Survival Analysis: Time To Event Modelling

Sudipta Das

Assistant Professor,
Department of Computer Science,
Ramakrishna Mission Vivekananda Educational & Research Institute

Outline I

1. Non Parametric Estimation

Non Parametric Estimation: Introduction I

- In this section, we shall examine techniques for drawing an inference about the distribution of the time to some event X, based on a sample of *right-censored survival data.*
- We assume that the potential censoring time is unrelated to the potential event time.
  - This assumption would be violated, for example, if patients with poor prognosis were routinely censored.
- The methods are appropriate for Type I, Type II, progressive or random censoring.

Non Parametric Estimation: Introduction II

- Notations:
  - Suppose that the events occur at D distinct times

$$0 = t_0 \leq t_1 < t_2 < \ldots < t_D < \infty$$

  - Let d_i be the number of events occurs at time t_i.
    - Events are sometimes simply referred to as deaths
  - Let Y_i be the number of individuals who are at risk at time t_i.
    - Note that Y_i is a count of the number of individuals with a time on study of t_i or more
    - Equivalently, this is the number of individuals who are alive at t_i or experience the event of interest at t_i

Non Parametric Estimation: Introduction III

- Objective: To model the time to event/ survival time (T) by
  - Modeling the survival function

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

  - Modeling the cumulative hazard function

$$H(t) = \int_0^t \frac{f(u)}{P(T > u)} du$$

- Note: We are not assuming any structural form for the survival time distribution

Modeling Survival Function I

- **Kaplan-Meier estimator** to model/estimate the survival function

$$
\begin{aligned}
\hat{S}(t) &= \prod_{t_i \le t,\ i=1}^{n} \left(1 - \frac{d_i}{Y_i}\right)^{\delta_i} \\
&= \prod_{i:\ t_i \le t} \left(\frac{Y_i - d_i}{Y_i}\right)
\end{aligned}
$$

  - $d_i =$ Number of failures/deaths at t_i
  - $Y_i =$ Number at risk of dying or failure at t_i
  - $\delta_i = \begin{cases} 1 & \text{if } t_i \text{ is observed failure,} \\ 0 & \text{if } t_i \text{ is censoring time.} \end{cases}$
  - Product of conditional survivals

Modeling Survival Function II

- Example: Consider the data on the time to relapse of patients in a clinical trial of 6-MP against a placebo. We shall consider only the 6-MP patients.

$$6, 6, 6, 6+, 7, 9+, 10, 10+, 11+, 13, 16, 17+,$$

$$19+, 20+, 22, 23, 25+, 32+, 32+, 34+, 35+$$

Modeling Survival Function III

Table: Construction of the Product-Limit Estimator for the 6-MP Group

Time t_i	Number of events d_i	Number at risk Y_i	KM estimate $\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{Y_i}\right)$
6	3	21	$\left[1 - \frac{3}{21}\right] = 0.857$
7	1	17	$0.857 \times \left[1 - \frac{1}{17}\right] = 0.807$
10	1	15	$0.807 \times \left[1 - \frac{1}{15}\right] = 0.753$
13	1	12	$0.753 \times \left[1 - \frac{1}{12}\right] = 0.690$
16	1	11	$0.690 \times \left[1 - \frac{1}{11}\right] = 0.628$
22	1	7	$0.628 \times \left[1 - \frac{1}{7}\right] = 0.538$
23	1	6	$0.538 \times \left[1 - \frac{1}{6}\right] = 0.448$

Modeling Survival Function IV

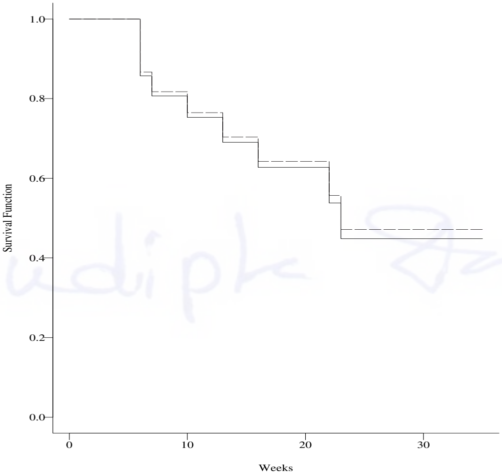

Figure 4.1A *Comparison of the Nelson–Aalen (------) and Product-Limit (————) estimates of the survival function for the 6-MP group.*

Modeling Survival Function V

- The Product-Limit estimator was constructed by using a reduced-sample approach.
- In this approach, note that, because events are only observed at the times t_i,
  - $S(t)$ should be a step function with jumps only at these times,
  - there being no information on events occurring at other times.
- We will estimate $S(t)$ by a discrete distribution with mass at the time points $t_1, t_2, \ldots, t_D$.

Modeling Survival Function VI

- We can estimate the $Pr[T > t_i | T \geq t_i]$ as the fraction of individuals who are at risk at time t_i but who do not die at this time, that is

$$
\begin{aligned}
\hat{P}r[T > t_i | T \geq t_i] &= \frac{Y_i - d_i}{Y_i}, \text{ for } i = 1, 2, \ldots, D. \\
&= 1 - \frac{d_i}{Y_i} \\
&= 1 - \hat{P}r[T = t_i | T \geq t_i]
\end{aligned}
$$

Modeling Survival Function VII

- To estimate $S(t_i)$, recall that

$$
\begin{aligned}
S(t_i) &= \frac{S(t_i)}{S(t_{i-1})} \times \frac{S(t_{i-1})}{S(t_{i-2})} \times \cdots \times \frac{S(t_2)}{S(t_1)} \times \frac{S(t_1)}{S(t_0)} \times S(t_0) \\
&= P[T > t_i | T > t_{i-1}] \times P[T > t_{i-1} | T > t_{i-2}] \cdots P[T > t_2 | T > t_1] \times P[T > t_1 | T > t_0] \times 1 \\
&= P[T > t_i | T \geq t_i] \times P[T > t_{i-1} | T \geq t_{i-1}] \cdots P[T > t_2 | T \geq t_2] \times P[T > t_1 | T \geq t_1]
\end{aligned}
$$

- Thus,

$$
\hat{S}(t) = \prod_{i:\ t_i \leq t} \left(\frac{Y_i - d_i}{Y_i} \right)
$$

- **Kaplan-Meier estimator**
  - It is also known as product-limit estimator
  - It can be shown to be nonparametric MLE of survival function, under certain regularity conditions
  - In the absence of censoring, it reduces to complement of the *empirical distribution function* (EDF):

$$\hat{S}(t) = 1 - \frac{\text{Number of obs } \leq t}{\text{Total Number of obs}}$$

Modeling Survival Function IX

- Kaplan-Meier estimators of either the survival function or the cumulative hazard rate are consistent.
- For values of t beyond the largest observation time this estimator is not well defined
    - Efron (1967) suggests estimating $\hat{S}(t)$ by 0 for $t > t_{\max}$. (This leads to a negatively biased estimator)
    - Gill (1980) suggests estimating $\hat{S}(t)$ by $\hat{S}(t_{\max})$ for $t > t_{\max}$. (This leads to a positively biased estimator)
    - Although both estimators have the same large-sample properties and converge to the true survival function for large samples

Modeling Survival Function X

- Variance of KM Estimate (Greenwood's formula):

$$\hat{Var}(\hat{S}(t)) = \hat{S}^2(t) \sum_{i:t_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)} = \hat{S}^2(t)\sigma_S^2(t),$$

where $\sigma_S^2(t) = \sum_{i:t_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)}$

  - It underestimate the true variance of the Kaplan-Meier estimator for small to moderate samples.

Table: Construction of the variance of KM Estimator for the 6-MP Group

Time	# of events	# at risk	KM est.		Variance of $\hat{S}(t)$
t_i	d_i	Y_i	$\hat{S}(t)$	$\sigma_S^2(t) = \sum_{t_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)}$	$\hat{V}(\hat{S}(t)) = \hat{S}^2(t) \sum_{t_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)}$
6	3	21	0.857	$\frac{3}{21\times18} = 0.0079$	$0.857^2 \times 0.0079 = 0.0058$
7	1	17	0.807	$0.0079 + \frac{1}{17\times16} = 0.0116$	$0.807^2 \times 0.0116 = 0.0076$
10	1	15	0.753	$0.0116 + \frac{1}{15\times14} = 0.0164$	$0.753^2 \times 0.0164 = 0.0093$
13	1	12	0.690	$0.0164 + \frac{1}{12\times11} = 0.0240$	$0.690^2 \times 0.0240 = 0.0114$
16	1	11	0.628	$0.0240 + \frac{1}{11\times10} = 0.0330$	$0.628^2 \times 0.0330 = 0.0130$
22	1	7	0.538	$0.0330 + \frac{1}{7\times6} = 0.0569$	$0.538^2 \times 0.0569 = 0.0164$
23	1	6	0.448	$0.0569 + \frac{1}{6\times5} = 0.0902$	$0.448^2 \times 0.0902 = 0.0181$

Modeling Survival Function XII

- The variance was constructed by the help of delta method.
- Recall that

$$\hat{S}(t_i) = \prod_{j=1}^{i} \left[\frac{Y_j - d_j}{Y_j} \right] = \prod_{j=1}^{i} \hat{p}_j$$

- Thus,

$$\log \left[\hat{S}(t_i) \right] = \sum_{j=1}^{i} \log \left[\hat{p}_j \right]$$

Modeling Survival Function XIII

- Hence,

$$
\begin{aligned}
\hat{Var}\left[\log\left[\hat{S}(t_i)\right]\right] &= \sum_{j=1}^{i} \hat{Var}\left\{\log\left[\hat{p}_j\right]\right\} \\
&= \sum_{j=1}^{i} [\hat{p}_j]^{-1}\left\{\hat{Var}\left[\hat{p}_j\right]\right\}[\hat{p}_j]^{-1} \\
&= \sum_{j=1}^{i} [\hat{p}_j]^{-2}\left\{\frac{\hat{p}_j\left[1-\hat{p}_j\right]}{Y_j}\right\} \\
&= \sum_{j=1}^{i} [\hat{p}_j]^{-1}\left[1-\hat{p}_j\right]\frac{1}{Y_j} \\
&= \sum_{j=1}^{i} \left[\frac{Y_j-d_j}{Y_j}\right]^{-1}\left[\frac{d_j}{Y_j}\right]\frac{1}{Y_j} = \sum_{j=1}^{i}\frac{d_j}{Y_j(Y_j-d_j)}
\end{aligned}
$$

Modeling Survival Function XIV

- Therefore,

$$\begin{aligned}
\hat{Var}\left[\hat{S}(t_i)\right] &= \hat{Var}\left[e^{\log\left[\hat{S}(t_i)\right]}\right] \\
&= \left[e^{\log\left[\hat{S}(t_i)\right]}\right] \hat{Var}\left[\log\left[\hat{S}(t_i)\right]\right] \left[e^{\log\left[\hat{S}(t_i)\right]}\right] \\
&= \left[\hat{S}(t_i)\right]^2 \sum_{j=1}^{i} \frac{d_j}{Y_j(Y_j - d_j)}
\end{aligned}$$

- And

$$\hat{Var}\left[\hat{S}(t)\right] = \left[\hat{S}(t)\right]^2 \sum_{i:t_i \le t} \frac{d_i}{Y_i(Y_i - d_i)}$$

Modeling Survival Function XV

- Standard error of KM Estimate:

$$\hat{SE}(\hat{S}(t)) = \hat{S}(t)\sqrt{\sum_{t_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)}}$$

- Asymptotic property of KM Estimate
  - Under suitable regularity conditions, the Product-Limit estimator converges weakly to Gaussian process.
  - This fact means that for fixed t, the estimator has an approximate normal distribution.

Modeling Survival Function XVI

- Thus, the $100(1-\alpha)\%$ **point-wise** Confidence Interval of $S(t)$:

$$\left[\hat{S}(t) - z_{1-\alpha/2} \times \hat{SE}(\hat{S}(t)), \hat{S}(t) + z_{1-\alpha/2} \times \hat{SE}(\hat{S}(t))\right] \tag{1}$$

  - Called linear confidence interval
  - Appropriate for large sample

Modeling Survival Function XVII

Table: The Product-Limit Estimator and Its Estimated Standard Error for the 6-MP Group

Time on Study t	KM Estimator $\hat{S}(t)$	Standard Error $\hat{SE}(\hat{S}(t))$	95% CI
$0 \le t < 6$	1.000	0.000	$[1, 1]$
$6 \le t < 7$	0.857	0.076	$[0.708, 1.006]$
$7 \le t < 10$	0.807	0.087	$[0.636, 0.978]$
$10 \le t < 13$	0.753	0.096	$[0.565, 0.941]$
$13 \le t < 16$	0.690	0.107	$[0.480, 0.900]$
$16 \le t < 22$	0.628	0.114	$[0.405, 0.851]$
$22 \le t < 23$	0.538	0.128	$[0.287, 0.789]$
$23 \le t < 35$	0.448	0.135	$[0.183, 0.713]$

Modeling Cumulative Hazard Function from Survival Function

- Cumulative Hazard Function $H(t)$ can be constructed from Survival Function $S(t)$
  - Cumulative Hazard Function:

$$\hat{H}(t) = -\log \hat{S}(t)$$

Modeling Cumulative Hazard Function I

- **Nelson-Aalen estimator** to model/estimate the cumulative hazard function

$$\tilde{H}(t) = \sum_{i:\ t_i \leq t} \frac{d_i}{Y_i}.$$

  - d_i = Number of failured/death at t_i
  - Y_i = number at risk of dying or failure at t_i

- Its variance

$$\hat{Var}(\tilde{H}(t)) = \sum_{t_i \leq t} \frac{d_i}{Y_i^2} = \sigma_H^2(t).$$

Modeling Cumulative Hazard Function II

Table: Construction of the variance of Nelson–Aalen Estimator for the 6-MP Group

Time t_i	# of events d_i	# at risk Y_i	Nalson-Aalen estimator $\tilde{H}(t) = \sum_{t_i \le t} \frac{d_i}{Y_i}$	Variance of $\tilde{H}(t)$ $\sigma_H^2(t) = \hat{V}(\tilde{H}(t)) = \sum_{t_i \le t} \frac{d_i}{Y_i^2}$
6	3	21	$\frac{3}{21} = 0.1428$	$\frac{3}{21^2} = 0.0068$
7	1	17	$0.1428 + \frac{1}{17} = 0.2017$	$0.0068 + \frac{1}{17^2} = 0.0103$
10	1	15	$0.2017 + \frac{1}{15} = 0.2683$	$0.0103 + \frac{1}{15^2} = 0.0147$
13	1	12	$0.2683 + \frac{1}{12} = 0.3517$	$0.0147 + \frac{1}{12^2} = 0.0217$
16	1	11	$0.3517 + \frac{1}{11} = 0.4426$	$0.0217 + \frac{1}{11^2} = 0.0299$
22	1	7	$0.4426 + \frac{1}{7} = 0.5854$	$0.0299 + \frac{1}{7^2} = 0.0503$
23	1	6	$0.5854 + \frac{1}{6} = 0.7521$	$0.0503 + \frac{1}{6^2} = 0.0781$

Modeling Cumulative Hazard Function III

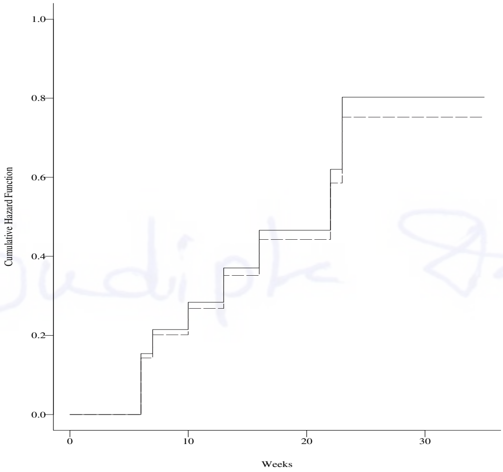

Figure 4.1B *Comparison of the Nelson–Aalen (------) and Product-Limit (———) estimates of the cumulative hazard rate for the 6-MP group.*

Modeling Cumulative Hazard Function IV

- Note
  - The Nelson-Aalen estimator of the cumulative hazard rate is the first term in a Taylor series expansion of minus the logarithm of the Product-Limit estimator.
  - Under certain regularity conditions, one can show that the Nelson-Aalen estimator is nonparametric maximum likelihood estimator.
  - Nelson-Aalen estimators of either the survival function or the cumulative hazard rate are consistent.

Modeling Cumulative Hazard Function V

- Standard error of NA estimator:

$$\hat{SE}(\tilde{H}(t)) = \sqrt{\sum_{t_i \le t} \frac{d_i}{Y_i^2}} = \sigma_H(t).$$

- Asymptotic property of NA Estimate
  - Under suitable regularity conditions, the Nelson-Aalen estimator converges weakly to Gaussian process.
  - This fact means that for fixed t, the estimator has an approximate normal distribution.
- Thus, the $100(1-\alpha)\%$ Confidence Interval of $H(t)$:

$$\left[\tilde{H}(t) - z_{1-\frac{\alpha}{2}} \times \hat{SE}(\tilde{H}(t)), \tilde{H}(t) + z_{1-\frac{\alpha}{2}} \times \hat{SE}(\tilde{H}(t))\right]$$

Modeling Cumulative Hazard Function VI

Table: The Nelson-Aalen Estimator, Its estimated standard error and 95% CI for the 6-MP Group

Time on Study t	NA Estimator $\tilde{H}(t)$	Standard Error $\hat{SE}(\tilde{H}(t))$	95% CI
$0 \leq t < 6$	0.0000	0.0000	$[1, 1]$
$6 \leq t < 7$	0.1428	0.0068	$[0.1295, 0.1561]$
$7 \leq t < 10$	0.2017	0.0103	$[0.1815, 0.2219]$
$10 \leq t < 13$	0.2683	0.0147	$[0.2395, 0.2971]$
$13 \leq t < 16$	0.3517	0.0217	$[0.3092, 0.3943]$
$16 \leq t < 22$	0.4426	0.0299	$[0.3840, 0.5012]$
$22 \leq t < 23$	0.5854	0.0503	$[0.4868, 0.6840]$
$23 \leq t < 35$	0.7521	0.0781	$[0.5990, 0.9052]$

Modeling Survival Function from Cumulative Hazard Function

- Survival Function $S(t)$ can be constructed from Cumulative Hazard Function $H(t)$
  - Survival Function:

$$\tilde{S}(t) = e^{-\tilde{H}(t)}$$

Non Parametric Estimation: Example in R

- Example: Bank Credit Data
  - Data read
  - Data preparation
  - Kaplan-Meier Estimator/ Product Limit Estimator for Survival Function
    - **FIGURE 5A**
  - Nelson-Aalen Estimator for Cumulative Hazard Function
    - **FIGURE 5B**
  - Nelson-Aalen Estimator for Survival Function
    - **FIGURE 5C**