제 7 강. 생존분석1

이재원 교수
고려대학교 통계학과

주요내용

1. 생존분석의 기초개념
   - 생존시간, 중도절단, 생존함수, 위험함수

2. 비모수적 방법을 이용한 생존함수의 추정
   - 생명표방법
   - 누적한계추정법

생존시간의 개념

◆ 생존시간

: 정의된 시점부터 특정한 사건(event)이 관측될 때까지의 시간

환자의 사망, 어떤 처리에 대한 반응이나 질병의 재발 등

예) 전구나 기계의 수명, 로봇이 미로에서 빠져 나올 때까지의 시간, 회사가 도산할 때까지의 시간, 강의법을 평가할 때 문제 풀 때까지의 시간, 결혼기간, 회사 복무기간 등.

생존분석에서의 관심문제

- 생존시간의 분포에 관한 정보
  (예) 특정시점에서의 생존율, 환자의 50%가 생존하는 시간(median survival time)
  - 실험군과 대조군이 있을 경우 두 처리군의 생존분포의 비교
- 예후인자들의 영향 파악
  (예) 나이, 병력, 백혈구 수 등

중도절단 자료

- 암 환자의 생존시간을 관측하는 임상시험에서 몇몇 환자들이 연구종료 시까지 계속 생존하면 정확한 생존시간을 알 수 없음

- 암 연구에서 환자가 연구대상인 질병인 암 외에 교통사고 등의 다른 원인으로 사망하게 된 경우에 이를 정확한 생존시간이라 할 수 없음

중도절단 자료

중도절단의 형태

1.중도절단의 제 1 형태(Type I censoring)

: 연구자에 의해 중도절단 시간이 고정됨

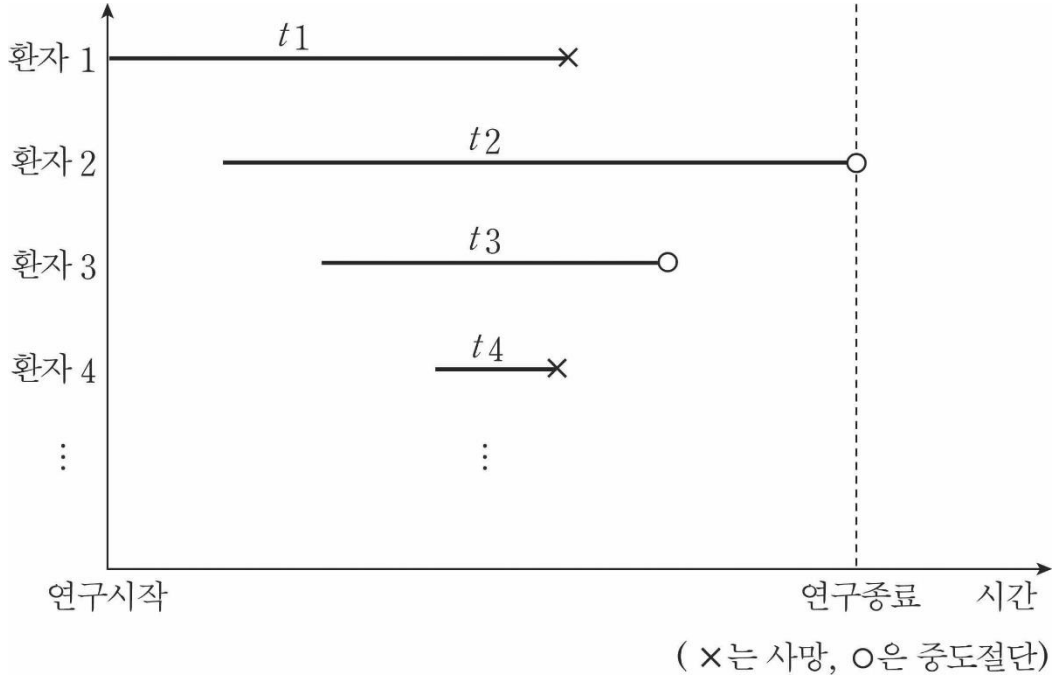
2.중도절단의 제 2 형태(Type II censoring)

: 사전에 관측할 사망자수를 정해 놓는 경우에 발생

3.임의중도절단(random censoring)

: 연구자가 제어하지 못하는 이유로 중도절단 되는 경우

임의중도절단 자료의 예 <그림 6.1>



(×는 사망, ○은 중도절단)

임의 중도 절단된 자료의 경우

가 정

$T_1, T_2, \cdots, T_n \overset{iid}{\sim} F$

$C_1, C_2, \cdots, C_n \overset{iid}{\sim} G$

T_i 와 C_i 는 서로 독립.

$X_i = \min(T_i, C_i)$

$\delta_i = 1,$ T_i가 중도절단되지 않은 경우

$= 0,$ T_i가 중도절단된 경우

로 정의되는 $(X_1, \delta_1), (X_2, \delta_2), \cdots, (X_n, \delta_n)$을 대신 관측

생존함수와 위험함수

- 생존시간 ➡ $T(\geq 0), \qquad T \sim f(t)$
- T의 생존함수(survival function)

$$S(t) = P(T > t) = 1 - F(t)$$

- T의 위험함수(hazard function)

$$h(t) = \lim_{dt \to 0} p(t < T \leq t + dt \mid T \geq t) \; / \; dt$$
$$= f(t) \; / \; S(t) \qquad (6.1)$$

생존함수의 추정

- 비모수적 방법
- 모수적 방법
- 준모수적 방법

비모수적 방법을 이용한 생존함수의 추정

비모수적 방법을 이용한 생존함수의 추정

- 중도절단이 없는 경우

생존시간 $x_1, x_2, \cdots, x_n$을 순서대로 배열

➡ $t_1 < t_2 < \cdots < t_n$

- 경험적 분포함수(empirical distribution function)

$$S(t_i) = P(T > t_i) \qquad (6.2)$$

$$= (\text{생존시간이 } t_i \text{ 이상인 관측치수})/\text{총 관측치수}$$

$$= 1 - (t_i \text{까지의 사망자수})/n$$

비모수적 방법을 이용한 생존함수의 추정

- 중도절단이 있는 경우

➡ 생명표(life table)를 이용한 방법

: 자료를 시간단위로 묶어서 생존함수를 추정

➡ 누적한계추정법(product-limit method)

: 환자 개개인의 생존(또는 중도절단)시간을 관측하여 각 시점에서의 생존함수를 추정

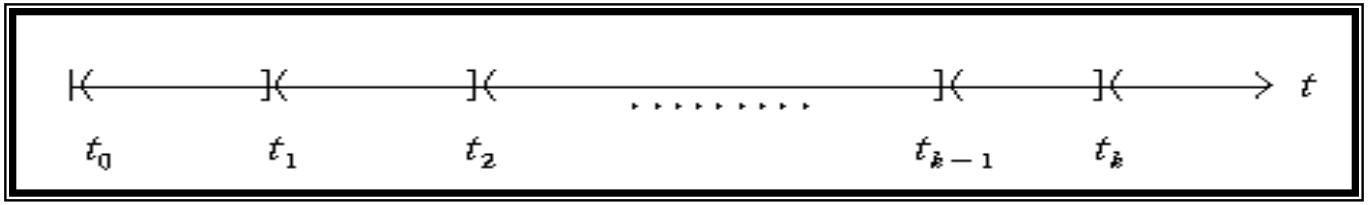
생명표 방법

< 질문>

◆ 환자 100명이 임상시험에 들어왔는데 3년 생존율을 알고 싶다. 3년 후에 처음으로 살펴본다고 가정하자.
3년 후에 살펴보니 20명은 죽었고, 20명은 중도절단이 되었고, 60명은 생존해있었다.

- 이런 경우에 3년 생존율은 얼마인가?

생명표 방법
(Life-table method)



$t_0 = 0,\ t_{k+1} = \infty$

n_i : i 번째 구간의 시작점에서의 생존자수

d_i : i 번째 구간에서의 사망자수

c_i : i 번째 구간에서 중도절단된 인원수

$p_i = P(i$ 번째 구간에서 내내 생존 $|\ i$ 번째 구간의 시작점에서 생존$)$

$= P(T > t_i \mid T > t_{i-1})$: 구간 내 생존율

$q_i = 1 - p_i$: 구간 내 사망률

생명표의 형태 <표 6.1>

구간	사망자수	중도절단수	위험그룹 인원수	유효인원수	사망률	생존율	생존함수
$(t_0, t_1]$	d_1	c_1	n_1	n_1'	$\hat{q}_1$	$\hat{p}_1$	$\hat{S}(t_1)$
$(t_1, t_2]$	d_2	c_2	n_2	n_2'	$\hat{q}_2$	$\hat{p}_2$	$\hat{S}(t_2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$(t_i, t_{i+1}]$	d_i	c_i	n_i	n_i'	$\hat{q}_i$	$\hat{p}_i$	$\hat{S}(t_i)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

생명표의 형태<표 6.1>

- 위험그룹 인원수 : $n_i = n_{i-1} - d_{i-1} - c_{i-1}$
- 유효 인원수 : $n_i' = n_i - c_i / 2$

→ 중도절단 사례가 발생한 경우에는
그들이 위험에 절반 가량 노출되었다고
생각하여 이와 같이 인원수를 조정

- i번째 구간 사망률 q_i 의 추정치: $\hat{q}_i = d_i / n_i'$
- i번째 구간 생존율 p_i 의 추정치: $\hat{p}_i = 1 - \hat{q}_i = 1 - d_i / n_i'$

생존함수

$$S(t_i) = P(T > t_i)$$
$$= P(T > t_1)P(T > t_2 \mid T > t_1)\cdots P(T > t_i \mid T > t_{i-1}) \qquad (6.3)$$
$$= p_1 p_2 \cdots p_i$$

$$\hat{S}(t_i) = \prod_{k=1}^{i} \hat{p}_k = \prod_{k=1}^{i} (1 - d_k / n_k') \qquad (6.4)$$

$$\hat{Var}(\hat{S}(t_i)) = \hat{S}^2(t_i) \sum_{k=1}^{i} \frac{d_k}{n_k'(n_k' - d_k)} \qquad (6.5)$$

Greenwood의 공식

<예제 6.1>

진단 후 경과기간 (단위:년)	사망자수	중도절단수
(0 - 1]	456	0
(1 - 2]	226	39
(2 - 3]	152	22
⋮	⋮	⋮
(13 - 14]	9	27
(14 - 15]	6	23
(15 -	0	30

1년: 위험그룹 인원수 = 유효인원수 = 2418

$$\text{생존율} = 1 - \frac{456}{2418} = 0.8114$$

1년~2년: 위험그룹 인원수 = 2418 - 456 = 1962

$$\text{유효인원수} = 1962 - \frac{39}{2} = 1942.5$$

$$\text{사망률} = \frac{226}{1942.5} = 0.1163$$

생존율 = 1 - 0.1163 = 0.8836

환자가 2년 이상 생존할 확률 $= 0.8114 \times 0.8836 = 0.7170$

R Program

```
library(KMsurv)
library(reshape)
협심증환자자료=read.table("e:₩₩WORK₩₩협심증환자자료.txt",header=T)
head(협심증환자자료)

# 생존자료 정리
생존자료 = cast(협심증환자자료, time ~ delta)
colnames(생존자료) <- c("시간","중도절단","사망")
생존자료$시간 <- floor(생존자료$시간)
생존자료
```

- cast 함수 (reshape 패키지) : 자료의 형태를 <표 6.2>와 같게 변경. 시간을 행 변수로, 중도절단 여부를 열 변수로 가지고, 각 칸에 해당 인원수를 채움.

R Program (계속)

협심증 환자 자료 형태

time	censor	freq
0.5	1	456
0.5	0	0
1.5	1	226
1.5	0	39
...		
15.5	1	0
15.5	0	30

time : $(t_{k-1}+t_k)/2$
censor : 1=사건발생, 0=중도절단
freq : 해당 구간의 사건발생 수
또는 중도절단 수

사망

time	censor	freq
0.5	1	456
1.5	1	226
...		
15.5	1	0

중도절단

time	censor	freq
0.5	0	0
1.5	0	39
...		
15.5	0	30

R Program (계속)

```
# 생명표 작성
attach(생존자료)
생명표 = lifetab(c(시간,NA), sum(c(중도절단,사망)), 중도절단, 사망)
생명표
```

- **lifetab 함수 (KMsurv 패키지)** : 생명표를 생성하는 함수. 함수에 차례대로 tis, ninit, nlost, nevent를 입력한다. lifetab(tis, ninit, nlost, nevent)
  - tis : 구간의 종료점을 입력한 벡터, 첫 성분으로 t_0가 입력되므로, nlost, nevent 벡터보다 길이가 1만큼 크다
  - ninit : 각 구간의 위험그룹 인원수를 입력한 벡터
  - nlost : 각 구간에서 중도절단된 인원수를 입력한 벡터
  - nevent : 각 구간에서 사건이 발생한 인원수를 입력한 벡터

R Program (계속)

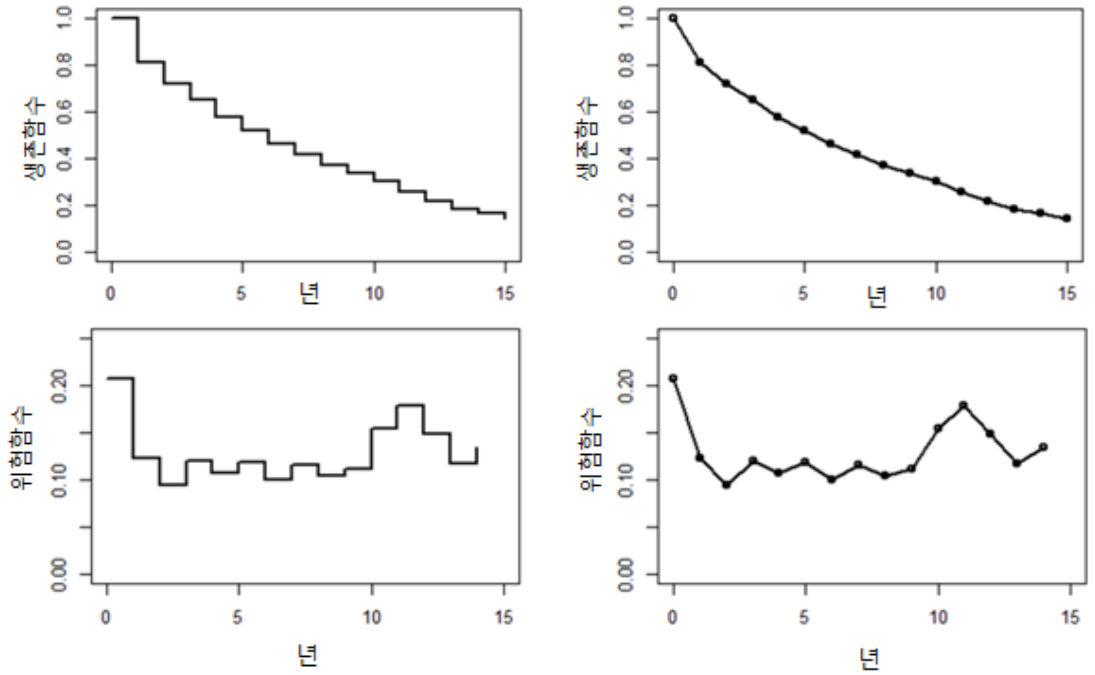
```
# 생명표를 이용한 생존함수 추정값 그림
par(mfrow=c(1,2))
종료시점=length(시간)
plot(시간, 생명표[,5], type="s",xlab="년",ylab="생존함수", ylim=c(0,1))
plot(시간, 생명표[,5], type="o",xlab="년",ylab="생존함수", ylim=c(0,1))

# 생명표를 이용한 위험함수 추정값 그림
plot(시간[-종료시점],생명표[-종료시점,7],type="s",xlab="년",ylab="위험함수"
, xlim=c(0,15),ylim=c(0,0.25))
plot(시간[-종료시점],생명표[-종료시점,7],type="o",xlab="년",ylab="위험함수"
, xlim=c(0,15),ylim=c(0,0.25))
```

- 출력되는 생명표의 5번째 열(surv)은 각 구간의 생존함수의 추정치이며, 7번째 열(hazard)은 각 구간의 위험함수의 추정치이므로, plot함수의 x축에 시간을 입력하고, y축에 생명표[,5], 생명표[,7]를 입력하면 각각 생존함수 그래프와 위험함수 그래프를 출력할 수 있다.

협심증 자료의 생명표
R 출력결과

```
> 생명표
      nsubs nlost  nrisk nevent      surv        pdf     hazard     se.surv      se.pdf   se.hazard
0-1    2418     0 2418.0    456 1.0000000 0.18858561 0.20821918 0.000000000 0.007955134 0.009697769
1-2    1962    39 1942.5    226 0.8114144 0.09440394 0.12353102 0.007955134 0.005975178 0.008201472
2-3    1697    22 1686.0    152 0.7170105 0.06464151 0.09440994 0.009179397 0.005069200 0.007649121
3-4    1523    23 1511.5    171 0.6523689 0.07380423 0.11991585 0.009734736 0.005428013 0.009153696
4-5    1329    24 1317.0    135 0.5785647 0.05930618 0.10804322 0.010138361 0.004945997 0.009285301
5-6    1170   107 1116.5    125 0.5192585 0.05813463 0.11859583 0.010304216 0.005033980 0.010588867
6-7     938   133  871.5     83 0.4611239 0.04391656 0.10000000 0.010379949 0.004690538 0.010962697
7-8     722   102  671.0     74 0.4172073 0.04601094 0.11671924 0.010450930 0.005175094 0.013545211
8-9     546    68  512.0     51 0.3711964 0.03697464 0.10483042 0.010578887 0.005024599 0.014659017
9-10    427    64  395.0     42 0.3342218 0.03553750 0.11229947 0.010717477 0.005307615 0.017300846
10-11   321    45  298.5     43 0.2986843 0.04302654 0.15523466 0.010890741 0.006269963 0.023601647
11-12   233    53  206.5     34 0.2556577 0.04209376 0.17941953 0.011124244 0.006847514 0.030646128
12-13   146    33  129.5     18 0.2135639 0.02968456 0.14937759 0.011396799 0.006682743 0.035110295
13-14    95    27   81.5      9 0.1838794 0.02030570 0.11688312 0.011765989 0.006514794 0.038894448
14-15    59    23   47.5      6 0.1635737 0.02066194 0.13483146 0.012259921 0.008035120 0.054919485
15-NA    30    30   15.0      0 0.1429117         NA         NA 0.013300258          NA          NA
```

생존함수 그래프와 위험함수 그래프 <그림 6.2 ~ 6.3>



누적한계추정법

생존함수의 누적한계추정치

생존시간 또는 중도절단 $x_1, x_2, \cdots, x_n$ 을 순서대로 배열 ➡ $t_1 < t_2 < \cdots < t_n$

δ_i : 생존시간 t_i 의 중도절단 여부

(중도절단인 경우 $\delta_i = 0$, 그렇지 않은 경우 $\delta_i = 1$)

생존함수 $S(t)$ 의 누적한계추정치 :

$$\hat{S}(t) = \prod_{i:t_i \le t} \left(1 - \frac{d_i}{n_i}\right)^{\delta_i} \qquad (6.6)$$

예제 <6.2>

<표 6.3> 신장이식 환자들의 호전기간 (단위 : 일)

3.0	4.0+	4.5	4.5	5.5	6.0	6.4	6.5
7.0	7.5	8.4+	10.0	10.0 +	12.0	15.0	

신장이식수술을 받은 15명 환자들의 호전기간(remission duration)이 <표 6.3>과 같다고 하자. 여기서 +로 표시된 것은 중도절단된 자료를 가리킨다. 각 시점에서의 생존함수를 누적한계추정법을 이용하여 추정해 보자.

예제 <6.2>

- $S(3.0) = P(t > 3.0) = p_1$, $\hat{p}_1 = 1 - \frac{1}{15} = 0.933$
- $\hat{S}(4.0) = 0.933 \times (1 - \frac{0}{14}) = 0.933$ (∵사망자수 0명)
- $\hat{S}(4.5) = 0.933 \times (1 - \frac{0}{14}) \times (1 - \frac{2}{13})$

$= 0.933 \times 0.846 = 0.789$

생존함수의 누적한계추정치

$$\hat{S}(t) = \prod_{i:t_i \le t} \left(1 - \frac{d_i}{n_i}\right)^{\delta_i} \qquad (6.6)$$

$$\hat{Var}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{i:t_i \le t} \frac{d_i}{n_i(n_i - d_i)} \qquad (6.7)$$

-Greenwood의 공식-

R Program

```
library(survival)
신장이식환자자료=read.table("c:₩₩WORK₩₩신장이식환자자료.txt",header=T)
```

- **survival 패키지** : 생존분석을 위한 R 패키지
- 신장이식 환자 자료 형태

time	Status
3.0	1
4.0	0
…	
12.0	1
15.0	1

time : 사건발생 또는 중도절단 시간
status : 1=사건발생, 0=중도절단

R Program (계속)

```
# 누적한계추정값 (Kaplan-Meier 추정값)
적합1=survfit(Surv(time,status)~1,conf.type="log-log",data=신장이식환자자료)
summary(적합1)

# 사망시점의 사분위수 추정값와 그의 신뢰구간
quantile(적합1,probs=c(0.25, 0.5, 0.75), conf.int=T)

# 누적한계추정값와 95% 신뢰구간 그래프
plot(적합1,xlab="시간",ylab="생존함수",mark.time=T)
legend(0.5,0.2,c("누적한계추정값","95% 신뢰구간"),lty=c(1,2))
```

- **survfit 함수 (survival 패키지)** : ~ 좌변에 Surv(시간, 절단상태)를 입력하고 ~ 우변에는 생존시간에 영향을 미치는 공변량을 고려할 경우 공변량들을 입력하고, 공변량을 고려하지 않을 경우에는 1을 입력한다.
- **quantile** : 입력된 확률에 대한 분위수를 계산. conf.int = T라고 입력하면 해당 분위수에 대한 신뢰구간을 제공한다.

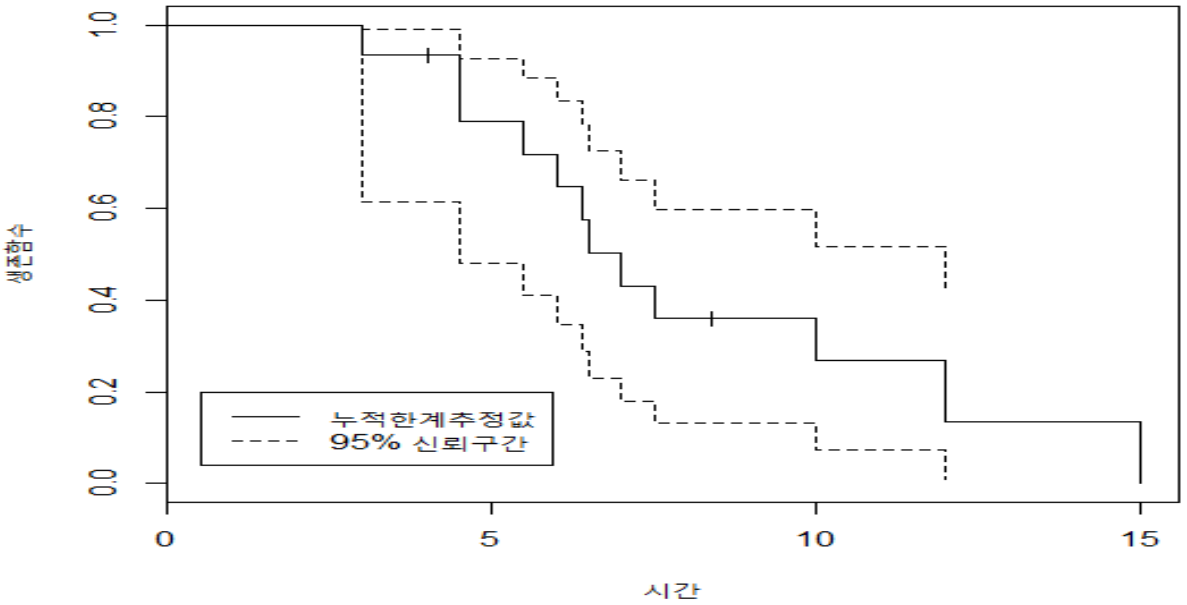
생존함수 추정을 위한 R 출력결과

```
> 적합1=survfit(Surv(time,status)~1,conf.type="log-log",data=신장이식환자자료)
> summary(적합1)
Call: survfit(formula = Surv(time, status) ~ 1, data = 신장이식환자자료,
    conf.type = "log-log")

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  3.0     15       1    0.933  0.0644       0.6126         0.990
  4.5     13       2    0.790  0.1081       0.4791         0.927
  5.5     11       1    0.718  0.1198       0.4111         0.884
  6.0     10       1    0.646  0.1275       0.3468         0.835
  6.4      9       1    0.574  0.1320       0.2866         0.782
  6.5      8       1    0.503  0.1336       0.2305         0.724
  7.0      7       1    0.431  0.1324       0.1786         0.662
  7.5      6       1    0.359  0.1283       0.1313         0.596
 10.0      4       1    0.269  0.1237       0.0738         0.517
 12.0      2       1    0.135  0.1135       0.0103         0.415
 15.0      1       1    0.000     NaN           NA            NA
```

생존함수 추정을 위한 R 출력결과 (계속)

```
> # 사망시점의 사분위수 추정값와 그의 신뢰구간
> quantile(적합1,probs=c(0.25, 0.5, 0.75),conf.int=T)
$quantile
  25  50  75
 5.5 7.0 12.0

$lower
 25  50  75
3.0 4.5 6.5

$upper
  25  50  75
 6.5 12.0  NA
```

생존함수추정치의 그래프
<그림 6.4>



핵심정리

1. 생존분석의 기초개념
   - 생존시간, 중도절단, 생존함수, 위험함수

2. 비모수적 방법을 이용한 생존함수의 추정
   - 생명표 방법
   - 누적한계추정법

제7강

수고하셨습니다!