Kaplan-Meier and Nelson-Aalen with right-censored and left truncated data

What can we estimate with right censored and left truncated data? To answer this, consider what we know at each moment of time during follow-up.

With right censored data, we know which observations are still being followed and we can observe which of them have an event. For now, we are concerned with how many have an event (for estimating the survival curve), rather than which ones have an event (for regression). What we know is the number of subjects being followed and the number with an event at each moment in follow-up time.

Denote a data set with n observations including right censored observations with follow-up times and status indicators by (x_i, δ_i) for $i = 1, \ldots, n$. This information can be organized several ways.

The information could be grouped according to a division of the time axis into disjoint subintervals with the following information:

- The number of observations being followed at the start of each interval.
- The number of subjects with an event during the interval.
- A lifetable also records the number of subjects censored (lost) during each interval.

Example: 6-mp data (section 1.2) with time intervals of 10 units.

4 time intervals		
Interval	Number starting	Number of events
(0 - 10]	21	5
(11-20]	13	2
(21-30]	7	2
(31+]	4	0

Note that the grouping of data into time intervals does not retain all of the information in the original data set. All the information is kept by recording the information at each time point (as a function of time, rather than in a table). This information can be recorded at each instant in time, t, during follow-up as follows.

- The number of observations being followed at time t, $Y(t)$.
- The number of subjects having an event at time t, $d(t)$.

It is sufficient to know this information only at the distinct times of death, t_i, during follow-up where $t_1 < t_2 < \ldots, < t_p$. This loses the timing of the censored observations between the event times, but this information is not used in non-parametric estimates.

- The number of observations being followed at (just before) time t_i, $Y_i = Y(t_i)$.
- The number of subjects having an event at time t, $d_i = d(t_i)$.

Example: 6-mp data (section 1.2)

$n = 21$ Observations	
x_i	δ_i
6	1
6	1
6	1
6	0
7	1
9	0
10	1
10	0
11	0
13	1
16	1
17	0
19	0
20	0
22	1
23	1
25	0
32	0
32	0
34	0
35	0

$p = 7$ Death Times			
i	t_i	d_i	Y_i
1	6	3	21
2	7	1	17
	9	0	16
3	10	1	15
4	13	1	12
5	16	1	11
6	22	1	7
7	23	1	6
	35	0	1

Notice that Y_i gives the number under observation just before time t_i and d_i is the number of deaths at that time. The censored observations at time t_i are removed after the deaths.

This information yields the following intuitive estimator of the fraction having an event at each time among those being followed at that time.
$\hat{h}(t) = 0$ if no events are observed at time t.
$\hat{h}(t) = d_t/n_t$ if d_t events are observed among n_t subjects being observed at time t.
This can be interpreted as the discrete time probability of an event at that time, $h(t_i)$, or as the probability of an event during an interval starting at that time, $h(t_i)\,dt$.

We can estimate the survival function by $\hat{S}(x) = \prod_{t_i \le x} \left(1 - \hat{h}(t_i)\right)$. This is the Kaplan-Meier estimate for the survival curve. It is also called the product-limit estimator. Note that it is the same as the life table estimator when the intervals of time are taken to be

arbitrarily short (thus the term limit above).

We can estimate the cumulative hazard $\hat{H}(x) = \sum_{t_i \le x} \hat{h}(t_i) \approx \int_0^x h(t)\,dt$. This is called the Nelson-Aalen estimator for the cumulative hazard.

Two of the most important functions for survival analysis are the survival function and the cumulative hazard function. These two quantities can be estimated with "intuitive" hazard estimator and product limit estimators above with both right-censored and left truncated data.

These estimator rely strongly upon the assumption that the death rate among those subjects under observation (not truncated and not right censored) is the same as the death rate in the population of interest.

Relationships

The negative log of the Kaplan-Meier estimator approximates the Nelson-Aalen cumulative hazard estimator.

$$\begin{aligned} -\ln(\widehat{S}_{KM}(x)) &= -\sum_{t_i \le x} ln\left(1 - \frac{d_{t_i}}{Y_{t_i}}\right) \\ &\approx -\sum_{t_i \le x} -\frac{d_{t_i}}{Y_{t_i}} \\ &= \widehat{H}_{NA}(x) \end{aligned}$$

Similarly, the Kaplan-Meier estimator can be approximated as exponential of the negative Nelson-Aalen cumulative hazard estimator.

$$\begin{aligned} \exp\left(-\widehat{H}_{NA}(x)\right) &= \exp\left(-\sum_{t_i \le x}\left(\frac{d_{t_i}}{Y_{t_i}}\right)\right) \\ &= \prod_{t_i \le x} \exp\left(-\frac{d_{t_i}}{Y_{t_i}}\right) \\ &\approx \prod_{t_i \le x}\left(1 - \frac{d_{t_i}}{Y_{t_i}}\right) \\ &= \widehat{S}_{KM}(x)) \end{aligned}$$

The next term in the approximation shows that $\widehat{S}_{KM}(x) < \exp\left(-\widehat{H}_{NA}(x)\right)$.

Variances The variance of the Nelson-Aalen estimator is estimated based on a Poisson

approximation

$$\begin{aligned}
\text{Var}\left(\frac{d_{t_i}}{Y_{t_i}}\right) &= \frac{\text{Var}(d_{t_i})}{Y_{t_i}^2} \\
&= \frac{E(d_{t_i})}{Y_{t_i}^2} \quad \text{Since the mean = variance for a Poisson quantity, } d_{t_i} \\
&\approx \frac{d_{t_i}}{Y_{t_i}^2} \quad \text{Estimate the mean by the observed count, since only one instance}
\end{aligned}$$

Since the terms in the sum are uncorrelated.

$$\text{Var}\left(\widehat{H}_{NA}(x)\right) = \sum_{t_i \le x}\left(\frac{d_{t_i}}{Y_{t_i}^2}\right)$$

For any random quantity, Z,

$$\text{Var}\left(\exp(Z)\right) \approx \exp(2\mu_Z)\text{Var}(Z).$$

$$\text{Var}\left(\ln(Z)\right) \approx \text{Var}(Z)/\mu_Z^2.$$

Thus, based on the variance of $\widehat{H}_{NA}(x) = \ln\left(\widehat{S}_{KM}(x)\right)$, a (not commonly used) approximate variance for $\widehat{S}_{KM}(x)$, is

$$\widehat{S}_{KM}^2(x)\sum_{t_i \le x}\left(\frac{d_{t_i}}{Y_{t_i}^2}\right)$$

Greenwood's formula is commonly used and estimates the variance of $\widehat{S}_{KM}(x)$ as

$$\text{Var}\left[\widehat{S}_{KM}(x)\right] = \widehat{S}_{KM}^2(x)\sum_{t_i \le x}\left(\frac{d_{t_i}}{Y_{t_i}(Y_{t_i} - d_{t_i})}\right)$$

Confidence intervals for $\widehat{S}_{KM}(x)$:

$$\widehat{S}_{KM}(x) \pm Z_{1-\alpha/2}\sqrt{\text{Var}\left[\widehat{S}_{KM}(x)\right]}.$$

Other confidence interval calculations are based on the sum in Greenwood's formula:

$$\sigma_S^2(t) = \text{Var}[\widehat{S}(t)]/\widehat{S}^2(t) = \sum_{t_i \le x}\left(\frac{d_{t_i}}{Y_{t_i}(Y_{t_i} - d_{t_i})}\right).$$

Starting with a confidence interval for the $\ln H(x)$ yields

$$[\hat{S}(x)^{1/\theta}, \hat{S}(x)^{\theta}]$$

where

$$\theta = \exp\left[\frac{Z_{1-\alpha/2}\ \sigma_S(x)}{\ln[\hat{S}(x)]}\right]$$

Confidence Bands for $S(x)$ are not widely used but can be computed as in the text.

Mean lifetime over the interval $[0, \tau]$ is often estimated as $\hat{\mu}_\tau = \int_0^\tau \hat{S}(u)\, du$ with variance

$$\hat{V}[\hat{\mu}_\tau] = \sum_{t_i}^{\tau} \left[\int_{t_i \le \tau} \hat{S}(u)\, du\right]^2 \frac{d_i}{Y_i(Y_i - d_i)}.$$

Percentiles are estimated by $\hat{x}_p = \inf\{t : \hat{S}(t) \le 1 - p\}$. Confidence intervals for the p'th percentile, x_p, is often calculated as the set of all t satisfying

$$\left|\frac{\hat{S}(t) - (1-p)}{\hat{V}^{1/2}[\hat{S}(t)]}\right| \le Z_{1-\alpha/2}.$$

The confidence interval based on the log transform is given in the book.