



# 통계실습 II. 생존분석

---

가천대 길병원 고급통계교육

방태모

2022-06-21

# 목차

- 기초개념
- 생존 함수 추정
  - Kaplan-Meier method
  - 실습
- 생존 함수 비교
  - Log rank test
  - 실습
- 위험인자 분석
  - Cox proportional hazard model
  - Time dependent Cox proportional hazard model
  - 실습

# 기초개념

## 생존 분석

- 생존자료(survival data)를 분석하는 통계적 방법

## 생존 자료

- 관심 사건의 발생 여부(이진형 변수)와 생존시간(연속형 변수)을 관측한 자료
  - 생존자료를 사건-시간 자료(time-to-event data)라 표현하기도 함
  - 실제 생존 시간(true survival time,  $T$ ) = 추적 시작 시점 ~ 사건 발생 지점
- 생존 자료의 가장 큰 특징: 중도 절단(censored)
  - 중도 절단 시간 (censoring time,  $C$ )
    - 연구기간 내 사건 미발생 시 마지막 추적 시점까지의 시간
  - 관측 생존 시간 (observed survival time) =  $\min(T, C)$ 
    - 연구기간 내 생존 시간 또는 중도 절단 시간까지의 시간

# 생존 자료의 특징

- 생존 시간은 정규분포를 따르지 않음
  - 항상 양의 값
  - 매우 치우친 분포
- 중도 절단
  - Type I censoring
  - Type II censoring
  - Type III censoring
    - 연구종료
    - loss to follow up
    - drop out
- 이진형 변수와 연속형 변수를 다루어야 함
  - 생존 분석 필요

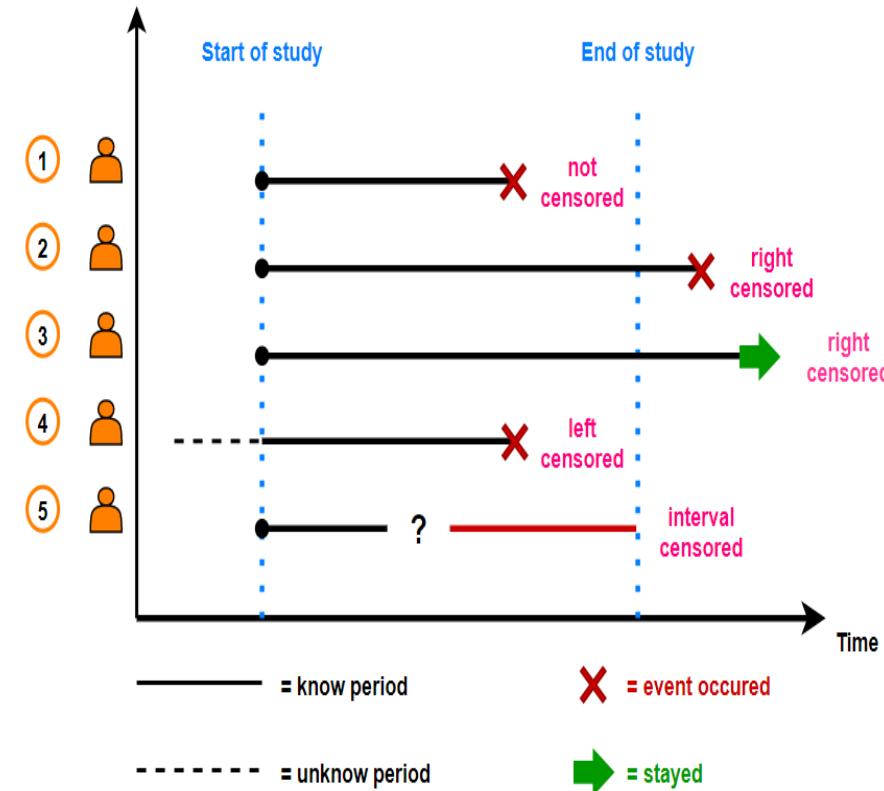


fig 1. Survival data (Jim Gruman)

# 생존 분석의 목적

- 추정(estimation)
  - 생존시간의 분포에 관한 정보
  - Kaplan-Meier Method
- 그룹간 비교(group comparison)
  - 실험군과 대조군이 있는 경우 두 처리군의 생존 분포 비교
  - Log-rank test
- 위험인자 분석(risk-factor analysis)
  - 유의한 예후인자(prognostic factor) 파악
  - Cox ph model
  - Time dependent Cox ph model

# 생존 함수 추정

## Kaplan-Meier method

# 생존 함수 $S(t)$

$$S(t) = Pr(T > t) \quad (1)$$

- 관심사건이  $t$  시점까지 일어나지 않을 확률
  - 관심사건: 사망
  - 환자가  $t$ 시간 이상 생존할 확률
- $S(t)$ : 생존율, 생존 곡선
- $F(t) = 1 - S(t)$ : 누적 발생률
- e.g.  $S(10) = 0.1$  ( $t$  in years)
  - 10년 생존율 =  $S(10) = 10\%$
  - 10년 누적 사망률 =  $F(10) = 90\%$

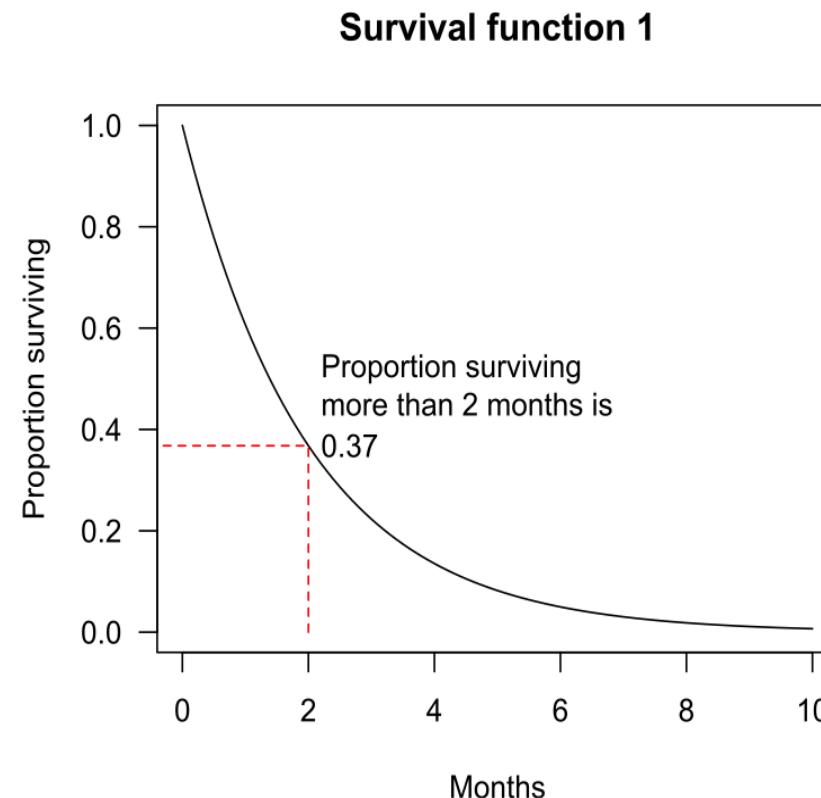


fig 2. Survival function (Wikipedia)

# Kaplan-Meier method

- 생존함수를 추정하는 대표적인 비모수적 방법론:

$$\hat{S}(t_j) = \prod_{i=1}^j \left(1 - \frac{d_i}{n_i}\right)^{\delta_i} \quad (2)$$

- $t_j$ :  $j$ -번째 사망 시간
- $d_i$ :  $t_i$  시점 사망자 수
- $n_i$ :  $t_i$  시점 바로 직전 사망자 수
- $\delta_i$ : 중도절단여부(중도절단 = 0, 아니면 1)
- 즉, 사건 발생 시점에서만 생존율 변화

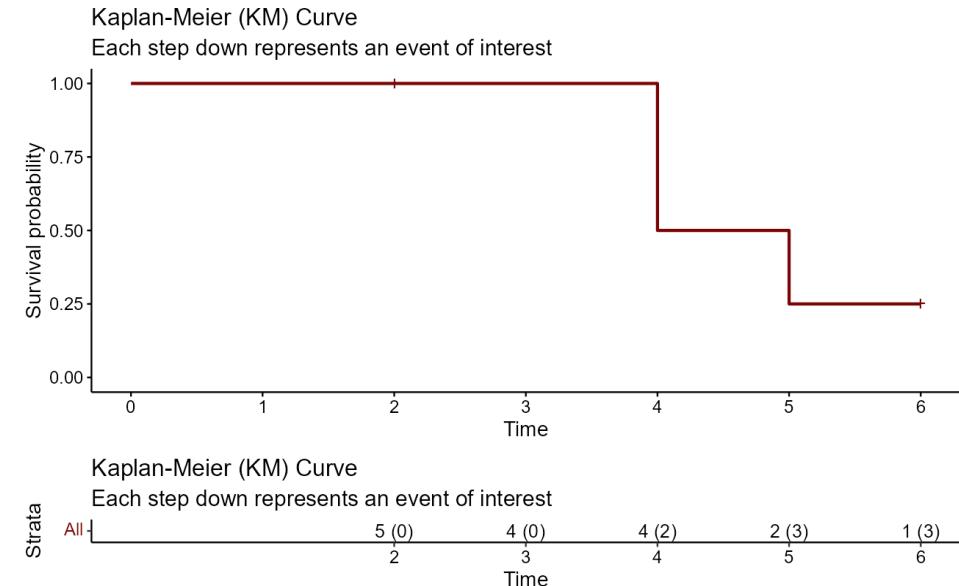


fig 3. KM method (Jim Gruman)

# 실습 - Kaplan-Meier method

## 패키지 설치

```
install.packages("dplyr")
install.packages("survival")
install.packages("survminer")
```

## 패키지 로딩

```
library(dplyr)
library(survival)
library(survminer)
```

## 데이터 소개

- 228명 말기 암환자들의 생존 시간(time), 사망 여부(status) 데이터
- 1022일간 추적 연구(follow-up study)

# 실습 - Kaplan-Meier method

## 생존 자료 객체 만들기

```
surv_lung <- Surv(lung$time, lung$status)[1:10]  
head(surv_lung)
```

```
## [1] 306 455 1010+ 210 883 1022+
```

## Kaplan-Meier method을 이용한 생존 곡선 추정

```
method_km <- survfit(Surv(time, status) ~ 1, data = lung)  
glimpse(method_km)
```

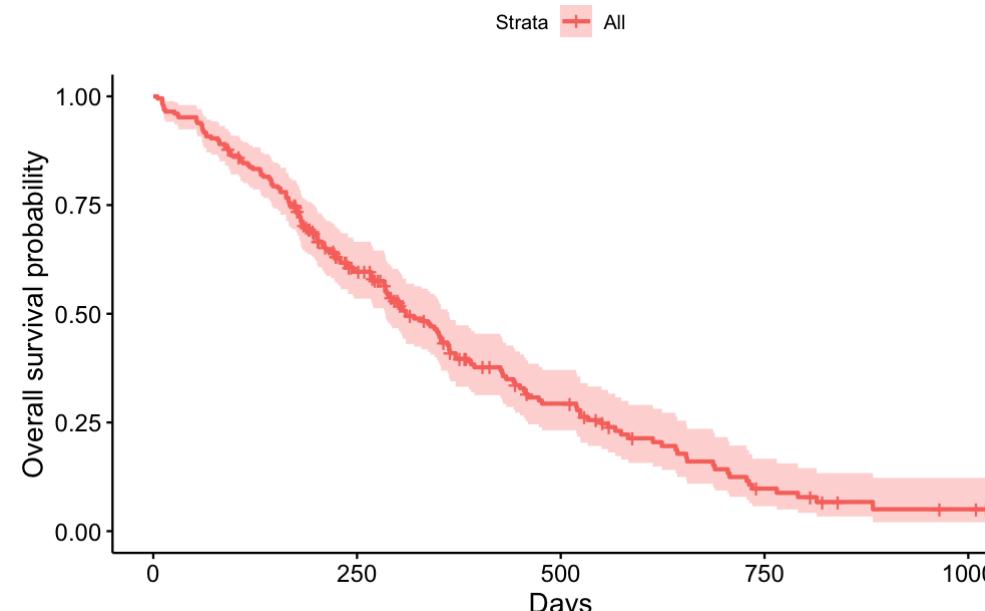
```
## #> #> ## List of 16  
## #> $ n : int 228  
## #> $ time : num [1:186] 5 11 12 13 15 26 30 31 53 54 ...  
## #> $ n.risk : num [1:186] 228 227 224 223 221 220 219 218 217 215 ...
```

# 실습 - Kaplan-Meier method

## Kaplan-Meier plot 시각화

- {survminer} 패키지의 `ggsurvplot()` 이용
- 🔗 [cheatsheet](#)

```
ggsurvplot(method_km, xlab = "Days", ylab = "Overall survival probability")
```



# 실습 - Kaplan-Meier method

## 1년 생존율 추정하기

- 추정된 KM curve에 의하면 말기 암환자의 추정된 1년 생존율은 41%에 해당:

```
summary(method_km, times = 365.25)
```

```
## Call: survfit(formula = Surv(time, status) ~ 1, data = lung)
##
##    time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    365     65      121     0.409  0.0358      0.345      0.486
```

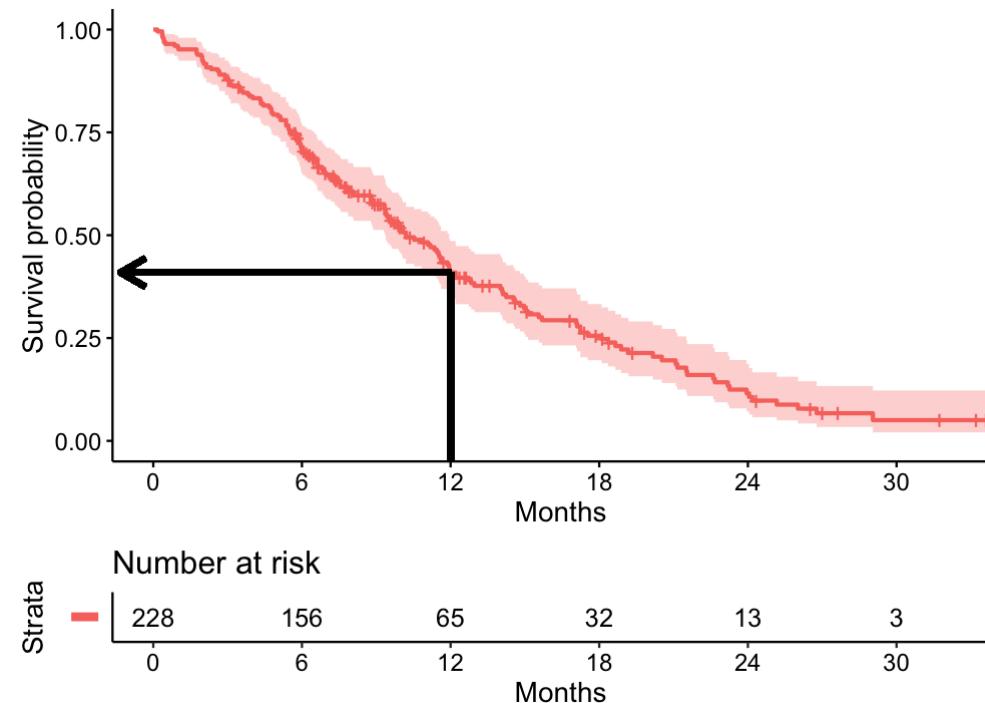
- censoring 고려없이 해당 시점의 사망자 수만 고려하여 생존율을 계산하면 생존율이 과대추정 됨:

$$\left(1 - \frac{121}{228}\right) \times 100 = 47\% \quad (3)$$

# 실습 - Kaplan-Meier method

## 1년 생존율 추정하기 - 시각화

- Number at risk: 해당 시점의 생존자 수



- 총 환자 수 228명 - 사망 121명 - 중도절단 42명 = 65명 생존 (1년 생존자 수)

# 실습 - Kaplan-Meier method

## 1년 생존율 추정하기 - 시각화 코드

```
plot_main <-
  ggsurvplot(
    data = lung,
    fit = method_km,
    xlab = "Months",
    legend = "none",
    xscale = 30.4,
    break.x.by = 182.4,
    risk.table = TRUE,
    risk.table.y.text = FALSE)
plot1 <- plot_main
plot1$plot <- plot1$plot +
  geom_segment(x = 365.25, xend = 365.25, y = -0.05, yend = 0.4092416,
               size = 1.5) +
  geom_segment(x = 365.25, xend = -40, y = 0.4092416, yend = 0.4092416,
```

# 실습 - Kaplan-Meier method

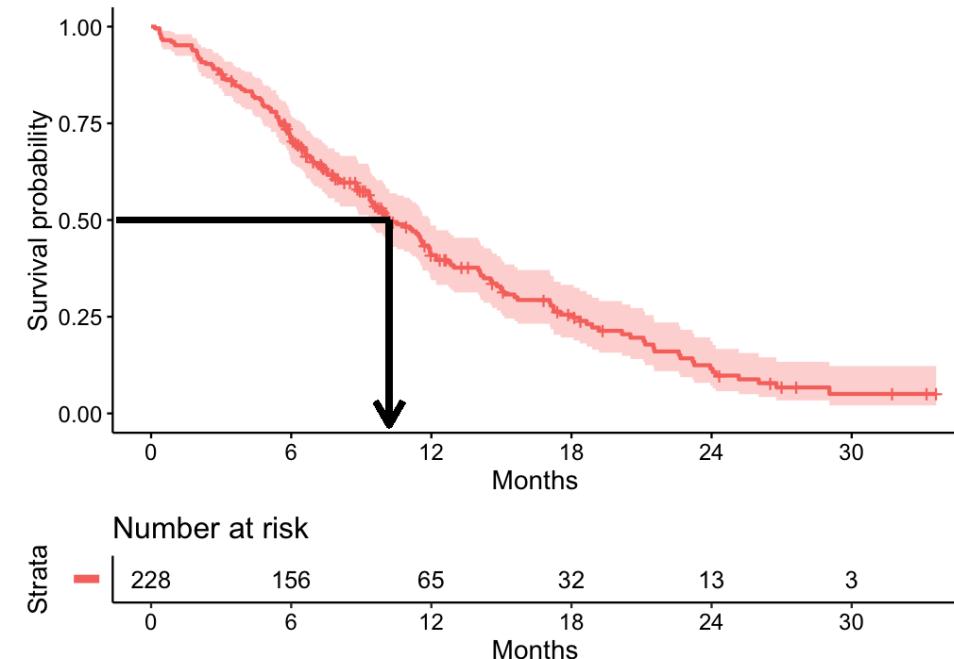
## 중위 생존시간 추정

method\_km

```
## Call: survfit(formula = Surv(time, status))
## 
##      n  events median 0.95LCL 0.95UCL
## [1,] 228     165     310     285     363
```

- 말기 암환자 중위 생존시간: 310일

## 중위 생존시간 추정 - 시각화



# 실습 - Kaplan-Meier method

## 중위생존시간 - 시각화 코드

```
plot2 <- plot_main
plot2$plot <- plot2$plot +
  geom_segment(x = -45, xend = 310, y = 0.5, yend = 0.5, size = 1.5) +
  geom_segment(x = 310, xend = 310, y = 0.5, yend = -0.03, size = 1.5,
               arrow = arrow(length = unit(0.2, "inches")))
plot2
```

# 생존 함수 비교

Log-rank test

# Log-rank test

- 그룹 간 생존 곡선을 비교하는 대표적인 비모수 검정 방법
- 두 그룹 간 생존 함수에 차이가 있는지 살펴보자 함
- 가설

$$H_0 : S_1(t) = S_2(t) \quad H_1 : S_1(t) \neq S_2(t)$$

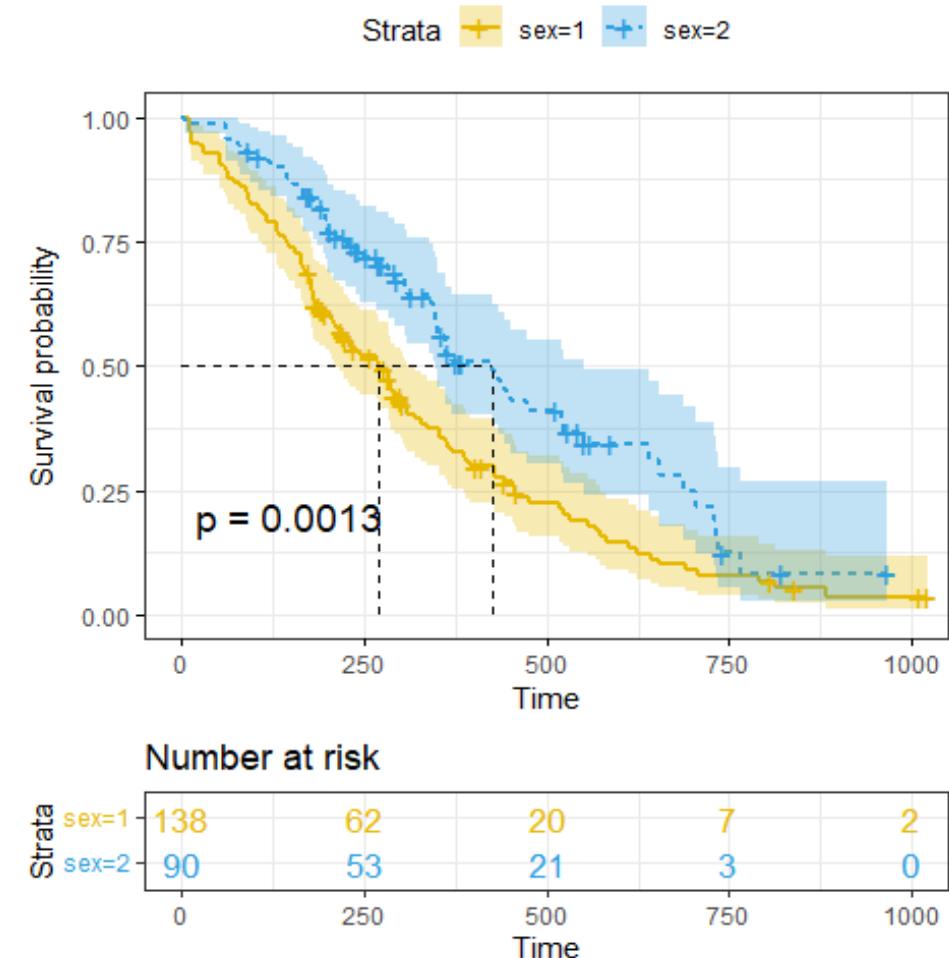


fig 4. Log-rank test

## 실습 - Log-rank test

- 앞서 이용한 패키지 {survival}, {survminer} 그대로 이용
- KM method를 통해 추정한 말기 암환자들의 생존 곡선을 성별 그룹에따라 비교

```
method_km2 <- survfit(Surv(time, status) ~ sex, data = lung)  
method_km2
```

```
## Call: survfit(formula = Surv(time, status) ~ sex, data = lung)  
##  
##      n events median 0.95LCL 0.95UCL  
## sex=1 138     112     270     212     310  
## sex=2   90      53     426     348     550
```

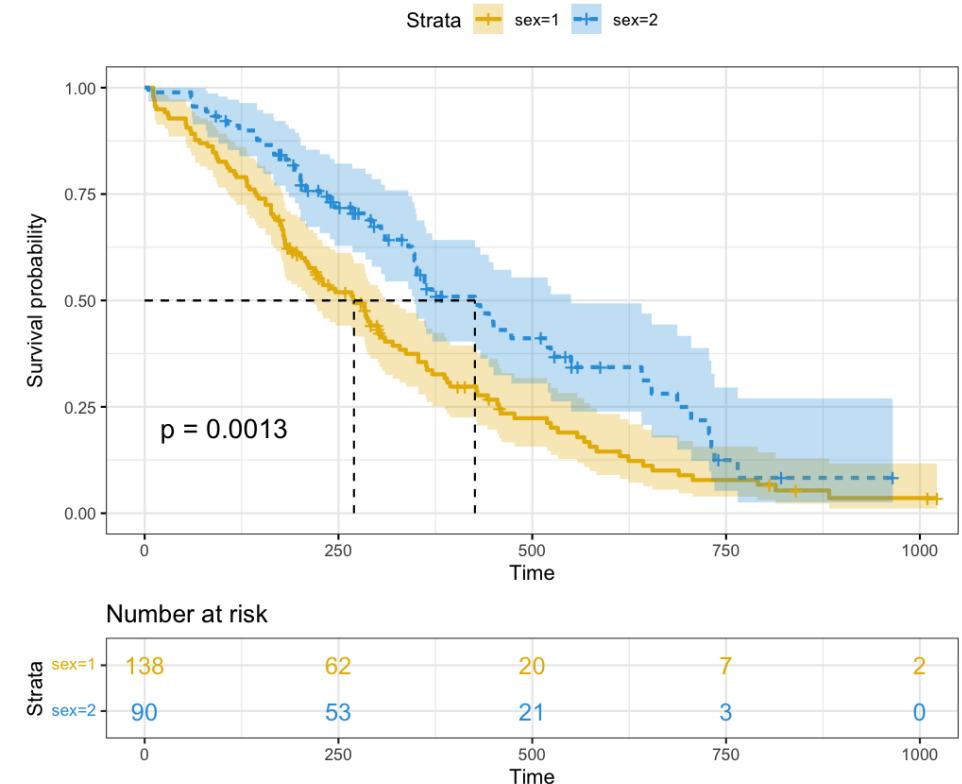
- 중위생존시간은 여성이 훨씬 긴

# 실습 - Log-rank test

## 시각화

```
ggsurvplot(method_km2,
            pval = TRUE, conf.int = TRUE,
            risk.table = TRUE, # Add risk table
            risk.table.col = "strata", # Change color
            linetype = "strata", # Change line type
            surv.median.line = "hv", # Specified median line
            ggtheme = theme_bw(), # Change ggplot2 theme
            palette = c("#E7B800", "#2E9FDF")
```

- 두 점선은 각 그룹의 중위생존시간을 의미
- $p < 0.01$



# 위험인자 분석

Cox proportional hazard model

Time dependent Cox proportional hazard model

## 위험함수 $h(t)$

$$h(t) = \lim_{dt \rightarrow 0} P(t < T \leq t + dt) \quad (4)$$

- 환자가  $t$  시점까지 생존했다가  $t$  시점 바로 직후에 사망하게 되는 순간 위험률

## Cox ph model

- 생존 자료에 대한 대표적인 준모수적 회귀 분석 방법론
  - Cox regression 이라고도 부름
- 비례 위험 가정 (proportinal hazard assumption)
  - 각 설명변수의 효과는 시간에 관계없이 일정

$$h(t) = h_0(t)\exp(\beta_1x_1 + \beta_2x_2 + \cdots + \beta_kx_k) \quad (5)$$

$$h_0(t) = \exp(\beta_0) : \text{baseline hazard} \quad (6)$$

# Cox ph model

- 위험비 HR (Hazard ratio)
  - 다른 설명변수들의 값이 고정되어 있을 때,  $x_k$ 의 효과
  - 시간에 대해 상수이므로  $x_k$ 에 대해서만 비례하여(proportional) 증가
  - 즉,  $x_k$ 가 한 단위 증가 시 사건 발생 위험은  $\exp(\beta_k)$ 배가 됨

$$\text{HR}_k = \frac{h(t|X_k = 1)}{h(t|X_k = 0)} = \frac{h_0(t)\exp(\beta_1x_1 + \cdots + \beta_kx_k)}{h_0(t)\exp(\beta_1x_1 + \cdots + \beta_{k-1}x_{k-1})} = \exp(\beta_k) \quad (7)$$

- Cox ph model은 생존 자료에서 다른 변수들의 효과를 보정한, treatment 효과를 볼 수 있는 가장 대표적인 통계 모형에 해당
- Crude HR: Univariable Cox ph model에서 구한 HR
- Adjusted HR: Multivariable Cox ph model에서 구한 HR

## 비례위험가정 검토

- Schoenfeld test
  - Schoenfeld residuals와 시간  $t$  간에 어떤 패턴이 존재하는가?
  - 시간을 설명변수, 잔차를 종속변수로 두고 단순회귀분석을 수행하는 개념
  - 설명변수별로 검토
  - 귀무가설(  $H_0$  ): 잔차와 시간 간 패턴이 존재하지 않음
  - 즉, 귀무가설을 기각할 수 없다면 비례위험가정을 만족함

# 실습 - Cox ph model

## 모형 적합

- {survival} 패키지의 `coxph()`를 이용해 univariable cox ph model 적합

```
mod_cox <- coxph(Surv(time, status) ~ sex, data = lung)  
mod_cox
```

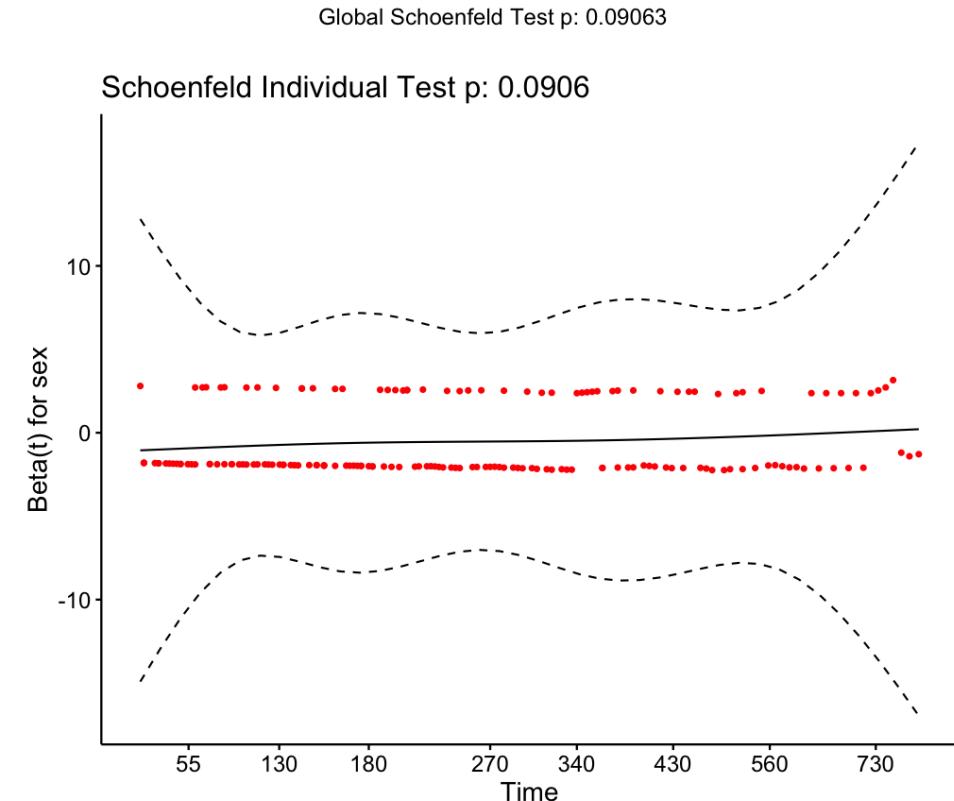
```
## Call:  
## coxph(formula = Surv(time, status) ~ sex, data = lung)  
##  
##      coef exp(coef) se(coef)     z      p  
## sex -0.5310    0.5880   0.1672 -3.176 0.00149  
##  
## Likelihood ratio test=10.63  on 1 df, p=0.001111  
## n= 228, number of events= 165
```

# 실습 - Cox ph model

## 비례위험가정 검토

- cox.zph() 함수를 이용하면 비례위험가정 검토를 손쉽게 수행할 수 있음
- sex의 경우 비례위험가정을 만족한다고 볼 수 있음

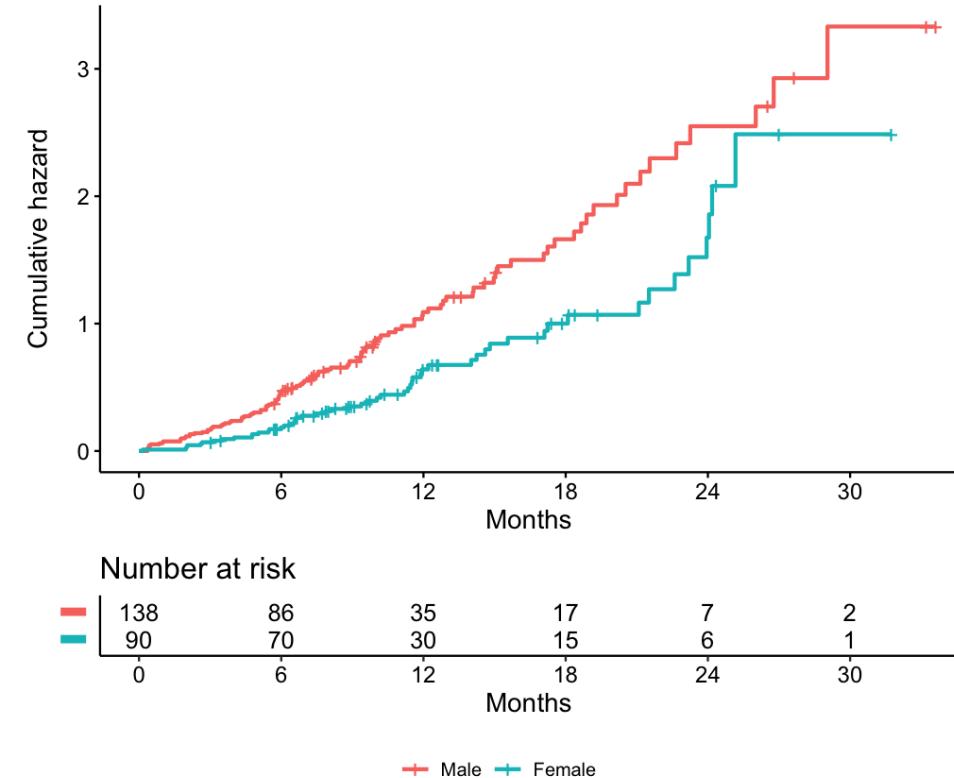
```
mod_cox |>
  cox.zph() |>
  ggcoxzph()
```



# 실습 - Cox ph model

## 누적 위험 시각화

```
mod_hr <- survfit(Surv(time, status) ~ sex,
ggsurvplot(data = lung,
            fit = mod_hr,
            xlab = "Months",
            xscale = 30.4,
            break.x.by = 182.4,
            fun = "cumhaz",
            legend.title = "",
            legend.labs = c("Male", "Female",
            legend = "bottom",
            risk.table = TRUE,
            risk.table.y.text = FALSE)
```



# Time dependent Cox ph model

- Cox ph model의 비례위험가정 위배시 고려해야하는 모형
- 편의상 1개의 설명변수를 갖는다고 가정

$$h(t) = h_0(t)\exp(\beta_1x + \beta_2xt) \quad (8)$$

- Time dependent hazard ratio

$$\text{HR}(t) = \frac{h(t|X = x+1)}{h(t|X = x)} = \frac{\exp(\beta_1(x+1) + \beta_2((x+1)t))}{\exp(\beta_1x + \beta_2(xt))} \quad (9)$$

$$= \exp(\beta_1 + \beta_2t) \quad (10)$$

- $x$ 가 한단위 증가시 HR은  $\exp(\beta_1 + \beta_2t)$ 배가 됨
- 즉, Time dependent Cox ph model의 위험비는  $x$  뿐만이 아닌 시간에도 depend

# 실습 - Time dependent Cox ph model

## 데이터 불러오기

- {SemiCompRisks} 패키지의 137명의 골수 이식 환자 데이터 BMT 이용

```
install.packages("SemiCompRisks")
data(BMT, package = "SemiCompRisks")
glimpse(BMT)
```

```
## #> #> Rows: 137
## #> #> Columns: 22
## #> #> $ g      <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## #> #> $ T1     <int> 2081, 1602, 1496, 1462, 1433, 1377, 1330, 996, 226, 1199, 1111, ...
## #> #> $ T2     <int> 2081, 1602, 1496, 1462, 1433, 1377, 1330, 996, 226, 1199, 1111, ...
## #> #> $ delta1 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, ...
## #> #> $ delta2 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, ...
## #> #> $ delta3 <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, ...
## #> #> $ TA     <int> 67, 1602, 1496, 70, 1433, 1377, 1330, 72, 226, 1199, 1111, 38, ...
```

# 실습 - Time dependent Cox ph model

## 전처리

- aGVHD 여부를 Time dependent covariate으로 모형 적합을 수행할 예정
- R에서 Time dependent Cox ph의 적합에는 특별한 형태의 데이터를 요구함
  - 1 행 번호 생성
  - 2 tmerge() 이용 특별한 형태의 데이터 생성

---

```
install.packages("tibble")
bmt2 <- BMT2 |>
  tibble::rowid_to_column("my_id")
bmt_time <- tmerge(
  data1 = bmt2 |> select(my_id, T1, delta1),
  data2 = bmt2,
  id = my_id,
  death = event(T1, delta1),
  agvhd = tdc(TA)
```

# 실습 - Time dependent Cox ph model

## 전처리

- 원 데이터

```
head(bmt2, 5)
```

```
##   my_id    T1 delta1    TA deltaA
## 1     1 2081      0    67      1
## 2     2 1602      0 1602      0
## 3     3 1496      0 1496      0
## 4     4 1462      0    70      1
## 5     5 1433      0 1433      0
```

- tmerge()에 의해 변환된 데이터

```
head(bmt_time, 5)
```

```
## # A tibble: 5 × 7
##   my_id    T1 delta1 tstart  tstop death
##   <int> <int>  <dbl>  <dbl> <int> <dbl>
## 1     1 2081      0      0    67     0
## 2     1 2081      0     67 2081     0
## 3     2 1602      0      0 1602     0
## 4     3 1496      0      0 1496     0
## 5     4 1462      0      0    70     0
```

# 실습 - Time dependent Cox ph model

## 모형 적합

- Cox ph model과 동일한 함수 `coxph()`로 적합 가능

```
coxph(  
  Surv(time = tstart, time2 = tstop, event = death) ~ agvhd,  
  data = bmt_time  
 ) %>%  
  gtsummary::tbl_regression(exp = TRUE)
```

Characteristic	HR <sup>1</sup>	95% CI <sup>1</sup>	p-value
agvhd	1.40	0.81, 2.43	0.2

<sup>1</sup> HR = Hazard Ratio, CI = Confidence Interval