

Survival Analysis

Lecture 5

Marta Fiocco^{1,2} & Hein Putter¹

(1) Department of Medical Statistics and Bioinformatics
Leiden University Medical Center

(2) Mathematical Institute Leiden University

Outline

Left truncation

- Left truncation

- Example

- Product limit estimator

- Using the survival package in R

Left and right censoring

- Left censoring

- Turnbull algorithm

- Example: marijuana data (Section 1.17)

Outline

Left truncation

- Left truncation

- Example

- Product limit estimator

- Using the survival package in R

Left and right censoring

- Left censoring

- Turnbull algorithm

- Example: marijuana data (Section 1.17)

Left truncation

- ▶ Left truncation happens when patients do not enter the study from the very beginning of the disease (late entry)
- ▶ The first time doctor sees them, the disease is already several weeks old
- ▶ The idea is that if they die within one week (say) then they will never enter the study
- ▶ The observation we have is conditional on the fact that they at least survive beyond the first week (or whatever the entering time)
- ▶ Terminology: left truncation or delayed entry
- ▶ When the truncation/entering time is 0, then there is no truncation

t_i	At risk(y_i)	Events(d_i)	Hazard(d_i/y_i)	Survival
13	2	1	1/2	1/2
16	2	1	1/2	$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$
17	1	1	1/1	$\frac{1}{4} \times 0 = 0$

- ▶ Left truncation can happen together with right censoring
- ▶ Example: *event/censoring time*

enter time event/censoring time

$(y_i, x_i) = (\underline{6}, 17), (3, 13), (2, 9^+), (0, 16)$

$(14, 17)$

- ▶ This means the first subject enters the study at 6 month after infection and die at 17 month after infection
- ▶ The third subject enters the study at 2 month and is right censored at 9 month
- ▶ Notice the observations must have $x_i > y_i$

if all $y_i = 0$, then only censored time.

- ▶ For each individual j known:
- ▶ L_j : random age at which he/she enters the study
- ▶ T_j : censored or death time
- ▶ $t_1 < t_2 \dots < t_D$: distinct death times
- ▶ d_i : number of individuals who experience the event of interest at time t_i
- ▶ Y_i : number at risk

- ▶ Y_i for right-censored data: number of individuals on study at time 0 with a study time of at least t_i $t_i \leq T_i$
- ▶ For left-truncated data, Y_i is the number of individuals who entered the study prior to time t_i and who have a study time of at least t_i ; i.e.
 - ▶ Y_i : number of individuals with $L_j \leq t_i \leq T_j$
- ▶ Use Y_i redefined for left-truncated data

- ▶ Product-Limit estimator of the survival function at a time t for the left truncated data is now an estimator of the probability of survival beyond t , conditional on survival to the smallest of the entry times L

$$P(X > t | X \geq L) = \frac{S(t)}{S(L)}$$

- ▶ Note that the number at risk could be quite small for small values of t_i (why?)
- ▶ If for some t_i we have $Y_i = d_i$ then, the Product-Limit estimator will be zero for all t beyond this point
- ▶ This happens although there are survivors and deaths beyond this point

Example: data channing library(KMsurv)

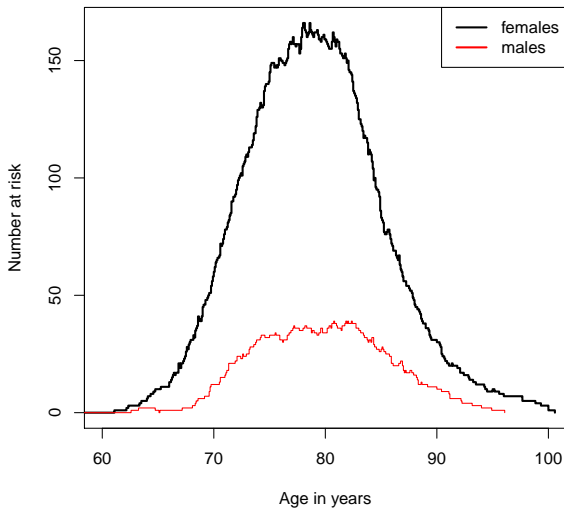
```
> library(KMsurv)
> attach(channing); ?channing
> head(channing)
```

	obs	death	age	entry	age	time	gender
1	1	1	1042	1172	130	2	
2	2	1	921	1040	119	2	
3	3	1	885	1003	118	2	
4	4	1	901	1018	117	2	
5	5	1	808	932	124	2	
6	6	1	915	1004	89	2	

Example

- ▶ *death*: Death status (1=dead, 0=alive)
 - ▶ *ageentry*: Age of entry into retirement home, months
 - ▶ *age*: Age of death or left retirement home, months
 - ▶ *time*: Difference between the above two ages, months
 - ▶ *gender*: Gender (1=male, 2=female)
-
- ▶ What is the truncation time? *Ages in months at which individuals enters the community*
 - ▶ Look at the number of individuals at risk as a function of the age at which individuals die (for males and females)
 - ▶ What do you expect?

The graph displays the number of individuals at risk for two groups: females (black line) and males (red line). The x-axis represents age in years, ranging from 60 to 100. The y-axis represents the number at risk, ranging from 0 to 150. The female curve starts near zero at age 60, rises sharply to a peak of about 160 around age 78, and then declines to near zero by age 100. The male curve also starts near zero at age 60, rises more gradually to a peak of about 40 around age 82, and then declines to near zero by age 100.



► Consider data only for males

```
> library(KMsurv)
> attach(channing)
> index <- which(gender==1) #male
> tmp <- channing[index,]
```

► Sort by age entry to see when people start entering the risk set

```
sort(tmp[,3])
[1] 751 759 782 806 817 820 821 823 830 835 835 836
[13] 836 837 843 846 847 847 852 853 854 856 856 856
[25] 863 865 865 866 871 871 875 876 878 878 879 883
[37] 885 886 890 891 893 894 898 900 906 906 909 915
[49] 919 919 921 923 925 926 936 936 938 943 943 946
[61] 953 953 955 955 956 959 960 962 962 964 966 967
[73] 967 969 969 971 978 978 981 982 984 984 988 1007
[85] 1010 1010 1016 1020 1021 1027 1036 1039 1041 1046 1051 1063
[97] 1073
```

Example

```

> head(sort(tmp[,3]))
[1] 751 759 782 806 817 820
> head(round(sort(tmp[,3]/12),0))
[1] 63 63 65 67 68 68
  > which(tmp[,3]==751)
[1] 86

```

- ▶ The risk set is empty until 751 months when the first individual enters the risk set
- ▶ A second individual enters the risk set at 759 months a third at 782 ...

- Find individuals who entered the risk set at time 751, 759, 782 months

```
> tmp[which(tmp[,3]==751),]
  obs death ageentry age time gender
451 451     1      751 777   26     1
> tmp[which(tmp[,3]==759),]
  obs death ageentry age time gender
455 455     1      759 781   22     1
> tmp[which(tmp[,3]==782),]
  obs death ageentry age time gender
366 366     1      782 909  127     1
```

- ▶ Recall

$$\prod_{i:t_i \leq t} \left(1 - \frac{d_i}{Y_i}\right) \text{ if } t_1 \leq t$$

- ▶ Compute the product limit estimator based on this data
- ▶ The estimates are as follows

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < 777 \\ 1/2 & \text{if } 777 \leq t < 781 \\ 0 & \text{if } t \geq 781. \end{cases}$$

- ▶ The estimated survival function computed in this way has no meaning since the majority of the males in the study survive beyond 781 months

- ▶ Estimate the **conditional probability** of surviving beyond age t , given survival up to age a

$$S_a(t) = P(X > t | X \geq a)$$

- ▶ Do **not** estimate the unconditional survival function
- ▶ To estimate the conditional probability consider **only** those deaths that occur after age a

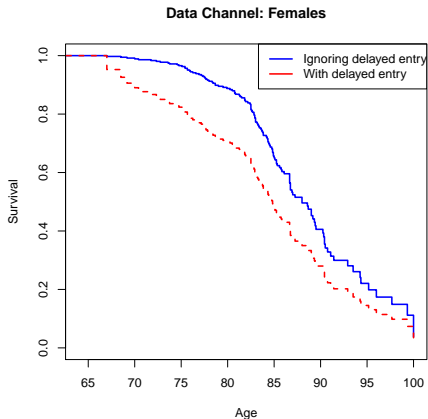
$$\hat{S}_a(t) = \prod_{a \leq t_i \leq t} \left(1 - \frac{d_i}{Y_i}\right), \quad t \geq a$$

- ▶ Note that only deaths beyond time a are considered

What do you expect in terms of estimator for $S(t)$ for data
Channing house if delayed entry is ignored?

What is important?

It is important to keep track on who is at risk!!!



- KM curve ignoring delayed entry overestimates the survival

```
> data(psych) #data from Section 1.15
> psych$time2 <- psych$age + psych$time
> head(psych)
```

	sex	age	time	death	time2
1	2	51	1	1	52
2	2	58	1	1	59
3	2	55	2	1	57
4	2	28	22	1	50
5	1	21	30	0	51
6	1	19	28	1	47

- ▶ Left truncation time is entered first as the variable `time`
- ▶ The event time (or censoring time) is `time2`
- ▶ The indicator variable δ_i for whether the event was observed is assigned to `event`

The data format is $(t_{\text{entry}}, t_{\text{exit}}, \delta)$

► R code:

```
> Surv(psych$age, psych$age+psych$time, psych$death)
```

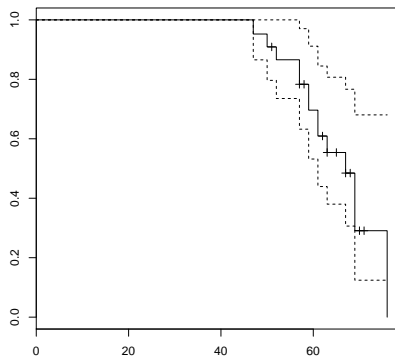
```
[1] (51,52 ] (58,59 ] (55,57 ] (28,50 ] (21,51+] (19,47 ] (25,57 ]
[8] (48,59 ] (47,61 ] (25,61+] (31,62+] (24,57+] (25,58+] (30,67+]
[15] (33,68+] (36,61 ] (30,61+] (41,63 ] (43,69 ] (45,69 ] (35,70+]
[22] (29,63+] (35,65+] (32,67 ] (36,76 ] (32,71+]
```

Using the survival package in R

```
> res <- survfit(Surv(age, time2, death) ~ 1, data=psych)
> res1 <- summary(res)
> res1
```

time	n.risk	n.event	entered	censored	survival	std.err
47	21	1	1	0	0.952	0.0465
50	22	1	0	0	0.909	0.0613
52	21	1	0	0	0.866	0.0721
57	21	2	0	1	0.783	0.0856
59	18	2	0	0	0.696	0.0957
61	16	2	0	2	0.609	0.1016
63	11	1	0	1	0.554	0.1064
67	8	1	0	1	0.485	0.1134
69	5	2	0	0	0.291	0.1261
76	1	1	0	0	0.000	NaN

```
> plot(res)
```



► Kaplan Meier for *psych* data

Outline

Left truncation

Left truncation

Example

Product limit estimator

Using the survival package in R

Left and right censoring

Left censoring

Turnbull algorithm

Example: marijuana data (Section 1.17)

Left censoring

- ▶ Left censored: data observed on the individual can be recorded as (T, δ) where $T = \max(X, C_l)$ and δ : indicator variable

$$\delta = \begin{cases} 1 & \text{if } T = X \\ 0 & \text{if } T = C_l \end{cases}$$

- ▶ Ex: childhood learning: Time-to-event: age at which a child learns to accomplish certain tasks in children learning centers
- ▶ Left censoring occurs if children can already perform the tasks when they start their study at the centers
- ▶ Examples of pure left censoring are rare; more common are samples which include both left and right censoring

- ▶ Turnbull (1974) proposed an algorithm to estimate the product limit estimator which has no closed form and it is based on an iterative procedure
- ▶ Let $0 = t_0 < t_1 < \dots < t_m$ be the grid of time points at which subjects are observed
- ▶ d_i : number of deaths at time t_i
- ▶ NB: t_i 's are not event times, this implies that d_i may be zero for some points
- ▶ r_i : number of individuals right-censored at time t_i (subjects withdrawn from the study without experiencing the event at t_i)
- ▶ c_i : number of left-censored observations at time t_i (number for which the only information is that they experienced the event prior to t_i)

- ▶ Use information from left-censored observation (event has occurred at some $t_j \leq t_i$)
- ▶ Estimates the probability that this event occurred at each possible $t_j < t_i$ based on an initial estimate of the survival function
- ▶ Compute an expected number of deaths at t_j (E-step) which is then used to update the estimate of the survival function
- ▶ Repeat the procedure until the estimated survival function stabilizes

- ▶ **Step 0:** $S_0(t_j)$: initial estimate of the survival function at t_j
 - ▶ Turnbull suggests the Product-limit estimate obtained by ignoring the left-censored data as initial value
- ▶ **Step (K+1) 1:** using the current estimate of S , estimate

$$p_{ij} = P(t_{j-1} < X \leq t_j | X \leq t_j)$$

as follows

$$\hat{p}_{ij} = \frac{S_k(t_{j-1}) - S_k(t_j)}{1 - S_k(t_j)}, \quad j \leq i$$

- ▶ **Step (K+1) 2:** use the results of the previous step to estimate the number of events at time t_j :

$$\hat{d}_j = d_j + \sum_{i=j}^m c_i p_{ij}$$

- ▶ **Step (K+1) 3:** Compute the Product-Limit estimator based on the estimated right-censored data with \hat{d}_j events and r_j right-censored observations at t_j by ignoring the left-censored data
 - ▶ if $|S_{K+1}(t) - S_K(t)| \leq \epsilon$ for all t_j stop the procedure otherwise go to step 1
- ▶ Apply the algorithm to example Section 1.17

Example: marijuana data (Section 1.17)

- ▶ In this study, 191 California high school boys were asked: When did you first use marijuana? The answers were
 - ▶ The exact ages (*uncensored observations*);
 - ▶ I never used it: *right-censored observations at the boys' current ages*;
 - ▶ I have used it but can not recall just when the first time was: *left-censored* observation

Example: marijuana data (Section 1.17)

► Data table 1.8 page 17

```
> mar
```

	<i>Age</i>	<i>N.ExactOb</i>	<i>N.YetToSmoke</i>	<i>N.StartedToSmoke</i>
1	10	4	0	0
2	11	12	0	0
3	12	19	2	0
4	13	24	15	1
5	14	20	24	2
6	15	13	18	3
7	16	3	14	2
8	17	1	6	3
9	18	0	0	1
10	19	4	0	0

- ▶ The initial Product-Limit estimator S_0 is obtained by ignoring the left-censored observations
- ▶ We need the following quantities for each time t_i :
 1. Number Left-Censored: c_i
 2. Number of events: d_i
 3. Number Right-Censored: r_i
 4. Number at risk: $Y_i = \sum_{j=1}^m (d_j + r_j)$
 5. compute the product limit estimator $S_0(t_i)$

► Reconstruct table 5.1 page 142

```

> # Age
> ti<-c(0,mar$Age)
> # Number left censored
> ci<-c(0,mar$N.StartedToSmoke)
> # Number of events
> di<-c(0,mar$N.ExactOb)
> # Number of right censored
> ri<-c(0,mar$N.YetToSmoke)
> n<-length(ti)
>
> ti
[1] 0 10 11 12 13 14 15 16 17 18 19
> ci
[1] 0 0 0 0 1 2 3 2 3 1 0
> di
[1] 0 4 12 19 24 20 13 3 1 0 4
> ri
[1] 0 0 0 2 15 24 18 14 6 0 0
>

```


Example: marijuana data (Section 1.17)

```

> Yi<-numeric(11)
> Si<-numeric(11)
> Yi[1]<-0
> # number at risk in time t1=1
> Yi[2]<-sum(di)+sum(ri)
> Yi
[1] 0 179 0 0 0 0 0 0 0 0 0

> Si[1]<-1
> Si[2]<-1-di[2]/Yi[2]
> Si
[1] 1.0000000 0.9776536 0.0000000 0.0000000 0.0000000
[6] 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
[11] 0.0000000

> for(j in 2:(n-1))
+ {
+ # keep track of the number at risk; rj: number right censored
+ Yi[j+1]<-sum(di)-sum(di[1:j])+sum(ri)-sum(ri[1:j])
+ Si[j+1]<-Si[j]*(1-di[j+1]/Yi[j+1])
+ }
> table5.1<-cbind(ti,ci,di,ri,Yi,Si)
> table5.1<-data.frame(table5.1)

```

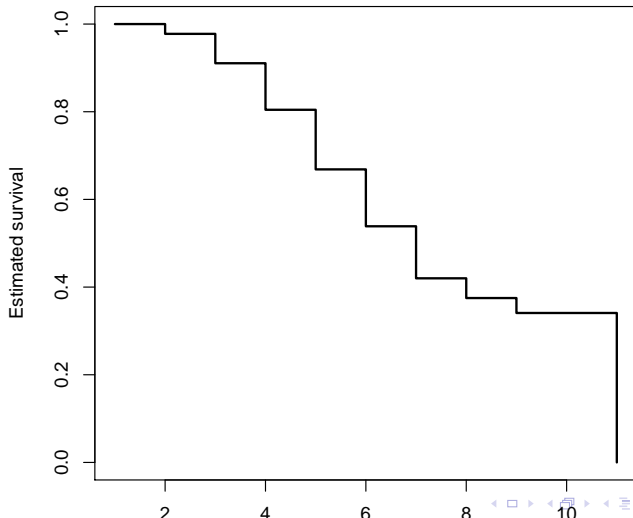
Example: marijuana data (Section 1.17)

```

> table5.1
  ti ci di ri  Yi      Si
1   0  0  0  0   0 1.0000000
2  10  0  4  0 179 0.9776536
3  11  0 12  0 175 0.9106145
4  12  0 19  2 163 0.8044693
5  13  1 24 15 142 0.6685026
6  14  2 20 24 103 0.5386963
7  15  3 13 18  59 0.4200005
8  16  2  3 14  28 0.3750005
9  17  3  1  6  11 0.3409095
10 18  1  0  0   4 0.3409095
11 19  0  4  0   4 0.0000000
> sum(table5.1$di)
[1] 100
> sum(table5.1$ri)
[1] 79

> # plot survival S(t)
plot(table5.1$Si, type="s", xlab="Age", ylab="Estimated survival"
+ lwd=2, main="Estimated S(t) by ignoring the left-censored
+ observations")

```

Estimated $S(t)$ by ignoring the left-censored observations

Example: marijuana data (Section 1.17)

- ▶ Use the table obtained to estimate $p_{ij} = P(t_{j-1} < X \leq t_j | X \leq t_i)$ as $\hat{p}_{ij} = \frac{S_k(t_{j-1}) - S_k(t_j)}{1 - S_k(t_i)}$
- ▶ We need to estimate only for i such that $c_i > 0$ (c_i : left-censored observation)
- ▶ For the left-censored observation at time t_4 we have (see table 5.1 first column next slide)

$$p_{41} = \frac{1 - 0.978}{1 - 0.669} = 0.067; \quad p_{42} = \frac{0.978 - 0.911}{1 - 0.669} = 0.202$$

$$p_{43} = \frac{0.911 - 0.804}{1 - 0.669} = 0.320; \quad p_{44} = \frac{0.804 - 0.669}{1 - 0.669} = 0.410$$

- ▶ Perform similar computations to estimate values of p_{ij}

Example: marijuana data (Section 1.17)

```
> table5.1
```

	<i>ti</i>	<i>ci</i>	<i>di</i>	<i>ri</i>	<i>Yi</i>	<i>Si</i>
1	0	0	0	0	0	1.0000000
2	10	0	4	0	179	0.9776536
3	11	0	12	0	175	0.9106145
4	12	0	19	2	163	0.8044693
5	13	1	24	15	142	0.6685026
6	14	2	20	24	103	0.5386963
7	15	3	13	18	59	0.4200005
8	16	2	3	14	28	0.3750005
9	17	3	1	6	11	0.3409095
10	18	1	0	0	4	0.3409095
11	19	0	4	0	4	0.0000000

Example: marijuana data (Section 1.17)

```
> # Find position left censored observations
> val <- which(ci>0)-1
> val
[1] 4 5 6 7 8 9
> table5.2<-matrix(rep(0,9*length(val)),nrow=9, ncol=length(val))
> table5.2<-data.frame(table5.2)
> colnames(table5.2)<-val
```

► estimate p_{ij} by $\hat{p}_{ij} = \frac{S_k(t_{j-1}) - S_k(t_j)}{1 - S_k(t_i)}$ for $j \leq i$

```
> for(j in val)
+ {
+ i<-which(val==j)
+ for(k in 1:j)
+ {
+ # estimate p_{ij}
+ table5.2[k,i]<-(Si[k]-Si[k+1])/(1-Si[j+1])
+ }
+ }
```

Example: marijuana data (Section 1.17)

```

> table5.2
      4      5      6      7      8      9
1 0.0674104 0.04844177 0.03852826 0.03575422 0.03390486 0.03390486
2 0.2022312 0.14532532 0.11558477 0.10726265 0.10171457 0.10171457
3 0.3201994 0.23009842 0.18300921 0.16983252 0.16104807 0.16104807
4 0.4101590 0.29474430 0.23442544 0.21754678 0.20629434 0.20629434
5 0.0000000 0.28139019 0.22380422 0.20769029 0.19694767 0.19694767
6 0.0000000 0.00000000 0.20464810 0.18991341 0.18009028 0.18009028
7 0.0000000 0.00000000 0.00000000 0.07200014 0.06827599 0.06827599
8 0.0000000 0.00000000 0.00000000 0.00000000 0.05172423 0.05172423
9 0.0000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000

```

- ▶ Column 1: values of p_{ij} for $i = 4$ and $j = 1, \dots, 4$ ($j \leq i$)
($p_{41}, p_{42}, p_{43}, p_{44}$)
- ▶ ⋮
- ▶ Column 9: values of p_{ij} for $i = 9$ and $j = 1, \dots, 9$ ($j \leq i$)
($p_{91}, p_{92}, \dots, p_{99}$)

Example: marijuana data (Section 1.17)

- ▶ Using the result from Step 1 estimate number of events at time t_j by: $\hat{d}_j = d_j + \sum_{i=j}^m c_i p_{ij}$

$$\hat{d}_1 = 4 + 0.067 \times 1 + 0.048 \times 2 + 0.039 \times 3 + 0.036 \times 2 + 0.034 \times 3 + 0.034 \times 1 = 4.487 = d_1 + \sum_{i=4}^9 p_{i1} c_i$$

- ▶ Values p_{ij} are found in Table 5.2; values c_i are the number of left observations given in the data
- ▶ Use these values to compute the updated estimate of the survival function $S_1(t)$
- ▶ Repeat the procedure until difference is small

Example: marijuana data (Section 1.17)

► Estimate $\hat{d}_j = d_j + \sum_{i=1}^m c_i p_{ij}$

```
> tj<-ti # age
> dj<-di # number of events
> rj<-ri # number right censored
> # find c_i
> cj<-ci[val+1]
>
> for(j in 2:(n-1))
+ { # estimate d_j
+ dj[j]<-dj[j]+sum(cj*table5.2[j-1,])
+ }
> dj
[1] 0.000000 4.487007 13.461020 21.313281 26.963195
[6] 22.437364 14.714132 3.417104 1.206897 0.000000
[11] 4.000000
```

Example: marijuana data (Section 1.17)

```

> Yj<-numeric(11)
> Sj<-numeric(11)
>
> Yj[1]<-0
> Yj[2]<-sum(dj)+sum(rj)
> Sj[1]<-1
> Sj[2]<-1-dj[2]/Yj[2]
> for(j in 2:(n-1))
+ {
+ # keep track of the number at risk; dj: events at time tj
+ # obtained with formula given in Step 2 of the algorithm;
+ # rj: number right censored at time tj (column 5 table 5.1);
+ Yj[j+1]<-sum(dj)-sum(dj[1:j])+sum(rj)-sum(rj[1:j])
+ # estimate S(t) with product limit estimator
+ Sj[j+1]<-Sj[j]*(1-dj[j+1]/Yj[j+1])
+ }
>

```

- Crucial ingredients: the **risk set** Y_i and the **number of events** d_i at every time point t_i

Example: marijuana data (Section 1.17)

- ▶ The estimated survival $\hat{S}(t)$ is computed with the usual product limit estimator $\prod_{t_i \leq t} (1 - d_i / Y_i)$ based on the estimated right-censored data with: \hat{d}_i events; r_i : right-censored observations at time t_i
- ▶ The computations are done by **ignoring** the left-censored data
- ▶ The values $\hat{d}_j = d_j + \sum_{i=j}^m c_i p_{ij}$ are then used in the code to compute the updated estimate of the survival function $S_1(t)$
- ▶ If this estimate, $\hat{S}_1(t)$ is close to $\hat{S}_0(t)$ for all t_i , stop the procedure; if not, go to step 1.

Example: marijuana data (Section 1.17)

```

> table5.3<-cbind(tj,dj,rj,Yj,Sj)
> table5.3<-data.frame(table5.3)
> table5.3

```

	tj	dj	rj	Yj	Sj
1	0	0.000000	0	0.000000	1.0000000
2	10	4.487007	0	191.00000	0.9765078
3	11	13.461020	0	186.51299	0.9060313
4	12	21.313281	2	173.05197	0.7944434
5	13	26.963195	15	149.73869	0.6513893
6	14	22.437364	24	107.77550	0.5157791
7	15	14.714132	18	61.33813	0.3920511
8	16	3.417104	14	28.62400	0.3452485
9	17	1.206897	6	11.20690	0.3080679
10	18	0.000000	0	4.00000	0.3080679
11	19	4.000000	0	4.00000	0.0000000

Example: marijuana data (Section 1.17)

- Plot estimated survival $S_0(t)$ obtained in Table 5.1 together with estimated $S_1(t)$ from table 5.3

```
> plot(table5.1$Si, type="s", xlab="Age", ylab="Estimated survival",
+      lwd=2, col="red")
> lines(table5.3$Sj, type="s", col="blue", lwd=2)
> legend("bottomleft", c("Estimated initial S(t)", "Estimated S(t)
+ first step"), lwd=2, col=c("blue", "red")) #lty=1:2)
```

