2. ESTIMATING THE SURVIVAL AND HAZARD FUNCTIONS

This chapter's topics:

- Life table method
- Role of assumptions on censoring distributions
- Nonparametric maximum likelihood
- Derivation of Kaplan-Meier estimator: NPMLE
- Derivation of Greenwood's formula
- Nelson-Aalen estimator: Properties and derivation of variance
- Role of counting processes
- Approximating stochastic integrals

This chapter's goal:

1. Estimating the survival function (and other functions) when there are no covariates.
2. Both informally and with a connection to counting processes and martingales

References: TG Chapter 2; FH Chapter 0; KP Chapter 1

Distributional Assumptions

- Distributional assumptions on T_i:
  - need to balance objectives of data reduction (via assumptions) with robustness
  - i.e., the stronger and/or more numerous are the assumptions, the less robust will be the results
  - two ends of the continuum:

    nonparametric methods, . . . , (fully) parametric models

- Distributional assumptions on C_i:
  - since C_i is typically of no interest, distributional assumptions are best avoided
  - we illustrate this idea through an example

Life Table Methods

- *Life table* methods are often applied to grouped survival data (exact survival times unknown)
- the $(0, \tau]$ interval is divided into K non-overlapping subintervals:

$$(\tau_0, \tau_1], (\tau_1, \tau_2], \ldots, (\tau_{K-1}, \tau_K],$$

  where $\tau_0 \equiv 0$ and $\tau_K \equiv \tau$
- τ represents the end of the observation period, while τ_k is the end of the kth subinterval
- Label the kth subinterval $I_k = (\tau_{k-1}, \tau_k]$
- Assume (as usual) C_i is independent of T_i
- Assumptions on C_i: required

- i.e., how to account for censoring within intervals?

Life Table Methods: Example

- Life table: A summary of the survival data grouped into convenient intervals
- Oldest estimator of survival functions
- Example: Consider a prospective cohort study to evaluate a new treatment for hypertension
  - patients begin follow-up when treatment is assigned
  - patients are examined at the end of each year of follow-up
  - event of interest: death (all causes)
  - exact times of death/censoring are not known, although the year of occurrence is recorded
  - patients who do not show up for their annual follow-up visit are telephoned, to determine their status (dead, alive)
  - **Objective**: to estimate $S(t)$ for $t = 5$ years

Life Table Method: Observed Data

- The kth subinterval is labeled $I_k = (\tau_{k-1}, \tau_k]$ where
  - y_k = number at risk at start of I_k
  - d_k = number of events in I_k
  - c_k = number of subjects censored in I_k
- The following data are observed over the first 5 years

k	τ_{k-1}	τ_k	y_k	d_k	c_k
1	0	1	146	27	3
2	1	2	116	18	10
3	2	3	88	21	10
4	3	4	57	9	3
5	4	5	45	1	3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Note: Data beyond $t = 5$ are not required to estimate $S(5)$

Life Table Example: Censoring Assumption #A1-A2

Let us begin with some unrealistic censoring assumptions

A1. Assume that all censoring occurs at the end of 5 years

$$\begin{aligned}\widehat{S}(5) = 1 - \widehat{F}(5) &= 1 - \frac{\sum_{k=1}^{5} d_k}{y_1} \\ &= 1 - \frac{76}{146} = 0.479\end{aligned}$$

A2. Assume that all censoring occurs at the start of $(0, 5]$

$$\begin{aligned}\widehat{S}(5) = 1 - \widehat{F}(5) &= 1 - \frac{\sum_{k=1}^{5} d_k}{y_1 - \sum_{k=1}^{5} c_k} \\ &= 1 - \frac{76}{146 - 29} = 0.350\end{aligned}$$

- Both of these estimators are biased, since C_i clearly does not follow the assumed distribution (and we are ignoring information we have)
- A more realistic assumption on C_i would incorporate the fact that the interval in which C_i occurs is always observed
  - need only to make assumptions on C_i within each I_k
- Necessitates an estimator for $S(t)$ which combines interval-specific probabilities

- Set $q_k = P(T_i \leq \tau_k | T_i > \tau_{k-1}) = P(T_i \in I_k | T_i > \tau_{k-1})$, such that

$$S(\tau_k) = \prod_{j=1}^{k}(1 - q_k)$$

- The estimator for q_k will depend on the assumptions regarding C_i over I_k

Life Table Example: Censoring Assumption #A3 and #A4

A3. Consider a more realistic censoring assumption: censoring occurs at the start of each year

- Set $\widehat{q}_k = d_k/(y_k - c_k)$

k	τ_{k-1}	τ_k	y_k	d_k	c_k	$\widehat{q}_k$	$\widehat{S}(\tau_k)$
1	0	1	146	27	3	0.189	0.811
2	1	2	116	18	10	0.170	0.673
3	2	3	88	21	10	0.269	0.492
4	3	4	57	9	3	0.167	0.410
5	4	5	45	1	3	0.024	0.400

A4. Now, assume that censoring occurs at the *end* of each year

- Set $\widehat{q}_k = d_k/y_k$

k	τ_{k-1}	τ_k	y_k	d_k	c_k	$\widehat{q}_k$	$\widehat{S}(\tau_k)$
1	0	1	146	27	3	0.185	0.815
2	1	2	116	18	10	0.155	0.689
3	2	3	88	21	10	0.239	0.524
4	3	4	57	9	3	0.158	0.441
5	4	5	45	1	3	0.022	0.432

- Censoring assumptions #A3 and #A4 are clearly more realistic than #A3 and #A4 but is this enough?

Life Table Example: Censoring Assumption #A5 (Uniform)

- **Note**: Actual censoring time is not used and may not even be known.
  - seems incorrect to have them count as much as uncensored subjects
  - however, they do contribute some follow-up to I_k

A5. Most commonly, it is assumed that C_i is distributed uniformly within each I_k

- Average censoring time is $E[C_i|C_i \in I_k] = (\tau_{k-1} + \tau_k)/2$
- implies that each censored subject should count as 1/2 an observation
- which suggests $\widehat{q}_k = d_k/(y_k - c_k + c_k/2)$

k	τ_{k-1}	τ_k	y_k	d_k	c_k	$\widehat{q}_k$	$\widehat{S}(\tau_k)$
1	0	1	146	27	3	0.187	0.813
2	1	2	116	18	10	0.162	0.681
3	2	3	88	21	10	0.253	0.509
4	3	4	57	9	3	0.162	0.426
5	4	5	45	1	3	0.023	0.417

Standard One-Sample Set-Up

- We now return to the setting where exact observation times are available
- Specifically, we observe (X_i, Δ_i) for a sample of $i = 1, \ldots, n$ subjects, where (recall) $X_i = T_i \wedge C_i$ and $\Delta_i = I(T_i < C_i)$
- Assume that $\{(X_i, \Delta_i)\}_{i=1}^n \sim$ i.i.d.
- Objective: to estimate $S(t)$ for $t \in (0, \tau]$ where $\tau = \max\{X_1, \ldots, X_n\}$
- We wish to make no distributional assumptions on T_i or C_i, except that $T_i \perp T_j$ $(i \neq j)$ and $T_i \perp C_i$
- Set-up: the (unique) observed death times are given by:

$$t_1 < t_2 < \ldots < t_D$$

- Let d_j be the number of deaths observed at t_j
- Let c_j be the number of censored subjects in $(t_j, t_{j+1}]$, with the exact censoring times in $(t_j, t_{j+1}]$ given by

$$c_{j1} < c_{j2} < \ldots < c_{jc_j}$$

Kaplan-Meier Estimator

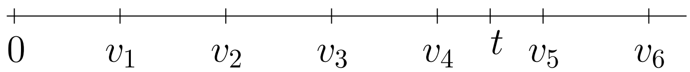
- Observed failure times: $t_1 < t_2 < \ldots < t_D$, where D is the number of unique times at which deaths are observed
- the Kaplan-Meier estimator of $S(t)$ is given by:

$$\widehat{S}(t) = \prod_{j:t_j \leq t} \left\{1 - \frac{d_j}{Y_j}\right\}$$

  - $Y_j = Y(t_j)$, number at risk at $t = t_j$ and
  - $d_j = dN(t_j)$, number of failures at $t = t_j$

- also known as *product limit* estimator
- could consider d_j/Y_j to be an estimator of $d\Lambda(t_j)$
- Intuitively why does the K-M estimator work?

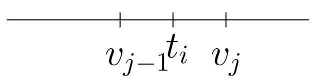
For any set of small intervals:



$$S(t) = P(T > t) = P(T > t|T > v_4)P(T > v_4|T > v_3)$$
$$\times P(T > v_3|T > v_2)P(T > v_2|T > v_1)P(T > v_1)$$

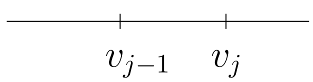
For very small intervals:

If this interval contains an observed event time:



$$\widehat{P}(T > v_j|T > v_{j-1}) \cong 1 - \frac{d_i}{N_i}$$

If the interval does not:



$$\widehat{P}(T > v_j|T > v_{j-1}) = 1$$

$$\widehat{S}(t) = 1 \cdot 1 \cdot 1 \cdot (1 - \frac{d_1}{N_1}) \cdot (1 - \frac{d_2}{N_2}) \cdot 1 \cdot 1 \cdots (1 - \frac{d_i}{N_i}) = \prod_{j: t_j \le t_i} (1 - \frac{d_j}{N_j})$$

where t_i is the largest observed event time $\le t$.

Example: K-M Procedure

- Example: The following data are observed, with + denoting censoring:

i	1	2	3	4	5	6	7	8	9	10
X_i	2	5^+	8	12^+	15	21^+	25	29	30^+	34

$$
\begin{aligned}
\widehat{S}(20) &= \{1 - d\widehat{\Lambda}(2)\}\{1 - d\widehat{\Lambda}(8)\}\{1 - d\widehat{\Lambda}(15)\} \\
&= \left\{1 - \frac{dN(2)}{Y(2)}\right\}\left\{1 - \frac{dN(8)}{Y(8)}\right\}\left\{1 - \frac{dN(15)}{Y(15)}\right\} \\
&= \{1 - 1/10\}\{1 - 1/8\}\{1 - 1/6\} \\
&= 0.656
\end{aligned}
$$

- We compute the KM estimator across the entire observation period

j	t_j	Y_j	d_j	$\widehat{\lambda}_j$	$\widehat{S}(t_j)$
–	0	10	0	0	1
1	2	10	1	0.1	0.9
2	8	8	1	0.125	0.788
3	15	6	1	0.167	0.656
4	25	4	1	0.25	0.492
5	29	3	1	0.333	0.328
6	34	1	1	1	0

Kaplan-Meier Estimator: Properties

- $\widehat{S}(t)$ has a total of D jumps; one at each unique failure time
- $\widehat{S}(t)$ is well-defined up to the last observation time, $\tau = \max\{X_1, \ldots, X_n\}$
- If last subject at risk is a death, $\widehat{S}(t)$ will drop to 0; if not, $\widehat{S}(t)$ will not reach 0

- Variance of K-M estimator can be estimated by Greenwood's formula:

$$\widehat{\sigma}_G^2(t) = \widehat{S}(t)^2 \sum_{j:t_j \le t} \frac{d_j}{Y_j(Y_j - d_j)}$$

Kaplan-Meier Estimator as a NPMLE

- We now derive the Kaplan-Meier estimator of $S(t)$ as a nonparametric MLE
- The likelihood contribution, under independent censoring,

$$\begin{aligned} L_i &= \{f(X_i)S_C(X_i)\}^{\Delta_i}\{S(X_i)f_C(X_i)\}^{1-\Delta_i} \\ &\propto f(X_i)^{\Delta_i}S(X_i)^{1-\Delta_i} = S(X_i)\lambda(X_i)^{\Delta_i} \end{aligned}$$

where S_C and f_C pertain to C_i

- We can build up L through the following probabilities,

$$\begin{aligned} P(T_i = t_j) &= S(t_j^-) - S(t_j) \\ P(T_i > t_j) &= S(t_j) \end{aligned}$$

- Likelihood is given by,

$$L(S) = L = \prod_{j=1}^{D} \left[\{S(t_j^-) - S(t_j)\}^{d_j} \prod_{\ell=1}^{c_j} S(c_{j\ell}) \right]$$

where we interpret $L = L(S)$ as a likelihood across the parameter space of all valid $S(t)$ functions

- To maximize L,
  - set $\widehat{S}(t)$ to be discontinuous at t_j for $j = 1, \ldots, D$
  - set $\widehat{S}(c_{j\ell}) = \widehat{S}(t_j)$ for $\ell = 1, \ldots, c_j$
- Therefore, $L(S)$ is maximized by $\widehat{S}$ which is cadlag (right continuous with left-hand limits), with jumps at t_j

- Re-expressing the likelihood, given our new information on its maximization,

$$L = \prod_{j=1}^{D} \{\lambda_j S(t_j^-)\}^{d_j} S(t_j)^{c_j} \tag{1}$$

where $\lambda_j = d\Lambda(t_j)$

- Now, the jumps at $t = t_j$ ($j = 1, \ldots, D$) in $S(t)$ will result in jumps in $\Lambda(t)$ at the same times
- Recall:

$$S(t) = \prod_{s \in (0,t]} \{1 - d\Lambda(s)\}$$

- We will now re-write the likelihood in terms of Λ increments, λ_j ($j = 1, \cdots, D$), then compute the MLEs, $\widehat{\lambda}_j$; after which, the following will be the maximizer of L,

$$\widehat{S}(t) = \prod_{j:t_j \in (0,t]} \{1 - \widehat{\lambda}_j\}$$

- We now compute the $\widehat{\lambda}_j$ for $j = 1, \ldots, D$
- Note that

$$\begin{aligned} S(t_j) &= \prod_{\ell=1}^{j} (1 - \lambda_\ell) \\ S(t_j^-) &= \prod_{\ell=1}^{j-1} (1 - \lambda_\ell) \end{aligned}$$

- Substituting these expressions into (1), we get

$$
\begin{aligned}
L &= \prod_{j=1}^{D} \left\{ \lambda_j \prod_{\ell=1}^{j-1} (1-\lambda_\ell) \right\}^{d_j} \left\{ \prod_{\ell=1}^{j} (1-\lambda_\ell) \right\}^{c_j} \\
&= \prod_{j=1}^{D} \lambda_j^{d_j} \prod_{\ell=1}^{j} (1-\lambda_\ell)^{d_j+c_j} (1-\lambda_j)^{-d_j} \\
&= \prod_{j=1}^{D} \lambda_j^{d_j} (1-\lambda_j)^{-d_j} \prod_{\ell=1}^{j} (1-\lambda_\ell)^{d_j+c_j} \\
&= \prod_{j=1}^{D} \lambda_j^{d_j} (1-\lambda_j)^{-d_j} \prod_{j=1}^{D} \prod_{\ell=1}^{j} (1-\lambda_\ell)^{d_j+c_j} \\
&= \prod_{j=1}^{D} \lambda_j^{d_j} (1-\lambda_j)^{-d_j} \prod_{\ell=1}^{D} \prod_{j=\ell}^{D} (1-\lambda_\ell)^{d_j+c_j} \\
&= \prod_{j=1}^{D} \lambda_j^{d_j} (1-\lambda_j)^{-d_j} \prod_{\ell=1}^{D} (1-\lambda_\ell)^{\sum_{j=\ell}^{D}(d_j+c_j)} \\
&= \prod_{j=1}^{D} \lambda_j^{d_j} (1-\lambda_j)^{-d_j} \prod_{\ell=1}^{D} (1-\lambda_\ell)^{Y_\ell} \\
&= \prod_{j=1}^{D} \lambda_j^{d_j} (1-\lambda_j)^{Y_j-d_j}
\end{aligned}
$$

where $Y_j = Y(t_j) = \sum_{k=j}^{D}(d_k + c_k)$

- We obtain the MLEs for λ_j,

$$\ell = \log L = \sum_{j=1}^{D} [d_j \log \lambda_j + (Y_j - d_j) \log(1-\lambda_j)]$$

$$\frac{\partial \ell}{\partial \lambda_j} = \frac{d_j}{\lambda_j} - \frac{Y_j - d_j}{1-\lambda_j}$$

$$\widehat{\lambda}_j = \frac{d_j}{Y_j}$$

- Through the invariance property, the NPMLE of $S(t)$ is given by

$$\widehat{S}(t) = \prod_{j:t_j \leq t} (1 - \widehat{\lambda}_j)$$

Cumulative Hazard Estimator

- Aside: Using the invariance property, an estimator for the cumulative hazard can be derived,

$$\Lambda(t) = \int_0^t d\Lambda(s)$$
$$\widehat{\Lambda}(t) = \int_0^t d\widehat{\Lambda}(s) = \sum_{j:t_j \leq t} \frac{d_j}{Y_j}$$

- Known as the Nelson-Aalen estimator

Kaplan-Meier Estimator: Deriving Variance

- Obtaining the (observed) information matrix,

$$I_{jj} = \frac{-\partial^2 \ell}{\partial \lambda_j^2} = \frac{d_j}{\lambda_j^2} + \frac{Y_j - d_j}{(1 - \lambda_j)^2}$$
$$I_{j\ell} = \frac{-\partial^2 \ell}{\partial \lambda_j \partial \lambda_k} = 0$$

- i.e., $\mathbf{I}$ matrix is diagonal, implying that $\widehat{\lambda}_j$ and $\widehat{\lambda}_k$ are uncorrelated
- Delta Method
  - Suppose that $\widehat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}_0$
  - First-order Taylor series, for $g(\cdot)$ invertible and differentiable,

$$g(\widehat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta}_0) = \left.\frac{\partial g}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}_0} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$$

as $n \to \infty$

– Then

$$V\{g(\widehat{\boldsymbol{\theta}})\} = \frac{\partial g}{\partial \boldsymbol{\theta}'} V(\widehat{\boldsymbol{\theta}}) \frac{\partial g}{\partial \boldsymbol{\theta}}$$

- With a view to estimating $V(\widehat{\lambda}_j)$, we estimate the diagonal elements of $\mathbf{I}$,

$$\begin{aligned}
\widehat{I}_{jj} = I_{jj}(\widehat{\lambda}_j) &= \frac{d_j Y_j^2}{d_j^2} + \frac{(Y_j - d_j)Y_j^2}{(Y_j - d_j)^2} \\
&= \frac{Y_j^2}{d_j} + \frac{Y_j^2}{(Y_j - d_j)} \\
&= Y_j^2 \left\{ \frac{1}{d_j} + \frac{1}{(Y_j - d_j)} \right\} = \frac{Y_j^3}{d_j(Y_j - d_j)}
\end{aligned}$$

- Therefore, an estimator of the variance is given by

$$\widehat{V}(\widehat{\lambda}_j) = \widehat{I}_{jj}^{-1} = \frac{d_j(Y_j - d_j)}{Y_j^3} = \frac{\widehat{\lambda}_j(1 - \widehat{\lambda}_j)}{Y_j}$$

- To avoid taking variance of a product, we work with

$$\log \widehat{S}(t) = \sum_{j:t_j \le t} \log(1 - \widehat{\lambda}_j)$$

knowing that, via the Delta method,

$$V\{\log \widehat{S}(t)\} = S(t)^{-2} V\{\widehat{S}(t)\}$$

- We now recover the variance as follows,

$$\begin{aligned}
V\{\widehat{S}(t)\} &= S(t)^2 V \left\{ \sum_{j:t_j \le t} \log(1 - \widehat{\lambda}_j) \right\} \\
&= S(t)^2 \sum_{j:t_j \le t} V\{\log(1 - \widehat{\lambda}_j)\} \\
&= S(t)^2 \sum_{j:t_j \le t} \frac{\lambda_j}{(1 - \lambda_j) Y_j}
\end{aligned}$$

- Since

$$V\{\log(1-\widehat{\lambda}_j)\} = \frac{1}{(1-\lambda_j)^2}V(\widehat{\lambda}_j)$$
$$= \frac{1}{(1-\lambda_j)^2}\frac{\lambda_j(1-\lambda_j)}{Y_j} = \frac{\lambda_j}{Y_j(1-\lambda_j)}$$

- Plugging in the MLEs,

$$\widehat{V}\{\widehat{S}(t)\} = S(t)^2 \sum_{j:t_j\leq t} \frac{d_j}{Y_j(Y_j-d_j)}$$

which is Greenwood's formula.

Notes on Kaplan-Meier Estimator

- Heuristic derivation of K-M estimator's properties is obtained by assuming that the death process behaves locally like a binomial variates; results obtained are equal to those just shown
- Greenwood's formula is consistent for $V\{\widehat{S}(t)\}$
- However, its justification via MLE theory is invalid
  - as $n \to \infty$, the parameter space also goes to ∞
  - violates regularity conditions required to employ standard MLE asymptotic results
  - in addition, the observed (instead of the expected) information matrix was used

Kaplan-Meier: Confidence Intervals

- $n^{1/2}\{\widehat{S}(t)-S(t)\}$ is asymptotically normal, meaning that a 95% CI could be estimated by,

$$\widehat{S}(t) \pm 1.96\, \widehat{\sigma}_G(t)$$

where $\widehat{\sigma}_G(t)$ is the standard deviation estimated by Greenwood's method

- CI bounds need not lie within $[0,1]$

- Often, CI's for $S(t)$ are based on a transformation
  - apply normal approximation to $g(\widehat{S}(t))$, rather than $\widehat{S}(t)$ itself, then back-transform
  - e.g., common choices for g are $g(x) = \log(x)$ and $g(x) = \log\{-\log(x)\}$
- Set $g(x) = \log(x)$
- using the Delta Method,

$$\begin{aligned}
\widehat{V}\{\log \widehat{S}(t)\} &= \widehat{S}(t)^{-2}\, \widehat{\sigma}_G^2(t) \\
\text{CI } \{\log S(t)\} &= \log \widehat{S}(t) \pm 1.96 \widehat{\sigma}_G(t) \widehat{S}(t)^{-1} \\
&= (\widehat{L}_\ell, \widehat{U}_\ell) \\
\text{CI } \{S(t)\} &= (\exp\{\widehat{L}_\ell\}, \exp\{\widehat{U}_\ell\})
\end{aligned}$$

- lower CI bound constrained to be non-negative; no constraint on upper CI bound
- Another option: $g(x) = \log\{-\log(x)\}$

$$g'(x) = \frac{1}{x \log x} \qquad\qquad g^{-1}(x) = \exp\{-e^x\}$$

- Using the Delta method,

$$\begin{aligned}
\widehat{V}\{\log(-\log \widehat{S}(t))\} &= \widehat{\sigma}_G^2(t)\{\widehat{S}(t) \log \widehat{S}(t)\}^{-2} \\
\text{CI } \{\log[-\log S(t)]\} &= \log\{-\log \widehat{S}(t)\} \pm 1.96 \frac{\widehat{\sigma}_G(t)}{\{\widehat{S}(t) \log \widehat{S}(t)\}} \\
&= (\widehat{L}_{\ell\ell}, \widehat{U}_{\ell\ell}) \\
\text{CI } \{S(t)\} &= (\exp\{-\exp(\widehat{U}_{\ell\ell})\}, \exp\{-\exp(\widehat{L}_{\ell\ell})\})
\end{aligned}$$

- Both bounds are constrained to lie in $[0,1]$ under this transform

Stochastic Integrals of Counting Processes

- $$\widehat{\Lambda}(t) = \int_0^t Y(s)^{-1} dN(s)$$

  : stochastic integral with respect to the counting process $N(t)$

- We will show that $\widehat{\Lambda}(t)$ is asymptotically unbiased for $\Lambda(t)$ under certain conditions.

- Recall: $N(t)$ is a counting process;
  - counts events in $(0, t]$
  - *cadlag*: right continuous with left hand limits
  - non-decreasing, piece-wise constant, increments of size 1

- $Y(t)$ is the at-risk process:
  - left-continuous: $Y(t) = Y(t^-)$
  - non-increasing, with decrements of size 1

Nelson-Aalen Estimator: Properties

- We now demonstrate that the Nelson-Aalen estimator is asymptotically unbiased over the $(0, \tau_*]$ where
  - τ_* is a constant chosen such that $P(X_i \geq \tau_*) > 0$;
  - the importance of this condition will become clear later

- We begin with,

$$\{\widehat{\Lambda}(t) - \Lambda(t)\} = \int_0^t Y(s)^{-1} dN(s) - \int_0^t d\Lambda(s)$$
$$= \int_0^t Y(s)^{-1}\{dN(s) - Y(s)d\Lambda(s)\}$$

- Now, break $(0, t]$ up into m non-overlapping subintervals of equal length
  - $0 \equiv t_0 < t_1 < t_2 < \ldots < t_{m-1} < t_m \equiv t$;
  - subintervals: $(0, t_1]$, $(t_1, t_2] \ldots, (t_{m-1}, t]$
  - $t_0 = 0$, $t_1 = t/m$, $t_2 = 2t/m$, …, $t_m = t$
- Define:

$$\Delta N_j = N(t_j) - N(t_{j-1})$$
$$\Delta(t_j) = t_j - t_{j-1}$$

- When $m \to \infty$, each $(t_{j-1}, t_j]$ can contain only 1 event

- We have,

$$\int_0^t Y(s)^{-1}\{dN(s) - Y(s)d\Lambda(s)\}$$
$$= \lim_{m\to\infty} \sum_{j=1}^{m} Y(t_j)^{-1}\{\Delta N_j - Y(t_j)\lambda(t_j)\Delta(t_j)\}$$

  where we would replace = with ≈ for finite m

- Let F_j denote all the death and censoring information up to time t_j
  - i.e., the process history up to time t_j

- Key points:

$$E[Y(t_j)|F_j] = Y(t_j)$$
$$\Delta N_j|F_j \sim \text{Binomial}\{Y(t_j), p_j\}$$
$$p_j = P(t_{j-1} < X_i \leq t_j, \Delta_i = 1)|X_i > t_{j-1})$$
$$\approx \lambda_1^{\#}(t_j)\Delta(t_j)$$

where

$$\lambda_1^{\#}(t_j) = \lim_{\delta \to 0} \frac{1}{\delta} P(t_j \le X_i < t_j + \delta, \Delta_i = 1 | X_i \ge t_j).$$

- Under the independent censoring, we have

$$p_j \approx \lambda(t_j)\Delta(t_j)$$

$$\begin{aligned}
\Big(\because \lambda_T^{\#}(t) &= \lim_{\delta \downarrow 0} \frac{1}{\delta} P(t \le X_i < t + \delta, \Delta_i = 1 | X_i \ge t) \\
&= \lim_{\delta \downarrow 0} \frac{1}{\delta} \frac{P(t \le X_i < t + \delta, \Delta_i = 1)}{P(X_i \ge t)} \\
&= \lim_{\delta \downarrow 0} \frac{1}{\delta} \frac{P(t \le T_i < t + \delta, C_i > T_i)}{P(T_i \ge t, C_i \ge t)} \\
&= \lim_{\delta \downarrow 0} \frac{1}{\delta} \frac{P(t \le T_i < t + \delta, C_i \ge t)}{P(T_i \ge t, C_i \ge t)}
\end{aligned}$$

If $C_i \perp T_i$, then

$$\begin{aligned}
\lambda_T^{\#}(t) &= \lim_{\delta \downarrow 0} \frac{1}{\delta} \frac{P(t \le T_i < t + \delta, C_i \ge t)}{P(T_i \ge t, C_i \ge t)} \\
&= \lim_{\delta \downarrow 0} \frac{1}{\delta} \frac{P(t \le T_i < t + \delta)P(C_i \ge t)}{P(T_i \ge t)P(C_i \ge t)} \\
&= \lim_{\delta \downarrow 0} \frac{1}{\delta} \frac{P(t \le T_i < t + \delta)}{P(T_i \ge t)} \\
&= \lambda_T(t)\Big)
\end{aligned}$$

- Taking the mean of our N-A integral,

$$E\left[\widehat{\Lambda}(t) - \Lambda(t)\right] = E\left[\sum_{j=1}^{m} \frac{1}{Y(t_j)} \{\Delta N_j - Y(t_j)\lambda(t_j)\Delta(t_j)\}\right]$$

- Consider each summand,

$$\begin{aligned}
& E\left[\frac{1}{Y(t_j)} \{\Delta N_j - Y(t_j)\lambda(t_j)\Delta(t_j)\}\right] \\
= \;& E\left[E\left[\frac{1}{Y(t_j)} \{\Delta N_j - Y(t_j)\lambda(t_j)\Delta(t_j)\} | F_j\right]\right] \\
= \;& E\left[\frac{1}{Y(t_j)} \{E[\Delta N_j | F_j] - Y(t_j)\lambda(t_j)\Delta(t_j)\}\right]
\end{aligned}$$

- The inner expectation equals 0 $\left(\because \Delta N_j | F_j \sim \text{Binomial}\{Y(t_j), \lambda(t_j)\Delta(t_j)\}\right)$

- Therefore, the marginal mean will also equal 0.

- Finally, summing the m zero-mean quantities,

$$E\left[\widehat{\Lambda}(t) - \Lambda(t)\right] \;=\; 0.$$

Technical Issues: Tail of Distribution

- Technical note: Even as $m \to \infty$ (meaning that $\Delta(t_j) \to 0$), $\widehat{\Lambda}(t)$ is still not unbiased.

- For t towards the tail of the X_i distribution, $P\{Y(t) = 0\} > 0$, leading to 0/0 terms in $\widehat{\Lambda}(t)$
  - It is typical to adopt the convention that $0/0 \equiv 0$;
  - It is also helpful to re-write the estimator as follows,

$$\widehat{\Lambda}(t) \;=\; \int_0^t I\{Y(s) > 0\} Y(s)^{-1} dN(s)$$

  which is consistent for $\int_0^t P\{Y(s) > 0\} d\Lambda(s)$

- We can then state the following,

$$E\left[\int_0^t \{Y(s)^{-1} dN(s) - \lambda(s)ds\} I\{Y(s) > 0\}\right] \;=\; 0$$

  implying that $E[\widehat{\Lambda}(t)] = E[\Lambda^*(t)]$ where $\Lambda^*(t) = \int_0^t I\{Y(s) > 0\} d\Lambda(s)$.

- Next, we derive the variance of the Nelson-Aalen estimator

Martingales: Introduction

- Define the Martingale and its corresponding increment as

$$\begin{aligned} M(t) &= N(t) - \int_0^t Y(s)\lambda(s)ds \\ dM(t) &= dN(t) - Y(t)\lambda(t)dt \end{aligned}$$

- For the purposes of approximating stochastic integrals with sums, define

$$\Delta M_j = \Delta N_j - Y(t_j)\lambda(t_j)\Delta(t_j)$$

- Since $E[\Delta N_j | F_j] = Y(t_j)\lambda(t_j)\Delta(t_j)$, we have $E[\Delta M_j | F_j] = 0$.

- Letting $m \to \infty$,

$$\begin{aligned} E[dN(t)|\mathcal{F}(t^-)] &= Y(t)\lambda(t)dt \\ \text{such that } E[dM(t)|\mathcal{F}(t^-)] &= 0 \end{aligned}$$

- In terms of martingale increments, we can write

$$\widehat{\Lambda}(t) - \Lambda(t) = \int_0^t \frac{dM(s)}{Y(s)} = \lim_{m\to\infty} \sum_{j=1}^m \frac{\Delta M_j}{Y(t_j)}.$$

- We now derive the variance.

Nelson-Aalen Estimator: Deriving Variance

- To begin, we work with the centered process,

$$\begin{aligned} V\{\widehat{\Lambda}(t) - \Lambda(t)\} = V\left\{\sum_{j=1}^{m} \frac{\Delta M_j}{Y(t_j)}\right\} &= E\left[\left\{\sum_{j=1}^{m} \frac{\Delta M_j}{Y(t_j)}\right\}^2\right] \\ &= \sum_{j=1}^{m} E\left[\left\{\frac{\Delta M_j}{Y(t_j)}\right\}^2\right] + 2\sum_{j=1}^{m}\sum_{k=j+1}^{m} E\left[\frac{\Delta M_j}{Y(t_j)}\frac{\Delta M_k}{Y(t_k)}\right]. \end{aligned}$$

- First, consider the second term, recalling that $k > j$,

$$\begin{aligned} E\left[\frac{\Delta M_j}{Y(t_j)}\frac{\Delta M_k}{Y(t_k)}\right] &= E\left[E\left[\frac{\Delta M_j}{Y(t_j)}\frac{\Delta M_k}{Y(t_k)}\middle| F_k\right]\right] \\ &= E\left[\frac{\Delta M_j}{Y(t_j)} E\left[\frac{\Delta M_k}{Y(t_k)}\middle| F_k\right]\right] \\ &= 0 \end{aligned}$$

- Now, consider the first term,

$$\begin{aligned} E\left[\left\{\frac{\Delta M_j}{Y(t_j)}\right\}^2\right] &= E\left[E\left[\left\{\frac{\Delta M_j}{Y(t_j)}\right\}^2\middle| F_j\right]\right] \\ &= E\left[Y(t_j)^{-2} E\left[(\Delta M_j)^2 | F_j\right]\right] \end{aligned}$$

- Now, since $\Delta N_j | F_j \sim \text{Binomial}\{Y(t_j), \lambda(t_j)\Delta(t_j)\}$,

$$E[\Delta N_j | F_j] \; = \; Y(t_j)\lambda(t_j)\Delta(t_j).$$

  such that

$$\begin{aligned} E\left[(\Delta M_j)^2 | F_j\right] &= E\left[\{\Delta N_j - E[\Delta N_j | F_j]\}^2 | F_j\right] \\ &= V\{\Delta N_j | F_j\} \\ &= Y(t_j)\lambda(t_j)\Delta(t_j)\{1 - \lambda(t_j)\Delta(t_j)\} \\ &= Y(t_j)\lambda(t_j)\Delta(t_j) \end{aligned}$$

  since $\Delta(t_j)^2 \to 0$ faster than $\Delta(t_j) \to 0$ as $m \to \infty$

- Collecting the preceding results,

$$V\left\{\sum_{j=1}^{m}\frac{\Delta M_j}{Y(t_j)}\right\} = \sum_{j=1}^{m} E\left[Y(t_j)^{-2}E\left[(\Delta M_j)^2|F_j\right]\right]$$
$$= \sum_{j=1}^{m} E\left[Y(t_j)^{-2}Y(t_j)\lambda(t_j)\Delta(t_j)\right]$$
$$= \sum_{j=1}^{m} E\left[\frac{\lambda(t_j)\Delta(t_j)}{Y(t_j)}\right]$$

- Taking the limit as $m \to \infty$,

$$V\left\{\widehat{\Lambda}(t) - \Lambda(t)\right\} = V\left\{\widehat{\Lambda}(t)\right\} = E\left[\int_0^t \frac{\lambda(s)}{Y(s)}ds\right]$$

- We can estimate this variance by

$$\widehat{\sigma}^2_\Lambda(t) = \int_0^t Y(s)^{-1}d\widehat{\Lambda}(s) = \int_0^t \frac{dN(s)}{Y(s)^2} \tag{2}$$

Note: This variance can also be derived using a Poisson process argument. (Virtually) Any counting process can be modeled as a Poisson process. Check page 11 of TG for more details.

Comparison to Greenwood's Formula

Note: $-\log \widehat{S}_{KM}(t)$ converges almost surely to $\Lambda(t)$

- This implies another variance estimator for $\widehat{\Lambda}(t)$ via the Delta method,

$$V\{\widehat{\Lambda}(t)\} = V\{-\log \widehat{S}_{KM}(t)\} = \widehat{S}_{KM}(t)^{-2}V\{\widehat{S}_{KM}(t)\}$$

Greenwood's formula: $\widehat{V}\{\widehat{S}_{KM}(t)\} = \widehat{S}_{KM}(t)^2 \sum_{j:t_j\le t}\frac{d_j}{Y_j(Y_j-d_j)}$.

Consequently, we have $\widehat{V}_1\{\widehat{\Lambda}(t)\} = \int_0^t \frac{dN(s)}{Y(s)\{Y(s)-\Delta N(s)\}}$