1.2. Event history analysis - Counting process formulation for survival models

September 01 - 25, 2015

Mathematical formulation

Cosnider a sample of n (uncensored) continuously distributed survival times $X_1, \ldots, X_n$ from survival functions $S()$ with hazard rate $\alpha()$

$$S(t) = P(X > t) = \mathcal{P}_0^t[1 - \alpha(s)ds] = \exp\Big(- \int_0^t \alpha(s)ds \Big).$$

Interpretation of the hazard rate α

$$P(t \leq X < t + dt \mid X \geq t) = \alpha(t)dt.$$

Cumulative hazard $A(t) = \int_0^t \alpha(s)ds$.
Interest is in estimation of $\alpha()$ or the cumulative hazard $A()$.

Survival data

- Data: $(\tilde{X}_i, D_i)$, $i = 1, \ldots, n$, $D_i =$ censoring indicator

$$\begin{cases} X_i = \tilde{X}_i & \text{if } D_i = 1, \text{ uncensored,} \\ X_i > \tilde{X}_i & \text{if } D_i = 0, \text{ censored.} \end{cases}$$

- All n survival periods start together at $t = 0$.
- Independent censoring: at any time t, the survival experience in the future is not statistically altered (from what it would have been without censoring) by censoring and survival experience in the past.

Filtration

Mathematically past is represented by so called history or filtration $(\mathcal{F}_t, t \geq 0)$ where $\mathcal{F}_t$ is the available data at time t and $\mathcal{F}_{t-}$ is the available data just prior to time t.

In survival data, $\mathcal{F}_t$ means the values of $(\tilde{X}_i, D_i)$ for all i such that $\tilde{X}_i \leq t$ otherwise just the information that $\tilde{X}_i > t$.

$\mathcal{F}_{t-} = \{(i : \tilde{X}_i < t, D_i) \text{ and } (i : \tilde{X}_i \geq t)\}$

$$P(t \leq \tilde{X}_i < t + dt, D_i = 1 \mid \mathcal{F}_{t-}) = \begin{cases} \alpha(t)dt & \text{if } \tilde{X}_i \geq t \\ 0 & \text{if } \tilde{X}_i < t \end{cases}$$

Expected failures

We have n individuals then the expectation of the sum of the indicator $1\{t \leq \tilde{X}_i < t + dt, D_i = 1\}$ is

$$
\begin{aligned}
&E(\sum_{i=1}^{n} 1\{t \leq \tilde{X}_i < t + dt, D_i = 1\} \mid \mathcal{F}_{t-}) \\
&= E(\#\{i : t \leq \tilde{X}_i < t + dt, D_i = 1\} \mid \mathcal{F}_{t-}) \\
&= \sum_{i=1}^{n} 1\{\tilde{X}_i \geq t\}\alpha(t)dt = \sum_{i=1}^{n} Y_i(t)\alpha(t)dt \\
&= Y(t)\alpha(t)dt \\
&= \lambda(t)dt,
\end{aligned}
$$

At risk process

For individual i, $Y_i(t) = 1\{\tilde{X}_i \geq t\}$ counts 1 if individual i is still at risk at time t and is 0 otherwise.

Summing over n such processes,
$Y(t) = \sum_{i=1}^{n} Y_i(t) = \sum_{i=1}^{n} 1\{\tilde{X}_i \geq t\}$ counts the number at risk at time t and gives the size of the risk set.

1. $Y(t)$ is a left-continuous process.
2. It is predictable with respect to the filtration $\mathcal{F}_{t-}$.
3. When the entry time is 0, the process $Y(t)$ is non-increasing with t.

Counting processes

A process counting the observed failures

$$N_i(t) = 1\{\tilde{X}_i \leq t, D_i = 1\},\ 0 \leq t \leq \tau$$

Proprties:

1. $N_i(0) = 0$
2. Right continuous process
3. Increments are $dN_i(t) = N_i(t) - N_i(t-)$ and is $+1$ in case of death

Counting process

- For the cohort of size n, the process $N = (N(t))_{t \geq 0}$ which counts the failures is

$$N(t) = \sum_{i=1}^{n} N_i(t) = \#\{i : \tilde{X}_i \leq t, D_i = 1\}$$

- Increment over the small interval $[t, t + dt)$ is $dN(t) = N((t + dt)-) - N(t-) = \#\{i : t \leq \tilde{X}_i < t + dt, D_i = 1\}$.
- The expectation of the increment given the history is

$$\begin{aligned} E(dN(t) \mid \mathcal{F}_{t-}) &= E(\sum_{i=1}^{n} 1\{t \leq \tilde{X}_i < t + dt, D_i = 1\} \mid \mathcal{F}_{t-}) \\ &= \lambda(t) dt. \end{aligned}$$

Intensity process

The intensity process $(\lambda(t))_{t\geq 0}$ is random, through dependence on the conditioning random variables in $\mathcal{F}_{t-}$.

Integrated or cumulative intensity process Λ is defined as

$$\Lambda(t) = \int_0^t \lambda(s)ds, t \geq 0$$

Counting process martingale (1)

The compensated counting process or counting process martingale M is defined as $M(t) = N(t) - \Lambda(t)$
difference between the counting process and its expectation!

$$E(dM(t) \mid \mathcal{F}_{t-}) = E(dN(t) - d\Lambda(t) \mid \mathcal{F}_{t-}) = 0.$$

The martingale property says that the conditional expectation of increments of M over small time intervals, given the past at the beginning of the interval, is zero.

This is heuristically equivalent to

$$E(M(t) \mid \mathcal{F}_s) = M(s), \ \forall\ s < t, \ \ E(M(t) \mid \mathcal{F}_0) = M(0) = 0.$$

Counting process martingale (2)

Martingale as a pure noise process - difference between the observed and expected number of death in the interval $[0, t]$

Method of moments

Example: Simulation of a counting process and its compensator

Predictable variation of a martingale (1)

Consider the process M^2 and note that

$$\begin{aligned} d(M^2)(t) &= M((t+dt)-)^2 - M(t-)^2 \\ &= (dM(t))^2 + 2dM(t)M(t-) \end{aligned}$$

$$\begin{aligned} E(d(M^2)(t) \mid \mathcal{F}_{t-}) &= E((dM(t))^2 \mid \mathcal{F}_{t-}) \\ &= var(dM(t) \mid \mathcal{F}_{t-}) = d < M > (t). \end{aligned}$$

Predictable variation of a martingale (2)

If M is a compensated counting process and the compensator Λ is continuous, then M's predictable variation process $< M >$ is simply Λ itself.
This can be seen by noting that

- No two uncensored failure times fall into the same small interval and hence, the increments of N over small time intervals are 0 or 1.
- $dN(t) = 1$ w.p. $d\Lambda(t)$ and $dN(t) = 0$ w.p. $1 - d\Lambda(t)$.
- $dM(t) = 1 - d\Lambda(t)$ w.p. $d\Lambda(t)$ and $dM(t) = 0 - d\Lambda(t)$ w.p. $1 - d\Lambda(t)$.
- $var(dM(t) \mid \mathcal{F}_{t-}) = (1 - d\Lambda(t))d\Lambda(t) \approx d\Lambda(t)$

Some insight

Conditional means and variances of increments of the counting process N over small intervals both coincide with the conditional local rate λ.

For a Poisson random variable, mean and variance coincide.

A counting process N behaves locally at time t, and conditional on the past, just like a Poisson process with rate $\lambda(t)$.

Estimation (1)

Statistical problem of nonparametric estimation of the cumulative hazard rate $A(t) = \int_0^t \alpha(t)$

$$\begin{aligned} dN(t) &= d\Lambda(t) + dM(t) = Y(t)\alpha(t)dt + dM(t) \\ \frac{dN(t)}{Y(t)} &= \alpha(t)dt + \frac{dM(t)}{Y(t)}, \text{ provided } Y(t) > 0, \\ \frac{J(t)}{Y(t)}dN(t) &= \alpha(t)dt + \frac{J(t)}{Y(t)}dM(t), \end{aligned}$$

where $J(t) = 1\{Y(t) > 0\}$.

Estimation (2)

$dM(t)$ is a pure noise and so as $(J(t)/Y(t))dM(t)$ and

$$\begin{aligned}
E(\frac{J(t)}{Y(t)}dM(t) \mid \mathcal{F}_{t-}) &= \frac{J(t)}{Y(t)}E(dM(t) \mid \mathcal{F}_{t-}) = 0, \\
var(\frac{J(t)}{Y(t)}dM(t) \mid \mathcal{F}_{t-}) &= \frac{J(t)}{Y(t)^2}\ var(dM(t) \mid \mathcal{F}_{t-}) \\
&= \frac{J(t)}{Y(t)^2} < M > (t).
\end{aligned}$$

Estimation (3)

Define

$$\begin{aligned}
\hat{A}(t) &= \int_0^t \frac{J(s)}{Y(s)} dN(s) \\
&= \underbrace{\int_0^t J(s)\alpha(s)ds}_{A^*(t)} + \underbrace{\int_0^t \frac{J(s)}{Y(s)} dM(s)}_{Z(t)}
\end{aligned}$$

$$\begin{aligned}
Z(t) &= \hat{A}(t) - A^*(t) \\
E(dZ(t) \mid \mathcal{F}_{t-}) &= 0, \\
var(dZ(t) \mid \mathcal{F}_{t-}) &= \frac{J(t)}{Y(t)^2} < M > (t).
\end{aligned}$$

Estimation (4)

Remarks:

- $\hat{A}(t)$ is indeed the Nelson-Aalen estimator - sum over the failure times up to and including t of the reciprocals of the corresponding risk set sizes.
- $A^*(t)$ is the same as $A(t)$, we only omit contributions $\alpha(s)ds$ where the risk set is empty.
- $\sqrt{n}(\hat{A}(t) - A^*(t)) = \sqrt{n}Z(t)$ goes to a zero-mean Gaussian martingale with variation process $< Z > (t)$, as $n \to \infty$ and $Y(t)/n \to y(t)$ where $y(t)$ a deterministic function.