}^2(s)I_{[|f_{ni}(s)|>\epsilon]}g_i(s)ds \to^P 0 . \tag{7}$$

A multivariate version of this theorem also hold.

The main reference for this section is **Theorem 5.1.1** of Fleming-Harrington p204.

10   Application in Survival Analysis

For T_i iid $F(t)$ with hazard h, and C_i iid $G(t)$, let $X_i = \min(T_i, C_i)$ and $\delta_i = I_{[T_i \leq C_i]}$. We have the basic martingale

$$M_n(t) = \sum M_i(t) = \sum_{i=1}^{n} \left(I_{[X_i \leq t,\ \delta_i = 1]} - \int_0^t h(s) I_{[X_i \geq s]} ds \right) \ .$$

10.1   The Nelson-Aalen and Kaplan-Meier estimators

The Nelson-Aalen estimator as stochastic integration of Martingales. Assume $R(s) > 0$ for $0 \leq s \leq t$, we have

$$\hat{H}(t) - H(t) = \int_0^t \frac{1}{R(s)} dM_n(s) \ . \tag{8}$$

We call this the martingale representation of the Nelson-Aalen estimator.

It is now easy to derive the asymptotic normality of the Nelson-Aalen estimator with the help of the CLT for martingales. By using the fact that $R(s)/n$ converge uniformly in probability to $P(X_i \geq s)$, (thus $n/R(s)$ converge to $1/P(X \geq s)$), we can verify (detail left as exercise) the two conditions in the martingale CLT and conclude

$$\sqrt{n}[\hat{H}(t) - H(t)] \xrightarrow{\mathcal{D}} BM(A(t))$$

with

$$A(t) = \int_0^t \frac{dH(s)}{P(X \geq s)} = \int_0^t \frac{dH(s)}{(1 - F(s-))(1 - G(s-))} \ .$$

The interval where convergence is guaranteed $0 \leq t \leq T$ must satisfy that $\sqrt{n}/R(T) \to^P 0$. This is true, for example, by assumeing $P(X \geq T) > 0$, since $n/R(T) \to 1/P(X \geq T) < \infty$, thus $\sqrt{n}/R(T) \to 0$.

The Kaplan-Meier estimator can also be represented as a martingale. There are two different approaches. The first is to directly represent $(\hat{F}_n - F)/(1 - F)$ as a martingale. The second is to approximate $\int_0^\infty g(t) d[\hat{F}_n(t) - F(t)]$ by an integral wrt Nelson Aalen estimator plus a small error. The integral wrt Nelson Aalen can easily be represented as martingales. For the second approach, please see Akritas (2000).

The representation is (for t such that $R(t) > 0$) slightly more complicated

$$\frac{\hat{F}(t) - F(t)}{1 - F(t)} = \int_0^t \frac{1 - \hat{F}(s-)}{1 - F(s)} \frac{1}{R(s)} dM_n(s) \ . \tag{9}$$

Assume $F(t)$ is continous, we can verify this representation easily in four steps: (1) both side starts at 0 when $t = 0$. (easy) (2) they have the same jump locations. (i.e. at the observed death times, easy) (3) their jump sizes are equal at the same location. The jump size of LHS is

$$\frac{\Delta \hat{F}(t)}{1 - F(t)} \ .$$

The jump size of RHS is

$$\frac{1 - \hat{F}(t-)}{1 - F(t)} \frac{1}{R(t)} \Delta \sum I_{[X_i \leq t, \delta_i = 1]} \ .$$

If t is a observed death time, make use of

$$\frac{1 - \hat{F}(t)}{1 - \hat{F}(t-)} = 1 - \frac{dN(t)}{R(t)}$$

to see they are equal.

(4) in between jumps, they have the same derivative, i.e. they grow at the same rate. The derivative of the LHS is

$$\frac{f(t)[\hat{F}(t) - 1]}{[1 - F(t)]^2}$$

The derivative of the RHS is

$$-\frac{1 - \hat{F}(t-)}{1 - F(t)} \frac{1}{R(t)} h(t) \sum I_{[X_i \geq t]} = -\frac{1 - \hat{F}(t-)}{1 - F(t)} h(t)$$

(notice, in between jumps, $\hat{F}(s-)$ is equal to $\hat{F}(s)$.)

Put all four points together you see that they are always equal. (this at least works for continous $F(t)$, for discrete $F()$ some modifications are needed). See homework for the case where the CDF $F(t)$ have finite number of jumps. The case where $F()$ have infinite and *not orderable* points of jumps is the most difficult case. We omit this and refer reader to FH book for a proof.

With this martingale representation (7) we have the following:

$$\sqrt{n} \frac{\hat{F}_n(t) - F(t)}{1 - F(t)} \xrightarrow{\mathcal{D}} BM(C(t))$$

under suitable conditions, where $C(t) = \int_0^t [P(X \geq s)]^{-1} dH(s)$. The proof proceeds by verify the two conditions in the martingale CLT.

We need to show

$$\int_0^t \left[\frac{1 - \hat{F}_n(s-)}{1 - F(s)}\right]^2 \frac{n}{R^2(s)} h(s) R(s) ds \xrightarrow{P} C(t)$$

and

$$\int_0^T \left[\frac{1 - \hat{F}_n(s-)}{1 - F(s)}\right]^2 \frac{n}{R^2(s)} I_{[|\cdot| > \epsilon]} h(s) R(s) ds \xrightarrow{P} 0$$

Again, uniform convergence of $R(t)/n$ and $\hat{F}(s)$ helps.

10.2 Two-sample Logrank and K-class of tests

The test statistic of Logrank can be written as (when $K \equiv 1$, and $t = +\infty$)

$$T_K = \int_0^t K(s) \frac{R_1(s)R_2(s)}{R_1(s) + R_2(s)} d\left[\hat{H}_1(s) - \hat{H}_2(s)\right]. \tag{10}$$

We shall keep the extra $K(s)$ to get a class of weighed logrank tests. Under null hypothesis, the two counting processes, one from each sample, have same compensator, therefore we have (after plus and minus the same compensator) the martingale representation

$$T_K = \int_0^t K \frac{R_1 R_2}{R_1 + R_2} d\left[\hat{H}_1(s) - \hat{H}_2(s)\right] = \int_0^t K \frac{R_1 R_2}{R_1 + R_2} \left[\frac{dM_{1n}(s)}{R_1(s)} - \frac{dM_{2n}(s)}{R_2(s)}\right]. \tag{11}$$

This imply the mean of T_K is zero under H_0.

Let us now consider power analysis, and how to chose the most powerful test among the K class. We shall work with a sequence of local alternative hypothesis. Let us make some assumptions to simplify notation/analysis: (1) WLOG assume the two hazard functions has the following relation

$$h_1(s) = h_2(s) + \frac{\gamma(s)h_2(s)}{\sqrt{n_1 n_2/(n_1 + n_2)}}, \quad \text{i.e.} \quad \frac{h_1(s)}{h_2(s)} = 1 + \frac{\gamma(s)}{\sqrt{n_1 n_2/(n_1 + n_2)}}$$

notice this include the H_0 with $\gamma(s) \equiv 0$ and local alternatives for a fixed $\gamma(s) \neq 0$. The interpretation of the $\gamma(s)$: it is the hazard ratio minus one.

By using the martingale CLT, we can show that the test statistic T_K, under H_0, has the following limit as $n \to \infty$

$$\sqrt{\frac{n_1 + n_2}{n_1 n_2}} T_K \xrightarrow{\mathcal{D}} N(0, \sigma^2)$$

where

$$\sigma^2 = \int_0^t K^2(s) \frac{P(Y_1 \geq s)P(Y_2 \geq s)}{P(Y_1 \geq s)\eta + P(Y_2 \geq s)(1 - \eta)} dH_2(s)$$

where $\eta = \lim n_1/(n_1 + n_2)$.

Under the local H_A assumed above, i.e. a fixed $\gamma(s) \neq 0$, we have

$$\sqrt{\frac{n_1 + n_2}{n_1 n_2}} T_K = M_n + B_n$$

where M_n is a mean zero martingale, and using the same martingale CLT as for H_0, it has the same limiting distribution as above (check), while B_n converge in probability to a constant μ. Thus by the Slutsky theorem the test statistic, under H_A, has the limiting distribution

$$\sqrt{\frac{n_1 + n_2}{n_1 n_2}} T_K \xrightarrow{\mathcal{D}} N(\mu, \sigma^2)$$

where

$$\mu = \int_0^t K(s) \frac{P(Y_1 \geq s)P(Y_2 \geq s)}{P(Y_1 \geq s)\eta + P(Y_2 \geq s)(1 - \eta)} \gamma(s) dH_2(s)$$

Now, the (asymptotic) power of the test can be calculated using the (limiting) normal distributions under H_0 and under H_A.

We claim that the power is determined by the quantity $|\mu|/\sigma$ under H_A (in fact, power is monotone with respect to this ratio). Denote the mean under H_A by μ_a. Call this ratio the efficacy.

Finally, we seek to maximize the above ratio by picking the right $K(s)$.

Theorem: To maximize the efficacy of the test, $|\mu_a|/\sigma$, we should set $K(s) = \gamma(s)$ (or a constant multiple of that).

Notice a constant multiple on $K(s)$ changes the variance and the mean, but do not change the efficacy and the power. Also WOLG we assume the numerator μ_a is positive. (otherwise multiple by -1)

To maximize the ratio $|\mu|/\sigma$ is equivalent to hold the denominator fixed while maximize the numerator, because we can always adjust the variance by adjusting the constant multiple of the $K(s)$ function. Therefore WLOG we can set the variance equal to a positive number c^2.

To maximize the μ while holding $\sigma = c$ or $\sigma^2 = c^2$, Lagrange multiplier then lead us to maximize

$$\mu_a + \lambda[\sigma^2 - c^2] = \int K(s) \frac{P()P()}{P()\eta + P()(1-\eta)} \gamma(s) dH_2(s) + \lambda \int K^2(s) \frac{P()P()}{P()\eta + P()(1-\eta)} dH_2(s) - \lambda c^2$$

or, maximize (wrt K and λ)

$$\int [K(s)\gamma(s) + \lambda K^2(s)] \frac{P(Y_1 \geq s)P()}{P()\eta + P()(1-\eta)} dH_2(s) - \lambda c^2 . \tag{12}$$

Notice we can maximize $[K(s)\gamma(s) + \lambda K^2(s)]$ **pointwise**.

Maximize under the integral sign (or pointwise at s, argue that if we can do that then the whole integral must also be maximized) the partial derivative wrt $K(s)$ under integral sign is

$$\gamma(s) + \lambda 2K(s)$$

setting the derivative to zero, we get $K(s) = -\gamma(s)/(2\lambda)$.

Finally we may solve the λ value by plug $K(s) = -\gamma(s)/(2\lambda)$ into the other partial derivative equation. But we do not need to since the information of $K(s) = C \times \gamma(s)$ is enough.

We point out, λ must be negative. For otherwise let $K(s) \to +\infty$ will make the target function (12) arbitrary large. This also imply the stationary point we solved is a max point in the Lagrange multiplier problem.

Discussion of how this result will guide us to a better test than logrank test.

In other words, the logrank test is equivalent to testing if the mean μ is zero under a normal distribution for T.

15

10.3 Counting processes in the Cox model

Score function as counting process martingale. Martingale residuals for Cox model.

Review of the setup of the Cox model: we observe

$$Y_i = \min(T_i, C_i), \quad \delta_i = I_{[T_i \leq C_i]}$$

here the lifetimes T_i are not iid but only independent, with hazard function

$$h_{T_i}(t) = h_0(t) \exp(\beta_0 z_i)$$

here z_i is the (observed) covariate for the ith patient, and $h_0(t)$ is the baseline hazard.

Define

$$M_i(t) = I_{[Y_i \leq t, \ \delta_i = 1]} - \int_0^t I_{[Y_i \geq s]} \exp(\beta_0 z_i) h_0(s) ds = I_{[Y_i \leq t, \ \delta_i = 1]} - \int_0^t I_{[Y_i \geq s]} \exp(\beta_0 z_i) d\Lambda_0(s)$$

In view of the results from previous sections and the Cox model assumptions, it is easy to see that $M_i(t)$ are martingales (in t, with a properly defined history filtration $\mathcal{F}_t$.

The martingale residuals after fitting the Cox model is defined similar to above except replace the parameter β, and unknown baseline by their estimator:

$$\hat{M}_i(t) = I_{[Y_i \leq t, \ \delta_i = 1]} - \int_0^t I_{[Y_i \geq s]} \exp(\hat{\beta} z_i) d\hat{\Lambda}_0(s)$$

where $\hat{\Lambda}(t)$ is usually the Breslow estimator, and the $\hat{\beta}$ the maximum partial likelihood estimator.

The martingale residules are defined as the $\hat{M}_i(\infty)$ $i = 1, 2, \cdots, n$. (well, instead of infitity, any time larger then all the observed times will work too).

Let $\Re_i = \{j : Y_j \geq Y_i\}$, the index of risk set at time Y_i. Define a function of β as (score function)

$$\ell(\beta) = \sum_{i=1}^n \delta_i z_i - \sum_{i=1}^n \delta_i \frac{\sum_{j \in \Re_i} z_j \exp(\beta z_j)}{\sum_{j \in \Re_i} \exp(\beta z_j)}, \tag{13}$$

If $\hat{\beta}_c$ is the solution of (13), i.e. $\ell(\hat{\beta}_c) = 0$, then $\hat{\beta}_c$ is called the Cox (maximum) partial likelihood estimate of regression coefficient β_0.

The proof of the asymptotic normality for $\hat{\beta}$ is very similar to the proof of an estimator $\hat{\theta}$ defined by estimating equation. After all, score function is an estimating equation. Please consult my note on Sta 701 for the general approach. We usually need to show two things: the score function at the true parameter value is asymptotic normal; and the derivative of the score function converge to a positive definite matrix in probability. We examine these two things below.

Claim The score function at the true parameter, $\ell(\beta_0)$, is a martingale, in t, introduced below. To introduce time t, define

$$\ell(\beta, t) = \sum_{i=1}^n \int_0^t \left(z_i - \frac{\sum_{j=1}^n z_j e^{\beta z_j} I_{[Y_j \geq s]}}{\sum_{j=1}^n e^{\beta z_j} I_{[Y_j \geq s]}} \right) dI_{[Y_i \leq s, \delta_i = 1]} \tag{14}$$

then it is easy to see $\ell(\beta, \infty) = (12)$; (actually, a time larger then any observed death times will be OK, no need to be true infinity).

When $\beta = \beta_0$, we have (please verify)

$$\ell(\beta_0, t) = \sum_{i=1}^{n} \int_0^t \left(z_i - \frac{\sum_{j=1}^n z_j e^{\beta_0 z_j} I_{[Y_j \geq s]}}{\sum_{j=1}^n e^{\beta_0 z_j} I_{[Y_j \geq s]}} \right) dM_i(s) \tag{15}$$

then the above is a martingale in t, since it is a sum of predictable integration of martingales.

Therefore the socre function at the true β_0 is a martingale in t, the usual Cox score function is when $t = \infty$.

Breslow estimator of the baseline hazard.

$$\sum_{Y_i \leq t} \frac{\delta_i}{\sum e^{\hat{\beta} z_j} I[Y_j \geq Y_i]} = \int_0^t \frac{\sum_{i=1}^n dN_i(s)}{\sum_j e^{\hat{\beta} z_j} I[Y_j \geq s]}$$

where $\hat{\beta}$ can be any consistant estimator of β, usually taken to be the maximum partial likelihood estimator.

Outline of the proof: $\hat{\beta}$ converges to β imply $e^{\hat{\beta} z_i}$ converges to $e^{\beta z_i}$ for all i, uniformlly. (need z_i to be bounded uniformly for all i).

imply $1/n \sum_{i=1}^n e^{\hat{\beta} z_i} I[Y_i \geq t]$ converge to $1/n \sum_{i=1}^n e^{\beta z_i} I[Y_i \geq t]$ uniformly for all t.

imply the ratio of $\sum_{i=1}^n e^{\hat{\beta} z_i} I[Y_i \geq t]$ and $\sum_{i=1}^n e^{\beta z_i} I[Y_i \geq t]$ converge to 1 uniformly for all t. (assume the denominator > 0)

This will imply the Breslow estimator minus the true is a martingale plus $o_p(1)$.

To handle the martingale term, we show that the quadratic variation process converges to 0. (notice this time we do not multiply by $\sqrt{n}$)

Need $\sum_{i=1}^n e^{\hat{\beta} z_i} I[Y_i \geq t] \to \infty$. Or, in words, the weighted risk set size must go to infinite.

This is true for $0 \leq t \leq T$ where T is such that $1/n \sum P(Y_i \geq T) > 0$.

Another useful fact to complete the proof of asymptotic distribution of Cox estimate $\hat{\beta}$ is that the derivative of the score function converge to a positive definite matrix (observed information matrix has a limit as n goes to infinite).

In other words, we need to show, when $\beta = \beta^*$, any consistent estimator of β_0,

$$-\frac{\partial \ell(\beta)}{\partial \beta} = \frac{1}{n} \sum_{i=1}^n \delta_i \frac{(\sum_j z_j^2 \exp(\beta z_j))(\sum_j \exp(\beta z_j)) - (\sum_j z_j \exp(\beta z_j))^2}{(\sum_j \exp(\beta z_j))^2}$$

converge to a positive definite matrix.

If we require the condition that, for $k = 0, 1, 2$

$$\frac{1}{n} \sum_{j=1}^n z_j^k \exp(\beta z_j) I[Y_i \geq s] \longrightarrow S^k(s)$$

in probability as $n \to \infty$, then apparently the converges of the information matrix holds. This final (three) convergence is similar to the empirical distribution, and can be proved by Hoeffding's inequality technique (Glivanko-Cantalli Lemma).

The standard reference for this proof is Andersen and Gill (1982).

11 Variance of the Kaplan-Meier mean

For positive random variables, the mean can also be obtained as

$$\mu = \int_0^\infty [1 - F(t)] dt .$$

When $F(t)$ is unknown and we need to estimate the mean, the natural thing is to replace $F(t)$ in the above by an estimator. For right censored data, the estimator of $F(t)$ is the so called Kaplan-Meier estimator: $\hat{F}_n(t)$. So our mean estimator is

$$\hat{\mu} = \int_0^\infty [1 - \hat{F}_n(t)]dt$$

Notice if there is no censoring this falls back to the sample mean.

After standardization, we have

$$\sqrt{n}(\hat{\mu} - \mu) = \int_0^\infty \sqrt{n}[F(t) - \hat{F}_n(t)]dt = -\int_0^\infty \sqrt{n}\frac{\hat{F}_n(t) - F(t)}{1 - F(t)}[1 - F(t)]dt$$

By the martingale CLT, we know that

$$\sqrt{n}\frac{\hat{F}_n(t) - F(t)}{1 - F(t)} \approx BM(C(t)) \ .$$

So in the limit we have that the distribution of $\sqrt{n}(\hat{\mu} - \mu)$ converge to the distribution of

$$-\int_0^\infty BM(C(t))[1 - F(t)]dt \ .$$

Finally we need to identify the distribution of the above integral.

It is obvious mean zero and normally distributed. (a linear combination of mean zero, jointly normal random variables). We now compute the variance.

$$Var = E\int_0^\infty BM(C(t))[1 - F(t)]dt \times \int_0^\infty BM(C(t))[1 - F(t)]dt$$

We can replace t by s, after all, it is only a dummy variable name.

$$Var = E\int_0 BM(C(t))[1 - F(t)]dt \times \int_0 BM(C(s))[1 - F(s)]ds \ .$$

Write $\int f(t)dt \int f(s)ds$ as $\int\int f(t)f(s)dtds$

$$Var = E\int\int BM(C(t))BM(C(s))[1 - F(t)][1 - F(s)]dtds \ ,$$

exchange the order of E and $\int$

$$Var = \int\int E\{BM(C(t))BM(C(s))\}[1 - F(t)][1 - F(s)]dtds \ .$$

Recall $E[BM(t)BM(s)] = \min(t, s)$ for Brownian Motion on the standard clock. For clock $C(t)$ we have $E[BM(C(t))BM(C(s))] = \min(C(t), C(s)) = C(\min(t, s))$ since $C(t)$ is an increasing function. Finally

$$Var = \int_0^\infty \int_0^\infty C(\min(t, s))[1 - F(t)][1 - F(s)]dtds$$

which is after some calculation (pure calculus, integration by parts) we get

$$Var = \int_0^\infty \frac{[\int_s^\infty 1 - F(u)du]^2}{P(Y \geq s)}dH(s) \ .$$

An alternative approach is taken by Akritas (2000) which results a different (but equivalent) variance formula.

18

12 Survival Analysis and Empirical Likelihood

NPMLE: there are examples of NPMLE working just like in the parametric case, and there are cases that NPMLE fail dramatically.

A general theory for the NPMLE is difficult and lacking. So we need to check each specific case case by case.

Empirical likelihood method can be used to obtain confidence interval for the Kaplan-Meier mean without estimating the variance (which we just derived in the previous section).

The first use of empirical likelihood is actually about confidence intervals with the Kaplan-Meier estimator (Thomas and Grunkmeier 1979), i.e. deals with right censored data. Owen recognized the generality of the method and proved it rigorously for the non-censored data (1989). He also coined the term "Empirical Likelihood", and applied it to many other cases. Since then, a lot of work on Empirical Likelihood appeared and established the empirical likelihood as a general method with many nice features. Most of the researches deals with non-censored data, however. See Owen (2001) Book "Empirical Likelihood".

We will mainly discuss the empirical likelihood in the context of survival analysis here.

12.1 What is the empirical likelihood method?

It is the non-parametric counter part of the Wilks likelihood ratio approach in the parametric likelihood inference. (if you are familiar with that). It allows a statistician to derive asymptotically correct tests without the need to pick a parametric family of distribution (– Owen).

The primary reference is Owen's Book of 2001 *Empirical Likelihood*.

It helps you to do statistical inference (calculate P-value/confidence intervals) with certain non-parametric tests/estimators without the need of calculating their variances.

12.2 Why empirical likelihood with survival analysis?

Because the variance is so much harder to estimate for NPMLE with censored/truncated data. Because empirical likelihood methods do not need a variance estimator. Because empirical likelihood inference is often (at least asymptotically) efficient. For specific examples demonstrating these points, see sections later.

12.3 NPMLEs and their empirical likelihood functions

The empirical likelihood method is intrinsically linked to the NPMLE. We first study the empirical likelihood (or nonparametric likelihood) function for right censored data.

The empirical distribution is the NPMLE for iid observations. (This is first proved by Kiefer Wolferwitz 1956. see Owen 2001 for a simple proof.) Here we establish the fact that the Kaplan-Meier estimator is NPMLE for the right censored data. The Nelson-Aalen estimator is the NPMLE for the cumulative hazard, but notice there are two difference ways of writing the likelihood function in terms of hazard.

Definition of Empirical Likelihood, The Nelson-Aalen, Kaplan-Meier estimator is NPMLE:

Let $(x_1, \delta_1), (x_2, \delta_2), \ldots, (x_n, \delta_n)$ be the i.i.d. right censored observations as defined in class: $x_i = \min(T_i, C_i), \delta_i = I[T_i \leq C_i]$, where T_i are lifetimes and C_i are censoring times.

Let the CDF of T_i be $F(t)$.

The (nonparametric) likelihood pertaining to $F(\cdot)$ based on the right censored samples (x_j, δ_j) is (a constant has been omitted)

$$L(F) = \prod_{\delta_i=1} \Delta F(x_i) \prod_{\delta_i=0} (1 - F(x_i)).$$

The Kaplan-Meier estimator (1958) maximizes the above likelihood among all the CDF's. Notice the Kaplan-Meier estimator is discrete.

Because the survival function $(1 - F(t))$ and hazard function $(\Lambda(t))$ are mathematically equivalent, inference for one can be obtained through a transformation of the other.

For discrete CDF, we have the relation

$$\begin{cases} \Delta F(x_i) & = \Delta\Lambda(x_i) \prod_{j:x_j<x_i}(1 - \Delta\Lambda(x_j)) \\ 1 - F(x_i) & = \prod_{j:x_j\leq x_i}(1 - \Delta\Lambda(x_j)) \end{cases},$$

therefore, while assuming F is discrete, the likelihood can be written as

$$L(\Lambda) = \prod_{i=1}^{n} \left\{ (\Delta\Lambda(x_i))^{\delta_i} \left(\prod_{j:x_j<x_i}(1 - \Delta\Lambda(x_j)) \right)^{\delta_i} \left(\prod_{j:x_j\leq x_i}(1 - \Delta\Lambda(x_j)) \right)^{1-\delta_i} \right\}.$$

Let $x_{k,n}$ be the m distinctive and ordered values of the x-sample above, where $k = 1, \cdots, m$ and $m \leq n$. Let us denote

$$D_k = \sum_{i:x_i=x_{k,n}} \delta_i \qquad \text{and} \qquad R_k = \sum_{i=1}^{n} I(x_i \geq x_{k,n}).$$

Then, $L(\Lambda)$ becomes

$$L = \prod_{k=1}^{m} (\Delta\Lambda(x_{k,n}))^{D_k} (1 - \Delta\Lambda(x_{k,n}))^{(R_k-D_k)}. \tag{16}$$

From this, we can easily get the log likelihood function:

$$\log L = \sum D_k \log \Delta\Lambda(x_{k,n}) + (R_k - D_k) \log(1 - \Delta\Lambda(x_{k,n})) .$$

Notice the similarity of this log likelihood to the log likelihood of binomial model.

Maximizing the log likelihood function with respect to $\Delta\Lambda(x_{k,n})$ gives the (nonparametric) maximum likelihood estimate of $\Delta\Lambda(x_{k,n})$ which is the Nelson-Aalen estimator,

$$\Delta\hat{\Lambda}(x_{k,n}) = \frac{D_k}{R_k}. \tag{17}$$

By the invariance property of the MLE this implies that the Kaplan-Meier estimator

$$1 - \hat{F}(t) = \prod_{x_{k,n}\leq t} (1 - \frac{D_k}{R_k})$$

is the MLE of $1 - F(t) = \prod_{s\leq t}(1 - \Delta\Lambda(s))$ (i.e. it maximizes the likelihood $L(F)$ above.)

If we use a 'wrong' formula connecting the CDF and cumulative hazard: $1 - F(t) = \exp(-\Lambda(t))$, (wrong: in the sense that the formula works for continuous F but here we have a discrete F) then the likelihood would be

$$AL = \prod_{i=1}^{n}(\Delta\Lambda(x_i))^{\delta_i}\exp\{-\Lambda(x_i)\} = \prod_{i=1}^{n}(\Delta\Lambda(x_i))^{\delta_i}\exp\{-\sum_{j:x_j \leq x_i}\Delta\Lambda(x_j)\}.$$

If we let $w_i = \Delta\Lambda(x_i)$ then

$$\log AL = \sum_{i=1}^{n}\delta_i \log w_i - \sum_{i=1}^{n}\sum_{j:x_j \leq x_i} w_j . \tag{18}$$

It is worth noting that among all the cumulative hazard functions, the Nelson-Aalen estimator also maximizes the $\log AL$,

$$w_i = w_i^* = \frac{D_i}{R_i} . \tag{19}$$

This can be verified easily by taking derivative of $\log AL$ with respect to w_i and solving the equation.

This version of the likelihood is sometimes called the 'Poisson' version of the empirical likelihood. The other version, 'binomial'. Though asymptotically equivalent, they are more convenient respectively for a particular type of constraint.

Notice that, if we view the *vector* (x_i, δ_i) as iid, then the empirical likelihood function is simply

$$L = \prod p_i$$

where $p_i = P(\{(x_i, \delta_i)\})$. This likelihood leads to the NPMLE of 2-dim empirical distribution function, which just puts $1/n$ probability at each (x_i, δ_i).

12.4 The property of the NPMLE. Computation. Self-consistent algorithm.

Many of the properties of NPMLE can be obtained by using the tools of counting process martingales and empirical process theory.

In simple cases (like in the previous section) the NPMLE has explicit formula. In more complicated cases (more complicated censoring, or with some extra constraints) no explicit formula exists and the computation of the NPMLE can be done by self-consistency/EM algorithm. See Zhou 2002 for details.

In particular, using the counting process/martingale theory, we can show that

Lemma 1 Let $\hat{\Lambda}(t)$ denote the Nelson-Aalen estimator. Under the conditions that make the Nelson-Aalen estimator $\hat{\Lambda}(t) - \Lambda(t)$ a martingale, we have

$$\sqrt{n}\int_0^\infty g_n(t)d[\hat{\Lambda}(t) - \Lambda(t)] \xrightarrow{\mathcal{D}} N(0, \sigma^2) \tag{20}$$

where $g_n(t)$ is a sequence of $\mathcal{F}_t$ predictable functions that converge (in probability) to a (non-random) function $g(t)$, provided the variance at right hand side is finite.

The variance σ^2 above is given by

$$\sigma^2 = \int_0^\infty \frac{g^2(s)d\Lambda(s)}{(1 - F(s-))(1 - G(s))} \tag{21}$$

Also, the variance σ^2 can be consistently estimated by

$$\sum_{i=1}^{n} n \left[\frac{g_n^2(x_i)}{R_i} \frac{D_i}{R_i} \right] = \int \frac{n g_n^2(t)}{R(t)} d\hat{\Lambda}(t) \ . \tag{22}$$

Proof: If we put a variable upper limit, v, of the integration in (14), then it is a martingale in v. By the martingale CLT, we can check the required conditions and obtain the convergence to normal limit at any fixed time v. Finally let v equal to infinity or a large value.

To get the variance estimator, we view the sum as a process (evaluated at ∞) and then compute its compensator, which is also the predictable variation process. Then the compensator or intensity is the needed variance estimator. Finally, use Lenglart inequality to show the convergence of the variance estimator. $\diamondsuit$

12.5 Maximization of the empirical likelihood under a constraint

We now consider maximization of the empirical likelihood functions with extra estimating equation constraint. The easiest case is the iid data with no censoring (where NPMLE = empirical CDF). This is first solved by Owen (1988).

For the right censored observations, with likelihood function in terms of hazard function (where NPMLE = Nelson-Aalen), this is solved by Pan and Zhou (1999).

For the right censored observation, with likelihood function in terms of CDF (where NPMLE = Kaplan-Meier), this is a much harder problem.

Exercise: Without extra constraint, the maximization of the nonparametric likelihood function with right censored data, is solved by Kaplan-Meier (1958). Try to solve it yourself, without consult the paper. There is an explicit solution given by Kaplan-Meier, there is also a recursive solution.

We will maximize the $\log AL$ among all cumulative hazard functions that satisfy

$$\int g(t) d\Lambda(t) = \sum g(x_j) \Delta\Lambda(x_j) = \mu \ . \tag{23}$$

Detailed proof is only given in this case here.

We may also maximize the $\log L$ among all cumulative hazard functions that satisfy another type of constraint

$$\sum_j g(x_j) \log(1 - \Delta\Lambda(x_j)) = \mu \ .$$

For detailed proof of this case, see Fang (2000). (Add some comments/examples about this equation.)

Theorem 12.1 *The maximization of the log likelihood* $\log AL$ *under the extra equation (17) is achieved when*

$$\Delta\Lambda(x_k) = \tilde{w}_k = \frac{D_k}{R_k + \lambda g(x_k)} \tag{24}$$

with λ *obtained as the solution of the equation*

$$\sum_{k=1}^{n} g(x_k) \frac{D_k}{R_k + \lambda g(x_k)} = \mu \ . \tag{25}$$

Proof: Use the Lagrange multiplier to compute the constrained maximum. Notice the equation for λ is monotone so there is a unique solution. $\diamondsuit$

Sometimes we may want to use the following estimating equation instead:

$$\sum g(x_j)\Delta F(x_j) = \mu \ .$$

i.e. maximize the $\log L(F)$ among all CDF that satisfy this (mean type) equation. Similar results also hold, but are harder to proof. Computation is also harder in this case. No explicit formulae exist for the $\Delta F(x_j)$ that achieve the max under the constraint equation (as far as I know). We have to use a modified self-consistent/EM algorithm to compute the NPMLE under this constraint, available in the R package `emplik`.

12.6 Wilks Theorem For Censored Empirical Likelihood Ratio

The (ordinary) Wilks theorem in parametric estimation says:

$$-2\log \frac{\max\limits_{\theta \in \Theta_0} L(\theta)}{\max\limits_{\theta \in \Theta} L(\theta)} \approx \chi_p^2$$

where Θ is the whole parameter space, and Θ_0 is the subspace of Θ, obtained by imposing p equations for θ to satisfy.

The right-hand side is the usual chi-square if the true θ is inside Θ_0 (i.e. when the null hypothesis is true), and is a non-central chi-square when the true θ is not inside Θ_0.

The likelihood function $L(\theta)$ above is the ordinary (parametric) likelihood function.

There is a similar Wilks theorem where the likelihood function above is replaced by the (censored) empirical likelihood function. The whole parameter space Θ will be equal to "all the CDFs" or "all the cumulative hazard functions". and the subspace Θ_0 is obtained by imposing equations like (17) or (20) on Θ.

We only prove a Wilks theorem for the censored empirical likelihood $\log AL$ and estimation equation (17).

Theorem 12.2 *Assume we have iid right censored observations as before.*

Then, as $n \to \infty$

$$-2\log \frac{AL(w_i = \tilde{w}_i)}{AL(w_i = w_i^*)} \xrightarrow{\mathcal{D}} \chi_1^2$$

where w_i^* is the Nelson-Aalen estimator, and $\tilde{w}_i$ is given by (18) with $\mu = \mu_0 = \int g(t)d\Lambda_0$.

We first prove a lemma.

Lemma 11.2 *The solution λ^* for the equation (19) above has the following asymptotic distribution.*

$$\frac{\lambda^*}{\sqrt{n}} \xrightarrow{\mathcal{D}} N(0, var = \frac{1}{\sigma^2})$$

with σ^2 defined in (15).

PROOF: (Basically, we do a one step Taylor expansion.) Assume λ is small (which can be shown), we expand (around $\lambda = 0$) the left hand side of the equation (19) that defines λ^*, we can re-write the equation as

$$\sum \frac{g(x_k)D_k}{R_k} - \lambda \sum \frac{g^2(x_k)D_k}{(R_k)^2} + O(\lambda^2) = \mu$$

23

Ignoring the higher order term, this equation can be solved easily, we therefore have:

$$\lambda^* = \frac{n \sum_k \frac{g(x_k)D_k}{R_k} - \mu}{n \sum_k \frac{g^2(x_k)D_k}{(R_k)^2}} + O_p(\lambda^2) \ .$$

Multiply $1/\sqrt{n}$ throughout, we notice that the numerator is same as the one treated in Lemma 1, and thus have an asymptotic normal distribution for numerator. The denominator is seen to converge in probability to σ^2 also by Lemma 1. Thus by Slutsky theorem we have the conclusion. $\diamondsuit$

Now we are ready to prove theorem 12.2.

PROOF OF THEOREM 12.2: Plug the expressions of $\tilde{w}_i$ and w_i^* into $AL(w_i)$ or $\log AL$.

$$-2 \log \frac{AL}{AL} = 2[\log AL(\lambda = 0) - \log AL(\lambda = \lambda^*)]$$

Now we use Taylor expansion (up to second order) on the second term above: (expand at $\lambda = 0$). $\log AL(\lambda^*) = \log AL(0) + \lambda^*(first\ derivative) + 1/2(\lambda^*)^2(second\ derivative)$.

The first term in the expansion will cancel with the other term.

The second term in the expansion is zero. Because the (first order) partial derivative wrt λ at $\lambda = 0$ is zero. This is also obvious, since $\lambda = 0$ gives rise to the Nelson-Aalen estimator, and the Nelson-Aalen estimator is the maximizer. The derivative at the maximizer must be zero.

The third term in the expansion is equal to

$$(\lambda^*)^2 \frac{\partial}{\partial \lambda^2} \log AL(\lambda) = (\lambda^*)^2 \cdot \sum \frac{D_k g^2(x_k)}{R_k^2}$$

which can be shown to be approximately chi-square by Lemma 2 and the variance estimator. (plug in the asymptotic representation of λ^*).

We see that the limiting distribution is chi-square. $\diamondsuit$

Multivariate version of this theorem exists. Zhou (2002)

This theorem can be used to test the hypothesis $H_0 : \int g(t)d\Lambda(t) = \mu$. We reject the null hypothesis when the $-2 \log$ empirical likelihood ratio is too large.

12.7 Confidence interval/region by using Wilks theorem

The above section points out a way to do testing hypothesis. To obtain confidence intervals we need to invert the testing.

The software we use in the example is a user contributed package `emplik` of R. They can be obtained from http://www.cran-us.org

Obtain confidence interval for one of the two parameters.

Example: See my notes "Survival Analysis using R".

12.8 Empirical likelihood for regression models

There are two views of a linear model (or may be regression model?): we call them regression models and correlation models; after Freedman (1981), Owen (1991).

In the regression model view, the predictors X_i are considered fixed, and the responses Y_i are considered sampled from the conditional distribution $F_{Y|X=x_i}$. (and the mean of this conditional distribution is $\beta_0 + \beta_1 X_i$ etc.)

In the correlation model view, the pair (X_i, Y_i) are both random and sampled from the joint distribution $F_{X,Y}$. And the parameter $\beta = \beta_{LS}$ is the one that minimizes $f(\beta) = E(Y - X^\top \beta)^2$.

This is also equivalent to the estimating equation $E[X(Y - X^\top \beta_{LS})] = 0$.

We comment that these two different model interpretations also leads to two different bootstrap approaches to linear model: re-sample the residuals or re-sample the (x, y) pairs.

The empirical likelihood for the correlation model, assume all n pairs (X, Y) are observed is $\prod p_i$, where the probability p_i is the probability we put on the pair (the (p+1)-dim) (x_i, y_i).

The empirical likelihood for the regression model, assume all n data $(X, Y$ are observed is again $\prod p_i$, where the probability p_i is the probability we put on the ith (1-dim) residual e_i.

13   The AFT models

An AFT model is just a usual linear regression model for the log of the survival times.

Suppose T_i are the survival times. Let $Y_i = \log T_i$. Suppose also X_i are the covariates associated with T_i. An AFT model is to suppose

$$Y_i = \log T_i = \beta^T X_i + \epsilon_i$$

where ϵ_i is iid. If the model include an intercept term (α or β_0) then the mean of the ϵ_i is zero. If the model do not have an intercept term then the mean of ϵ_i do not need to be zero (or it absorbs the intercept term). Parameter β needs to be estimated. The responses Y_i or T_i may be right censored by C_i.

Compared to the Cox model, AFT model is easier to interpret (similar to location models). AFT model and Cox model are both flexible semiparametric models.

The responses Y may be right censored, so we only observe

$$Y_i^* = \min(Y_i, C_i), \qquad \delta_i = I_{[Y_i \leq C_i]}$$

along with covariate X_i.

13.1   The rank based estimator

One estimation procedure is the rank based estimate. Define the (censored) shifted residules (for b)

$$e_i(b) = \min(Y_i, C_i) - bX_i \ .$$

Define

$$\bar{X}(t, b) = \frac{\sum X_j I_{[e_j \geq t]}}{\sum I_{[e_j \geq t]}} = average \ of \ (X_i \ such \ that \ e_i \geq t)$$

i.e. it is the average of X_j that are still "at risk" at time t in the residual scale.

The rank based estimator of β is the solution of the estimating equation

$$0 = \sum_{i=1}^{n} \phi(e_i(b))\delta_i[X_i - \bar{X}(e_i(b), b)] = v(b) \tag{26}$$

where $\phi(\cdot) \equiv 1$ (logrank) or $\phi(\cdot) = \sum I_{[e_j > \cdot]}$ (Gehan) or some other predictable function.

Intuitively, if the e_i's are the true residuals ($= \epsilon_i$), then the order of the e_i should have nothing to do with the covariates X_i. The quantity $\bar{X}(e_i(b))$ can be thought of as a conditional expectation of X, condition on $e_i > t$. But if e_i has nothing to do with X_i then the conditional expectation is the same as the unconditional expectation. Thus the difference should have mean zero.

The only software I know that produce this estimator is the R package `rankreg`.

Example For two sample problem, testing that the two samples are following the same distribution. This is a special case of the AFT model, when $X_i = 1$ or 0 (indicating the sample), and testing $\beta = 0$. Some algebra show this is the same as the Log rank (or Gehan) test. QED

Testing the hypothesis $H_0 : \beta = \beta_0$ can be done by using the following χ-square statistic:

$$v(\beta_0)^T \Sigma^{-1} v(\beta_0)$$

where Σ can be estimated. This statistic is asymptotically distributed as a central χ-square when H_0 is true. The function `RankRegV()` inside the R package `rankreg` will compute this χ-square statistic and give a P-value.

We can use the counting process martingale result discussed before to show the asymptotic normality of $v(\beta_0)$.

Remark: Since the above estimation uses only rank, the intercept (a shift) term in the linear model is not identifiable by the ranks. If one really needs to estimate the intercept, this can be done as follows: Assume in addition that the mean of the $\epsilon_i = 0$. Form the 'shifted residuals' $(=e_i(\hat{\beta}))$ with $\hat{\beta}$ from rank estimator. Estimate the intercept by the mean of the Kaplan-Meier estimator based on the residuals (and the original censoring indicators).

The next estimation procedure (B-J) usually assumes an intercept term in the linear model and the intercept can be estimated together with the slopes.

13.2 The Buckley-James estimator

The Buckley-James (1979) estimator is an iterative estimator for the censored AFT model (regression model). The available of cheap, fast computer and ever-improving software in the last 10 to 15 years made the calculation of the Buckley-James estimator a routing business. For example, the program `bj( )` within the `Design` library of Harrell (available for both S-plus and R). The program `BJnoint( )` within the `emplik` library can also be used.

The B-J estimator of β in the AFT model is a generalization to the least squares estimator.

The idea is that for uncensored observations we can define residuals, and for censored observations we can substitute it with the expected value, condition on covariates and the censored value. However, this conditional expectation depends on the (unknown) CDF of the ϵ and needs to be estimated. The estimation of this CDF is by the Kaplan-Meier estimator based on the residuals. Then use an E-M type iteration to find the β estimator.

The Buckley-James estimating equation is

$$0 = \sum_{i=1}^{n} \left(\delta_i X_i e_i(b) + (1 - \delta_i) X_i \sum_{j:e_j > e_i} e_j(b) \hat{P}(e_i^* = e_j | e_i^* > e_i) \right) \qquad (27)$$

where e_i^* denotes the (un-observable) residual before been censored. And $\hat{P}$ is based on the Kaplan-Meier estimator. Compare this to the least squares normal equation.

Formally, the iteration goes like this: (1) pick an initial estimator $\beta^{(0)}$. (2) Use this $\beta^{(0)}$ to obtain residuals. (3) From the residuals, obtain the estimator $\hat{P}$ (Kaplan-Meier). (4) Plug the residuals and $\hat{P}$ into the right hand of above equation and solve the equation to obtain a new estimator $\beta^{(1)}$.

But the variance estimation of the Buckley-James estimator remains very difficult. The function `bj()` uses a variance estimation formula given by the BJ's original paper which do not have a rigorous justification and as Lai and Ying (1991) have pointed out this formula may not be correct.

Zhou and Li (2004) propose to use an empirical likelihood approach to test hypothesis about β. This approach avoid the need to estimate the variance of the Buckley-James estimator. The P-value of the test is obtained from the standard chi-square quantile. The calculation is available in `emplik` package for R. See the function `bjtest()`.

Model assesment by residual sum of squares.

Residual sum of squares is a useful measure of how good the model fits the data. It needs to be generalized since the data we have are censored.

One such generalization is implemented by the R function `myRSS( )`. Please verify that (a) the B-J estimator $\hat{\beta}$ actually minimizes this generalized RSS. (b) the sum of the residuals (not squared) for the B-J estimator is zero.

To use the `myRSS( )` with the rank estimator, we need to first get an estimator of the intercept term per the remark above.

14   References

Owen, A. (1988), "Empirical Likelihood Ratio Confidence Intervals for a Single Functional," *Biometrika*, **75** 237-249.

Owen, A. (1990), "Empirical Likelihood Confidence Regions," *The Annals of Statistics*, **18**, 90-120.

Owen, A. (2001). *Empirical Likelihood.* Chapman and Hall, New York.

Pan, X. and Zhou, M. (2002). "Empirical likelihood ratio in terms of cumulative hazard function for censored data" *Journal of Multivariate Analysis.* 166-188.

Kalbfleisch, J. and Prentice, R. (2002) *The Statistical Analysis of Failure Time Data* 2nd Edition, Wiley, NewYork

Zhou, M. and Li, G. (2004). Empirical likelihood analysis of the Buckley-James estimator. Preprint.

Lai T.L. and Ying, Z. (1991). Large sample theory of a modified Buckley-James estimator for regression analysis with censored data. *Ann. Statist*, 19, 1370-402.

Fleming and Harrington Book outline:

Chap 1. Definitions (filtration, martingale, conditional expectation)

Example 1.2.1 (and continuation)

$M(t) = N(t) - A(t)$ is a martingale

Theorem 1.3.1

Doob-Meyer decomposition

Martingale $\int H(t)dM(t)$

Chap 2. N and $(N - A)^2$ has same compensator when A is cont.

Discrete compensator A. $(N - A)^2$ has compensator ?

Chap 3. Nelson-Aalen estimator, Log-rank test, Kaplan-Meier estimator martingale representations.

Chap 4. Martingales associated with Cox model, residules

Chap 5. Martingale CLT.

Chap 6. Kaplan-Meier estimator CLT, confidence band.

Chap 7. Weighted log-rank tests, class K

Chap 8. Cox model CLT

CLT for Poisson processes:

As $\lambda \to \infty$, for $t \in [0,1]$ (or $t \in [0,u]$,

$$\left(\frac{N(t) - \lambda t}{\sqrt{\lambda}}\right) \to B(t) .$$

At least for each fixed t we can show the convergence in distribution by checking the moment generating function. Any finite dimensional joint distribution can be proved by using the independent increment property.

A Poisson process $N(t)$ with $\lambda = 2$ can be viewed as the sum of two independent Poisson processes $N_1(t)$ and $N_2(t)$ each with $\lambda = 1$.

This also indicates a Poisson process $N(t)$ with a huge, integer valued λ can be viewed as the sum of many independent Poisson processes each with $\lambda = 1$.

Therefore, when $\lambda \approx k$

$$\left(\frac{N(t) - \lambda t}{\sqrt{\lambda}}\right) \approx \frac{\sum_{i=1}^{k}[N_i(t) - t]}{\sqrt{k}}$$

which for fixed t can use the CLT for iid sum to show this convergence. (at fixed t)

What about the tightness?

This section is included here mainly for reference. (Page 340 of FH book.) We are going to avoid the need of directly verify tightness by focusing on the stochastic processes that are martingales. For martingales, tightness reduces to a condition similar to the Lindeberg condition, which is much easier to check.

In the $D[0,u]$ space, (for processes that may have discontinuous sample paths) we can verify tightness by verify the following.

$W_n(t), 0 \leq t \leq u$ is tight, if

For any $\epsilon > 0$, and any $0 \leq s, t \leq u$,

$$\lim_{\delta \to 0} \limsup_{n \to \infty} P\{\sup_{|s-t|<\delta} |W_n(s) - W_n(t)| > \epsilon\} = 0$$

i.e. for any $\epsilon > 0, \eta > 0$ there exist $\delta^* > 0$ such that

$$\limsup_{n \to \infty} P\{\sup_{|s-t|<\delta^*} |W_n(s) - W_n(t)| > \epsilon\} < \eta$$

How to show the Poisson process is tight?

1. Exponential probability bound: For a Poisson r.v. N, with mean $\lambda > 0$,

$$P(|N - \lambda| > a) < \exp(?)$$

Center the process, sup becomes n points.

Bound the sup by the sum.

Example 2 The empirical process.

The fi-di is easy to proof.

The tightness is a little harder: in one of the proof the basic steps are

1. Center the process, sup becomes n points.
2. Bound the sup by sum of n probabilities.
3. each probability is bounded by an exponential inequality, the sum is also small.

CLT for empirical processes

$$W_n(t) = \sqrt{n}[\hat{F}_n(t) - F(t)]$$

where

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} I_{[X_i \leq t]} \ .$$

1. It is easy to prove the fidi convergence. Using multinomial distribution and multivariate CLT convergent (or Cremer-Wold device).

2. proof tightness.

a. How to bound the probability $P\{\sup_{|s-t|<\delta} |W_n(t) - W_n(s)| > \epsilon\}$

$$W_n(t) - W_n(s) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{[s < X_i \leq t]} - [F(t) - F(s)]$$

Bound the sup by polynomial many simple probabilities. And each probability is exponentially small. So the sum is also small.

For details, see the paper by Zhou *Closeness of empirical type processes*

15 Conditional Expectation

For a random variable X, a sigma algebra $\mathcal{F}$, what is the conditional expectation

$$E(X|\mathcal{F}) =?$$

Think of $\mathcal{F}$ as a partition of the probability space Ω. (the collection of sets made of those atoms and all the different unions of those atoms is $\mathcal{F}$.) This partition can be very coarse or very fine. (larger sigma algebra is a finer partition, i.e. it contain more sets).

Then the conditional expectation $E(X|\mathcal{F})$ is just a local average: within each atom, it is a constant: the average value of the X over this small piece of land (atom).
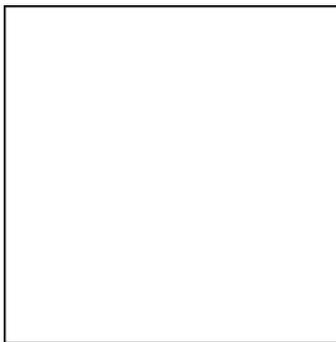

Figure 1: Space Ω.

Figure 2: The random variable X live on Ω. Darker = larger value.

Four special cases.

Case 1: When the partition is the whole space. $\mathcal{F} = \{\emptyset, \Omega\}$ then there is just one (non-empty atom, the whole space) so $E(X|\mathcal{F}) = $ constant $= \mu = EX$ on the whole space.

Case 2: Every different value of X sets (i.e. sets of the form $\{\omega|X(\omega) = constant\}$) are part of $\mathcal{F}$. Then $E(X|\mathcal{F}) = X$, i.e. $\mathcal{F}$ is so fine, there is no average at all.

Case 3 (the one pictured): the space Ω is partitioned into k atoms, then the sigma algebra $\mathcal{F}$ is all the combination/unions of those atoms.

Case 4 (independent) If X and $\mathcal{F}$ are independent, then $E(X|\mathcal{F}) = EX$

The conditional expectation with respect to this $\mathcal{F}$ is the "Pixelation" of the original random variable, i.e. the local average over each atoms. (Think of the example of "Ma-Sai-Ke".)
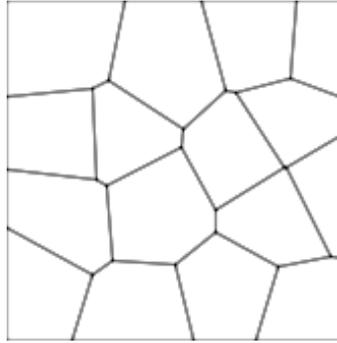
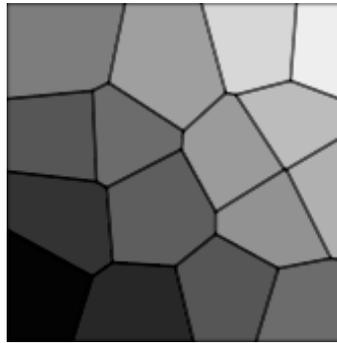*Figure 3: The σ-algebra $\mathcal{F}$ of subsets of Ω.*

Figure 4: The conditional expectation $E(X|\mathcal{F})$.

To show that independent information is irrelevant: Use one dimensional Ω to show conditional expectation. Then two information/random variables are independent when we use the product space as the new Ω. and the partition on one dimension did not affect the partition on the other dimension.

Also, by this interpretation it is clear we have

$$E(X|\mathcal{F}) = E[E(X|\mathcal{G})|\mathcal{F}]$$

when $\mathcal{G}$ is larger then $\mathcal{F}$, (average on a coarse partition is equal to average first on a finer partition and then on the coarse partition).

But the above equality is not true if $\mathcal{G}$ is smaller then $\mathcal{F}$ (in this case the right hand side is equal to $E(X|\mathcal{G})$).

If $\mathcal{G}$ is neither smaller nor larger than $\mathcal{F}$ then we do not know what equality holds.

Warning: unfortunately, this intuition only works for finite partitions (finite atoms). For (uncountable) infinite many atoms, the structure of $\mathcal{F}$ can be more complicated.

16   Homework Problems

(1). Suppose $N(t)$ is a Poisson process.

Verify the following is a (cont. time) martingale.

$$M(t) = N(t) - \lambda t$$

for t in $[0, T]$.

Be sure to define your filtration (increasing family of sigma algebras) $\mathcal{F}_t$.

(2). Suppose $N_i(t)$, $i = 1, 2, ..., n$ are iid Poisson(λ) processes. $0 < t < 8$. (or some other fixed finite number).

Show that

$$M_n(t) = \sum [N_i(t) - \lambda t]/\sqrt{(n\lambda)}$$

as $n \to \infty$ has the following property

(a) for fixed t the distribution of $M_n(t)$ converges to Normal (identify the variance).

(b) the sample path of $M_n(t)$ has jumps of size converge to 0.

Note: $\sum N_i(t)$ is also a Poisson process with parameter $(n\lambda)$

(3). If $N(t)$ is a Poisson process, show that $N(3t)$ is also a Poisson process. Identify the rateof this new poisson process.

(4). Suppose $X_1, X_2, ... X_n$ are iid r.v.s with a cont. CDF $F(t)$.

Let

$$U_n(t) = \sqrt{n}[\hat{F}_n(t) - F(t)]$$

Compute the covariance of $U(t)$ and $U(s)$, i.e. $EU(t)U(s) =$? (where $\hat{F}_n(t)$ is the empirical distribution based on the X_i's).

(5). Suppose $M(t)$ is the Poisson martingale we defined in the first problem. Show that

$$M^2(t) - \lambda t$$

is a (cont. time) $\mathcal{F}_t$ martingale. (same filtration we used there).

(6) Let $N(t) = I_{[X \leq t]}$, where X is a positive, cont. rv. Define the filtration $\mathcal{F}_t = \sigma\{N(s); 0 \leq s \leq t\}$.

(a) show that $h(t)I[X \geq t]$ is predictable. where $h(t)$ is the hazard function of X.

(b) verify the conditional expectation of the increment of $N(t)$ over $(t - dt, t]$, given $\mathcal{F}_{t-dt}$ (i.e. either given $I[X \geq t] = 0$ or given $I[X \geq t] = 1$) is $h(t)I[X \geq t]dt$.

Note this shows that $N(t) - \int_0^t h(s)I[X \geq s]ds$ is an $\mathcal{F}_t$ martingale.

(7) We proved in class the martingale representation of the Kaplan-Meier estimator for continuous CDF $F()$. Now Suppose that the CDF $F(\cdot)$ hava a jump at t, verify the martingale representation of the Kaplan-Meier estimator is still valid.

For simplicity, assume all the observed death times do not equal to t. Also, when the CDF $F(\cdot)$ is not continuous, the compensator of the counting preocess needs to be written as

$$\int_0^t I_{[X \geq s]}dH(s);$$

instead of

$$\int_0^t I_{[X \geq s]}h(s)ds.$$

The Cox partial likelihood is defined as

$$\prod_{i:\delta_i=1} \frac{\exp(\beta z_i)}{\sum_{j:y_j \geq y_i} \exp(\beta z_j)}$$

Sometimes, both the numerator and denominator carry a term like $h_0(y_i)$, they cancel anyway, but it may aid in understanding the meaning of the partial likelihood.