Survival Analysis Biostatistics 140.641 Zebin WANG

100

Problem Set 2.

\# 1. failure time goes with exponential distribution. plug in the equation (λ is the HAZARD parameter)

Use $\mathcal{L} = \prod_{i=1}^{n} \{[f(y_i;\theta)^{\delta_i} S(y_i;\theta)^{1-\delta_i}][g(y_i)^{1-\delta_i} G(y_i^+)^{\delta_i}]\}$

$T \sim Exp(\lambda)$

$\Rightarrow f_T(t,\lambda) = \lambda e^{-\lambda t} \mathbb{1}\{t \geq 0\}$ $\quad S(t,\lambda) = e^{-\lambda t} \mathbb{1}\{t \geq 0\}$

(a) If we assume the independency between the failure and censoring times, we may reduce the likelihood to a more simple case:

$L = \prod_{i=1}^{n} [f(y_i;\lambda)^{\delta_i} S(y_i;\lambda)^{1-\delta_i}]$ (with "+" $\to \delta_i = 0$ $\to$ data censored)

$= (\lambda e^{-0.24\lambda})(\lambda e^{-0.47\lambda})(\lambda e^{-0.81\lambda})(e^{-1.22\lambda})(\lambda e^{-1.75\lambda})(e^{-2.53\lambda})$

$= \lambda^4 \cdot [e^{-0.24\lambda - 0.47\lambda - 0.81\lambda - 1.22\lambda - 1.75\lambda - 2.53\lambda}] = \lambda^4 \cdot e^{-7.02\lambda}$

Or alternatively, if we have NO such assumption, we should do

$L = \prod_{i=1}^{n} [f(y_i;\lambda)^{\delta_i} S(y_i;\lambda)^{1-\delta_i}][g(y_i)^{1-\delta_i} G(y_i^+)^{\delta_i}]$

instead, yet we have no idea on a possible close form of $g(y_i)$ (pdf for censoring) or $G(y_i)$ ($\mathbb{P}(C_i > y_i)$), so we can simply put it here without further calculation.

(b). Use the most straightforward way to do this:

$L = L(\lambda) = \lambda^4 \cdot e^{-7.02\lambda}$

$\log L = 4\log\lambda - 7.02\lambda$ and L & $\log L$ has the same maximum as $\log(\cdot)$ is monotonic.

$\Rightarrow \frac{\partial \log L}{\partial \lambda} = \frac{4}{\lambda} - 7.02$ $\quad \hat{\lambda} = \frac{4}{7.02} = 0.57$ is the MLE of λ; ($\hat{\lambda}_{MLE} = \sum_{i=1}^{n} \delta_i / \sum_{i=1}^{n} y_i$)

□

#2. (a). for placebo group:

1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

Empirical Survival distribution:

$$\hat{S}(t) = \frac{\#\{t_i \geq t\}}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{t_i \geq t\}$$

$t \in [0,1]$ $\hat{S}(t) = 1$ $t \in (1,2]$ $\hat{S}(t) = 19/21$

$t \in (2,3]$ $\hat{S}(t) = 17/21$ $t \in (3,4]$ $\hat{S}(t) = 16/21$

$t \in (4,5]$ $\hat{S}(t) = 14/21 = 2/3$ $t \in (5,8]$ $\hat{S}(t) = 12/21 = 4/7$

$t \in (8,11]$ $\hat{S}(t) = 8/21$ $t \in (11,12]$ $\hat{S}(t) = 6/21 = 2/7$

$t \in (12,15]$ $\hat{S}(t) = 4/21$ $t \in (15,17]$ $\hat{S}(t) = 3/21 = 1/7$

$t \in (17,22]$ $\hat{S}(t) = 2/21$ $t \in (22,23]$ $\hat{S}(t) = 1/21$

$t \in (23,+\infty)$ $\hat{S}(t) = 0$

Kaplan - Meier estimator:

first, get $N_{(i)}$ & $d_{(i)}$ & $y_{(i)}$

$y_{(1)} = 1$ $y_{(2)} = 2$ $y_{(3)} = 3$ $y_{(4)} = 4$ $y_{(5)} = 5$ $y_{(6)} = 8$ $y_{(7)} = 11$

$y_{(8)} = 12$ $y_{(9)} = 15$ $y_{(10)} = 17$ $y_{(11)} = 22$ $y_{(12)} = 23$.

$N_{(1)} = 21$ $N_{(2)} = 19$ $N_{(3)} = 17$ $N_{(4)} = 16$ $N_{(5)} = 14$ $N_{(6)} = 12$

$N_{(7)} = 8$ $N_{(8)} = 6$ $N_{(9)} = 4$ $N_{(10)} = 3$ $N_{(11)} = 2$ $N_{(12)} = 1$

$d_{(1)} = 2$ $d_{(2)} = 2$ $d_{(3)} = 1$ $d_{(4)} = 2$ $d_{(5)} = 2$ $d_{(6)} = 4$ $d_{(7)} = 2$

$d_{(8)} = 2$ $d_{(9)} = 1$ $d_{(10)} = 1$ $d_{(11)} = 1$ $d_{(12)} = 1$

$\Rightarrow$ K-M estimate $\hat{S}(t) = \prod_{y_{(j)} < t} \left(1 - \frac{d_{(j)}}{N_{(j)}}\right)$.

$\Rightarrow$ $t \in [0,1]$ $\hat{S}(t) = 1$ $t \in (1,2]$ $\hat{S}(t) = 1 - \frac{2}{21} = 19/21$

$t \in (2,3]$ $\hat{S}(t) = (1-\frac{2}{21})(1-\frac{2}{19}) = 17/21$

$t \in (3,4]$ $\hat{S}(t) = (1-\frac{2}{21})(1-\frac{2}{19})(1-\frac{1}{17}) = 16/21$

$t \in (4,5]$ $\hat{S}(t) = 16/21 \cdot (1 - 2/16) = 14/21 = 2/3$

$t \in (5,8]$ $\hat{S}(t) = 14/21 \cdot (1 - 2/14) = 12/21 = 4/7$

$t \in (8,11]$ $\hat{S}(t) = 12/21 \cdot (1 - 4/12) = 8/21$

$t \in (11,12]$ $\hat{S}(t) = 8/21 \cdot (1 - 2/8) = 6/21 = 2/7$

$t \in (12,15]$ $\hat{S}(t) = 6/21 \cdot (1 - 2/6) = 4/21$

$t \in (15,17]$ $\hat{S}(t) = 4/21 \cdot (1 - 1/4) = 3/21 = 1/7$

$t \in (17,22]$ $\hat{S}(t) = 3/21 \cdot (1 - 1/3) = 2/21$

$t \in (22,23]$ $\hat{S}(t) = 2/21 \cdot (1 - 1/2) = 1/21$

$t \in (23,+\infty)$ $\hat{S}(t) = 1/21 \cdot (1 - 1/1) = 0$

From the calculation, I can find out that
1° Both of the estimates happen to be exactly the same: they have the same value in all the $t \ge 0$
2° Both of the estimates hit 0 when time t is beyond the time for the last patient in the group to terminate remission time (targeted event occurs).

(b). For all uncensored survival data:
1° The empirical survival distribution is compatible with the Kaplan-Meier estimate; they happen to be exactly the same.

2°. In all uncensored data, the empirical survival distribution and the Kaplan-Meier estimator will reach $\hat{S}(t)=0$ after the last event occurs.

(c). K-M for 6MP..
$y_{(1)}=6$ $y_{(2)}=7$ $y_{(3)}=10$ $y_{(4)}=13$ $y_{(5)}=16$ $y_{(6)}=22$ $y_{(7)}=23$
$N_{(1)}=21$ $N_{(2)}=17$ $N_{(3)}=15$ $N_{(4)}=12$ $N_{(5)}=11$ $N_{(6)}=7$ $N_{(7)}=6$
$d_{(1)}=3$ $d_{(2)}=1$ $d_{(3)}=1$ $d_{(4)}=1$ $d_{(5)}=1$ $d_{(6)}=1$ $d_{(7)}=1$

$\Rightarrow$ $t\in[0,6]$ $\hat{S}(t)=1$ $t\in(6,7]$ $\hat{S}(t)=(1-\frac{3}{21})=6/7=18/21$
$t\in(7,10]$ $\hat{S}(t)=18/21\cdot(1-\frac{1}{17})=0.81$
$t\in(10,13]$ $\hat{S}(t)=0.81\cdot(1-1/15)=0.75$
$t\in[13,16]$ $\hat{S}(t)=0.75\cdot(1-1/12)=0.69$
$t\in(16,22]$ $\hat{S}(t)=0.69\cdot(1-1/11)=0.63$
$t\in(22,23]$ $\hat{S}(t)=0.63\cdot(1-1/7)=0.54$
$t\in(23,+\infty)$ $\cdot\hat{S}(t)=0.54\cdot(1-1/6)=0.45$

(d). Variance estimate for Placebo data: $n=21$
$t\in[0,1]$ $\hat{S}(t)=1$; 95% CI: {1}. $(t=1)$
$t=2$ $\hat{S}(t)=19/21=0.90$; 95% CI: $(0.78,1]$
$t=3$ $\hat{S}(t)=17/21=0.81$; 95% CI: $(0.64,0.98)$

$t=4$	$\hat{S}(t) = 16/21 = 0.76$;	95% CI: (0.58, 0.94)
$t=5$	$\hat{S}(t) = 14/21 = 0.67$;	95% CI: (0.47, 0.87)
$t=8$	$\hat{S}(t) = 12/21 = 0.57$;	95% CI: (0.36, 0.78)
$t=11$	$\hat{S}(t) = 8/21 = 0.38$;	95% CI: (0.17, 0.59)
$t=12$	$\hat{S}(t) = 6/21 = 0.29$;	95% CI: (0.09, 0.48)
$t=15$	$\hat{S}(t) = 4/21 = 0.19$;	95% CI: (0.02, 0.36)
$t=17$	$\hat{S}(t) = 3/21 = 0.14$;	95% CI: [0, 0.29)
$t=22$	$\hat{S}(t) = 2/21 = 0.10$;	95% CI: [0, 0.22)
$t=23$	$\hat{S}(t) = 1/21 = 0.05$;	95% CI: [0, 0.14)
$t>23$	$\hat{S}(t) = 0$	95% CI: {0}

Comment:

#1. The CI's are symmetric.

#2. When $\hat{S}(t)$ close to 0.5, the CI tend to be wider; When $\hat{S}(t)$ close to 0, (or 1), the CI tend to be shallow.

#3. The CI's use the Central Limit theorem, which may not be perfect in cases that $n=21$ may not be large enough for CLT to work well.

R: Somethimes the CI will go ant of bound (>1 or <0), which shall not happen. We need to draw the 95% CI back to where $\hat{S}(t)$ is defined.

(e). Calculate the 95% CI for the K-M at each uncensored time for 6MP data.

Apply the Greenwood's formula.

$$\hat{var}(\hat{S}(t)) \approx [\hat{S}(t)]^2 \sum_{y_{(j)}<t} \frac{d_{(j)}}{N_{(j)}(N_{(j)}-d_{(j)})}$$

$t \in [0,6]$ $\hat{S}(t)=1$. yet all $y_{(j)} \geq 6$ => 95% CI: {1}

→ $t=6$	$\hat{S}(t)=1$;	95% CI, {1}.	
$t=7$	$\hat{S}(t)=0.87$;	95% CI: [0.72, 1]	$\hat{se}=0.078$
$t=10$	$\hat{S}(t)=0.81$;	95% CI: (0.64, 0.98)	$\hat{se}=0.087$
$t=13$	$\hat{S}(t)=0.75$;	95% CI: (0.56, 0.94)	$\hat{se}=0.096$
$t=16$	$\hat{S}(t)=0.69$;	95% CI: (0.48, 0.90)	$\hat{se}=0.107$
$t=22$	$\hat{S}(t)=0.63$;	95% CI: (0.41, 0.85)	$\hat{se}=0.115$

$t = 23$ $\quad \hat{S}(t) = 0.54$; $\quad$ 95% CI : $(0.29, 0.79)$ $\hat{se} = 0.129$
$t > 23$ $\quad \hat{S}(t) = 0.45$; $\quad$ 95% CI : $(0.19, 0.71)$ $\hat{se} = 0.135$

Comments:
1° All the CI's are symmetric
2° We shall rule out the impossible values for a survival function ($S<0$ or $S>1$). and fix the CI a little bit.
3°. When time rises, the $\hat{S}(t)$ drops but the

$$\sum_{y_{(j)}<t} \frac{d_{(j)}}{N_{(j)}(N_{(j)}-d_{(j)})}$$

grows, so the overall $\hat{se} = \text{sqrt}(\hat{var})$ grows slightly over time.
4° In the beginning, the

$$\sum_{y_{(j)}<t} \frac{d_{(j)}}{N_{(j)}(N_{(j)}-d_{(j)})}$$ keeps to be 0

$\Rightarrow \hat{se} = 0 \quad \Rightarrow \{1\}$ is the 95% CI.
5° Proper results rely on large size of uncensored data.

(f): Compare the estimates for placebo & 6MP groups.

For all the time when $t>8$, the 95% CI for the $\hat{S}(t)$ of placebo group has no overlap with the 95% CI for the $\hat{S}(t)$ of 6MP group, and the estimate $\hat{S}(t)$ for 6MP group is always higher than the $\hat{S}(t)$ estimate for placebo group.

For $t>8$, there is evidence from the data showing that the survival function for 6MP group has a higher value than the survival function for placebo group. i.e., there is evidence supporting the proposition / assumption that 6MP helps in maintaining remission status for patients with acute leukemia after 8 units of time after the beginning of remission.

#3. a) The two data sets have the same data that are censored (all with $2^+, 5^+$). For the uncensored data, the only difference is that one case (uncensored data) 5 is Data A has been moved to 7 in Data B.

better state a conclusion here, like B

A: 0 1 1 2^+ 3 5 6^+ 9 10

B: 0 1 1 2^+ 3 6^+ 7 9 10

(b). for A.

$t=0 \quad \hat{S}(t)=1$

$t\in(0,1] \quad \hat{S}(t)=(1-\frac{1}{9}) = 8/9$

$t\in(1,3] \quad \hat{S}(t)= 8/9 \cdot (1-2/8) = 6/9 = 2/3$

$t\in(3,5] \quad \hat{S}(t) = 2/3 \cdot (1-\frac{1}{5}) = 2/3 \cdot 4/5 = 8/15$

$t\in(5,9] \quad \hat{S}(t) = 8/15 \cdot (1-1/4) = 8/15 \cdot 3/4 = 2/5$

$t\in[9,10] \quad \hat{S}(t) = 2/5 \cdot (1-1/2) = 1/5$

$t>10\ (t\in(10,+\infty)) \quad \hat{S}(t) = 1/5 \cdot (1-1/1) = 0.$

as $y_{(1)}=0$ $y_{(2)}=1$ $y_{(3)}=3$ $y_{(4)}=5$ $y_{(5)}=9$ $y_{(6)}=10.$

$N_{(1)}=9$ $N_{(2)}=8$ $N_{(3)}=5$ $N_{(4)}=4$ $N_{(5)}=2$ $N_{(6)}=1$

$d_{(1)}=1$ $d_{(2)}=2$ $d_{(3)}=1$ $d_{(4)}=1$ $d_{(5)}=1$ $d_{(6)}=1$

(c). for B.

$t=0 \quad \hat{S}(t)=1$

$t\in(0,1] \quad \hat{S}(t) = (1-1/9) = 8/9$

$t\in(1,3] \quad \hat{S}(t) = (1-1/9)(1-2/8) = 6/9 = 2/3$

$t\in(3,7] \quad \hat{S}(t) = 2/3 \cdot 4/5 = 8/15$

$t\in(7,9] \quad \hat{S}(t) = 8/15 \cdot (1-1/3) = 8/15 \cdot 2/3 = 16/45$

$t\in(9,10] \quad \hat{S}(t) = 16/45 (1-1/2) = 8/45$

$t\in(10,+\infty) \quad \hat{S}(t) = 8/45 (1-1/1) = 0$

as $y_{(1)}=0$ $y_{(2)}=1$ $y_{(3)}=3$ $y_{(4)}=7$ $y_{(5)}=9$ $y_{(6)}=10$

$N_{(1)}=9$ $N_{(2)}=8$ $N_{(3)}=5$ $N_{(4)}=3$ $N_{(5)}=2$ $N_{(6)}=1$

$d_{(1)}=1$ $d_{(2)}=2$ $d_{(3)}=1$ $d_{(4)}=1$ $d_{(5)}=1$ $d_{(6)}=1$

(d). Results of Comparison:

1° for $t\in[0,5]$, the $\hat{S}(t)$ (K-M estimate) for both of the groups stays the same. This is before any difference between the groups.

2° for $t\in(5,10]$, the K-M estimate $\hat{S}(t)$ for the two groups are different. This difference holds even after $t \geq 9$, where Group A & B have exactly the same data after that time.

3°. for $t > 10$, since both the data ends with a non-censor observation, the K-M estimate backs to be the same as $\hat{S}(t) = 0$ for $t > 10$ for both Data A & Data B.

#4 M1. $h(t) = \theta \mathbb{1}\{t>0\}$ where $\theta > 0$.

M2 $h(t) = \theta e^{\theta t} \mathbb{1}\{t>0\}$ where $\theta > 0$

M3 $h(t) = 5e^{-\theta t} \mathbb{1}\{t>0\}$ where $\theta > 0$.

(a) $h(t) = f(t)/S(t)$ for all the continuous cases

and $f(t) = -\dfrac{dS(t)}{dt}$.

M1: $\Rightarrow \theta S(t) \mathbb{1}\{t>0\} = -\dfrac{dS(t)}{dt}$

$\Rightarrow \theta t \mathbb{1}\{t>0\} + C = -\log S(t)$

$\Rightarrow S(t) = e^{-\theta t} \mathbb{1}\{t>0\} \cdot e^{c}$ $S(0)=1$ $\Rightarrow$

$S(t) = e^{-\theta t} \mathbb{1}\{t>0\}$.

$f(t) = \theta e^{-\theta t} \mathbb{1}\{t>0\}$. is the pdf of T.

$\mathbb{P}(0 \leq T < +\infty) = \int_0^{+\infty} f(t)dt = \lim_{k \to +\infty} \int_0^k \theta e^{-\theta t} dt = \lim_{k \to +\infty} (1 - e^{-\theta k})$

$= 1 \Rightarrow \mathbb{P}(0 \leq T < +\infty) = 1$ So it is a regular model.

$S(+\infty) = 0$ as $e^{-\theta t} \to 0$ for $\theta > 0$, $t \to +\infty$.

$\Rightarrow$ M1 is a regular model.

M2: $\theta e^{\theta t} S(t) \mathbb{1}\{t>0\} = -\dfrac{dS(t)}{dt}$

$\Rightarrow \int \theta e^{\theta t} dt \, \mathbb{1}\{t>0\} = -\log S(t)$

$[e^{\theta t} + C] \mathbb{1}\{t>0\} = -\log S(t)$

$S(t) = e^{-[e^{\theta t} + C]} \mathbb{1}\{t>0\}$ $S(0)=1 \Rightarrow e^{-[1+C]} \mathbb{1}\{t>0\} = 1$

$\Rightarrow C = -1$

$S(t) = e^{-[e^{\theta t} - 1]} \mathbb{1}\{t>0\} = e \cdot e^{-e^{\theta t}} \mathbb{1}\{t>0\}$

$$f(t) = -\frac{dS(t)}{dt} = (-e)\cdot \frac{d}{dt}\left[e^{-e^{\theta t}}\right] \mathbb{1}\{t>0\}$$

$$= (-e)\cdot e^{-e^{\theta t}} \cdot \frac{d}{dt}(-e^{\theta t}) \mathbb{1}\{t>0\}$$

$$= e\cdot \theta e^{\theta t}\cdot e^{-e^{\theta t}} \mathbb{1}\{t>0\}$$

and $\mathbb{P}(0 \le T < +\infty) = \lim_{k\to+\infty} \mathbb{P}(0\le T<k) = \int_0^{+\infty} e\cdot\theta e^{\theta t} e^{-e^{\theta t}}\,dt$

$$= -e\cdot e^{-e^{\theta t}}\Big|_0^{+\infty} = 1-0=1$$

And $S(+\infty) = \lim_{t\to+\infty} S(t) = 0$ (Just take the limit $t\to+\infty$).

$\Rightarrow$ this model is a regular model with $S(+\infty)=0$

$\Rightarrow$ M_2 is a regular model.

M3. $5e^{-\theta t} S(t)\mathbb{1}\{t>0\} = -\frac{dS(t)}{dt}$

$$\Rightarrow \int 5e^{-\theta t}\,dt\, \mathbb{1}\{t>0\} = -\log S(t).$$

$$\Rightarrow \left[-\frac{5e^{-\theta t}}{\theta} + c\right]\mathbb{1}\{t>0\} = -\log S(t)$$

$$\Rightarrow S(t) = e^{\frac{5e^{-\theta t}}{\theta} - c}\mathbb{1}\{t>0\} \quad S(0)=1 \Rightarrow$$

$$S(0) = e^{\frac{5}{\theta}-c} = 1 \Rightarrow c = \frac{5}{\theta}$$

$$\Rightarrow S(t) = e^{\frac{5}{\theta}[e^{-\theta t}-1]}\mathbb{1}\{t>0\}$$

$$f(t) = -\frac{dS(t)}{dt} = -\frac{d}{dt}\left[e^{\frac{5}{\theta}e^{-\theta t}}\right]e^{-\frac{5}{\theta}} = -e^{\frac{5}{\theta}e^{-\theta t}}\frac{d}{dt}\left[\frac{5}{\theta}e^{-\theta t}\right]\cdot e^{-\frac{5}{\theta}}$$

$$= 5e^{-\theta t}e^{\frac{5}{\theta}e^{-\theta t}}\cdot e^{\frac{5}{\theta}} = 5e^{-\theta t}e^{\frac{5}{\theta}[e^{-\theta t}-1]}\mathbb{1}\{t>0\}.$$

$$\mathbb{P}(0\le T<+\infty) = \int_0^{+\infty} f(t)\,dt = -e^{\frac{5}{\theta}[e^{-\theta t}-1]}\Big|_0^{+\infty} = -e^{-\frac{5}{\theta}}+1 = 1-e^{-\frac{5}{\theta}} < 1$$

$S(+\infty) = e^{-\frac{5}{\theta}} > 0 \Rightarrow$ M3 is a cure model.

(b). pdf for T.

M1: pdf for T is $f(t) = \theta e^{-\theta t}\mathbb{1}\{t>0\}$

M2: pdf for T is $f(t) = e\cdot\theta e^{\theta t}e^{-e^{\theta t}}\mathbb{1}\{t>0\}$

M3: pdf for T is $f(t) = 5e^{-\theta t}\cdot e^{\frac{5}{\theta}[e^{-\theta t}-1]}\mathbb{1}\{t>0\}$.

$\Rightarrow f(t) = \mathbb{1}\{0<t<+\infty\}\, 5e^{-\theta t}e^{\frac{5}{\theta}[e^{-\theta t}-1]} + \mathbb{1}\{t=+\infty\}\cdot e^{-\frac{5}{\theta}}$ □