생존분석

b.a.f 김영석

생존분석이란?

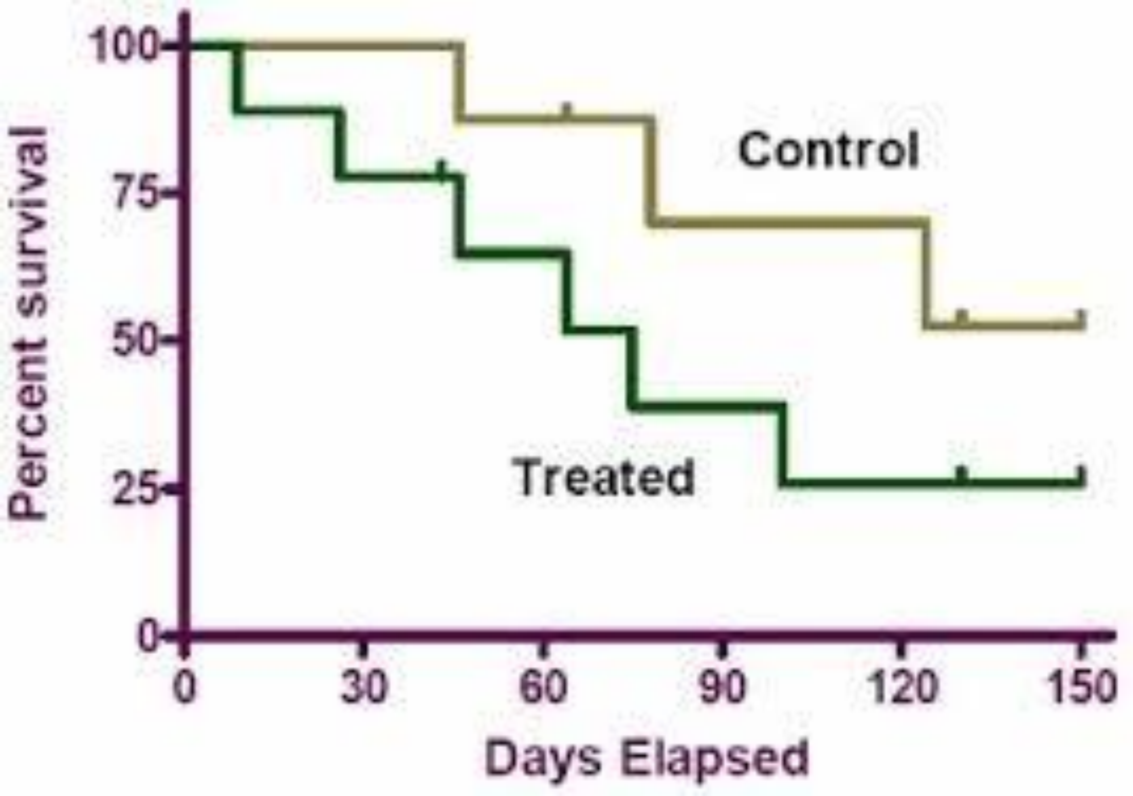
PEEDORI

소개 | 생존함수 | 모형 | 잔차

어떤 연구에 들어온 시간부터 어떤 사건이 발생 할 때 까지의 시간구간 데이터에 관심.

반응변수(Y) : 사건이 발생할 때까지 걸린 시간.

예)
의학: 질병완치까지 호르몬 요법의 비교,
다리 골절이 완치되기까지의 시간
수술 후 생존기간 등
산업 : 기계가 고장 날 때 까지의 시간
사회과학: 결혼의 지속시간,
실업자들의 구직까지 실업기간
범죄학 : 수감자들이 출소 후 재범까지 걸리는 시간
마케팅 : 잡지 구독기간, 고객 이탈 예측
상품 재구매까지 걸리는 시간 등



문제점

1. 동시에 실험에 참여해도 사건이 발생하는데 까지 걸리는 시간이 다르므로 정규분포를 따른다고 가정할 수 없다.

2. 연구기간 내 추적을 실패할 수 있다. 연구가 제한된 기간 내에 이루어 지므로 연구마감 시점에 환자가 살아남은 경우 언제 사건이 발생했는지 알 수없다.

3. 중간에 환자가 병원을 옮기는 등 여러가지 이유로 연락이 안될 수 있다.

→ **중도절단 (CENSORED)**

가정

1. 추적에 실패한 환자도 예후를 갖는다
2. 중도절단과 사건발생은 관련이 없다.

모수

소개 | 생존함수 | 모형 | 잔차

특정 사건 발생까지 시간 : T T의분포를 알고자 한다.
(사건이란 : 사망, 재발, 장비의 고장 등)

중도절단자료로 인해 분포 추정에 있어 평균, 분산 등의 통계량이 아닌
생존함수, 위험함수, 누적위험함수, 평균잔여수명. 이 4개의 통계량을 이용하며 한 개만 알면
나머지는 유일하게 결정 됨.

생존함수

생존함수 : t 시점 이후 사건이 발생할 확률

$$S(t) = P(T > t) = \int_t^\infty f(x)\,dx = 1 - P(T \le t) = 1 - F(t)\,, \qquad S(0) = 1, S(\infty) = 0, t \in [0, \infty)$$

확률이므로 $0 \le S(t) \le 1$

생존함수를 이용, 확률밀도함수 $f(t) = -\dfrac{dS(t)}{dt}$

위험함수

위험함수 : t 시점에서 생존한 조건 하에서 t시점 바로 직후 사건이 발생할 조건부 확률

$$h(t) = \lim_{\Delta t \to 0} P(t \le T < t + \Delta t \mid T \ge t) = \frac{f(t)}{S(t)} = \frac{-S`(t)}{S(t)} = -\frac{d}{dt}\log(S(t)) \ , f(t) = \lim_{x \to 0} p(t \le T < t + x)$$

① $h(t)$ 가 커지면 $0 \le S(t) \le 1$ 이므로 $S(t)$의 값은 작아진다. 즉, 위험함수 값이 커지면 생존시간이 대체로 작아지는 경향이 있다. But, 생존함수 값이 크다고 위험함수 값이 반드시 작진 않다.

② 확률론 표현 되므로 일반적으로 관측이 불가능하다. 데이터에 근거하여 추정해야 한다.

누적위험함수, 평균잔여수명

누적위험함수 : t 시점 까지의 누적 위험률

$$\int_0^t h(u)\ du$$

평균잔여수명 : x 시점까지 생존한 조건에서 x시간 이후 생존 가능한 시간의 기댓값

$$f(x) = E(T - x \mid T > x)$$

생존데이터에 대한 모수적 분포

비모수적 모형 (흔히 사용) - Kaplan-Meier 곡선
- Cox proportional hazard model

모수적 모형 - 지수분포, 와이블분포, 감마분포, 로그-정규분포 등등

=> 생존데이터의 상황을 잘 반영하는 분포 선택

중도절단 (우중도절단, 좌중도절단, 구간중도절단)

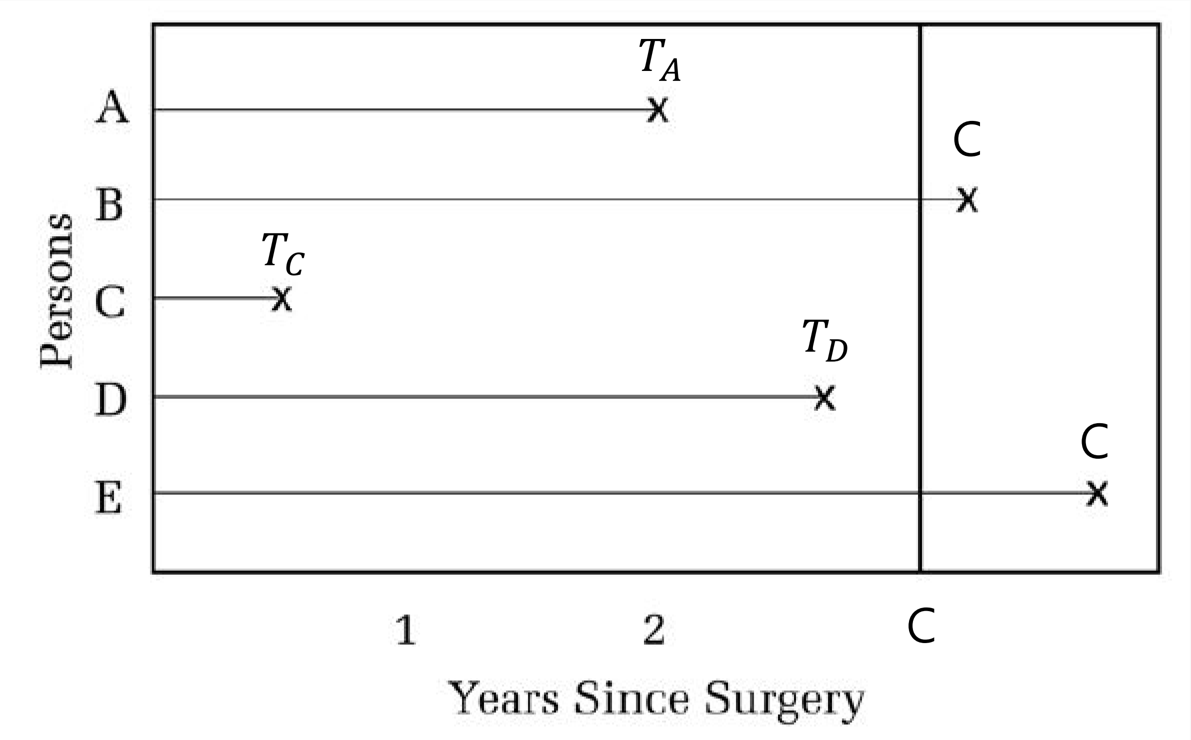
i번째 개체의 생존시간 : T_i , 중도 절단 여부 : δ_i , 관측중단시점 : C_i , 연구종료시점: C , $i = 1,2,\cdots,n$

제 1유형 우중도절단 : 모든 실험 개체에 대한 우중도절단 시간 동일

$$C_1 = C_2 = \cdots = C_n = C$$

중도절단 시점 전에 사건 발생 : T_i
중도절단 시점 후에 사건 발생 : $T_i = C$

예) **장기 이식 수술 후 63개월 생존**
but 연구기간이 60개월, 생존기간은 60+로 기록



중도절단 (우중도절단, 좌중도절단, 구간중도절단)

제 2유형 우중도절단 : 전체 실험 개체들 중 미리 정해놓은 시점 발생률 까지 관측 후 중지

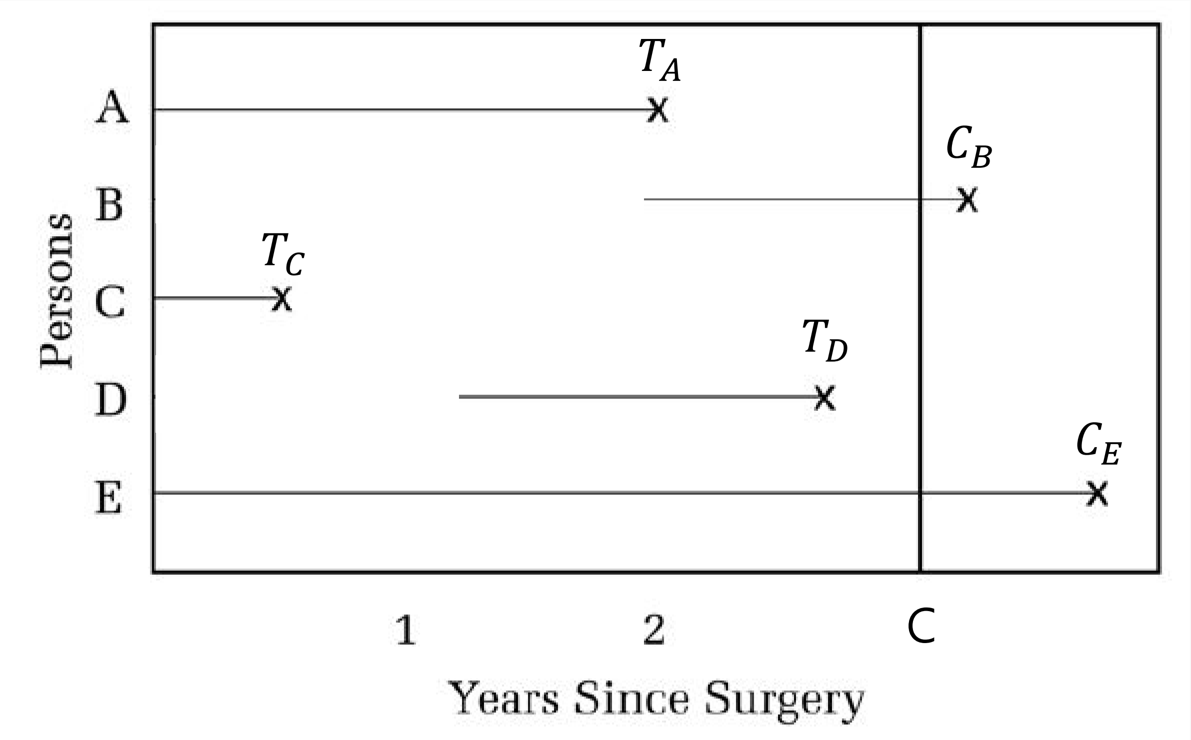
예) **전구 수명 실험 : 전구 100개를 켜놓고 5개가 꺼질 때 연구 종료**

=> 생존 시간에 대해 순서통계량으로 다룬다.

중도절단 (우중도절단, 좌중도절단, 구간중도절단)

i번째 개체의 생존시간 : T_i , 중도 절단 여부 : δ_i , 관측중단시점 : C_i , $i = 1,2,\cdots,n$

임의 우중도절단 : 실제 생존데이터를 얻을 때 서로 다른 시점에 연구에 참여 할 수 있고 연구가 종료되지 않아도 여러 이유로 더 이상 연구에 참여하지 못하는 경우



예) 질병이 아닌 교통사고에 의한 사망, 환자가 병원을 옮겨 연락두절 등

중도절단 (우중도절단, 좌중도절단, 구간중도절단)

구간 중도절단 : 생존시간이 어떤 구간내에서 발생하는 경우

예) 1. 감염연구에서 환자가 매달 병원을 방문한다. 지난 달에는 감염되지 않았으나, 이번 달 방문에 감염 되어있었다면 한 달 사이에 감염이 발생 했고 언제 발생했는지 정확히 알 수 없으므로 구간 중도절단.

2. 산업분야에서 장비에 대한 점검은 어떤 주기를 갖고 하므로 구간 중도절단 발생.

좌 중도절단 : 사건이 연구시작 이전에 발생

예) 고등학생 흡연 연구 => 처음 흡연시기를 물었는데 기억이 안 난다고 한다.
=> 흡연 학생이지만 처음 흡연시기를 알 수 없으므로 좌 중도절단.

우도함수

생존시간과 중도절단 시간이 독립임을 이용해 중도절단 유형을 고려해 생존데이터의 우도함수를 구함 => PASS

비모수 방법에 의한 생존함수 추정

생명표 : 생존데이터를 그룹지어 얻는 경우 이용 (어떤 구간 내에 발생한 사건의 개수를 기록
*** 표본의 크기가 50 이상일 때 적용한다.**

예) 중도절단이 구간 끝에서 일어났다고 가정한 생명표

Year of entry $[t_{i-1}, t_i)$	구간초기 위험집합 Y_i	사건발생 건수 d_i	중도절단 건수 c_i	치사율 $\widehat{m_i} = \frac{d_i}{Y_i - c_i}$	구간 내 생존율 $1\text{-}\widehat{m_i}$	구간내 생존함수 $\widehat{S(t_i)} = \prod(1 - \widehat{m_i})$
[0,1)	146	27	3	$\frac{27}{146} = 0.185$	0.815	0.815
[1,2)	116	18	10	$\frac{18}{116} = 0.155$	0.845	$0.815 \times 0.845 = 0.689$
[2,3)	88	21	10	$\frac{21}{88} = 0.239$	0.761	$0.689 \times 0.761 = 0.524$
[3,4)	57	9	3	$\frac{29}{57} = 0.158$	0.842	$0.524 \times 0.842 = 0.441$
[4,5)	45	1	3	$\frac{1}{45} = 0.022$	0.972	$0.441 \times 0.972 = 0.432$

비모수 방법에 의한 생존함수 추정

예) 중도절단이 구간 초기에 일어났다고 가정한 생명표

Year of entry $[t_{i-1}, t_i)$	구간초기 위험집합 Y_i	사건발생 건수 d_i	중도절단 건수 c_i	치사율 $\widehat{m_i} = \frac{d_i}{Y_i - c_i}$	구간 내 생존율 1-$\widehat{m_i}$	구간내 생존함수 $\widehat{S(t_i)} = \prod(1 - \widehat{m_i})$
초기점	146	0	0	0	1	1
[0,1)	146	27	3	$\frac{27}{146-3} = 0.189$	0.811	0.811
[1,2)	116	18	10	$\frac{18}{116-10} = 0.170$	0.830	0.673
[2,3)	88	21	10	$\frac{21}{88-10} = 0.269$	0.731	0.492
[3,4)	57	9	3	$\frac{29}{57-3} = 0.167$	0.833	0.410
[4,5)	45	1	3	$\frac{1}{45-3} = 0.024$	0.976	0.400

비모수 방법에 의한 생존함수 추정

Kaplan-Meier 누적한계 추정량 : 사건(사망)이 발생한 시점마다 생존율을 계산

*** 표본의 크기가 50 이하일 때 적용한다.**

사건 발생시점 t_i를 순서대로 나열한다. $t_1 < t_2 < \cdots < t_n$

$$\widehat{S(t)} = \begin{cases} 1 & , t < t_i \\ \prod_{t_i \le t} 1 - \frac{d_i}{Y_i} & , t \ge t_i \end{cases}$$

Y_i : t_i 시점에 개체수 (만약 중도절단이 있다면 개체수에서 빼줌.

d_i : t_i 시점에 발생 사건 수

델타방법(Delta method)를 이용해 분산계산이 가능하다. 또한, 근사적으로 정규분포를 따라 신뢰구간 또한 계산이 가능하다.

비모수 방법에 의한 생존함수 추정

PEEDORI

예) 백혈병 데이터 그룹1에 대한 Kaplan-Meier 생존함수 추정

Ordered time	# of obs at risk	# of observed	no. of censored $[t_i, t_{i+1})$	$\prod_{t_j \le t} P(T > t_i \mid T \ge t_i)$ $=\prod_{t_i \le t}\left[1-\frac{d_i}{Y_i}\right]$	Standard error $se[\widehat{S(t)}]$	Lower 95% CI	Upper 95% CI
t_i	Y_i	d_i	c_i	$\widehat{S(t_i)}$			
0	21	0	0	1			
6	21	3	1	$1 \times (18/21) = 0.857$	0.076	0.720	1
7	17	1	1	$0.857 \times (16/17) = 0.807$	0.087	0.653	0.996
10	15	1	2	$0.807 \times (14/15) = 0.753$	0.096	0.586	0.968
13	12	1	0	$0.753 \times (11/12) = 0.690$	0.107	0.510	0.935

생존함수 동일성 검정: 모집단 위험함수가 적절한지 혹은 2개 이상의 집단의 생존함수에 차이가 있는지 검정

1. 카이제곱 검정
2. 로그-순위 검정 (log-rank test) => 가장 널리 쓰이는 방법

카이제곱검정처럼 생존함수에서 기댓값과 관측값을 이용해 관측값과 기댓값의 차이를 구해 이를 가중값으로 이용하여 검정한다.

H_o : 모집단 생존시간에 대한 분포함수는 $F_o(t)$ 이다.
H_a : 적어도 일부분에서 모집단 생존시간에 대한 분포함수는 $F_o(t)$가 아니다.

모형

PEEDORI

Cox 비례위험모형 : **비례위험모형**은 생존분석에서 쓰이는 통계 모형이다. 준모수적 방법을 이용하여 생존함수를 추정한다. 1972년 통계학자 데이비드 콕스에 의해 처음 개발되었다. 모형의 이름인 비례위험은 시간에 상관없이 어떤 변수의 위험비(hazard ratio, HR)는 항상 일정하다는 모형의 기본가정에서 비롯되었다.

$h(t \mid Z_i) = h_o(t) * \varphi(Z\,;\beta)$,

$h_o(t)$는 기저위험함수 (baseline hazard function) , Z_i : 변수
log (생존시간)=$h(t)$의 관계

$\varphi(Z\,;\beta)$는 일반적으로 지수함수$\exp(Z\beta)$ 를 고려한다. (Z_i가 한 단위 증가할 때 $\exp(\beta)$ 만큼 위험률 증가)

그 외 $\varphi(Z\,;\beta)$=$1+Z\beta$, $\varphi(Z\,;\beta)$=$\log\left(1+e^{Z\beta}\right)$ 등의 함수 고려 가능 , 이때 위험률은 음수가 될 수 없음.

회귀계수 β : 부분우도함수를 이용해 추정
회귀계수 검정 : 우도비, wald, score test를 이용.

모형

동점 처리 : 비례위험모형에서 유도된 부분우도함수 식은 모든 시점들이 서로 동일하지 않다는 가정하에서 유도된다. 그러나 실제 데이터에서 생존시간이 동일한 경우는 종종 발생하므로 이들을 처리하는 방법이 필요하다.

예) 정확성방법, Breslow 근사법, Efron 근사법 등등

모형 가정

PEEDORI

잔차 확인 이유

I. 비례위험모형의 가정인 위험률의 비례성 가정의 검토 필요
II. 이상점, 영향점 확인.

비례성 가정 확인 방법

① 비례성 가정에 대한 가설 검정
② 로그 누적 위험함수(log-log plot)을 그려 두 그룹의 선이 교차하면 비례성 만족 X