Lin_ST625_HW4

Frances Lin

1/28/2022

1a

$\frac{d_i}{n_i}$ estimates conditional failure probability at t_i given alive before t_i.

1b

$1 - \frac{d_i}{n_i}$ estimates conditional survival probability at t_i given alive before t_i.

1c

No, to maximize the likelihood, point mass must only be put at each observed time. KM estimator does not change at censoring times or between events.

1d

Yes, to maximize the likelihood, point mass must be put at each observed time. KM estimator changes value only at event times.

1e

Suppose we have a toy example given as $\{2, 3, 4, 8\}$, computing the KM estimate manually using the KM formula, it is easy to see that

t	d	n	1 - d/n	S_t
2	1	4	0.75	0.75
3	1	3	0.6667	0.5
4	1	2	0.5	0.25
8	1	1	0	0

$S(t)$ from the above matches if we were to calculate the survival probability using the empirical survival function, which is given as $S(t) = \frac{\#\ of\ individal\ with\ T \geq t}{total\ \#\ of\ individal}$.

For example, $S(t = 2) = 3/4 = 0.75, S(t = 3) = 2/4 = 0.5$, etc.

2a

At event time t_i, the KM estimator is calculated recursively by

$$\hat{S}(t_i) = \hat{S}(t_{i-1})(1 - \frac{d_i}{n_i}),$$

where d_i is the # of events at t_i, n_i is the # of individuals at risk just before t_i, $\frac{d_i}{n_i}$ is the conditional probability of failure at t_i given alive before t_i, and $1 - \frac{d_i}{n_i}$ is the conditional probability of surviving t_i given alive before t_i.

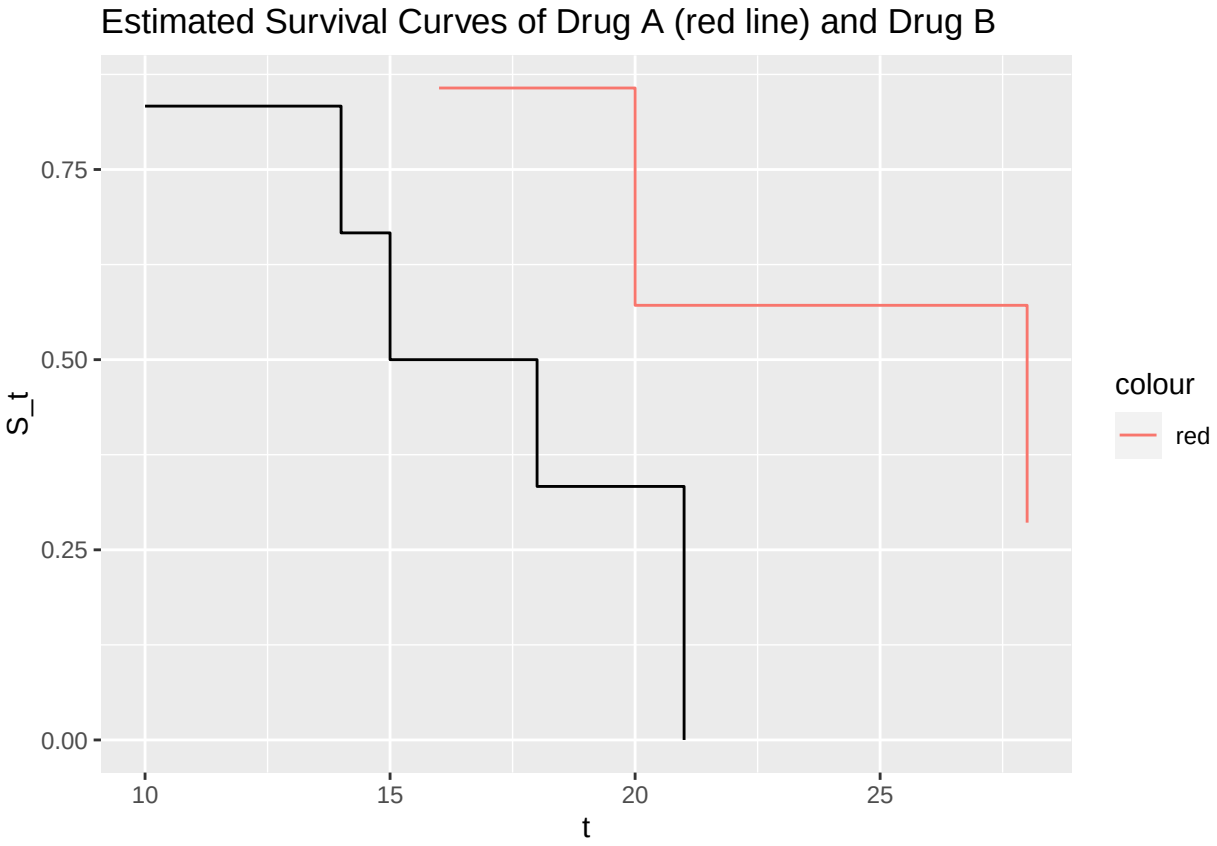
Drug A: $16, 16+, 18+, 19+, 20, 28, 32+$

t	d	n	1 - d/n	S_t
16	1	7	0.8571	0.8571
20	1	3	0.6667	0.5714
28	1	2	0.5	0.2857

1. List out the uncensored survival times (e.g. $t = 16, 20, 28$)
2. Count the # of event at each t_i (e.g. $d = 1, 1, 1, 1$)
3. Count the # of people still alive including t_i (e.g. $n = 7, 3, 2$)
4. Compute $1 - \frac{d}{n}$ for all i
5. Recursively compute $S(t_i) = S(t_{i-1})(1 - \frac{d_i}{n_i})$ (e.g. $S(t = 16) = 1 - \frac{d_1}{n_1}$, $S(t = 20) = S(t = 16)(1 - \frac{d_2}{n_2}) = 0.8571 * 0.6667 = 0.5714286$)

Drug B: $10, 14, 15, 18, 20+, 21$

t	d	n	1 - d/n	S_t
10	1	6	0.8333	0.8333
14	1	5	0.8	0.6667
15	1	4	0.75	0.5
18	1	3	0.6667	0.3333
21	1	1	0	0



2b

$t = 28$ is the median survival time for Drug A since $S(t = 28) \leq 0.5$.

$t = 15$ is the median survival time for Drug B since $S(t = 15) \leq 0.5$.

2c

For Drug A, since $S(t = 20) = 0.5714286, H(t = 20) = -log(S(t = 20)) = -log(0.5714286) =$

```
## [1] 0.5596157
```

For Drug B, since $S(t = 20) = 0.3333333, H(t = 20) = -log(S(t = 20)) = -log(0.3333333) =$

```
## [1] 1.098612
```

2d

Drug A is more effective, since the median survival time of Drug A is higher than that of Drug B. (For this data set, whenever times overlap, the estimated survival probability of Drug A is also higher than that of Drug B at each time point.)

3a

Life-Table method is calculated by

$$\hat{S}(b_j) = \hat{S}(b_{j-1})(1 - \frac{d_j}{n_j - c_j/2}),$$

where $m_j = \frac{d_j}{n_j - c_j/2}$ is the mortality rate, $e_j = n_j - c_j/2$ is the effective risk size, d_j is the # of event times, n_j is the # of individuals at risk at the start of the j interval, and c_i is the # of censored individuals.

age_1	age_2	d	c	n	e	m	S_t
45	50	17	29	1571	1556	0.01092	0.9891
50	55	36	60	1525	1495	0.02408	0.9653
55	60	62	83	1429	1388	0.04468	0.9221
60	65	76	441	1284	1064	0.07146	0.8562
65	70	50	439	767	547.5	0.09132	0.778
70	75	9	262	278	147	0.06122	0.7304
75	80	0	7	7	3.5	0	0.7304

1. Calculate n_j (# of individuals at risk at the start of the ith interval), which is = total # $-(d + c)$ (e.g. $n_2 = n_1 - (d_1 + c_1) = 1571 - (17 + 29) = 1525$)
2. Calculating e_j (effective risk size), which is $= n_j - c_j/2$
3. Obtain m_j (mortality rate), which is $= \frac{d_j}{e_j}$
4. Use m_j to get $\hat{S}(t)$ recursively (e.g. $\hat{S}(1) = 1 - m_1$, $\hat{S}(2) = \hat{S}(1) * (1 - m_2) = 0.9891 * (1 - 0.02408) = 0.9652825$, etc.)

3b

0.9221284 is the estimated survival probability for the third interval $[55 - 60]$.

```
## [1] 0.9221284
```

3c

The survival probability past 60 years (or past the third interval $[55 - 60]$) is $S(60) = 0.9221284$.

4

4a

Admitdate	Fdate	time
1997-01-13	2002-12-31	2178 days
1997-01-19	2002-12-31	2172 days
1997-01-01	2002-12-31	2190 days
1997-02-17	1997-12-11	297 days
1997-03-01	2002-12-31	2131 days
1997-03-11	1997-03-12	1 days

First few rows of the survival time of interest is given as:

```
## Time differences in days
##  [1] 2178 2172 2190  297 2131    1 2122 1496  920 2175 2173 1671 2192  865 2166
## [16] 2168  905 2353 2146   61
```

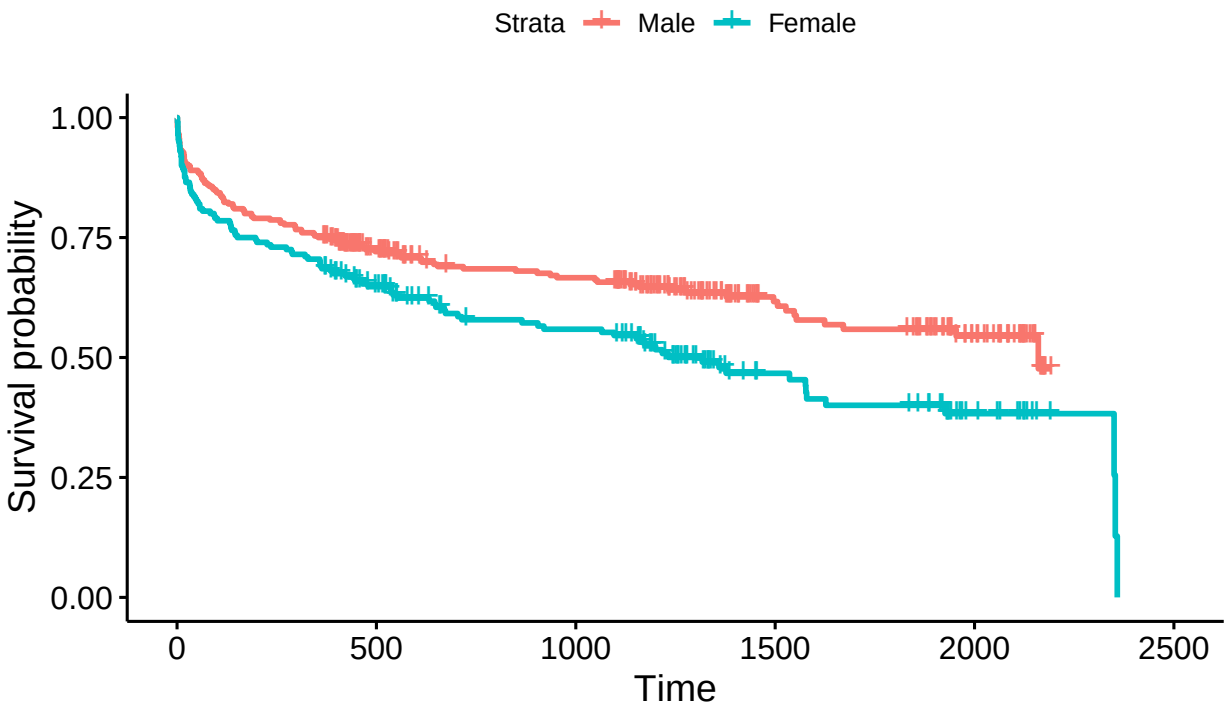
4b

First few rows of the survival probability (Male) is given:

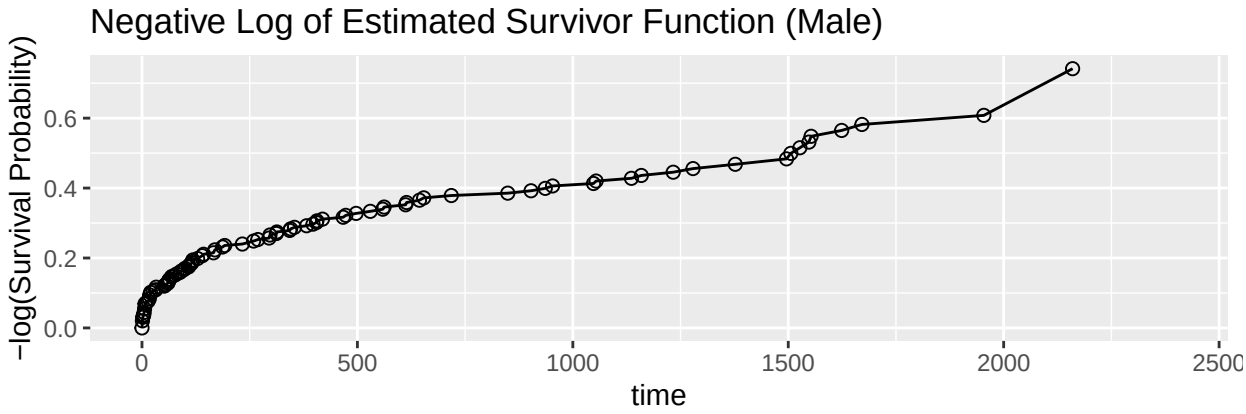
```
##  [1] 1.0000000 0.9800000 0.9700000 0.9666667 0.9633333 0.9566667 0.9466667
##  [8] 0.9333333 0.9300000 0.9266667 0.9200000 0.9100000 0.9033333 0.9000000
## [15] 0.8966667 0.8900000 0.8866667 0.8833333 0.8800000 0.8766667
```

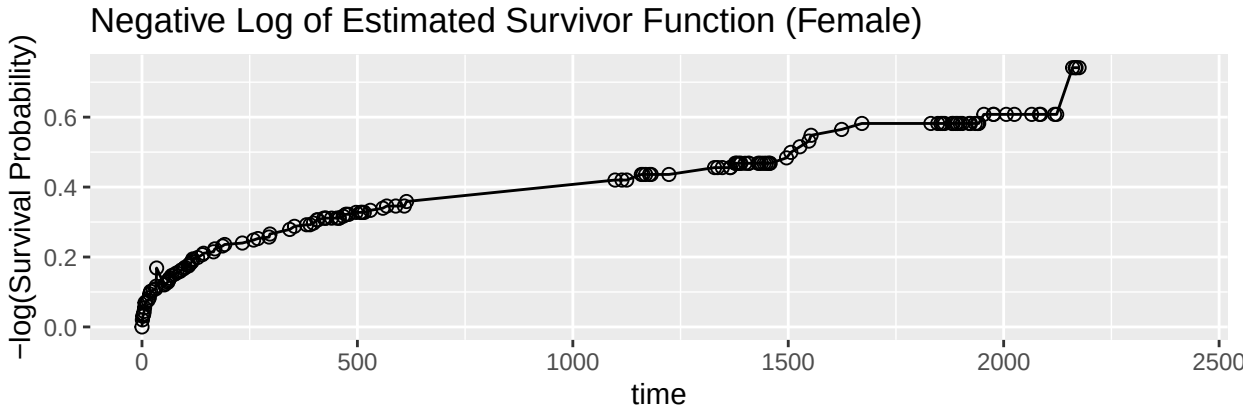
First few rows of the survival probability (Female) is given:

```
##  [1] 1.000 0.990 0.965 0.955 0.950 0.940 0.930 0.920 0.900 0.895 0.890 0.875
## [13] 0.865 0.860 0.855 0.850 0.845 0.840 0.835 0.830
```


Survival Curves of Male (gender = 0) and
Female (gender = 1)
Strata
Male
Female
1.00
0.75
0.50
0.25
0.00
Survival probability
0
500
1000
1500
2000
2500
Time

4c

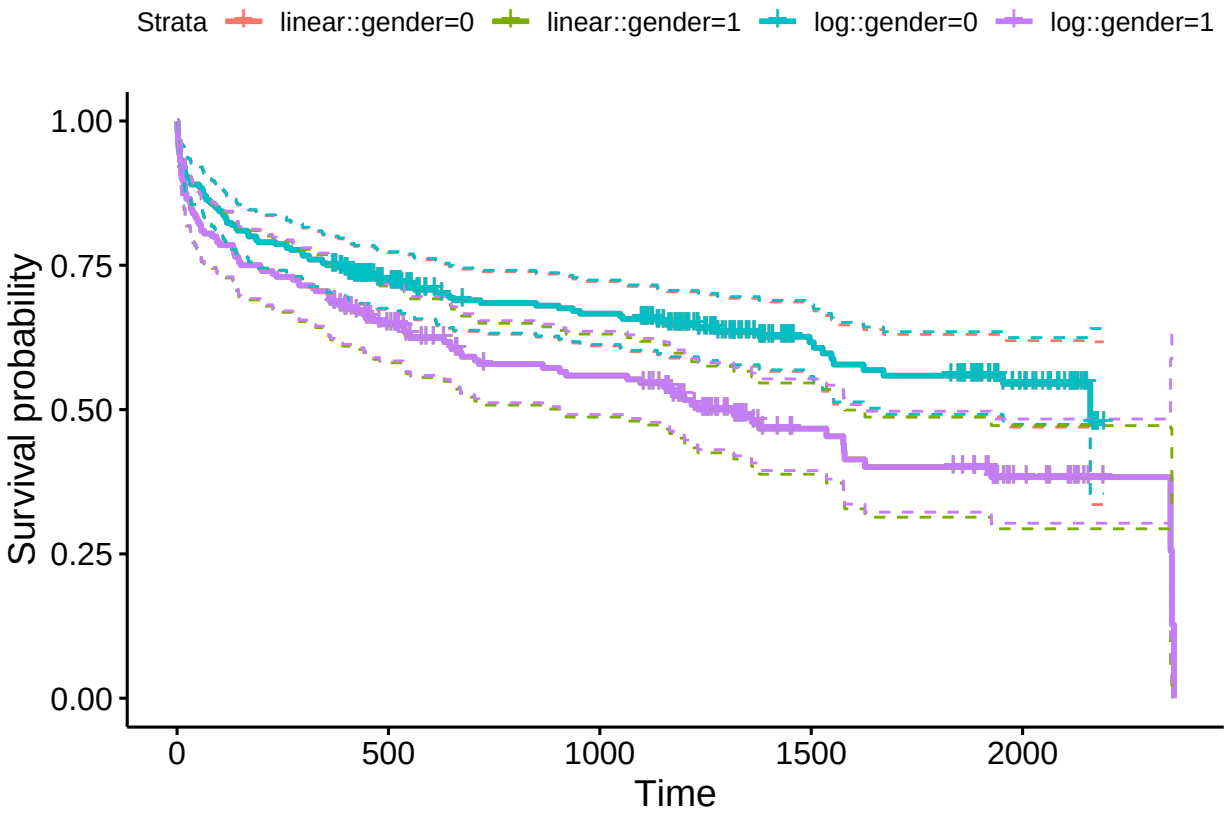




Exponential distribution does not appear to fit the survival time well for both groups, since if the survival time response (`time`) follows an exponential distribution, then the `-log(survival function)` is a linear function of `survival time`. They seem curved and are especially nonlinear at both ends.

4d

Kaplan-Meier plot with point-wise CIs for Male (gender = 0) and Female (gender = 1) is given:



4e

Based on the estimated survival curves, survival of male patients appears to be better than survival of female patients. Their difference widens as time increases.