5장 비모수 방법에 의한 생존함수 추정

2020년 가을학기

전북대학교 통계학과

Outline

생명표

Kaplan-Meier 누적한계추정량

Nelson-Aalen 누적위험함수추정량

- 생존분석의 주요 관심문제: 생존함수 추정$\Rightarrow$ 생존률과 위험률 정보 얻어냄
- 생존데이터에 대한 가정
    - 개개의 생존시간은 독립
    - 중도절단은 생존시간과 서로 독립

생명표

생명표

생존데이터를 그룹지어 즉 어떤 구간 내에 발생한 사건의 개수 기록한 경우 - 생명표 (life table) 방법 활용

예제 5.1

[표 5.1] 심근경색증 데이터

연 단위 구간 year of entry	구간초기 위험집합 number of alive	사건 발생수 (사망 환자수) number of dying during interval	중도절단된 수 number of censored during interval
[0,1)	146	27	3
[1,2)	116	18	10
[2,3)	88	21	10
[3,4)	57	9	3
[4,5)	45	1	3
[5,6)	41	2	11
[6,7)	28	3	5
[7,8)	20	1	8
[8,9)	11	2	1
[9,10)	8	2	6
합		86	60

5년 이상 생존확률 $P(T \geq 5)$

- 가장 간단한 추정량 (과대추정경향)
- 해당구간에서만 중도절단된 경우 제외0 (과소추정경향)
- 중도절단된 모든 경우 제외 (과소추정경향)

위험집합 (nmber at risk) 정하는 바에 따라 추정량이 달라짐 $\Rightarrow$ 생명표 방법

5년 이상 생존확률 $S(5)$

$$\begin{aligned} S(5) &= P(T \geq 5) = P(T \geq 5, T \geq 4) = P(T \geq 4)P(T \geq 5 | T \geq 4) \\ &= P(T \geq 4)[1 - P(4 \leq T < 5 | T \geq 4)] = P(T \geq 4)q_5 \\ &= P(T \geq 3)P(T \geq 4 | T \geq 3)q_5 = q_1 q_2 q_3 q_4 q_5 \end{aligned}$$

여기서 $q_i = 1 - P((i-1) \leq T < i | T \geq i-1)$

- $m_i = 1 - q_i$: $t = i - 1$시점에서의 치사율 (mortality)
- $Y_i - c_i/2$: 유효인원수(effective sample size)
- m_i추정: $\hat{m}_i = \dfrac{d_i}{Y_i - c_i/2}$

생명표

생명표가 $K+1$개의 구간 $[t_{i-1}, t_i),\ i = 1, 2, \ldots K+1$으로 나 누어져 정리된 경우

$$S(t_i) = P(T \geq t_i) = P(T \geq t_{i-1})P(T \geq t_i | T \geq t_{i-1}) = \cdots = q_1 q_2 \cdots q_i$$

$q_i = 1 - P(t_{i-1} \leq T < t_i | T \geq t_{i-1})$

생명표 방법 추정량 $\hat{S}(t_i)$

$$\begin{aligned}\hat{S}(t_i) &= \prod_{k=1}^{i} \hat{q}_k = \prod_{k=1}^{i} (1 - \hat{m}_k) \\ &= \prod_{k=1}^{i} \left(1 - \frac{d_k}{Y_k - c_k/2}\right)\end{aligned}$$

- $\hat{S}(0) = 1$
- 추정된 생존함수의 분산

$$\widehat{Var}(\hat{S}(t_i)) = \hat{S}^2(t_i) \sum_{k=1}^{i} \frac{d_k}{(Y_k - c_k/2)(Y_k - c_k/2 - d_k)}$$

Kaplan-Meier 누적한계추정량

중도절단자료가 없는 경우

- 생존시간 $\{T_i, i = 1, 2, \ldots, n\}$
  - 동점이 없는 경우 : 관측값의 순서통계량 $t_1 < t_2 < \cdots < t_n$
  - 동점이 발생한 경우 (D개의 구별되는 값): $t_1 < t_2 < \cdots < t_D$, 사건발생시점 t_i 에서 d_i개 사건 발생

생존함수 추정량

$$\hat{S}(t) = \hat{P}(T > t) = 1 - F_n(t) = \frac{\text{number of } (t_i > t)}{n}$$

$\hat{S}(t)$: 우연속 계단함수

중도절단자료가 있는 경우

- 생존데이터 $\{(T_i, \delta_i), i = 1, 2, \ldots, n\}$: $T_i = \min(\tilde{T}_i, C_i), \delta_i = I(\tilde{T}_i < C_i)$
  - $\tilde{T}_1, \tilde{T}_2, \ldots, \tilde{T}_n \sim F$ 서로 독립
  - $C_1, C_2, \ldots, C_n \sim F_C$ 서로 독립
- d_i : t_i시점에서 발생한 사건 수
- Y_i : t_i시점에서 위험에 놓인 개체수 (risk number)
- $p_i = P\{\text{surviving through } (t_{i-1}, t_i] | \text{ alive at the beginning of } (t_{i-1}, t_i]\} \Rightarrow \hat{p}_i = 1 - \dfrac{d_i}{Y_i}$

t_i시점까지 생존확률 $S(t_i)$

$$
\begin{aligned}
S(t_i) &= P(\tilde{T} > t_i) \\
&= P(\tilde{T} > t_i | \tilde{T} > t_{i-1}) P(\tilde{T} > t_{i-1}) \\
&= p_i \times S(t_{i-1}) \\
&= P(\tilde{T} > t_i | \tilde{T} > t_{i-1}) P(\tilde{T} > t_{i-1} | \tilde{T} > t_{i-2}) \cdots P(\tilde{T} > t_1) \\
&= p_1 p_2 \cdots p_i \Rightarrow \hat{S}(t_i) = \prod_{j : t_j \leq t_i} \left(1 - \frac{d_j}{Y_j}\right)
\end{aligned}
$$

Kaplan-Meier(1958) 추정량

$$\hat{S}(t) = \begin{cases} 1 & t < t_1 \\ \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{Y_i}\right) & t \geq t_1 \end{cases}$$

데이터 범위 내 모든 t 값에서 계산이 가능

Kaplan-Meier 추정량의 특징

- t_n시점 이후에 대해서는 잘 정의되지 않음
- 생존시간 관측값에서만 점프가 일어나는 계단함수
- 점프크기는 사건발생수와 중도절단개수에 의존

Kaplan-Meier 추정량 $\hat{S}(t)$

Kaplan-Meier 추정량의 분산

$$\hat{Var}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)}$$

(Greenwood 공식)

Kaplan-Meier 추정량를 이용한 누적위험함수

$$\hat{H}(t) = -\log \hat{S}(t)$$

t_0시점에서 생존함수에 대한 근사적인 $100 \times (1-\alpha)$% 신뢰구간

$$\left(\hat{S}(t_0) - z_{\alpha/2}\sqrt{\hat{Var}[\hat{S}(t_0)]}, \hat{S}(t_0) + z_{\alpha/2}\sqrt{\hat{Var}[\hat{S}(t_0)]}\right)$$

Nelson-Aalen 누적위험함수추정량

Nelson-Aalen 누적위험함수추정량

개수과정을 활용한 Nelson-Aalen(1969) 추정량

$$\hat{H}(t) = \int_0^t \frac{N(s+u) - N(s)}{Y(s)} ds = \int_0^t \frac{dN((s)}{Y(s)}$$

$t_1 < \cdots < t_D$에 대하여

$$\tilde{H}(t) = \begin{cases} 0 & t < t_1 \\ \sum_{t_i \le t} \frac{d_i}{Y_i} & t \ge t_1 \end{cases}$$

Nelson-Aalen 추정량의 특징

- 관측된 최대값까지 범위 내에서 정의됨
- $\tilde{H}(t) - H(t)$ 거의 마팅게일

Nelson-Aalen 누적위험함수추정량

Nelson-Aalen 추정량의 분산

$$\hat{Var}[\tilde{H}(t)] = \hat{\sigma}_H^2 = \sum_{t_i \le t} \frac{d_i}{Y_i^2}$$

Nelson-Aalen 추정량을 이용한 생존함수 추정

$$\tilde{S}(t) = \exp\{-\tilde{H}(t)\}$$

t_0**시점에서** $H(t)$**에 대한 근사적인** $100 \times (1-\alpha)$**% 신뢰구간**

$$\left(\tilde{H}(t_0) - z_{\alpha/2}\hat{\sigma}_H, \tilde{H}(t_0) + z_{\alpha/2}\hat{\sigma}_H\right)$$

NOTE Nelson-Aalen 추정량과 Kaplan-Meier 추정량은 근사적으로 동등하며 일치성을 가짐