Survival Analysis

Lecture 6

Marta Fiocco & Hein Putter

Department of Medical Statistics and Bioinformatics
Leiden University Medical Center

Outline

Interval-censored data

Interval-censored data

Self-consistency algorithm

Turnbull’algorithm (2)

Turnbull’algorithm (2)

Algorithm

Example Section 1.18

Time to Cosmetic Deterioration of Breast Cancer Patients

Right-truncated data

Right-truncated data

Example Section 1.19

AIDS data

Outline

Interval-censored data

- Interval-censored data are quite usual in longitudinal studies where individuals are not monitored continuously but scheduled to be inspected at certain times
- here the time to the event of interest is observed within consecutive visits
- Ex (De Gruttola and Lagakos (1989)) estimated the chronological time to HIV infection among haemophiliacs receiving contaminated blood factor between 1978 and 1988
- blood samples were periodically collected and stored and retrospectively tested to determine a time interval during which the infection occurred
- the infection was only known to be between the times specified by the last negative and the first positive assessment

Self-consistency algorithm

- One of the most popular methods to obtain a non-parametric estimator for the survival function under interval censoring is the use of self-consistency or Turnbull's algorithm (Turnbull, 1976)
- The search of the NPMLE (non-parametric maximum likelihood) of the survival function under interval censoring requires the definition of a set of intervals
- These intervals are obtained from the set of all left and right interval endpoints

Self-consistency algorithm

- Turnbull proved that a maximum likelihood estimator of the survival function under interval censoring concentrates its mass on this set of intervals
- Turnbull's approach for maximizing the likelihood is based on the solution of the self-consistent equations and is a special case of the **EM** algorithm.

Outline

Interval-censored data

- Interval-censored data
- Self-consistency algorithm

Turnbull'algorithm (2)

- Turnbull'algorithm (2)
- Algorithm

Example Section 1.18

- Time to Cosmetic Deterioration of Breast Cancer Patients

Right-truncated data

- Right-truncated data

Example Section 1.19

- AIDS data

Turnbull’algorithm (2)

- likelihood for interval-censored data

$$L = \prod_{i=1}^{n} [S(L_i) - S(R_i)]$$

- goal: find a monotonically decreasing function $\hat{S}_n(t)$ which maximizes the likelihood function
- several algorithms exist for obtaining the NPMLE of the survival function under interval censoring
- a popular method to obtain the survival function under interval censoring is the Turnbull’algorithm
- it requires the definition of a set of intervals (so-called Turnbull intervals)

Turnbull'algorithm (2)

- let $0 = \tau_0 < \tau_1 < \ldots < \tau_m$ be be a grid of time points which includes all the points $I = \{(\tau_0, \tau_1], (\tau_2, \tau_3], \ldots, (\tau_{m-1}, \tau_m]\}$
- these intervals are obtained from the set of all left and right interval endpoints in such a way that τ_{j-1} is a left endpoint, τ_j is a right endpoint and there is no other left or right endpoint between τ_{j-1} and τ_j
- for the i^{th} observation, define a weight α_{ij} to be 1 if the interval $(\tau_{j-1}, \tau_j]$ is contained in the interval $(L_i, R_i]$, and 0 otherwise

Turnbull'algorithm (2)

- indicator function α_{ij}

$$\alpha_{ij} = \begin{cases} 1 & \text{if } (\tau_{j-1}, \tau_j] \subset (L_i, R_i] \\ 0 & \text{otherwise} \end{cases}$$

- α_{ij}: indicator of whether the event which occurs in the interval $(L_i, R_i]$ could have occurred at time τ_j
- an initial guess of $S(\tau_j)$ is made

- **Step 1:** Compute the probability of an event occurring at time k_j $p_j = S(k_{j-1}) - S(k_j)$, $j = 1, \ldots, m$; this is the weight of the j^{th} Turnbull's interval
  - maximization of the likelihood function L defined before reduces to maximization of the following likelihood function

$$L_T(p_1, \ldots, p_m) = \prod_{i=1}^{n} (\sum_{i=1} \alpha_{ij} p_j)$$

- **Step 2:** Estimate the number of events occurred at τ_j

$$d_j = \sum_{i=1}^{n} \frac{\alpha_{ij} p_j}{\sum_l \alpha_{il} p_l}$$

  - $\sum_l \alpha_{il} p_l$: total probability assigned to possible event times in the interval $(L_i, R_i]$

Algorithm

- **Step 3:** Compute the estimated number at risk at time τ_j $Y_i = \sum_{l=j}^{m} d_l$
- **Step 4:** Compute the updated Product-Limit estimator using the pseudo data found in steps 2 and 3
  - If the updated estimate of S is close to the old version of S for all τ_i's, stop the iterative process, otherwise repeat steps 1-3 using the updated estimate of S

- set of six individuals with censoring intervals $\{(L_i, R_i], 1 \leq i \leq 6\} = \{(4,7], (3,5], (0,2], (1,4], (6,9], (8,10]\}$
- the corresponding Turnbull intervals are given by $I = \{(\tau_0, \tau_1] = (1,2], (\tau_2, \tau_3] = (3,4], (\tau_4, \tau_5] = (4,5], (\tau_6, \tau_7] = (6,7], (\tau_8, \tau_9] = (8,9]\}$
- Recall: these intervals are obtained from the set of all left and right interval endpoints in such a way that τ_{j-1} is a left endpoint, τ_j is a right endpoint and there is no other left or right endpoint between τ_{j-1} and τ_j

Outline

Interval-censored data

Interval-censored data

Self-consistency algorithm

Turnbull’algorithm (2)

Turnbull’algorithm (2)

Algorithm

Example Section 1.18

Time to Cosmetic Deterioration of Breast Cancer Patients

Right-truncated data

Right-truncated data

Example Section 1.19

AIDS data

Time to Cosmetic Deterioration of Breast Cancer Patients

- retrospective study to compare the cosmetic effects of radiotherapy alone versus radiotherapy and adjuvant chemotherapy on women with early breast cancer
- To compare the two treatment regimes, a retrospective study of 46 radiation only and 48 radiation plus chemotherapy patients was made
- Patients were observed initially every 4-6 months, but, as their recovery progressed, the interval between visits became bigger

Time to Cosmetic Deterioration of Breast Cancer Patients

- At each visit, the clinician recorded a measure of breast retraction on a three- point scale (none, moderate, severe)
- Event of interest: *time to first appearance of moderate or severe breast retraction*
- Patients were observed only at these random times then the exact time of breast retraction is known only to fall in the interval between visits: interval-censored data

Time to Cosmetic Deterioration of Breast Cancer Patients

- Data table 1.8 page 17

```
> library(KMsurv)
> data(bcdeter)
> head(bcdeter)
  lower upper treat
1     0     5     1
2     0     7     1
3     0     8     1
4     4    11     1
5     5    11     1
6     5    12     1
```

- $(a, b]$: interval (in months) in which deterioration took place
- treatment 1: Radiotherapy only
- treatment 2: Radiotherapy only + Chemotherapy

Time to Cosmetic Deterioration of Breast Cancer Patients

```
> bcdeter[22:29,]
   lower upper treat
22     0     5     2
23     0    22     2
24     4     8     2
25     4     9     2
26     5     8     2
27     8    12     2
28     8    21     2
29    10    17     2
> tail(bcdeter)
   lower upper treat
90    23    NA     2
91    31    NA     2
92    32    NA     2
93    34    NA     2
94    34    NA     2
95    35    NA     2
```

Time to Cosmetic Deterioration of Breast Cancer Patients

```
# select the 46 individuals given radiation therapy only
> dat <- bcdeter[bcdeter$treat == 1, ]

> # Grid points
> tau <- sort(unique(c(dat$lower,dat$upper)))
> tau:
> tau
 [1]  0  4  5  6  7  8 10 11 12 14 15 16 17 18 19 22 24 25 26
[20] 27 32 33 34 35 36 37 38 40 44 45 46 48
```

- Turnbull intervals

```
> n<-length(tau)
> turnbull<-matrix(c(tau[-n],tau[-1]),ncol=2)
> # ... and Turnbull intervals
> head(tau.mat)
     [,1] [,2]
[1,]    0    4
[2,]    4    5
[3,]    5    6
[4,]    6    7
[5,]    7    8
[6,]    8   10
```

Time to Cosmetic Deterioration of Breast Cancer Patients

```
> tail(tau.mat)
       [,1] [,2]
[26,]    37   38
[27,]    38   40
[28,]    40   44
[29,]    44   45
[30,]    45   46
[31,]    46   48
```

Time to Cosmetic Deterioration of Breast Cancer Patients

- Apply the algorithm to estimate the survival function
- keep track of the number at risk and number of death
- estimate the survival by the product limit estimator

```
> head(TABLE_5.4)
  tau Inizial_Survival  n.events   n.risk    Updated
1   0            1.000 0.0000000 46.00000 1.0000000
2   4            0.979 0.8417029 46.00000 0.9817021
3   5            0.955 1.1509271 45.15830 0.9566820
4   6            0.934 0.8518827 44.00737 0.9381628
5   7            0.905 1.4751998 43.15549 0.9060932
6   8            0.874 1.7420809 41.68029 0.8682219
        Change
1  0.000000000
2 -0.002354284
3 -0.001609492
4 -0.004467113
5 -0.001020741
6  0.006208760
```

Time to Cosmetic Deterioration of Breast Cancer Patients

- last part of the table

```
> tail(TABLE_5.4)
   tau Inizial_Survival  n.events   n.risk    Updated
27  38            0.439  1.996730 25.52406 0.5114637
28  40            0.385  2.294812 23.52733 0.4615764
29  44            0.328  2.358345 21.23252 0.4103081
30  45            0.284  1.329205 18.87417 0.3814123
31  46            0.229  1.850062 17.54497 0.3411936
32  48            0.000 15.694905 15.69490 0.0000000
         Change
27 -0.07257380
28 -0.07652735
29 -0.08211908
30 -0.09740226
31 -0.11223203
32  0.00000000
```

Outline

Interval-censored data

Interval-censored data

Self-consistency algorithm

Turnbull'algorithm (2)

Turnbull'algorithm (2)

Algorithm

Example Section 1.18

Time to Cosmetic Deterioration of Breast Cancer Patients

Right-truncated data

Right-truncated data

Example Section 1.19

AIDS data

Right-truncated data

- only individuals for which the event has occurred by a given date are included in the study
- Right truncation arises commonly in the study of infectious diseases
- T_i; chronological time at which the i^{th} individual is infected
- X_i time between infection and the onset of disease
- observe (T_i, X_i) for patients over the period $(0, \tau)$: only patients who have the disease prior to τ are included in the study

Right-truncated data

- the survival function is estimated by reversing the time axis
- define $R_i = \tau - X_i$; the R_i's are now left truncated since only individuals with values of $T_i \leq R_i$ are included in the sample (T_i: time at which individual i^{th} is infected)
- apply the same methodology introduced for left-truncated data
- the Product-Limit estimator of $P(R > t | R \geq 0)$ can be constructed
- in the original time scale, this is an estimator of $P(X < \tau - t | X \leq \tau)$

Outline

Interval-censored data

Interval-censored data

Self-consistency algorithm

Turnbull'algorithm (2)

Turnbull'algorithm (2)

Algorithm

Example Section 1.18

Time to Cosmetic Deterioration of Breast Cancer Patients

Right-truncated data

Right-truncated data

Example Section 1.19

AIDS data

AIDS data

- data on the infection and induction times for 258 adults and 37 children who were infected with the AIDS virus and developed AIDS by June 30, 1986
- data consists of the time in years, measured from April 1,1978, when adults were infected by the virus from a contaminated blood transfusion, and the *waiting time to development of AIDS*, measured from the date of infection
- children were infected in utero or at birth, and the infection time is the number of years from April 1, 1978 to birth

AIDS data

- **only** individuals who have developed AIDS prior to the end of the study period are included in the study
- Infected individuals who have yet to develop AIDS are not included in the sample: right-truncated data
- the data was based on an eight year observational window, $\tau = 8$ years

AIDS data

```
> data(aids)
> head(aids)
  infect induct adult
1   0.00   5.00     1
2   0.25   6.75     1
3   0.75   5.00     1
4   0.75   5.00     1
5   0.75   7.25     1
6   1.00   4.25     1
```

- infect: Infection time for AIDS, years
- induct: Induction time for AIDS, years
- adult: Indicator of adult (1=adult, 0=child)

AIDS data

- select children

```
data <- aids[aids$adult==0,]
head(data)
    infect induct adult
259   1.00   5.50     0
260   1.50   2.25     0
261   2.25   3.00     0
262   2.75   1.00     0
263   3.00   1.75     0
264   3.50   0.75     0
```

Steps needed to construct the estimate of the induction time for AIDS

- $R_i = \tau - X_i = 8 - X_i$;
  - X_i: induction time
  - R_i left truncated time
- d_i: number of individuals with the given value of R_i or, in the original time scale, the number with an induction time of X_i
- Y_i: number at risk give by
  - number of individuals with a value of R between X_i and R_i
  - in the original time scale: number of individuals with *induction times* no greater than X_i and *infection times* no greater than $8 - X_i$

AIDS data

- Ex.: $X_i = 1$ (in the original scale) $\Rightarrow R_i = 8 - 1 = 7$
- need to find which individuals satisfied this condition

```
> which(data$induct>1)
 [1]  1  2  3  5  9 10 11 12 14 15 17 20 21 24 26 27 28 30 32
> which(data$infect>7)
[1] 37
```

- 19 individuals with induction times greater than 1 and 1 individual with an infection time greater than 7
- number at risk: $Y_i = 37 - 19 - 1 = 17$
- we have to estimate: $Pr[X < x|X \leq 8]$ (estimate of the waiting time to AIDS)