R 기반 의학통계 및 머신러닝

박 승

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

CHAPTER

08

생존 분석 기법

08 생존 분석 기법

1. 생존분석의 개념

- 생존분석(Survival Analysis)은 특정 사건(Event)가 발생하기까지의 시간(Duration)을 분석하는 통계적 방법
- 시간이 중요한 변수이며, “언제” 발생하는지, 발생하기까지 얼마나 걸렸는지를 분석하는 것이 핵심
- 생존분석을 사용하는 대표적 분야
  - 의학연구
    ex) 환자의 치료 후 생존 시간, 재발까지 걸리는 시간 분석을 통해 특정 치료법이 미치는 영향 분석
  - 공학/신뢰성분석
    ex) 부품의 수명, 공정 오류가 발생하기까지의 시간을 분석하여 유지 보수 계획 수립
  - 마케팅/경제
    ex) 고객이 서비스를 해지할 때 까지의 시간을 분석하여 이탈 예측 및 마케팅 전략 수립

Medical

Engineering

Marketing

1. 생존분석의 개념

▪ 생존분석이 필요한 이유

- (1) 시간(time) 변수를 고려한 분석이 가능
  단순히 사건의 발생 여부, 발생률 뿐만 아니라 얼마나 지속성에 대해서도 분석 가능
  ex) 수술 후 5년간 생존률, 재발율 분석 등

- (2) 검열 데이터(Censored data) 처리 가능
  연구 종료 시점까지 사건이 발생하지 않은 데이터를 포함하여 분석 가능
  ex) 임상 실험 또는 비파괴 수명 검사 등

- (3) 생존 확률과 위험률(Hazard Rate) 동시 분석 가능
  단순 평균 비교 뿐만 아니라 시간의 변화에 따라 위험률의 변화도 분석 가능
  ex) 구매연도에 따른 가전제품 고장률 분석(위험률 점차 증가), 수술 환자 예후 분석(위험률 점차 감소) 등

1. 생존분석의 개념

- **생존함수(Survival Function) - $S(t)$**
  특정 시간 t 까지 생존할 확률을 나타내는 함수,
  $S(t) = P(T > t)$ 즉, 시간 t까지 사건이 발생하지 않을 확률을 의미함

- **위험함수(Hazard Function) - $h(t)$**
  특정 시점 t에서 사건이 발생할 "즉시 위험"을 나타내는 함수

$$h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1-F(t)}$$

- **누적위험함수(Cumulative Hazard Function) - $H(t)$**
  위험함수를 적분하여 누적된 위험을 측정한 함수

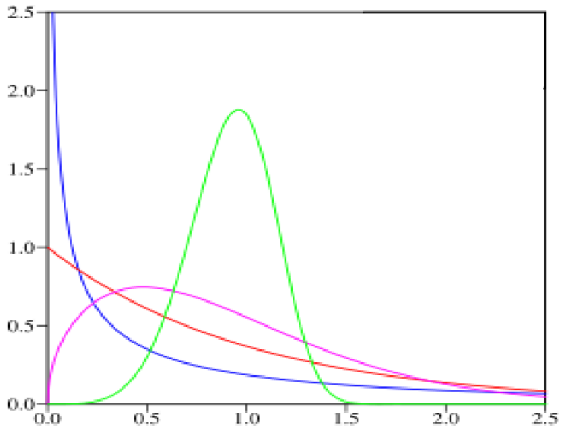
$$H(t) = \int_0^t h(u)du$$

생존함수와의 관계는 $S(t) = e^{-H(t)}$ 즉, 누적위험이 증가하면 생존 확률이 감소

1. 생존분석의 개념

- 와이블 분포(베이불 분포, Weibull Distribution)
  정의역 $t > 0$ 에서 정의되는 연속 확률 분포. 매우 유연하기 때문에 β의 값에 따라 가우시안분포나
  지수분포같은 다른 분포들을 흉내 낼 수 있으며, 주로 수명을 추정하는 데 사용됨
- 와이블분포의 pdf

$$f(t) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} e^{-\left(\frac{t}{\eta}\right)^{\beta}}, \; t, \eta, \beta > 0$$



- 와이블분포의 cdf

$$F(t) = 1 - e^{-\left(\frac{t}{\eta}\right)^{\beta}}, \qquad t, \eta, \beta > 0$$

- 이때, 와이블 분포의 위험 함수는 아래와 같이 간단하게 표현됨

$$h(t) = \frac{f(t)}{1-F(t)} = \frac{\frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} e^{-\left(\frac{t}{\eta}\right)^{\beta}}}{e^{-\left(\frac{t}{\eta}\right)^{\beta}}} = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1}$$

1. 생존분석의 개념

- 와이블 분포(베이불 분포, Weibull Distribution)
  - 와이블 분포의 위험함수 $h(t)$는 β의 값에 따라서 위험률이 증가하는 경우, 위험률이 일정한 경우, 위험률이 감소하는 경우를 모두 나타낼 수 있음
- 위험률이 증가하는 경우 : $\beta > 1,\ \ h(t)$는 점점 증가 ($\left(\frac{t}{\eta}\right)^{\beta-1}$ 이 점점 커짐)
  - ex) Wear-out problem
- 위험률이 일정한 경우 : $\beta = 1, h(t) = \frac{1}{\eta}$로 t에 관계없이 항상 일정
  - cf) 이때의 와이블 분포는 지수분포와 동일
- 위험률이 감소하는 경우 : $\beta < 1, h(t)$는 점점 감소 ($\left(\frac{t}{\eta}\right)^{\beta-1}$ 이 점점 작아짐)
  - ex) Infant Mortality

1. 생존분석의 개념

▪ 생존분석 데이터의 기본 구조

- 생존분석에서 사용되는 데이터는 일반적인 데이터셋과 다르게 시간(time) 변수를 필수적으로 포함해야함
- 일반적으로 생존 데이터는 다음과 같은 구조를 가짐

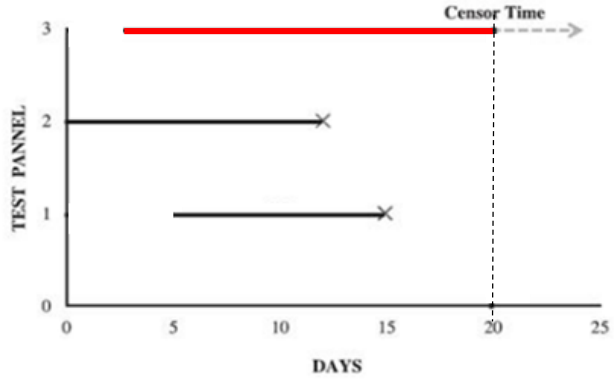
ID	Time	Event	Variables
1	10	1	A
2	11	1	A
3	8	0	B
⋮	⋮	⋮	⋮
n	20	0	A

Time: 생존 시간.
관찰 시점부터 사건(발병, 사망 등)이 발생하기까지 걸린 시간

Event: 사건 발생 여부
1 = 사건 발생, 0 = 관찰 종료 시점까지 생존

1. 생존분석의 개념

- 중도절단자료(Censored Data)_중도절단자료의 세가지 유형
  - (1) 우측 중도절단(Right censored) – 주로 다루는 내용
    연구 종료 시점까지 사건이 발생하지 않은 경우
    해당 개체가 사건을 언제 경험할지 모르지만, 연구 종료 시점까지는 생존한 상태
  - (2) 좌측 중도절단(Left censored)
    사건이 연구 시작 전에 이미 발생했지만, 정확한 시점을 모르는 경우
    ex)병원 방문시 이미 질병이 진행중이었지만, 발병 시점을 모르는 케이스
  - (3) 구간 중도절단(Interval censored)
    우측 중도절단과 좌측 중도절단이 모두 일어난 경우



#관찰 시작점이 다른 우측중도절단 자료의 예시

**관찰 종료 시점에 따라 Type I, II censored data로 나누어지지만 여기서는 다루지 않음

1. 생존분석의 개념

- 중도절단자료(Censored Data)_중도절단자료의 고려가 필요한 이유
  - (1) 검열 데이터를 고려하지 않으면 생존율이 왜곡됨

    연구 종료 시점까지 생존한 환자들이 많을 경우, 단순히 이들을 제외하면 생존률이 과소평가될 수 있음
  - (2) 중도 절단된 데이터가 많을수록 분석이 더 어려워짐

    중도절단되지 않은 데이터 즉, 사건(Event) 발생 데이터가 너무 적으면, 유의미한 결과를 얻기 어려움

1. 생존분석의 개념

- 인년 (Person-year)
  연구 대상자가 연구 기간 동안 얼마나 오랫동안 사건 없이 추적 관찰되었는지를 측정하는 단위
  사건이 발생하기 전까지 연구에 참여한 총 시간을 나타내며, 종단적(longitudinal) 연구에서의 분모로 사용됨

ex)

ID	Duration	사건	Censored
1	12개월	x	o
2	8개월	o	x
3	15개월	o	x
4	20개월	x	x
5	10개월	o	x
Sum	65개월	2명	

총 65개월, 약 5.42년 동안 추적 관찰이 이루어졌으며, 3명의 환자가 발생함

$$\text{Incidence} = \frac{\text{총 사건 수}}{\text{총 인년}} = \frac{3}{5.4166} \cong 0.554$$

1000인년 당 발생 수는 554명

2. 비모수 통계학

▪ 비모수 통계학이란?

- 데이터가 특정한 분포(정규분포 등)을 따른다고 가정하지 않고 수행하는 통계적 방법
  데이터의 분포에 대한 강한 가정 없이 데이터 자체의 순서나 순위 등을 활용하여 분석을 수행하는 방식

▪ 왜 생존분석에서 비모수 통계를 사용하는가?

- 생존 데이터는 중도절단된 자료(Censored data)가 포함되므로 정규성이 보장되지 않을 때가 많음
  특정한 분포를 따르는 것을 가정하기 어렵고, 검정하기도 어려움
  따라서 순위 기반 비교를 수행하는 비모수 검정 방법이 더 적절함
  대표적인 방법으로 카플란-마이어 그래프를 그릴 때 수행하는 로그-랭크 검정(Log-Rank Test)이 있으며
  윌콕슨 순위합 검정(두 집단의 비교, t-test에 대응), 크루스칼-왈리스 검정(세 개 이상의 집단의 비교,
  분산분석등의 F-test에 대응) 등이 있음

2. 비모수 통계학 – 로그 순위 검정

▪ 로그 순위 검정 (Log-Rank Test)

- 8명이 4명씩 팀을 이뤄 동시에 달리기를 해서 순위를 매긴다고 가정해보자.

#순위에 기반한 검정(로그 부호 검정)

- 철수: 1등은 1점, 2등은 2점 순으로 8등은 8점을 주고, 점수의 총합이 적은 쪽이 이기는 것으로 하자
- 영희: 무조건 1등이 짱짱맨이지. 1등한 사람이 있는 팀이 무조건 이기는 것으로 하자 #순서통계량에 기반한 검정
- 민수: 극단값은 지우는게 좋겠어. 각 팀의 1등은 지우고, 여섯명만 계산해서 철수의 방식을 따를래
- 도연: 난 각 선수의 달리는 시간을 계산 하고 평균이 더 적은 쪽을 고를래. #순위에 기반한 검정
  참고로 제한시간 내에 못들어오면 아웃이야. #시간의 개념이 포함된 로그 순위 검정

- 당신이 룰을 결정해야만 한다면, 어떻게 승리 팀을 구해야 합리적일까?
  (수많은 검정통계량 중에, 어떤 것이 가장 합리적인 검정통계량일까?)

2. 비모수 통계학 – 로그 순위 검정

▪ 로그 순위 검정 (Log-Rank Test)

– 두개의 그룹의 평균을 비교하는 비모수 통계 방법

Group	Time	Event(완주)
A	15.2	예
A	17.5	예
A	14.0	예
A	20.0	아니오
B	20.0	아니오
B	18.0	예
B	17.0	예
B	19.5	예

```
library(survival)

run_data <- data.frame(
  group = c("A", "A", "A", "A", "B", "B", "B", "B"),
  time = c(15.2, 17.5, 14.0, 20.0, 20.0, 18.0, 17.0, 19.5),
  event = c(1,1,1,0,0,1,1,1))

log_rank_test <- survdiff(Surv(time, event) ~ group,
                          data = run_data)
log_rank_test
```

2. 비모수 통계학 – 로그 순위 검정

- 로그 순위 검정 (Log-Rank Test)

```
> log_rank_test
Call:
survdiff(formula = Surv(time, event) ~ group, data = run_data)

        N Observed Expected (O-E)^2/E (O-E)^2/V
group=A 4        3     2.25     0.254     0.417
group=B 4        3     3.75     0.152     0.417

 Chisq= 0.4  on 1 degrees of freedom, p= 0.5
```

- 로그 순위 검정 결과 p-값은 0.5가 나왔다. 그럼, '두 그룹 간 달리기 성적은 차이가 없다' 라고 말하면 될까?

- 로그 순위 검정의 귀무가설과 대립가설은 뭐지?

2. 비모수 통계학 – 로그 순위 검정

▪ 로그 순위 검정 (Log-Rank Test)

- 로그 순위 검정은 사실 로그를 쓰지도 않고, 순위를 쓰지도 않음
  (시간 개념이 없고 순위 개념만 쓰는 검정 방법으로는 로그 부호 검정이 있음)
  로그 순위 검정의 본질은 시간이 지남에 따라 그룹간 사건이 발생하는 패턴(생존 곡선)이 같은지 검정하는 것
- 위험률, 사건발생률은 생존함수의 도함수와 비교되지만 로그 순위 검정은 위험률을 직접 비교하는 것이 아니고
  생존 데이터의 누적 분포를 비교하는 방식(후에 다룰 Cox 회귀 모델과의 혼동 방지)
  범주형 자료분석에서의 동일성, 동질성 검정의 개념과 조금 더 유사함(Fisher's Exact test, Chi-square Test)

H_0 : 두 그룹간 달리는 패턴에 차이가 없다. 즉, 시간이 지남에 따라 두 팀의 선수들이 완주하는 비율이 같음
vs. H_1 : 두 그룹간 달리는 패턴에 차이가 있다
라고 보는 것이 타당함

2. 비모수 통계학 – 로그 순위 검정

▪ 로그 순위 검정 (Log-Rank Test)

기대 사건 수 계산 $\left(E_{A_i} = d_i \times \frac{n_{A_i}}{n_i}, E_{B_i} = d_i \times \frac{n_{B_i}}{n_i}\right)$

분산 계산 $\left(V_A = E_A \times \left(1 - \frac{n_{A_i}}{n_{A_i}+n_{B_i}}\right), V_B = E_B \times \left(1 - \frac{n_{A_i}}{n_{A_i}+n_{B_i}}\right)\right)$

Group	Time	Event
A	15.2	예
A	17.5	예
A	14.0	예
A	20.0	아니오
B	20.0	아니오
B	18.0	예
B	17.0	예
B	19.5	예

시간(초)	그룹 A 잔여인원 (n_{A_i})	그룹 B 잔여인원 (n_{B_i})	그룹 A 기대 사건 수 E_{A_i}	그룹 B 기대 사건 수 E_{B_i}	V_A	V_B
14.0	4	4	$1 \times \frac{4}{(4+4)} = 0.5$	$1 \times \frac{4}{(4+4)} = 0.5$	$0.5 \times \left(1 - \frac{4}{8}\right) = 0.25$	$0.5 \times \left(1 - \frac{4}{8}\right) = 0.25$
15.2	3	4	$1 \times \frac{3}{(3+4)} = 0.43$	$1 \times \frac{4}{(3+4)} = 0.57$	$0.43 \times \left(1 - \frac{3}{7}\right) = 0.245$	$0.57 \times \left(1 - \frac{4}{7}\right) = 0.244$
17.0	2	4	$1 \times \frac{2}{(2+4)} = 0.333$	$1 \times \frac{3}{(2+4)} = 0.667$	$0.333 \times \left(1 - \frac{2}{6}\right) = 0.22$	$0.667 \times \left(1 - \frac{4}{6}\right) = 0.223$
17.5	2	3	$1 \times \frac{2}{(2+3)} = 0.4$	$1 \times \frac{3}{(2+3)} = 0.6$	$0.4 \times \left(1 - \frac{2}{5}\right) = 0.24$	$0.6 \times \left(1 - \frac{3}{5}\right) = 0.24$
18.0	1	3	$1 \times \frac{1}{(1+3)} = 0.25$	$1 \times \frac{2}{(1+3)} = 0.75$	$0.25 \times \left(1 - \frac{1}{4}\right) = 0.1875$	$0.75 \times \left(1 - \frac{3}{4}\right) = 0.1875$
19.5	1	2	$1 \times \frac{1}{(1+2)} = 0.333$	$1 \times \frac{2}{(1+2)} = 0.667$	$0.333 \times \left(1 - \frac{1}{3}\right) = 0.22$	$0.667 \times \left(1 - \frac{2}{3}\right) = 0.2233$
20.0	1	1	**합계 :** $\cong 2.2452$	**합계 :** $\cong 3.7547$	**합계 :** $\cong 1.367$	**합계 :** $\cong 1.367$

2. 비모수 통계학 – 로그 순위 검정

- 로그 순위 검정 (Log-Rank Test)

```
> log_rank_test
Call:
survdiff(formula = Surv(time, event) ~ group, data = run_data)

        N Observed Expected (O-E)^2/E (O-E)^2/V
group=A 4        3     2.25     0.254     0.417
group=B 4        3     3.75     0.152     0.417

 Chisq= 0.4  on 1 degrees of freedom, p= 0.5
```

검정통계량 계산

$$\chi^2 = \frac{\sum(O-E)^2}{\sum V_i}$$

따라서 $\chi^2 = \frac{(3-2.2452)^2+(3-3.7547)^2}{1.37+1.37} \cong 0.417$, p-값은 0.5로 귀무가설 기각 못함

일전에 이항분포에 대한 정규분포의 근사 및 카이제곱 분포를 다룬 적이 있으며,

해당 검정통계량의 식이 이전의 것과 다소 다른 것을 확인할 수 있음

2. 비모수 통계학 – 로그 순위 검정

- 로그 순위 검정 (Log-Rank Test)
- remind : 카이제곱과 정규분포

이항분포의 pmf는 $P(X=x)=\binom{n}{x}p^x(1-p)^{1-x}, x=0,1,2,\cdots,0\leq \mathrm{p}\leq 1$ 이다.

$\binom{n}{x}$에서 n이 커질 수록 연산량이 막대하게 늘어나기 때문에 계산의 편의를 위해 정규분포로 근사할 수 있다.

→ 이항분포를 정규분포로 근사시킨 뒤 제곱하여 카이제곱 분포를 따르도록 가정했기 때문

2. 비모수 통계학 – 로그 순위 검정

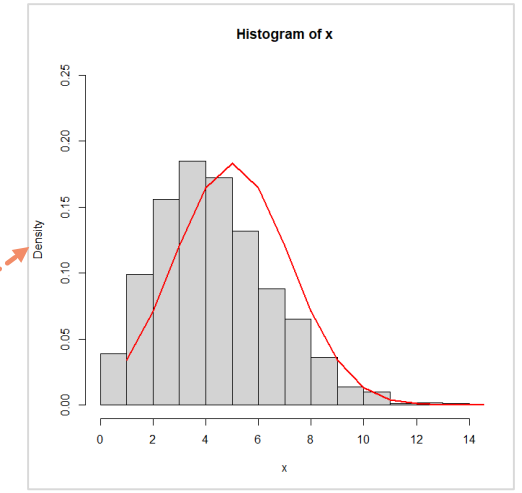
▪ 이항분포의 정규근사

확률변수 X가 모수가 n, p인 이항분포를 따를 때, X의 평균은 np, 분산은 $np(1-p)$.

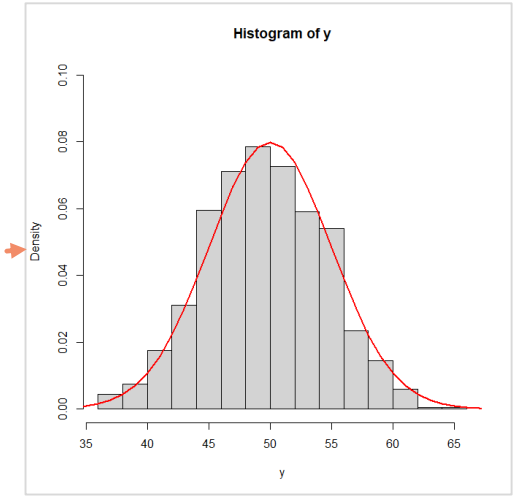
이를 정규화 하면 $\frac{X-np}{\sqrt{np(1-p)}}$는 평균이 0, 분산이 1인 표준정규분포를 따른다.

이때, n이 크고 p가 $\frac{1}{2}$에 가까울 수록 정규분포에 더 잘 맞는다.

```
x<-rbinom(1000, 100, 0.05)
hist(x, freq=F, ylim=c(0,0.25),breaks=20)
lines(1:1000,dnorm(1:1000, mean=5,
sd=sqrt(100*0.05*0.95)), col="red", lwd=2)
```




```
y<-rbinom(1000, 100, 0.5)
hist(y, freq=F, ylim=c(0,0.1), breaks=20)
lines(1:100,dnorm(1:100, mean=50,
sd=sqrt(100*0.5*0.5)), col="red", lwd=2)
```



2. 비모수 통계학 – 로그 순위 검정

- 이항분포의 정규근사

결론적으로, $\frac{X-np}{\sqrt{np(1-p)}}$가 정규분포를 따르므로, $\frac{(X-np)^2}{np(1-p)}$은 카이제곱분포를 따르는 것

따라서, $\frac{\sum_{i=1}(O-E)^2}{\sum_{i=1} V_i} \sim \chi^2$

3. 카플란 마이어 생존 곡선

- 카플란-마이어(Kaplan-Meier) 방법이란?
  - 생존함수(Survival Function, $S(t)$)를 추정하는 비모수적(non-parametric) 방법
  - 특정 시점에서 사건(사망, 재발 등)이 발생하지 않을 확률을 추정하는데 주로 사용
  - 중도절단(Censoring) 자료를 고려할 수 있기 때문에 현실 연구에서 자주 활용

- 카플란-마이어 추정량

$$\widehat{S(t)} = \prod_{t_i \le t} \left(1 - \frac{d_i}{n_i}\right)$$

d_i: 시간 t_i에서 사건이 발생한 개수

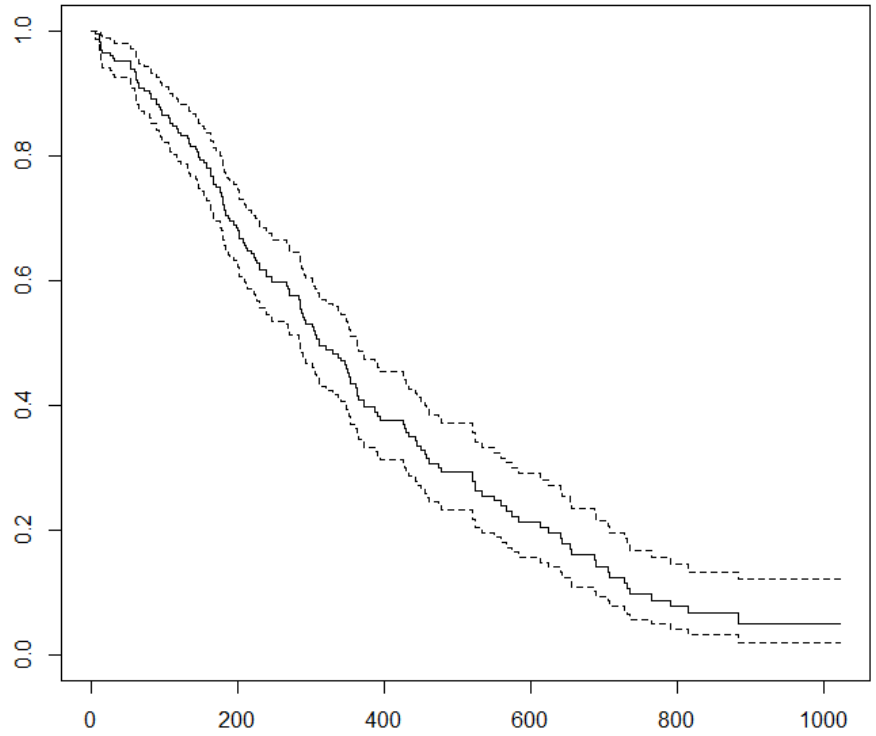
n_i: 시간 t_i에서 아직 위험(Risk Set) 상태에 있는 총 개체 수

- 각 사건 발생 시점에서 생존 확률을 계산하고, 이를 누적하여 전체 생존 확률을 추정

3. 카플란 마이어 생존 곡선

- 내장 데이터 lung을 활용한 카플란 마이어 생존 곡선 작성

```
library(survival)
library(survminer)
df<-lung                    #event 변수를 1,0으로 재 코딩
df$event<-ifelse(df$status==2, 1,0)
km_fit<- survfit(Surv(time, event) ~ 1, data=df)
plot(km_fit)
```



3. 카플란 마이어 생존 곡선

```
ggsurvplot(km_fit, data=df)
```

X축: 생존시간(Days)

특정 환자가 생존한 기간을 의미

Y축: 생존확률(Survival Probability)

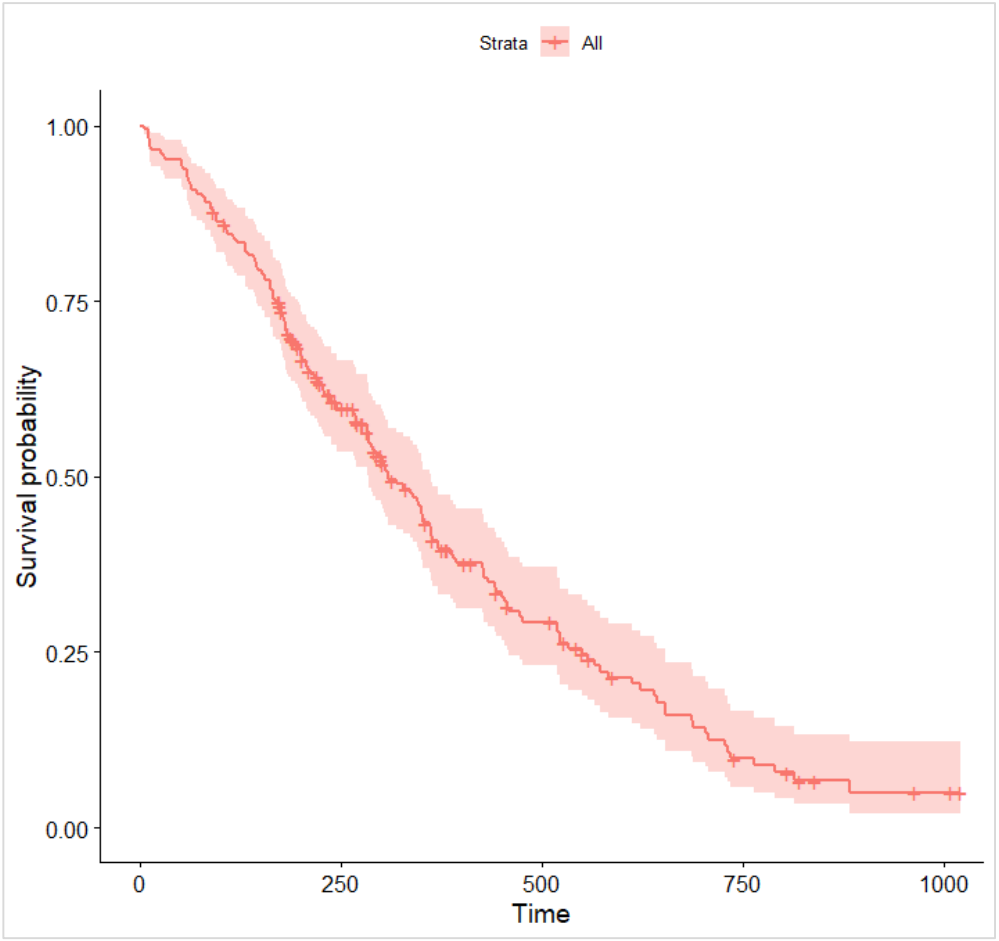
특정 시점에서 생존할 확률을 나타냄

붉은색 음영: 신뢰구간(Confidence Interval)

넓을 수록 추정이 불확실하다는 의미

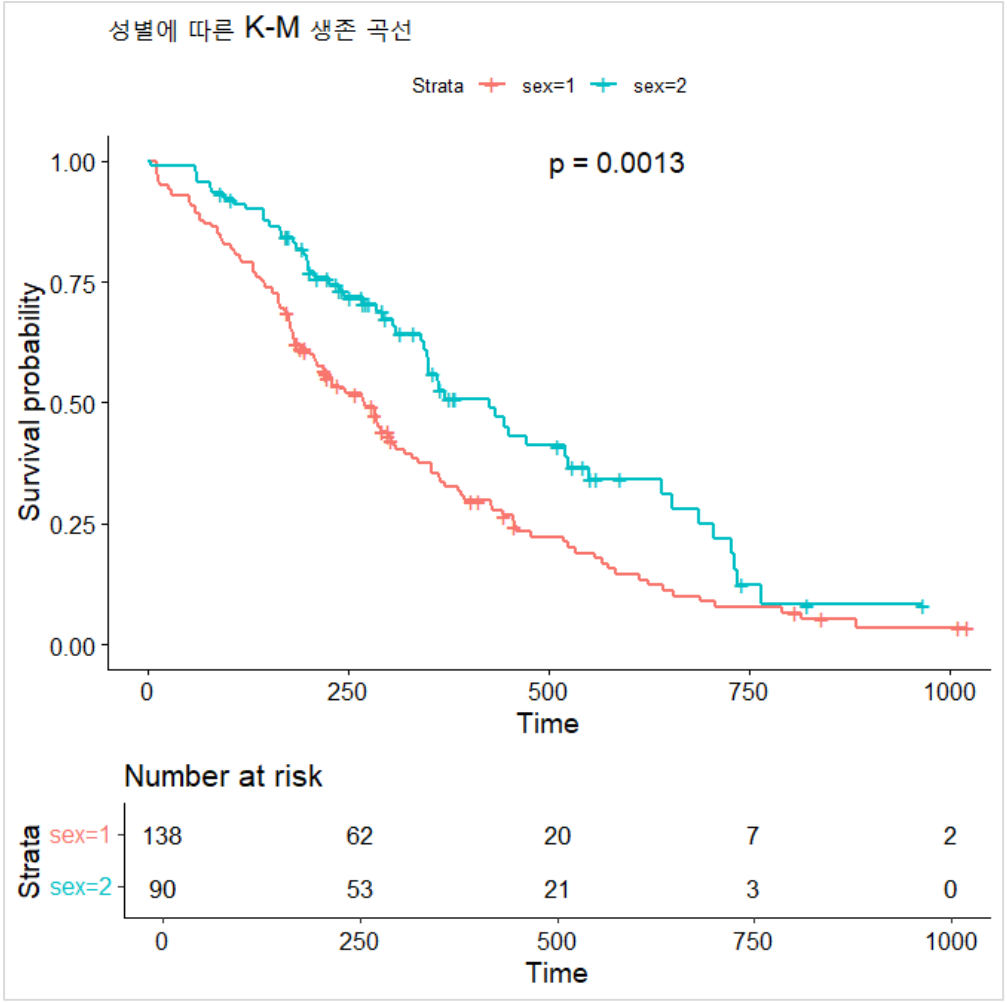
빨간색 실선: Kaplan-Meier 추정 생존 곡선

+ 표시: 중도절단 데이터(연구종료시점까지 생존 또는 중도탈락)



3. 카플란 마이어 생존 곡선

```
#성별을 strata로 놓고 그래프 작성
km_fit_sex<- survfit(Surv(time, event) ~sex, data=df)

ggsurvplot(km_fit_sex, data=df, pval=T, pval.coord=c(500,1),
           risk.table=T,
           title="성별에 따른 K-M 생존 곡선")
                                  #p-값 출력 위치 설정
#두 그룹간 로그랭크검정 결과 표시
```



- 그래프 해석

  - 전체적인 생존 경향은 시간이 지나면서 점차 감소

    남성(sex=1)의 경우 여성(sex=2)보다 빠르게 감소

  - Risk Table

    각 시점 별로 생존 상태를 유지하는 환자 수

    ex)500일 시점에서 남성은 20명, 여성은 21명 생존

  - 로그-랭크 검정 해석

    p-값은 0.0013으로, 성별에 따른 생존 곡선 차이가 통계적으로 유의미

3. 카플란 마이어 생존 곡선

- 실습1)
  lung data의 구조를 파악한 뒤 (?lung 및 str(lung))
  연속형 변수 age를 60세 미만, 60세 이상으로 나누어서
  두 그룹간 생존곡선의 차이를 비교하는 카플란 마이어 그래프를 작성해보자

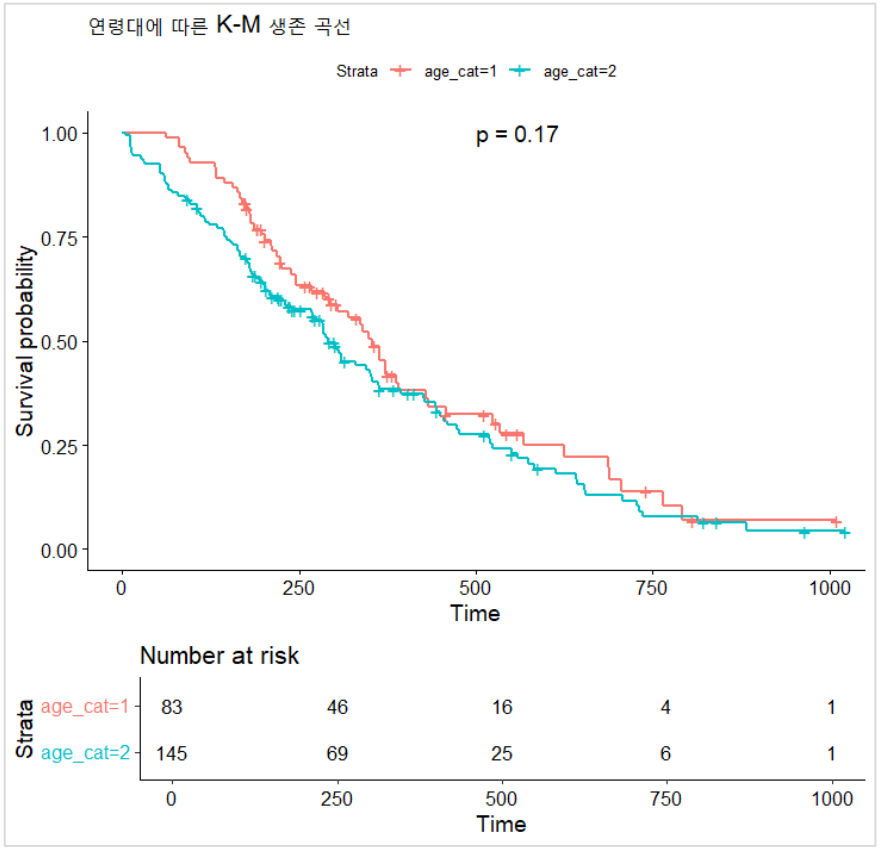
- 실습2)
  이번에는 age를 70세 미만, 70세 이상으로 나누어서
  두 그룹간 생존곡선의 차이를 비교하는 카플란 마이어 그래프를 작성해보자

- 두 결과를 비교할 때, 얻을 수 있는 결론은 무엇인가?

3. 카플란 마이어 생존 곡선

▪ 실습1)

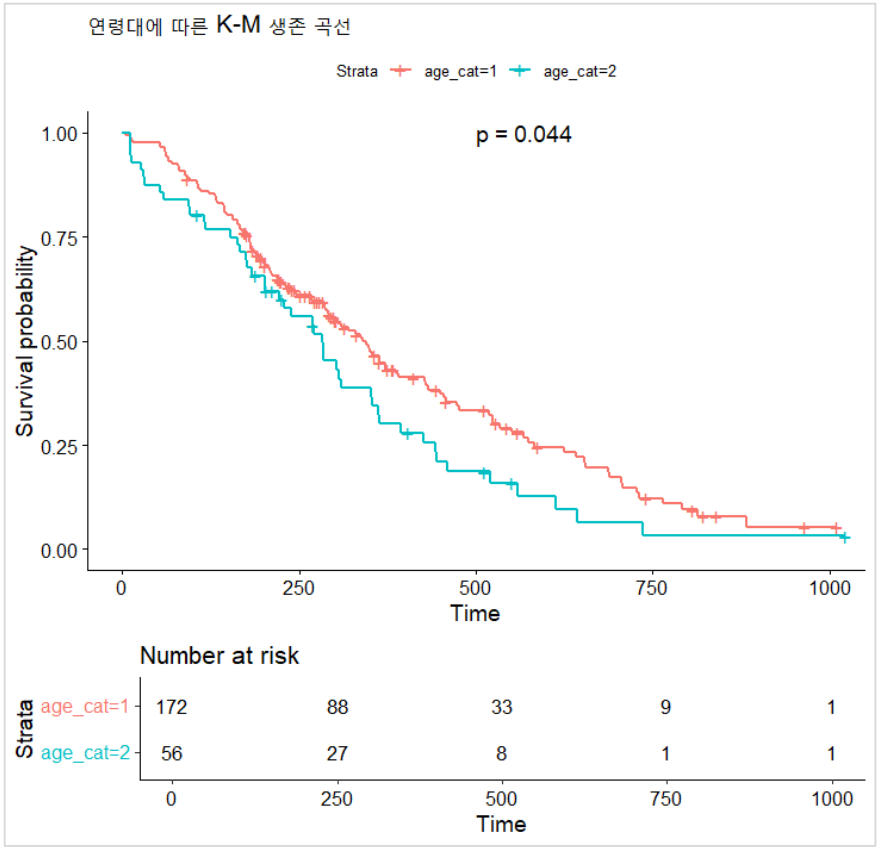
```
df$age_cat<-NA
df$age_cat<-ifelse(df$age<60,1,df$age_cat)
df$age_cat<-ifelse(df$age>=60,2,df$age_cat)
df$age_cat %>% table
km_fit_age<- survfit(Surv(time, event) ~ age_cat, data=df)
ggsurvplot(km_fit_age, data=df, pval=T, pval.coord=c(500,1),
           risk.table=T,
           title="연령대에 따른 K-M 생존 곡선")
```



3. 카플란 마이어 생존 곡선

- 실습2)

```
df$age_cat<-NA
df$age_cat<-ifelse(df$age<70,1,df$age_cat)
df$age_cat<-ifelse(df$age>=70,2,df$age_cat)
df$age_cat %>% table
km_fit_age<- survfit(Surv(time, event) ~ age_cat, data=df)
ggsurvplot(km_fit_age, data=df, pval=T, pval.coord=c(500,1),
           risk.table=T,
           title="연령대에 따른 K-M 생존 곡선")
```



4. Cox 비례 위험 모형(Cox Proportional Hazards Model)

- Cox 비례 위험 모형은 생존 시간과 여러 변수(공변량) 간의 관계를 분석하는 회귀 모델
- Kaplan-Meier 생존 곡선이 단순 그룹 비교라면, Cox 비례위험 모형은 다중 변수의 영향을 고려할 수 있음(따라서 Cox 회귀분석이라고도 함)

- 비례 위험 가정(Proportional Hazards Assumption)을 따름
- 위험비(Hazard Ratio, HR)를 통해 변수의 상대적 위험을 해석

- **Cox 비례 위험 모형의 가장 중요한 가정: 비례 위험 가정
- Cox 비례 위험 모형은 시간이 지나도 각 변수의 위험비가 일정하게 유지된다는 가정을 전제함
  즉, 모든 설명변수 X가 시간 t에 따라 위험에 미치는 영향이 일정해야 함

$$\frac{h_1(t)}{h_2(t)} = e^{\beta X}$$

- Cox 회귀를 수행 후 비례 위험 가정 검정이 필요

4. Cox 비례 위험 모형(Cox Proportional Hazards Model)

- Cox 비례 위험 모형은 결국 생존 분석에서 독립변수(설명변수)가 생존 시간에 미치는 영향을 평가하는 다변량 회귀 모델
- Cox 회귀 모델의 수학적 정의

Cox 회귀 모델에서 위험 함수(Hazard function, $h(t)$)는 다음과 같이 정의됨

$$h(t) = h_0(t) \times e^{(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}$$

$h(t)$: 시간 t에서 사건이 발생할 위험도

$h_0(t)$: 기준 위험 함수(Baseline Hazard Function)

$X_1, X_2, \cdots, X_p$: 설명 변수

$\beta_1, \beta_2, \cdots, \beta_p$: 회귀 계수

4. Cox 비례 위험 모형(Cox Proportional Hazards Model)

```
cox_model <- coxph(Surv(time, event) ~ sex+age+ph.ecog, data=df)
cox_model %>% summary
```

- 해석

- 계수 부분
  - sex:
    값이 1 증가 할 때 즉,
    여성이 남성에 비해 위험율이 0.575배 수준
  - age:
    값이 1증가 할 때, 위험율이 1.01배로 증가
    다만, 통계적으로 유의하지 않음
  - ph.ecog:
    값이 1증가 할 때, 위험율이 1.59배

```
>  cox_model %>% summary
Call:
coxph(formula = Surv(time, event) ~ sex + age + ph.ecog, data = df)

  n= 227, number of events= 164
   (결측으로 인하여 1개의 관측치가 삭제되었습니다.)

             coef exp(coef)  se(coef)      z Pr(>|z|)
sex     -0.552612  0.575445  0.167739 -3.294 0.000986 ***
age      0.011067  1.011128  0.009267  1.194 0.232416
ph.ecog  0.463728  1.589991  0.113577  4.083 4.45e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

         exp(coef) exp(-coef) lower .95 upper .95
sex         0.5754     1.7378    0.4142    0.7994
age         1.0111     0.9890    0.9929    1.0297
ph.ecog     1.5900     0.6289    1.2727    1.9864

Concordance= 0.637  (se = 0.025 )
Likelihood ratio test= 30.5  on 3 df,   p=1e-06
Wald test            = 29.93  on 3 df,   p=1e-06
Score (logrank) test = 30.5  on 3 df,   p=1e-06
```

4. Cox 비례 위험 모형(Cox Proportional Hazards Model)

```
cox_model <- coxph(Surv(time, event) ~ sex+age+ph.ecog, data=df)
cox_model %>% summary
```

- 해석
  - 여성이 위험률이 남성의 57.5% 수준일 때,
    1. 남성의 위험에 비해 42.5% 감소한다
    2. 남성의 위험에서 42.5%가 감소한다

    어떤 표현을 써야 할까?

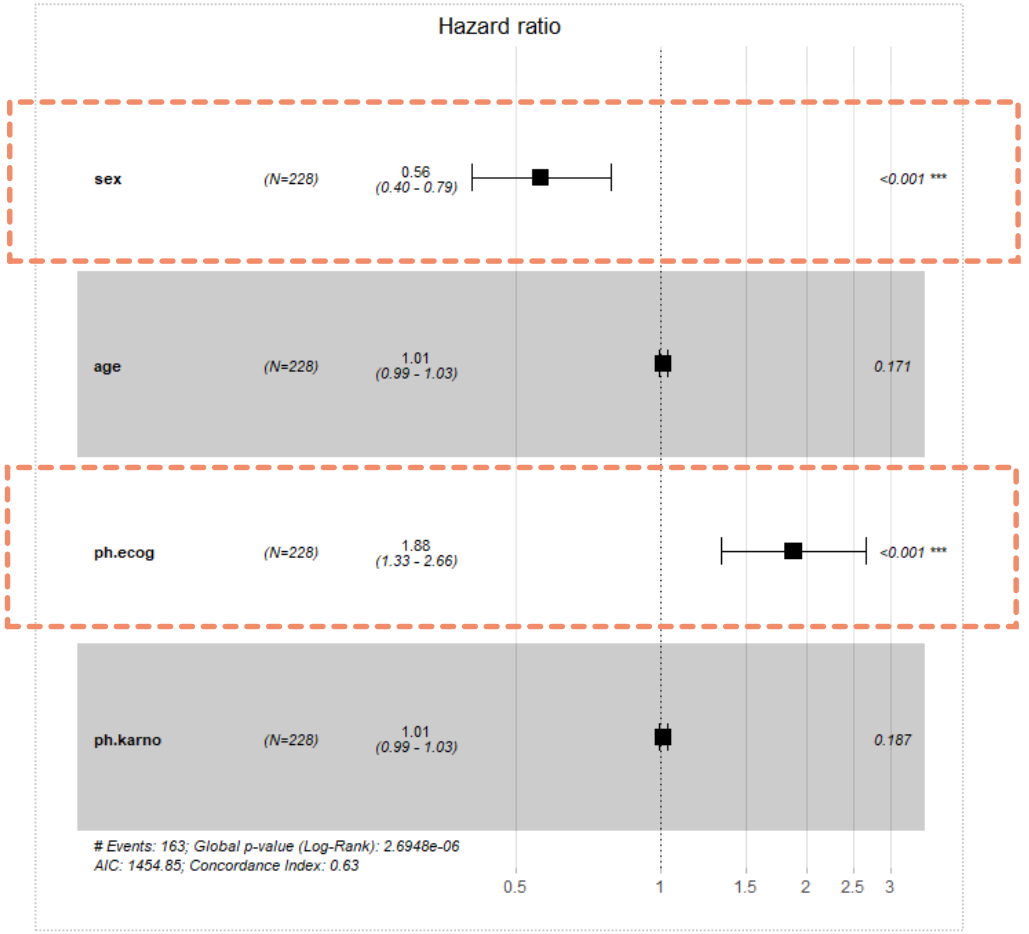
```
>  cox_model %>% summary
Call:
coxph(formula = Surv(time, event) ~ sex + age + ph.ecog, data = df)

  n= 227, number of events= 164
   (결측으로 인하여 1개의 관측치가 삭제되었습니다.)

             coef exp(coef)  se(coef)      z Pr(>|z|)
sex     -0.552612  0.575445  0.167739 -3.294 0.000986 ***
age      0.011067  1.011128  0.009267  1.194 0.232416
ph.ecog  0.463728  1.589991  0.113577  4.083 4.45e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

         exp(coef) exp(-coef) lower .95 upper .95
sex         0.5754     1.7378    0.4142    0.7994
age         1.0111     0.9890    0.9929    1.0297
ph.ecog     1.5900     0.6289    1.2727    1.9864

Concordance= 0.637  (se = 0.025 )
Likelihood ratio test= 30.5  on 3 df,   p=1e-06
Wald test            = 29.93  on 3 df,   p=1e-06
Score (logrank) test = 30.5  on 3 df,   p=1e-06
```

4. Cox 비례 위험 모형(Cox Proportional Hazards Model)

- 추가

계수의 hazard ratio는
forest plot으로 도식화하여 확인 가능

```
ggforest(cox_model,data=df)
```



4. Cox 비례 위험 모형(Cox Proportional Hazards Model)

```
cox_model <- coxph(Surv(time, event) ~ sex+age+ph.ecog, data=df)
cox_model %>% summary
```

- 해석

- 모형 적합도
  - Likelihood ratio test p-value <0.000
    (상수항만 있는 축소모형과 비교)
    Wald test p-value <0.000
    (모든 계수가 0인지 아닌지 검정)
    Score (log rank) test p value <0.000
    (우도함수의 기울기를 이용한 검정)

```
>  cox_model %>% summary
Call:
coxph(formula = Surv(time, event) ~ sex + age + ph.ecog, data = df)

  n= 227, number of events= 164
   (결측으로 인하여 1개의 관측치가 삭제되었습니다.)

              coef exp(coef)  se(coef)      z Pr(>|z|)
sex      -0.552612  0.575445  0.167739 -3.294 0.000986 ***
age       0.011067  1.011128  0.009267  1.194 0.232416
ph.ecog   0.463728  1.589991  0.113577  4.083 4.45e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

         exp(coef) exp(-coef) lower .95 upper .95
sex         0.5754     1.7378    0.4142    0.7994
age         1.0111     0.9890    0.9929    1.0297
ph.ecog     1.5900     0.6289    1.2727    1.9864

Concordance= 0.637  (se = 0.025 )
Likelihood ratio test= 30.5  on 3 df,   p=1e-06
Wald test            = 29.93  on 3 df,   p=1e-06
Score (logrank) test = 30.5  on 3 df,   p=1e-06
```

4. Cox 비례 위험 모형(Cox Proportional Hazards Model)

▪ 비례위험 가정의 검정

Cox 모형의 핵심 가정은 공변수의 효과 즉, 상대적 위험 HR이 시간에 따라 일정하다는 것임

$$h(t|X) = h_0(t)\exp(X\beta)$$

HR(e^{β})이 시간(t)에 의존하지 않아야 함

Schoenfeld 잔차의 시간과의 상관성 평가

$$Z_i = \frac{\sum_{j=1}^{n}(t_j - \bar{t}) \times r_{ij}(t_j)}{\sqrt{\mathrm{Var}\left(\sum_{j=1}^{n}(t_j - \bar{t}) \times r_{ij}(t_j)\right)}}$$

t_j : j번째 사건 발생 시점

$\bar{t}$: 사건 발생 시점들의 평균

r_j : 사건 발생 시점 j에서의 schoenfeld 잔차 $r_{ij} = X_{ij}(t_j) - E\left(X_{ij}(t_j)\right)$

4. Cox 비례 위험 모형(Cox Proportional Hazards Model)

▪ 비례위험 가정의 검정

cox.zph() 함수로 비례위험 가정에 대한 검정을 시행

```
>  cox.zph(cox_model)
          chisq df    p
sex       2.305  1 0.13
age       0.188  1 0.66
ph.ecog   2.054  1 0.15
GLOBAL    4.464  3 0.22
```

H_0: 변수의 위험비(HR)은 시간에 의존하지 않는다.
즉, 비례위험 가정을 만족한다. $HR(t) = HR(\text{constant})$

vs

H_1: 변수의 위험비는 시간에 따라 변한다.
즉, 비례위험 가정을 위반한다.

검정 결과 p-값은 전부 0.05 이상으로,

유의수준 0.05 하에서 모든 변수가 비례위험 가정을 만족한다고 말할 수 있음

4. Cox 비례 위험 모형(Cox Proportional Hazards Model)

- 비례위험 가정의 검정
- 비례위험 가정에 위배되는 경우

  이전 모형에 ph.karno 변수를 추가하였을 때 해당 변수가 비례위험 가정의 검정 결과에서 귀무가설을 기각함

- 해결방안은?

```
>  cox_model <- coxph(Surv(time, event) ~
+                     sex+age+ph.ecog+ph.karno, data=df)
>  cox_model %>% summary
Call:
coxph(formula = Surv(time, event) ~ sex + age + ph.ecog + ph.karno,
    data = df)

  n= 226, number of events= 163
   (결측으로 인하여 2개의 관측치가 삭제되었습니다.)

               coef exp(coef)  se(coef)      z Pr(>|z|)
sex       -0.572802  0.563943  0.169222 -3.385 0.000712 ***
age        0.012868  1.012951  0.009404  1.368 0.171226
ph.ecog    0.633077  1.883397  0.176034  3.596 0.000323 ***
ph.karno   0.012558  1.012637  0.009514  1.320 0.186842
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
>  cox.zph(cox_model)
            chisq df    p
sex        1.7133  1 0.19
age        0.0668  1 0.80
ph.ecog    1.6902  1 0.19
ph.karno   5.4535  1 0.02
GLOBAL     7.3678  4 0.12
```

4. Cox 비례 위험 모형(Cox Proportional Hazards Model)

- 비례위험 가정의 검정

```
cox_model <- coxph(Surv(time, event) ~
                   sex+age+ph.ecog+strata(ph.karno), data=df)
cox_model %>% summary
```

- 비례위험 가정을 깨는 변수를 층화

```
>  cox_model %>% summary
Call:
coxph(formula = Surv(time, event) ~ sex + age + ph.ecog + strata(ph.karno),
    data = df)

  n= 226, number of events= 163
   (결측으로 인하여 2개의 관측치가 삭제되었습니다.)

             coef exp(coef)  se(coef)      z Pr(>|z|)
sex     -0.591683  0.553395  0.173049 -3.419 0.000628 ***
age      0.014807  1.014918  0.009651  1.534 0.124972
ph.ecog  0.483879  1.622356  0.196558  2.462 0.013825 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
>  cox.zph(cox_model)
          chisq df    p
sex        1.52  1 0.22
age        0.46  1 0.50
ph.ecog    0.31  1 0.58
GLOBAL     2.05  3 0.56
```

#비례위험 가정은 지켰지만
해석은 어떻게 해야할까?

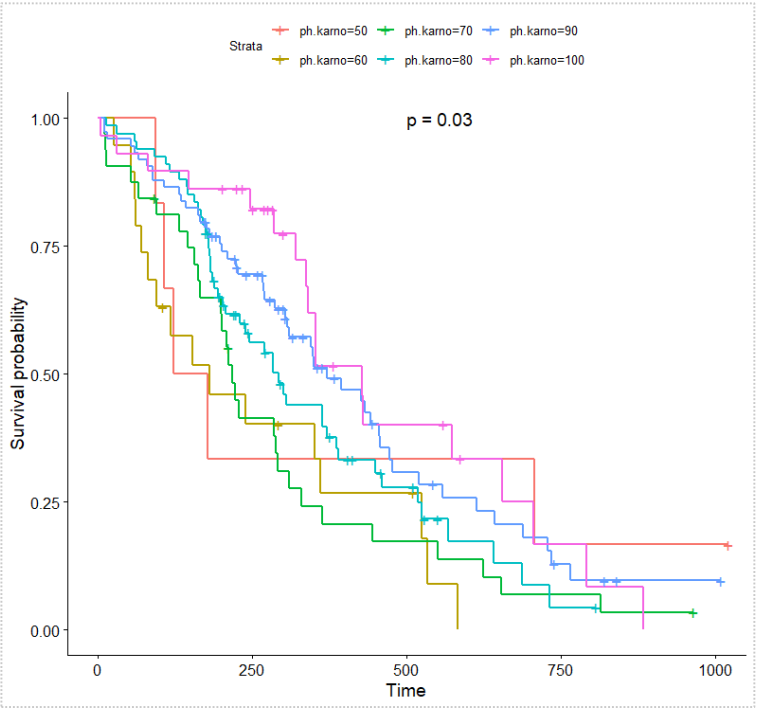
4. Cox 비례 위험 모형(Cox Proportional Hazards Model)

- 비례위험 가정의 검정

- 비례위험 가정을 깨는 변수를 층화

  본 모형에서 해당 변수는 “다른 변수들의 영향을 통제한 상태에서 각 층 별로 서로 독립된 기준 위험함수를 가진다” 로 해석
  해당 변수의 상대위험도(HR)를 수치로 명시하는 것은 불가능
  다만 K-M 그래프를 통해 각 층 별 시간에 따른 추이를 볼 수 있음

```
fit <- survfit(Surv(time, event) ~ ph.karno, data = df)
ggsurvplot(fit, data=df, pval=T, pval.coord=c(500,1))
```



감사합니다

Q&A

충북대학교
CHUNGBUK NATIONAL UNIVERSITY