Notes on Survival analysis

- Develop models for *survival times* Y.
- $f(y)$ is its pdf and $F(y)$ its distribution function.
- *Survivor function*: probability of survival beyond time y or $S(y) = P[Y > y] = 1 - F(y)$.
- *Hazard function*: probability of death in an infinitesimal interval given survival to time y
- We showed that the hazard $h(y)$ satisfies

$$h(y) = \frac{f(y)}{S(y)} = -\frac{d}{dy}[log(S(y))]$$

- May think of the survival function in terms of *cummulative hazard* $H(y)$.

- $S(y) = exp(-H(y))$ where

$$H(y) = \int_0^y h(t)dt$$

.

- Inference could be related to quantiles of $F(y)$ (or $S(y)$)
- *q-quantile* $y(q)$ solution to the equation

$$F(y(q)) = q \text{ or } S(y(q)) = 1 - q$$

- Simplest parametric model is $f(y|\theta) = exp(-\theta y); \theta > 0, y > 0$
- In particular its hazard function is simply θ (does not depend on y).

- Another one is the Weibull distribution.

$$f(y|\lambda,\phi) = \lambda\phi y^{\lambda-1} exp(-\phi y^{\lambda}); y, \lambda, \phi > 0$$

- More flexible than exponential.
- In particular, its hazard function is

$$h(y|\lambda,\phi) = \lambda\phi y^{\lambda-1}$$

  which is an ”accelerated failure time model”.
- It can be shown that for this distribution

$$log(-log(S(y|\lambda,\phi))) = log(\phi) + \lambda log(y)$$

- A plot of $log(y)$ values vs. *empirical* values of $log(-log(S(y|\lambda,\phi)))$ may suggest if the *Weibull* dist. is adequate.

Non-parametric estimation

- Empirical survivor function (without censoring).

$$\tilde{S}(y) = \frac{\text{number of subjects with time} \geq y}{\text{total no. of subjects}}$$

- Step/decreasing function that starts at 1.
- *Censoring:* survival times are incomplete. Individual left study.
- *Right censoring:* true survival time greater than the observed.
- Use *Kaplan-Meier* estimator to account for censoring.

- Arrange survival times in increasing order

$$y_{(1)} \leq y_{(2)} \leq y_{(3)} \cdots$$

- d_j number of deaths that occur in the interval $y_{(j)} - \delta, y_{(j)}$.
- n_j number of subjects alive just before $y_{(j)}$
- Probability of survival past $y_{(j)}$ is approximated as

$$\frac{n_j - d_j}{n_j}$$

- The *Kaplan Meier* estimate for $y_{(k)} \leq y < y_{(k+1)}$ is

$$\hat{S}(y) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right)$$

- **Example**: 21 remission times of Leukemia patients: $6, 6, 6, 6^*, 7, 9^*, 10, 10^*, 11^*, 13, 16, 17^*, 19^*, 20^*, 22, 23, 25^*, 32^*, 32^*, 34^*, 35^*$.

- $*$ means subject left study.

Interval	n_j	d_j	$\frac{n_j - d_j}{n_j}$	$S(t)$
$[0, 6)$	21	0	21/21=1	1
$[6, 7)$	21	3	18/21=0.857	0.857
$[7, 10)$	17	1	16/17=0.941	0.807
$[10, 13)$	15	1	14/15=0.933	0.753
$[13, 16)$	12	1	11/12=0.917	0.69
$[16, 22)$	11	1	10/11=0.909	0.627
$[22, 23)$	7	1	6/7=0.857	0.538
$[23, \infty)$	6	1	5/6=0.833	0.448

Life Table Estimate

- Considers the time interval $[t_j^{'}, t_{j+1}^{'}]$.
- d_j = number of deaths in interval, c_j = number of censored cases, n_j = number of alive at the start.
- Censored survival times occur uniformly through the $j-th$ interval (actuarial assumption).
- Average number of individuals at risk during this interval,

$$n_j^{'} = n_j - c_j/2 = n_j - 0.5c_j$$

- Probability of 'death' in the $j-th$ interval $\approx d_j/n_j^{'}$

- The probability of survival is

$$P(Y \geq t | Y \geq t_j^{'}) \approx 1 - d_j / n_j^{'} = \frac{n_j^{'} - d_j}{n_j^{'}}$$

  for $t_j^{'} \leq t < t_{j+1}^{'}$

- Suppose we now take t so that $t_k^{'} \leq t < t_{k+1}^{'}$
- Survivor function estimate based on conditional probability.

$$\begin{aligned} S(t) = P(Y \geq t) &= P(Y \geq t | Y \geq t_k^{'}) P(Y \geq t_k^{'}) \\ &\left(\frac{n_k^{'} - d_k}{n_k^{'}} \right) P(Y \geq t_k^{'}) \end{aligned}$$

- Apply the same idea so that

$$P(Y \geq t_k') = P(Y \geq t_k' | Y \geq t_{k-1}') P(Y \geq t_{k-1}')$$
$$\left(\frac{n_{k-1}' - d_{k-1}}{n_{k-1}'} \right) P(Y \geq t_{k-1}')$$

- An estimate of the survivor function is $S^*(t)$ where

$$S^*(t) = \left(\frac{n_k' - d_k}{n_k'} \right) \left(\frac{n_{k-1}' - d_{k-1}}{n_{k-1}'} \right) P(Y \geq t_{k-1}')$$

- Keep going until we get to, $P(Y \geq t_1') = \left(\frac{n_1' - d_1}{n_1'} \right)$.

- The Life table estimator is

$$S^*(t) = \prod_{j=1}^{k} \left(\frac{n_j' - d_j}{n_j'} \right) ; t_k' \leq t < t_{k+1}'$$

- Kaplan-Meier estimator follows same idea as the life table estimator, except
  - Each time interval (except initial) has 1 death.
  - Death time at start of interval.
  - No actuarial correction (use n_j in formulas).
- For large samples, survival times are near each other with few censored cases.

$$\hat{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right) ; t_k' \leq t < t_{k+1}'$$

- **Example**: Survival times of women with tumors that were negatively stained (Collett Ex. 1.2 pg. 7):
  $23, 47, 69, 70^*, 71^*, 100^*, 101^*, 148, 181, 198^*, 208^*, 212^*, 224^*$
- $*$ means censored observation.

Interval	n_j	d_j	c_j	n_j'	$\frac{n_j' - d_j}{n_j'}$	$S^*(t)$
$[0, 50)$	13	2	0	13	0.846	0.846
$[50, 100)$	11	1	2	10	0.900	0.761
$[100, 150)$	8	1	2	7	0.857	0652
$[150, 200)$	5	1	1	4.5	0.778	0.507
$[200, 250)$	3	0	3	1.5	1	0.501

- Kaplan-Meier estimate of $S(t)$ on the negatively stained data
$23, 47, 69, 70^*, 71^*, 100^*, 101^*, 148, 181, 198^*, 208^*, 212^*, 224^*$

Interval	n_j	d_j	$\frac{n_j - d_j}{n_j}$	$\hat{S}(t)$
$[0, 23)$	13	0	1	1
$[23, 47)$	13	1	12/13	0.923
$[47, 69)$	12	1	11/12	0.846
$[69, 148)$	11	1	10/11	0.769
$[148, 181)$	6	1	5/6	0.641
$[181, \infty)$	5	1	4/5	0.513

- Kaplan-Meier estimate ignoring censoring (empirical survivor function).

Interval	n_j	d_j	$\frac{n_j - d_j}{n_j}$	$\hat{S}(t)$
$[0, 23)$	13	0	1	1
$[23, 47)$	13	1	12/13	12/13
$[47, 69)$	12	1	11/12	11/13
$[69, 70)$	11	1	10/11	10/13
$[70, 71)$	10	1	9/10	9/13
$[71, 100)$	9	1	8/9	8/13
$[100, 101)$	8	1	7/8	7/13
$[101, 148)$	7	1	6/7	6/13
$[148, 181)$	6	1	5/6	5/13
$[181, 198)$	5	1	4/5	4/13
$[198, 208)$	4	1	3/4	3/13
$[208, 212)$	3	1	2/3	2/13
$[212, 224)$	2	1	1/2	1/13
$[224, \infty)$	1	1	0/1	0/13

Estimating the hazard function

- Recall that the hazard function is

$$h(t) = \frac{f(t)}{S(t)} \approx \frac{P(t \leq Y \leq t + \Delta | Y \geq t)}{\Delta}$$

- For $t_j' \leq t < t_{j+1}'$, if $\tau_j = t_{j+1}' - t_j'$ is the $j-th$ interval length.

$$h(t) \approx \frac{P(t_j' \leq Y \leq t_j' + \tau_j | Y \geq \tau_j)}{\tau_j}$$

- Since $n_j' = n_j - 0.5c_j$ is the average no. at risk in the $j-th$ interval, $n_j' - 0.5d_j$ is the average no. alive in the $j-th$ interval.

- Then

$$P(t_j' \leq Y \leq t_j' + \tau_j | Y \geq \tau_j) \approx d_j/(n_j' - 0.5d_j)$$

so the estimator for the hazard is

$$h^*(t) = \left(\frac{d_j}{n_j' - 0.5d_j}\right)\left(\frac{1}{\tau_j}\right)$$

with actuarial correction.

- The Kaplan-Meier estimator (no correction) is

$$\hat{h}^*(t) = \left(\frac{d_j}{n_j}\right)\left(\frac{1}{\tau_j}\right)$$

- Collett's book provides a a formula for SE's of estimated hazards.

$$SE(h^*(t)) = \frac{h^*(t)}{\sqrt{d_j}}\sqrt{1 - \left(\frac{h^*(t)\tau_j}{2}\right)^2}; t_j \leq t < t_{j+1}$$

- Kaplan-Meier estimate of $h(t)$ on the negatively stained data

Interval	n_j	d_j	$\frac{d_j}{n_j}$	τ_j	$\hat{h}^*(t)$
$[0, 23)$	13	0	0	23	0
$[23, 47)$	13	1	1/13	24	(1/13)(1/24)=0.32
$[47, 69)$	12	1	1/12	22	(1/12)(1/22)=0.038
$[69, 148)$	11	1	1/11	79	(1/11)(1/79)=0.001
$[148, 181)$	6	1	1/6	33	(1/6)(1/33)=0.005
$[181, \infty)$		undefined			

- Estimator of the hazard function $h(t)$ with actuarial correction.

Interval	n_j	d_j	c_j	n_j'	$n_j' - 0.5d_j$	τ_j	$h^*(t)$
$[0, 50)$	13	2	0	13	12	50	(2/12)(1/50)=0.0033
$[50, 100)$	11	1	2	10	9.5	50	(1/9.5)(1/50)=0.002
$[100, 150)$	8	1	2	7	6.5	50	(1/6.5)(1/50)=0.00307
$[150, 200)$	5	1	1	4.5	4.0	50	(1/4)(1/50)=0.005
$[200, 250)$	3	0	3	1.5	1.5	50	(0/1.5)(1/50)=0