Stat 344
Life Contingencies I

Estimating Survival Models

Incomplete or modified data

Observations may be incomplete because of censoring and/or truncation. An observation is:

- **left truncated** at *d* if when it is below *d*, it is not recorded, but when it is above *d*, it is recorded at its observed value.
- **right truncated** at *u* if when it is above *u* it is not recorded, but when it is recorded at its observed value.
- **left censored** at *d* if when it is below *d*, it is recorded as being equal to *d*, but when it is above *d*, it is recorded at its observed value.
- **right censored** at *u* if when it is above *u*, it is recorded as being equal to *u*, but when it is below *u*, it is recorded at its observed value.

It is most common in life contexts to find left truncated or right censored observations. Left truncation usually occurs because a person died before our study began. Right censoring occurs when the person is still alive as our study ends.

Interpreting the risk set

For survival/mortality data, the risk set r_j is the number of people observed alive at age y_j. If deaths and censoring occur at the same time, we assume the deaths occur first.

Example Dataset

Suppose we have a seriatim (one observation per person or policy) dataset that includes censored values. Assuming that $^+$ denotes a censored observation, the data is:

$$4^+, 5, 10, 10^+, 14, 15^+, 16^+, 17, 19^+, 20$$

j	$t_{(j)}$	r_j
0	0	10
1	5	9
2	10	8
3	14	6
4	17	3
5	20	1

Kaplan-Meier (product-limit) estimator

With d_i being the number of deaths at $t_{(i)}$, the Kaplan-Meier (product limit) estimate for the survival function is given by

$$S_n(t) = \begin{cases} 1, & 0 \leq t < y_1, \\ \prod_{i=1}^{j-1} \left(\frac{r_i - d_i}{r_i} \right), & y_{j-1} \leq t < y_j, j = 2, \ldots, k, \\ \prod_{i=1}^{k} \left(\frac{r_i - d_i}{r_i} \right) \text{ or } 0, & t \geq y_k. \end{cases}$$

Clearly when $d_k = r_k$, we have $S_n(t) = 0$ for $t \geq y_k$.

K-M Example

Data: $4^+, 5, 10, 10^+, 14, 15^+, 16^+, 17, 19^+, 20$

Summary based on death times:

j	$t_{(j)}$	r_j	d_j	$\hat{p}_{(j)}$	$\hat{S}_{KM}(t)$
0	0	10			1
1	5	9	1	8/9	8/9
2	10	8	1	7/8	7/9
3	14	6	1	5/6	35/54
4	17	3	1	2/3	70/162
5	20	1	1	0/1	0

Variance - Greenwood's Formula

We can approximate the variance of the Kaplan-Meier estimates using Greenwood's formula.

$$V\left[\hat{S}(t)\right] \approx \hat{S}(t)^2 \left(\sum_{j:t_{(j)} \leq t} \frac{d_j}{r_j(r_j - d_j)} \right)$$

K-M Confidence Intervals

Linear confidence intervals are made in the standard way:

$$\hat{S}(t) \pm 1.96\sqrt{V\left[\hat{S}(t)\right]}$$

Sometimes these intervals can provide bounds greater than 1 or less than 0. To correct this issue, you can use log-confidence intervals (non-linear).

$$(g_L, g_U) = \log(-\log \hat{S}(t)) \pm 1.96\sqrt{\left(\frac{1}{\hat{S}(t)\log \hat{S}(t)}\right)^2 V\left[\hat{S}(t)\right]}$$

$$(s_L, s_U) = \exp\{-exp\{(g_L, g_u)\}\}$$

SOA Exam Question

You are given:

- All members of a mortality study are observed from birth. Some leave the study by means other than death.
- $d_3 = 1$, $d_4 = 3$
- The following Kaplan-Meier product limit estimates were obtained:

$$S_n(y_3) = 0.65, \quad S_n(y_4) = 0.50, \quad S_n(y_5) = 0.25$$

- Between times y_4 and y_5, six observations were censored.
- Assume no observations were censored at the times of deaths.

Calculate the value of d_5.

Nelson-$\AA$alen estimator

First, derive the Nelson-$\AA$alen estimator for the cumulative hazard rate function as given by

$$\hat{H}(t) = \begin{cases} 0, & 0 \leq t < y_1, \\ \sum_{i=1}^{j-1} \frac{d_i}{r_i}, & y_{j-1} \leq t < y_j, j = 2, \ldots, k, \\ \sum_{i=1}^{k} \frac{d_i}{r_i}, & t \geq y_k. \end{cases}$$

Then use $\hat{S}(t) = e^{-\hat{H}(t)}$ to estimate the survival function.

Example Continued

Data: $4^{+}, 5, 10, 10^{+}, 14, 15^{+}, 16^{+}, 17, 19^{+}, 20$

Summary based on death times:

j	$t_{(j)}$	r_j	d_j	$\hat{p}_{(j)}$	$\hat{S}_{KM}(t)$	$\hat{S}_{NA}(t)$
0	0	10			1	1
1	5	9	1	8/9	0.89	$e^{-1/9} = 0.89$
2	10	8	1	7/8	0.78	$e^{-(1/9+1/8)} = 0.79$
3	14	6	1	5/6	0.67	0.67
4	17	3	1	2/3	0.43	0.48
5	20	1	1	0/1	0	0.18

Variance and Confidence Intervals

We can approximate the variance of the Nelson-$\AA$alen estimates using similar arguments as Greenwood's formula.

$$V\left[\hat{S}(x)\right] \approx \hat{S}(x)^2 \left(\sum_{j:t_{(j)}\leq x} \frac{d_j(r_j - d_j)}{r_j^3} \right)$$

Linear confidence intervals are made in the standard way:

$$\hat{S}(x) \pm 1.96\sqrt{V\left[\hat{S}(x)\right]}$$

(Modified) SOA Exam Question

You are studying the length of time attorneys are involved in settling bodily injury lawsuits. Let T represent the number of months from the time an attorney is assigned such a case to the time the case is settled.

Nine cases were observed during the study period, two of which were not settled at the conclusion of the study. For those two cases, the time spent up to the conclusion of the study, 4 months and 6 months, was recorded instead.

The observed values of T for the other seven cases are as follows:

$$1 \quad 3 \quad 3 \quad 5 \quad 8 \quad 8 \quad 9$$

Use the Nelson-$\mathring{A}$alen estimator to estimate $\Pr(3 \leq T \leq 5)$. Compare the estimate using Kaplan-Meier estimator.