Statistical Inference of Interval-censored Failure Time Data

Jinheum Kim (Univ. Suwon)

2015. 1. 22

OUTLINE

생존함수 추정
생존함수 비교
회귀모형
다변량 생존자료분석

생존분석이란?
중도절단과 절삭
누적한계추정량
좌절삭된 자료의 생존함수 추정
구간중도절단된 자료의 생존함수 추정
좌중도절단된 자료의 생존함수 추정

생존함수 추정

생존분석이란?
중도절단과 절단
누적한계추정량
좌절삭된 자료의 생존함수 추정
구간중도절단된 자료의 생존함수 추정
좌중도절단된 자료의 생존함수 추정

생존함수 추정

What is survival analysis?

- Outcome variable: Time until an event occurs (T)
- Time origin
  - eg, birth date, occurrence of entry into a study or diagnosis of a disease
- Time
  - eg, years, months, weeks or days
- Event
  - eg, death, disease incidence, relapse from remission…

생존분석이란?
중도절단과 절삭
누적한계추정량
좌절삭된 자료의 생존함수 추정
구간중도절단된 자료의 생존함수 추정
좌중도절단된 자료의 생존함수 추정

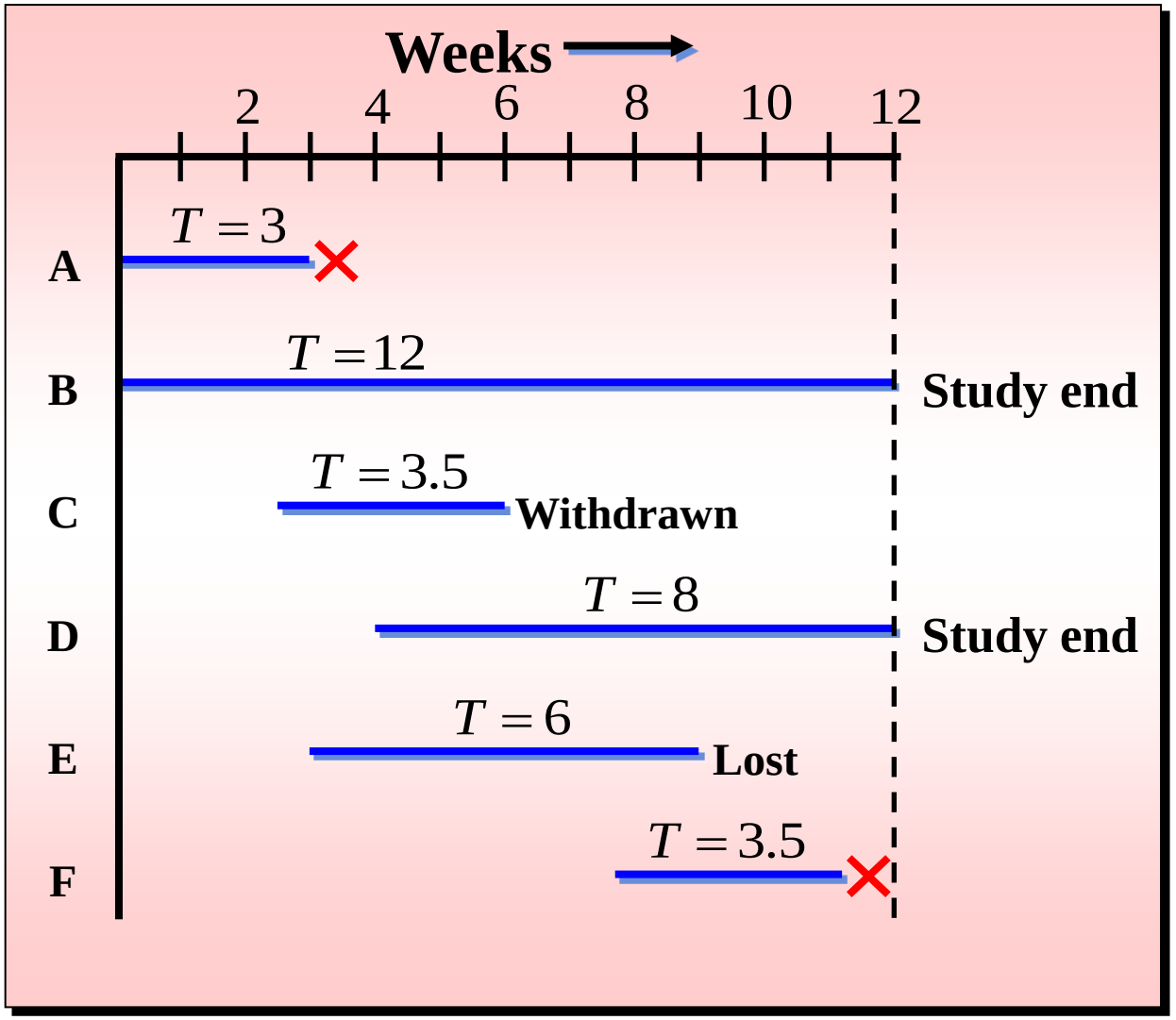
생존함수 추정

Censoring

- Censoring: Don’t know survival time exactly

- Why censoring may occur?
  - No event before the study ends
  - Lost to follow-up
  - Withdrawn from the study

A hypothetical example



Another types of censoring

- Left censoring: $T \in (0, c)$, observed to fail prior to c
  - eg, Time to first use of marijuana
    - Q: When did you first use marijuana?
    - A: exact age, "I never used it." or "I have used it but can not recall just when the first was."
- Double censoring
- Interval censoring: $T \in (a, b]$
  - eg, Time to cosmetic deterioration of breast cancer patients

Censoring *vs.* Truncation

- When occurs?
  - Only those individuals whose event time lies within a certain observational window (Y_L, Y_R) are observed
- In contrast to censoring where there is at least partial information on each subject
- Left truncation
  - When $Y_R = \infty$
  - eg, Life lengths of elderly residents of a retirement community
- Right truncation
  - When $Y_L = 0$
  - eg, Waiting time from infection at transfusion to clinical onset of AIDS (sampled on June 30, 1986)

생존분석이란?
중도절단과 절삭
누적한계추정량
좌절삭 된 자료의 생존함수 추정
구간중도절단된 자료의 생존함수 추정
좌중도절단된 자료의 생존함수 추정

생존함수 추정

Survivor function

- (definition) $S(t) = P(T > t)$
  - Probability that a person survives longer than t
- (properties)
  - Non-increasing
  - $S(0) = 1$
  - $S(\infty) = 0$
    - Eventually nobody would survive

Hazard function

- (definition) $\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$
  - instantaneous potential per unit time for the event to occur, given that the individual has survived up to t
- (properties)
  - Non-negative
  - No upper bound

Product-limit (or Kaplan-Meier) estimator

- $t_1 < t_2 < \cdots < t_k$: distinct observed failure times
  - Conventionally $t_0 = 0, t_{k+1} = \infty$
- $d_j (j = 0, \ldots, k)$: # of individuals who fail at t_j
- $m_j (j = 0, \ldots, k)$: # of individuals censored at $[t_j, t_{j+1})$
- $n_j = (d_j + m_j) + \cdots + (d_k + m_k)$: # of individuals at risk just prior to t_j
- $\hat{S}(t) = \prod_{j|t_j \le t} \left(1 - \frac{d_j}{n_j}\right)$: PL estimator

Remarks on PL estimator

- Never reduce to zero if $m_k > 0$
  - Not defined for $t >$largest time recorded
- (estimated asymptotic variance)
  - $\mathrm{V}\left(\hat{S}(t)\right) = \hat{S}(t)^2 \sum_{j|t_j \le t} \frac{d_j}{n_j(n_j - d_j)}$: Greenwood's formula
- Pointwise 95% confidence interval for $S(t)$
  - $\hat{S}(t) \pm 1.96 \times se\left(\hat{S}(t)\right)$
  - Linear and symmetric, but possibly lies out of (0,1) and low coverage rate with very small samples
- Nelsen-Aalen estimator for cumulative hazard rate $\Lambda(t)$
  - $\widehat{\Lambda}(t) = \sum_{j|t_j \le t} \frac{d_j}{n_j}$

AML data

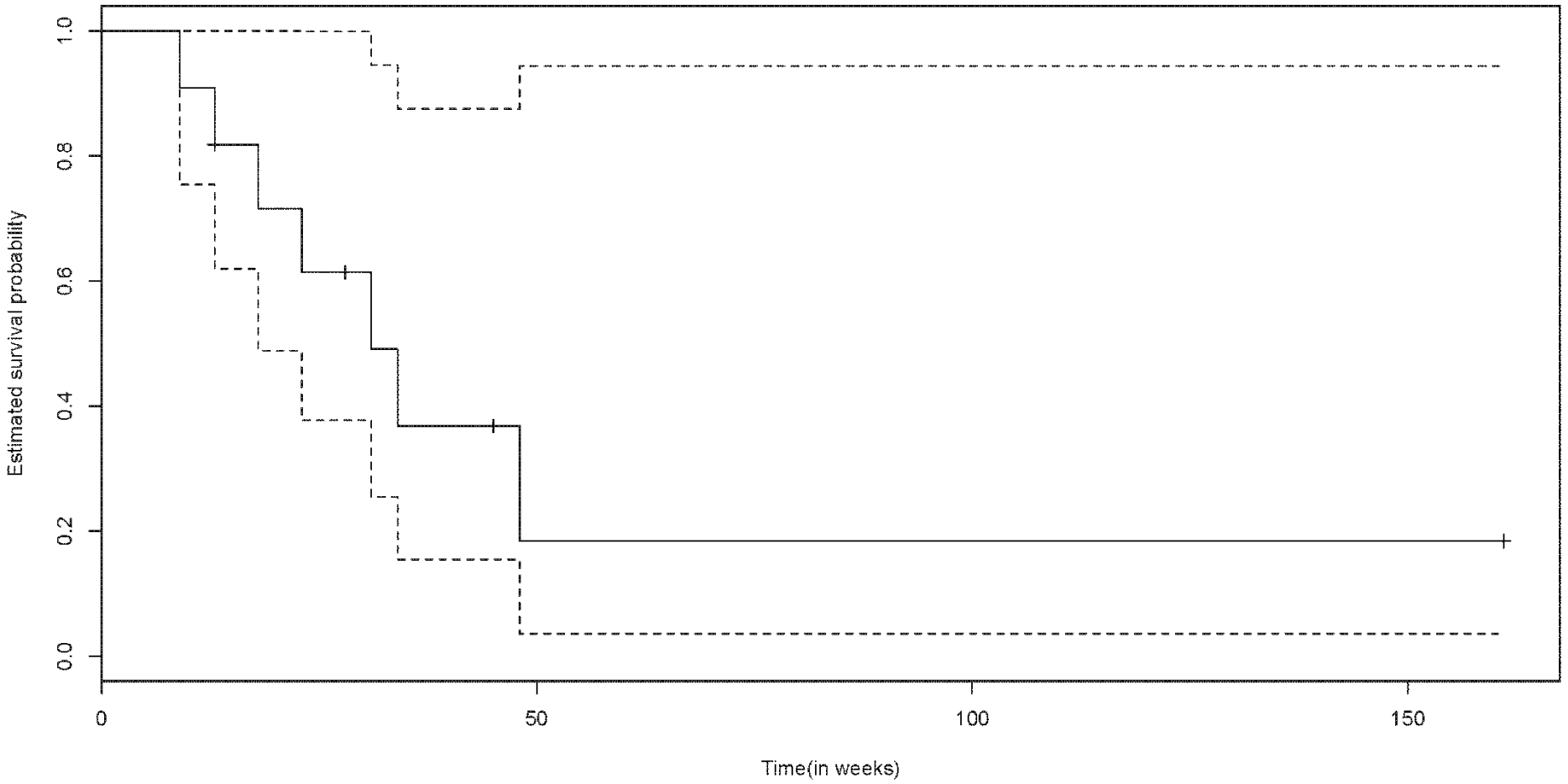
- 화학요법이 백혈병 환자들의 재발시간을 연장하는가? (Embury 등, WesternJMed, 1977)
- R codes
  - Chemo=c(9,13,13,18,23,28,31,34,45,48,161)
  - Non.Chemo=c(5,5,8,8,12,16,23,27,30,33,43,45)
  - Time=c(Chemo,Non.Chemo)
  - C.Status=c(1,1,0,1,1,0,1,1,0,1,0)
  - Non.C.Status=c(1,1,1,1,1,0,1,1,1,1,1,1)
  - Status=c(C.Status,Non.C.Status)
  - Group=rep(c("Chemo","Non.Chemo"),c(length(Chemo),length(Non.Chemo)))
  - aml=data.frame(Time=Time,Status=Status,Group=Group)
  - fit=survfit(Surv(Time,Status)~Group, data=subset(aml,Group=="Chemo"))
  - summary(fit)
  - plot(fit, xlab="Time(in weeks)", ylab="Estimated survival probability")

AML data

```
> summary(fit)
Call: survfit(formula = Surv(Time, Status) ~ Group, data = subset(aml,
    Group == "Chemo"))

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    9     11       1    0.909  0.0867        0.7541        1.000
   13     10       1    0.818  0.1163        0.6192        1.000
   18      8       1    0.716  0.1397        0.4884        1.000
   23      7       1    0.614  0.1526        0.3769        0.999
   31      5       1    0.491  0.1642        0.2549        0.946
   34      4       1    0.368  0.1627        0.1549        0.875
   48      2       1    0.184  0.1535        0.0359        0.944
```

AML data


Estimated survival probability
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
Time(in weeks)

생존분석이란?
중도절단과 절삭
누적한계추정량
좌절삭된 자료의 생존함수 추정
구간중도절단된 자료의 생존함수 추정
좌중도절단된 자료의 생존함수 추정

생존함수 추정

Bone marrow transplantation data

- 재발과 사망이 주요 관심사건인데 intermediate event (platelet recovery, acute/chronic GvHD)을 경험할 수도 있음
- 골수이식 후 혈소판이 정상수준으로 회복된 환자들의 생존시간에 관심
  - 중간사건을 경험하지 못한 환자는 분석에서 제외됨. 즉, 중간사건을 경험할 때까지 left-truncated 됨
  - 혈소판회복 때까지 시간이 delayed entry time이 됨

Channing house data

- 요양원에 거주하는 462명(남자: 97, 여자: 365)의 노인을 대상으로 사망시간을 관측
- 요양원에 들어온 나이와 사망한 나이 또는 요양원을 떠난 (right-censored) 나이를 포함
- 65세 이상 노인만이 요양원에 들어올 수 있기 때문에 65세 이전에 사망한 노인들은 분석에서 제외됨. 즉, left-truncated 됨

Channing house data

- library(KMsurv)
- library(survival)
- data(channing)
- fit=survfit(Surv(ageentry,age,death)~gender, data=subset(channing,gender==2))
- plot(fit$time, fit$n.risk, xlab="Time(in months)", ylab="Risk set size", type="l")
- fit2=survfit(Surv(ageentry,age,death)~gender, data=subset(channing,age>=816))
- plot(fit2, lty=1:2, xlab="Time(in months)", ylab="Estimated survival probability", xlim=c(804,1215))
- legend("topright",lty=1:2, c("Male","Female"))

Channing house data

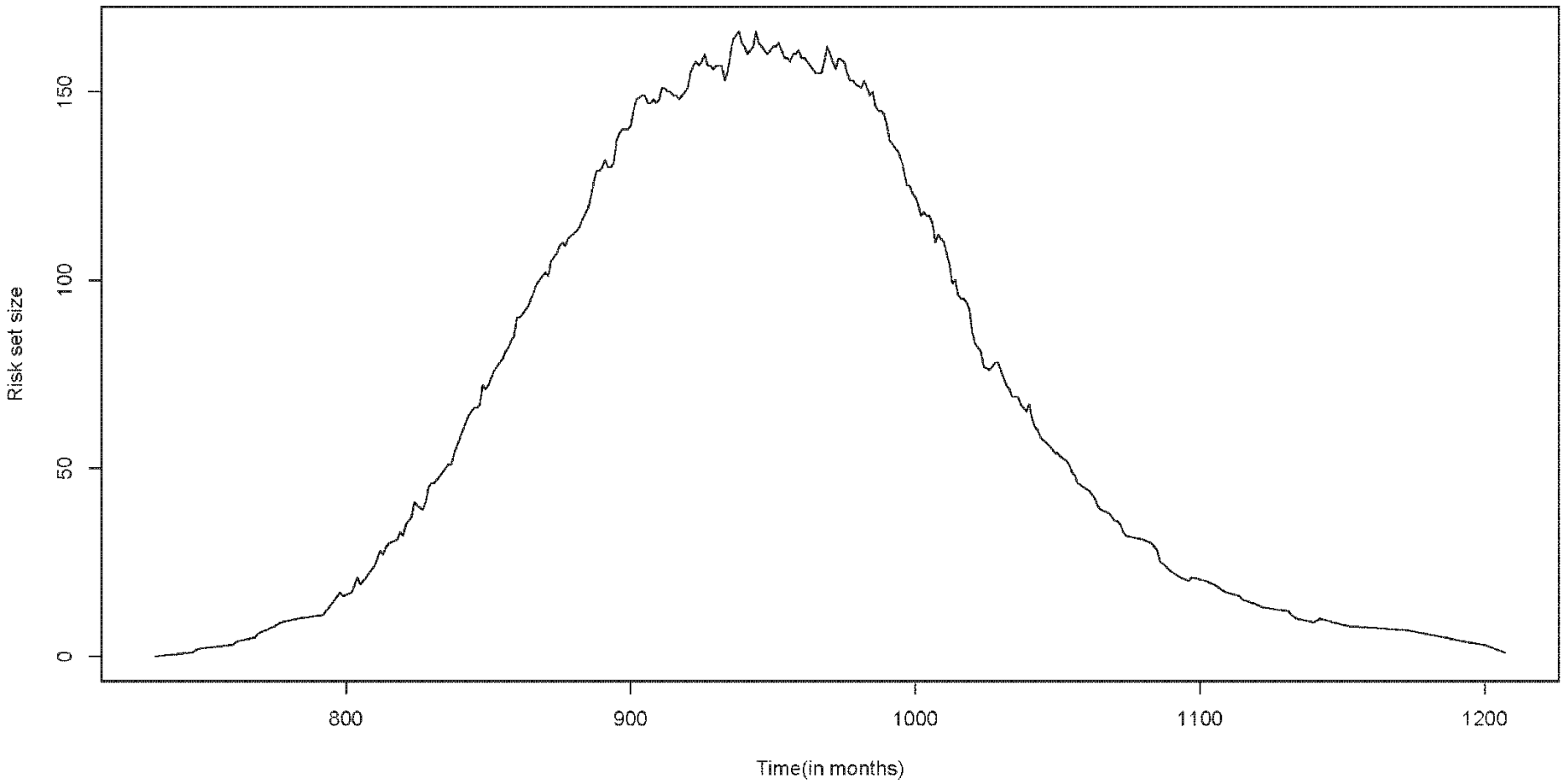
```
> head(channing)
  obs death ageentry  age time gender
1   1     1     1042 1172  130      2
2   2     1      921 1040  119      2
3   3     1      885 1003  118      2
4   4     1      901 1018  117      2
5   5     1      808  932  124      2
6   6     1      915 1004   89      2
```

Channing house data

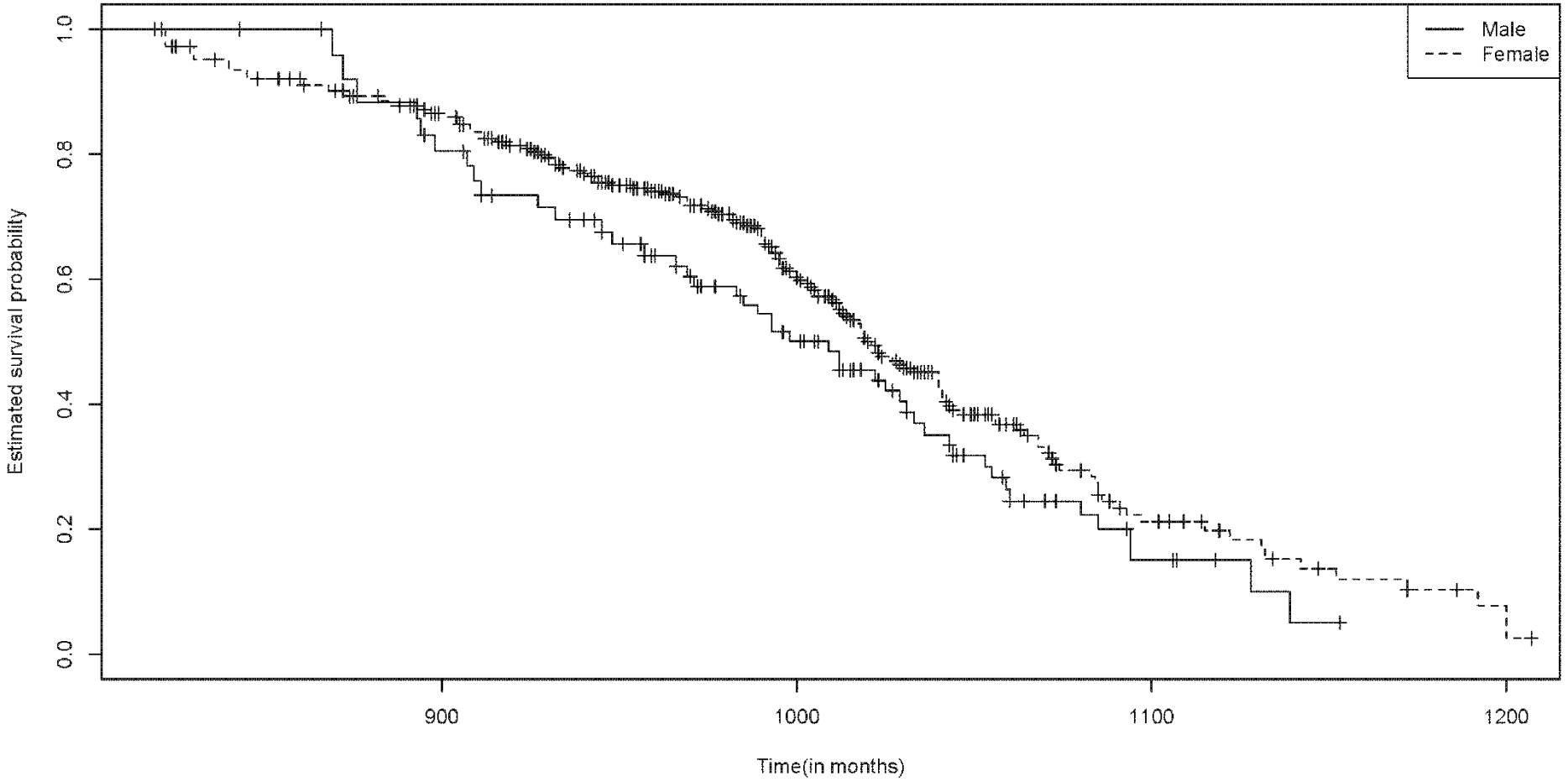
```
> summary(fit)
Call: survfit(formula = Surv(ageentry, age, death) ~ gender, data = subset(channing,
    gender == 2))

3 observations deleted due to missingness
 time n.risk n.event entered censored survival std.err lower 95% CI upper 95% CI
  804     21       1       0        1   0.9524  0.0465      0.86552        1.000
  822     36       1       2        0   0.9259  0.0522      0.82912        1.000
  830     46       1       1        0   0.9058  0.0548      0.80455        1.000
  840     58       1       3        0   0.8902  0.0560      0.78689        1.000
  845     66       1       1        0   0.8767  0.0568      0.77220        0.995
  861     90       1       4        1   0.8670  0.0570      0.76218        0.986
  868    100       1       2        0   0.8583  0.0571      0.75343        0.978
  873    106       1       2        0   0.8502  0.0571      0.74534        0.970
  883    116       1       4        0   0.8429  0.0571      0.73811        0.962
  885    119       1       4        0   0.8358  0.0570      0.73116        0.955
  895    137       1       4        1   0.8297  0.0569      0.72526        0.949
  897    140       1       2        1   0.8237  0.0568      0.71956        0.943
  901    145       1       4        0   0.8181  0.0567      0.71411        0.937
  905    149       2       3        3   0.8071  0.0565      0.70362        0.926
  908    148       2       1        0   0.7962  0.0563      0.69322        0.914
```

Channing house data



Channing house data



생존분석이란?
중도절단과 절삭
누적한계추정량
좌절삭된 자료의 생존함수 추정
구간중도절단된 자료의 생존함수 추정
좌중도절단된 자료의 생존함수 추정

생존함수 추정

Types of interval-censored data

- Case I interval-censored data or current status data
  - T is only known to be larger or smaller than an observed monitoring time C
  - Either $L = 0$ or $R = \infty$
  - Observed data: $\{(C_i, \delta_i = I(T_i \leq C_i)), i = 1, \ldots, n\}$
  - eg, Cross-sectional studies or tumourigenicity experiments
- Case II interval-censored data
  - Include at least one interval $(L, R]$ with both L and R
  - In experiments with two monitoring times, U and V, with $U \leq V$,
    - $T \leq U, U < T \leq V$, or $T > V$
- Case K interval-censored data
  - In longitudinal studies with periodic follow-up and K monitoring times, $M_1, \ldots, M_K$, the event is only observed between two consecutive inspecting times, M_l and M_{l+1}, and the observed data reduced to $(M_l, M_{l+1}]$

Notation

- Observed data: $\mathcal{F} = \{(l_i, r_i], i = 1, \dots, n\}$
- $0 = t_0 < t_1 < \cdots < t_m < t_{m+1} = \infty$: unique ordered elements of $\mathcal{F}$
- Define $\alpha_{ij} = I\left(\left(t_{j-1}, t_j\right] \subset (l_i, r_i]\right)$ and $p_j = S\left(t_{j-1}\right) - S\left(t_j\right), j = 1, \dots, m+1$

Non-parametric MLE

- Likelihood function for $\boldsymbol{p} = (p_1, \ldots, p_{m+1})'$
  - $L(\boldsymbol{p}) = \prod_{i=1}^{n}[S(L_i) - S(R_i)] = \prod_{i=1}^{n}\sum_{j=1}^{m+1}\alpha_{ij}p_j$
- NPMLE, $\hat{S}$, of S
  - Maximize $L(\boldsymbol{p})$ under $\sum_{j=1}^{m+1} p_j = 1$ and $p_j \geq 0$
  - $\hat{S}$: Right-continuous step function, i.e., $\hat{S}(t) = \hat{S}(t_{j-1}), t_{j-1} \leq t < t_j$
- Remarks
  - Some elements of $\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_{m+1})'$ could be 0 and it could help to know these zero components before running a determination process
  - $\hat{p}_j$ could be non-zero only if $t_{j-1} = L_i$ for some i and $t_j = R_k$ for some k, for $i, k = 1, \ldots, n$ (Turnbull (JRSSB, 1976)'s approach)

Illustrative example

i	L_i	R_i
1	0	7
2	0	8
3	6	10
4	7	16
5	7	14
6	17	∞
7	37	∞
8	45	∞
9	46	∞
10	46	∞

Turnbull intervals

Sub. #	t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}
1	0		7										
2	0			8									
3		6			10								
4			7				16						
5			7			14							
6								17					∞
7									37	44			
8											45		∞
9												46	∞
10												46	∞
		(6	7](7	8]					(37	44]		(46	∞)

Breast cosmesis data

- Two treatments for breast cancer, radiation (Rad, n=46), and radiation with chemotherapy (RadChem, n=48)
- Response: Time in months until breast retraction (Finkelstein & Wolfe, BCS, 1985)
- Use R package **interval: icfit** function
  - **icfit** function calculates NPMLE by EM algorithm
  - download package **Icens** from bioconductor
- R codes
  - > library(interval)
  - > data(bcos)
  - > fit=icfit(Surv(left,right,type="interval2")~treatment, data=bcos)
  - > summary(fit)
  - > plot(fit)

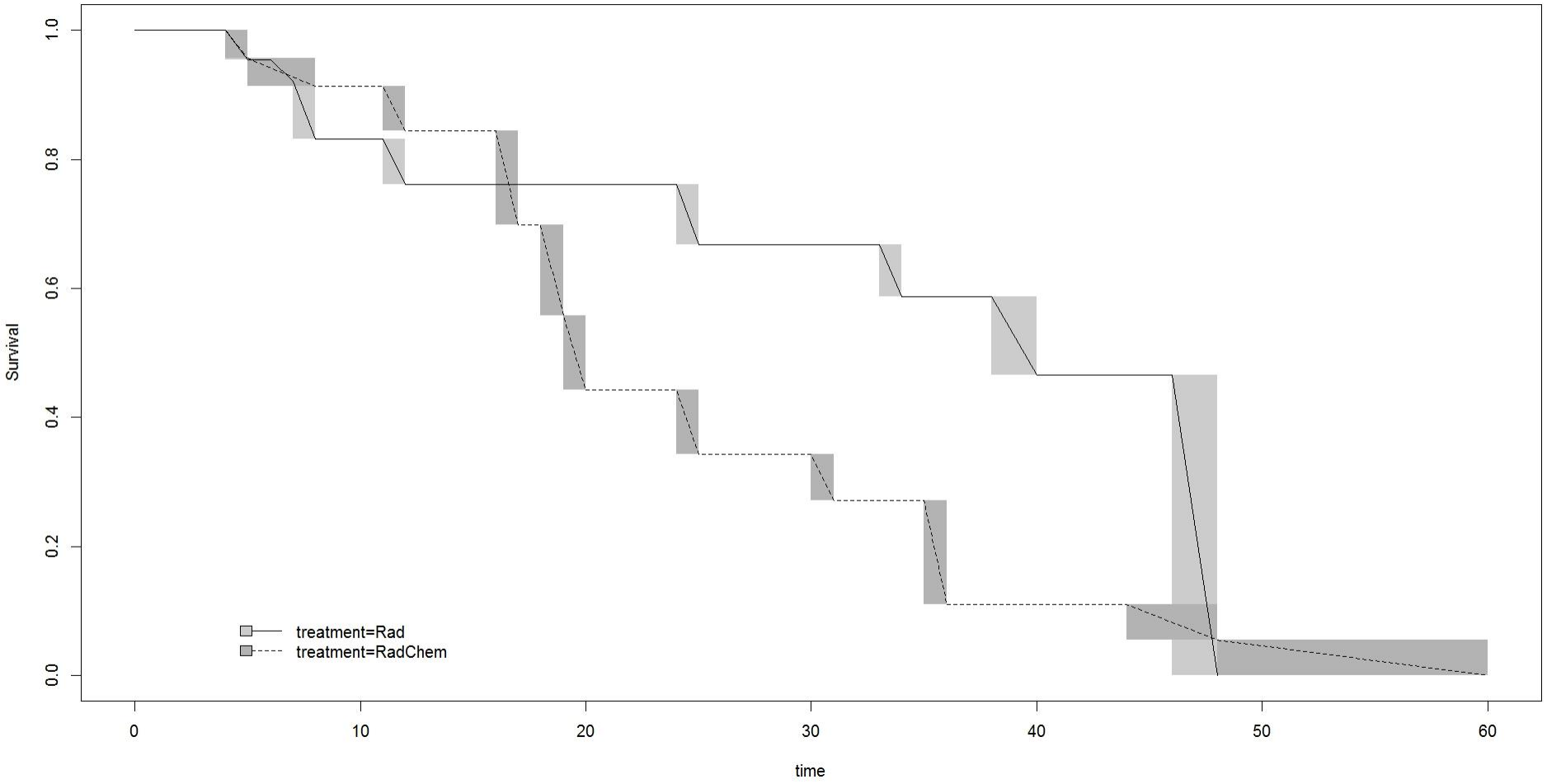
Breast cosmesis data

```
> head(bcos)
  left right treatment
1   45   Inf       Rad
2    6    10       Rad
3    0     7       Rad
4   46   Inf       Rad
5   46   Inf       Rad
6    7    16       Rad
```

Breast cosmesis data

```
> summary(fit)
treatment=Rad:
  Interval Probability
1    (4,5]      0.0463
2    (6,7]      0.0334
3    (7,8]      0.0887
4   (11,12]     0.0708
5   (24,25]     0.0926
6   (33,34]     0.0818
7   (38,40]     0.1209
8   (46,48]     0.4656
treatment=RadChem:
   Interval Probability
1     (4,5]      0.0433
2     (5,8]      0.0433
3    (11,12]     0.0692
4    (16,17]     0.1454
5    (18,19]     0.1411
6    (19,20]     0.1157
7    (24,25]     0.0999
8    (30,31]     0.0709
9    (35,36]     0.1608
10   (44,48]     0.0552
11   (48,60]     0.0552
```

Breast cosmesis data



생존분석이란?
중도절단과 절삭
누적한계추정량
좌절삭절단된 자료의 생존함수 추정
구간중도절단된 자료의 생존함수 추정
좌중도절단된 자료의 생존함수 추정

생존함수 추정

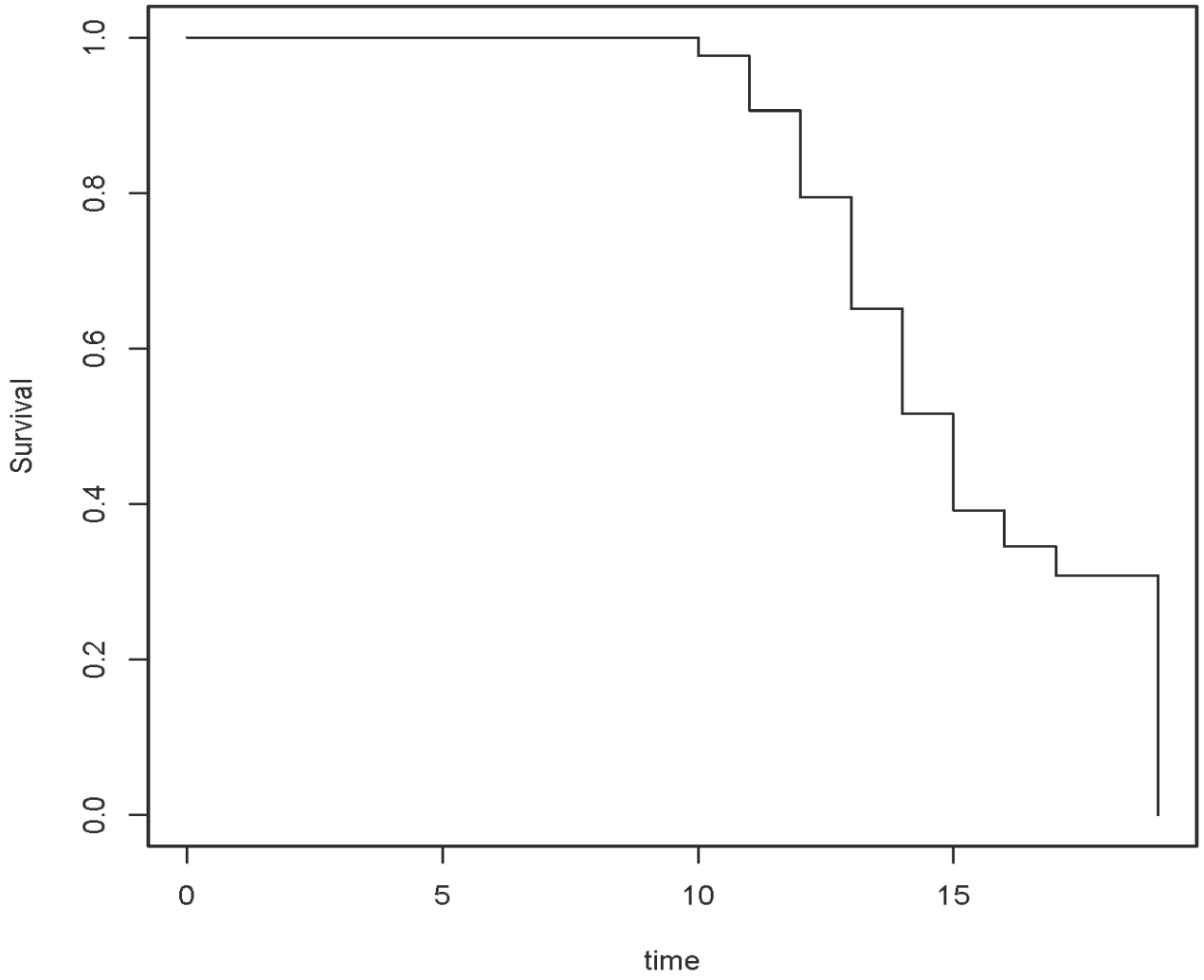
Marijuana data

- library(interval)
- Age=10:19
- Exact=c(4,12,19,24,20,13,3,1,0,4)
- L.Exact=rep(Age,Exact)
- R.Exact=rep(Age,Exact)
- R.Censored=c(0,0,2,15,24,18,14,6,0,0)
- L.R.Censored=rep(Age,R.Censored)
- R.R.Censored=rep(Inf,sum(R.Censored))
- L.Censored=c(0,0,0,1,2,3,2,3,1,0)
- L.L.Censored=rep(0,sum(L.Censored))
- R.L.Censored=rep(Age,L.Censored)
- marijuana=data.frame(left=c(L.Exact,L.R.Censored,L.L.Censored),right=c(R.Exact,R.R.Censored,R.L.Censored))
- fit=icfit(Surv(left,right,type="interval2")~1, data=,marijuana)
- summary(fit)
- plot(fit)

Marijuana data

```
> head(marijuana)
  left right
1   10    10
2   10    10
3   10    10
4   10    10
5   11    11
6   11    11
```

Marijuana data



로그순위검정

구간중도절단자료에서 생존함수 비교

생존함수 비교

로그순위검정

구간중도절단자료에서 생존함수 비교

생존함수 비교

Comparison of survivor functions

- Test whether or not the survivor functions for two or more groups are equivalent
  - $H_o: S_1(\mathrm{t}) = \cdots = S_g(t), \forall t > 0$
- $t_1 < \cdots < t_D$: distinct survival times by pooling all the sample from two groups (set $g = 2$)
- d_{ij}: observed # of failures at t_j in group $i, i = 1,2; j = 1, \ldots, D$
  - $d_j = \sum_{i=1}^{2} d_{ij}$
- n_{ij}: # of individuals at risk just prior to t_j in group i
  - $n_j = \sum_{i=1}^{2} n_{ij}$

Comparison of survivor functions

- Idea: Based on

$$Z_i = \sum_{j=1}^{D} \left(d_{ij} - d_j \frac{n_{ij}}{n_j} \right) = \sum_{j=1}^{D} \left(O_{ij} - \widehat{E}_{ij} \right)$$

- Log-rank statistic
  - $X^2 = \left[\sum_{j=1}^{D} \left(O_{ij} - \widehat{E}_{ij} \right) \right]^2 \Big/ \sum_{j=1}^{D} d_j \frac{n_{ij}}{n_j} \left(1 - \frac{n_{ij}}{n_j} \right) \left(\frac{n_j - d_j}{n_j - 1} \right) \sim \chi^2(1)$ under H_0
  - If $X^2 > \chi^2_\alpha(1)$, reject a test for equality of the survivor functions at level α

Remarks for log-rank test

- Choice of weight function: $W(t_j)$, specially $W(t_j) = 1$ for log-rank test
  - $W(t_j) = \hat{S}(t_j)^p(1 - \hat{S}(t_j)^q$: Fleming-Harrington (CommStat, 1981)
  - $p = q = 0$: log-rank test, $p = 1, q = 0$: Peto-Peto test
- Extension to three or more groups
- Stratification on a set of covariates
- Trend test for ordered alternatives: plugging in any set of scores

Illustration (revisited)

- AML 자료에서 화학요법 지속그룹과 비지속그룹의 재발시간분포가 동일한가?
- Three types of test statistic
  - Log-rank=3.4 (*p*-value=0.0653) with $W(t_j) = 1$
  - Peto-Prentice=2.8 (*p*-value=0.0955) with $W(t_j) = \hat{S}(t_j)$
  - Flaming-Harrington=4.2 (*p*-value=0.041) with $W(t_j) = \hat{S}(t_j)^{-1}$

Illustration (revisited)

- R codes
  - library(survival)
  - fit=survfit(Surv(Time,Status)~Group, data=aml)
  - plot(fit, lty=1:2, xlab="Time(in weeks)", ylab="Estimated survival probability")
  - legend("topright",lty=1:2, c("Chemo","Non-Chemo"))
  - lr.fit=survdiff(Surv(Time,Status)~Group, data=aml)
  - pp.fit=survdiff(Surv(Time,Status)~Group, data=aml, rho=1)
  - fh.fit=survdiff(Surv(Time,Status)~Group, data=aml, rho=-1)

Illustration (revisited)

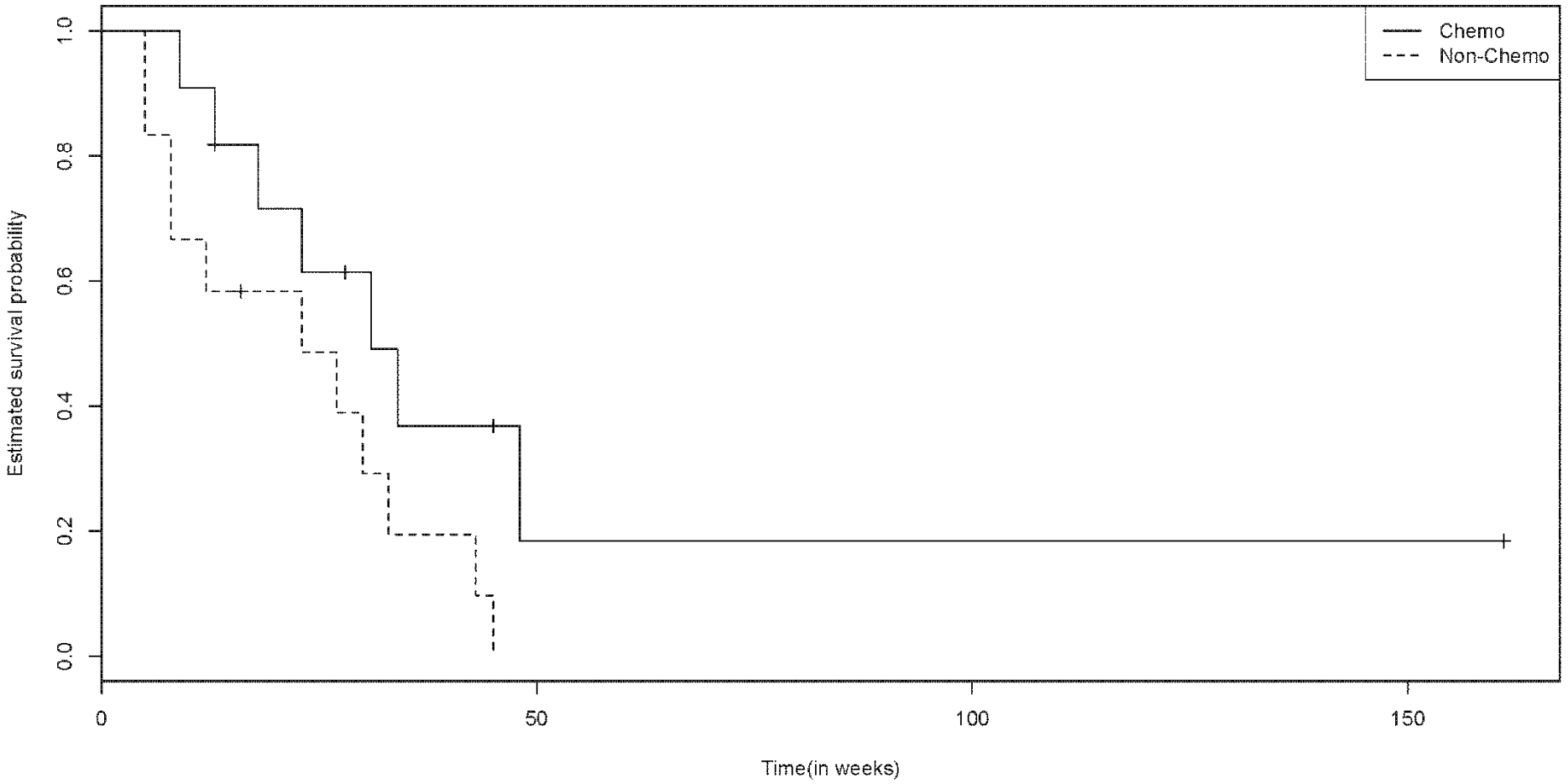


Illustration (revisited)

```
> lr.fit
Call:
survdiff(formula = Surv(Time, Status) ~ Group, data = aml)

                 N Observed Expected (O-E)^2/E (O-E)^2/V
Group=Chemo     11        7    10.69      1.27       3.4
Group=Non.Chemo 12       11     7.31      1.86       3.4

 Chisq= 3.4  on 1 degrees of freedom, p= 0.0653
```

로그순위검정

구간중도절단자료에서 생존함수 비교

생존함수 비교

Notation

- $\hat{S}$: NPMLE of S_i's under $H_0: S_1 = \cdots = S_g$
- $\delta_i = 0$ if right-censored and otherwise, 1
- $\rho_{ij} = I(\delta_i = 0, L_i \geq t_j)$, i.e., $\rho_{ij} = 1$ if T_i is right-censored and subject i is still at risk at t_i-
- Define the estimates of the total observed failures and risk numbers, respectively, as
  - $d_j = \sum_{i=1}^{n} \delta_i \frac{\alpha_{ij}\left[\hat{S}\left(t_j^-\right) - \hat{S}(t_j)\right]}{\sum_{l=1}^{m+1} \alpha_{il}\left[\hat{S}\left(t_l^-\right) - \hat{S}(t_l)\right]}, j = 1, \ldots, m$
  - $n_j = \sum_{i=1}^{n} \delta_i \frac{\sum_{r=j}^{m+1} \alpha_{ir}[\hat{S}(t_r^-) - \hat{S}(t_r)]}{\sum_{l=1}^{m+1} \alpha_{il}\left[\hat{S}\left(t_l^-\right) - \hat{S}(t_l)\right]} + \sum_{i=1}^{n} \rho_{ij}, j = 1, \ldots, m$
- Similarly, define d_{jk} and n_{jk} from subjects in group $k, k = 1,2$

Weighted log-rank tests

- Define $U_1 = \sum_{j=1}^{m} \left(d_{j1} - d_j \frac{n_{j1}}{n_j} \right)$
- Estimation of the variance of U_1 : employ resampling methods such as multiple imputation, bootstrap, and permutation procedures
- Remark
  - With $W(t) = 1$,
    $U_1 = \int_0^{\infty} W(t) \frac{Y_1(t)Y_2(t)}{Y_1(t)+Y_2(t)} \left[d\widehat{\Lambda}_1(t) - d\widehat{\Lambda}_2(t) \right]$
    - $Y_k(t) = \sum_{j|t_j \le t} n_{jk}$ and $\widehat{\Lambda}_k(t) = \sum_{j|t_j \le t} \frac{d_{jk}}{n_{jk}}$

Illustration (revisited)

- 유방암 자료에서 두 가지 치료방법(Rad , RadChem)에 띠라 유방 함몰 시점까지의 분포가 동일한가?
- Three types of test statistic
  - Log-rank=-2.67 (*p*-value=0.008) with $W(t_j) = 1$
  - Finkelstein(BCS, 1986)=-2.69 (*p*-value=0.007) with $W(t_j) \approx 1$
  - Wilcoxon-Mann-Whitney=-2.16 (*p*-value=0.03) with $W(t_j) = \hat{S}(t_j^-)$

Illustration (revisited)

- R codes
  - library(interval)
  - data(bcos)
  - lr.fit=ictest(Surv(left,right,type="interval2")~treatment, data=bcos)
  - f.fit=ictest(Surv(left,right,type="interval2")~treatment, data=bcos, score="logrank2")
  - wmw.fit=ictest(Surv(left,right,type="interval2")~treatment, data=bcos, score="wmw“)

Illustration (revisited)

```
> lr.fit

  Asymptotic Logrank two-sample test (permutation form), Sun's scores

data:  Surv(left, right, type = "interval2") by treatment
Z = -2.6684, p-value = 0.007622
alternative hypothesis: survival distributions not equal

                    n Score Statistic*
treatment=Rad      46        -9.141846
treatment=RadChem 48         9.141846
* like Obs-Exp, positive implies earlier failures than expected
```

콕스비례위험모형

구간중도절된 자료의 콕스비례위험모형

회귀모형

Cox proportional hazards model

- Why regression models need?
  - To predict covariates(or explanatory variables, risk factors) for time to event
- Data: $\{(t_j, \delta_j, z_j); t_j = \min(x_j, c_j), \delta_j = I(t_j = x_j), j = 1, \dots, n\}$
- Cox model: $\lambda(t|z) = \lambda_0(t) \exp\{\beta' z\}$
  - Hazard rate at t for an individual with risk vector z
- A sort of semiparametric model
  - parametrically for the covariate effect + nonparametrically for baseline hazard function
- Why PH is called?
  - RR(or HR)=$\frac{\lambda(t|z)}{\lambda(t|z^*)} = \exp\{\sum_k \beta_k (z_k - z_k^*)\}$ is constant against t

Partial likelihood

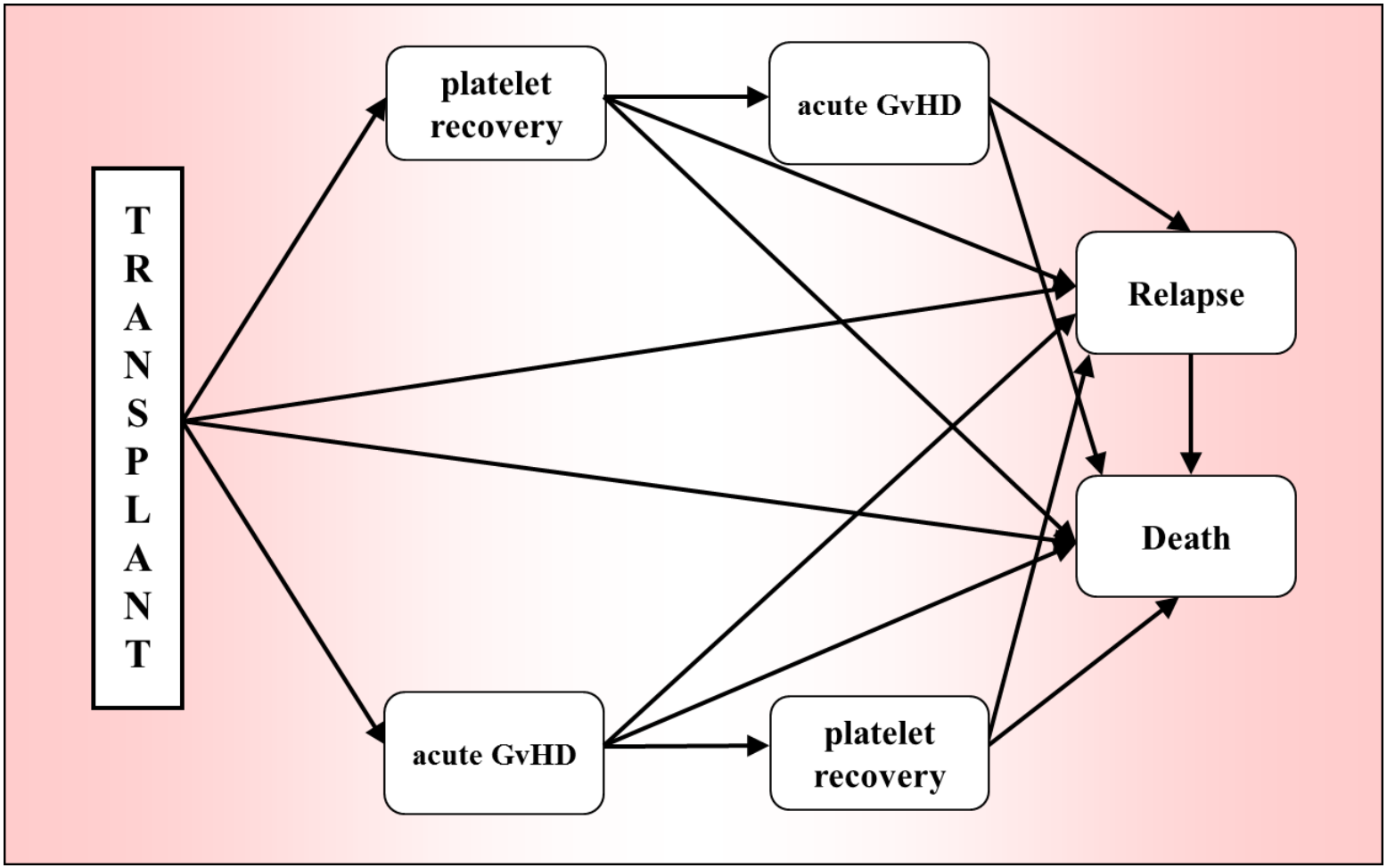
- $t_1 < \cdots < t_D$: ordered event times
- $z_{(i)k}$: kth covariate associated with the individual whose failure time is t_i
- $R(t_i)$: set of individuals who are still under study at a time prior to t_i
- Partial likelihood:

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp\{\sum_k \beta_k z_{(i)k}\}}{\sum_{j \in R(t_i)} \exp\{\sum_k \beta_k z_{jk}\}}$$

BMT data (revisted)

- Data on 137 bone marrow transplant patients
- Risk factors: patient and donor age, sex, and CMV status, waiting time from diagnosis to transplantation, FAB, MTX
- Three groups: AML low risk(54), AML high risk(45), ALL(38)
- Survival times
  - T_1: time(in days) to death or end of study
  - T_2: disease-free survival time(time to relapse, death or end of study)
  - T_A: time to acute GvHD
  - T_C: time to chronic GvHD
  - T_P: time to return of platelets to normal levels

Simplified recovery process from BMT data



Fixed risk factors

- z_1=1 if AML low-risk, z_2=1 if AML high-risk
- z_3=waiting time
- z_4=FAB(binary)
- z_5=MTX(binary)
- z_6=1 if donor: male; z_7=1 if patient: male; $z_8 = z_6 \times z_7$=1 if donor & patient: male
- z_9=1 if donor: CMV positive; z_{10}=1 if patient: CMV positive; $z_{11} = z_9 \times z_{10}$=1 if donor & patient: CMV positive
- z_{12}=donor age-28; z_{13}=patient age-28; $z_{14} = z_{12} \times z_{13}$

BMT data

```
> head(bmt)
  DG   TD  TDF DI RI DFI   TA AI  TC CI TP PI PAGE DAGE PSEX DSEX PCMV DCMV   WT FAB
1  1 2081 2081  0  0   0   67  1 121  1 13  1   26   33    1    0    1    1   98   0
2  1 1602 1602  0  0   0 1602  0 139  1 18  1   21   37    1    1    0    0 1720   0
3  1 1496 1496  0  0   0 1496  0 307  1 12  1   26   35    1    1    1    0  127   0
4  1 1462 1462  0  0   0   70  1  95  1 13  1   17   21    0    1    0    0  168   0
5  1 1433 1433  0  0   0 1433  0 236  1 12  1   32   36    1    1    1    1   93   0
6  1 1377 1377  0  0   0 1377  0 123  1 12  1   22   31    1    1    1    1 2187   0
  H MTX p28 d28
1 1   0  -2   5
2 1   0  -7   9
3 1   0  -2   7
4 1   0 -11  -7
5 1   0   4   8
6 1   0  -6   3
>
```

Local tests

- $H_0: \beta_1 = \beta_{10}, \beta = (\beta_{10}, \beta_{20})$
  - Wald test: $X_W^2 = (b_1 - \beta_{10})'[I^{11}(b)]^{-1}(b_1 - \beta_{10}) \sim \chi^2(q)$
  - Likelihood ratio test: $X_{LR}^2 = 2\{LL(b) - LL[\beta_{10}, b_2(\beta_{10})]\} \sim \chi^2(q)$
  - Score test: $X_{SC}^2 = U_1[\beta_{10}, b_2(\beta_{10})]' I^{11}(\beta_{10}, b_2(\beta_{10})) U_1[\beta_{10}, b_2(\beta_{10})] \sim \chi^2(q)$
- 공변량 z_1, z_2를 포함하여 유의한 공변량을 전진선택법으로 선택하면, $z_4 \rightarrow (z_{12}, z_{13}, z_{14})$ 순으로 선택됨

Tentative model (fixed only)

- R codes
  - library(survival)
  - bmt=read.csv(file="bmt.csv",header=T)
  - fit.fab=coxph(Surv(TDF,DFI)~factor(DG)+FAB, data=bmt)
    # Cox model after adjusting for the risk groups
  - df=1
  - pos=3
  - coef=fit.fab$coefficients[pos]
  - var=fit.fab$var[pos,pos]
  - wald=t(coef)%*%solve(var)%*%coef
  - fab.pvalue=pchisq(wald,df,lower.tail=FALSE)

Tentative model (fixed only)

- bmt$p28=bmt$PAGE-28
- bmt$d28=bmt$DAGE-28
- fit.age=coxph(Surv(TDF,DFI)~factor(DG)+FAB+p28+d28+p28*d28, data=bmt) # Cox model after adjusting for the risk groups and FAB
- df=3
- pos=4:6
- coef=fit.age$coefficients[pos]
- var=fit.age$var[pos,pos]
- wald=t(coef)%*%solve(var)%*%coef
- age.pvalue=pchisq(wald,df,lower.tail=FALSE)

Tentative model (fixed only)

```
> fit.age
Call:
coxph(formula = Surv(TDF, DFI) ~ factor(DG) + FAB + p28 + d28 +
    p28 * d28, data = bmt)


                coef exp(coef) se(coef)      z       p
factor(DG)2 -1.09065     0.336 0.354279 -3.078 0.00210
factor(DG)3 -0.40391     0.668 0.362777 -1.113 0.27000
FAB          0.83742     2.310 0.278464  3.007 0.00260
p28          0.00682     1.007 0.019683  0.347 0.73000
d28          0.00387     1.004 0.018256  0.212 0.83000
p28:d28      0.00316     1.003 0.000951  3.323 0.00089

Likelihood ratio test=32.8  on 6 df, p=1.14e-05  n= 137, number of events= 83
```

Other regression models

- Additive hazards model: $\lambda(t|z) = \lambda_0(t) + \beta' z$
- Accelerated failure time model: $\log T_i = \beta' z + \epsilon$
  - Focus on direct relationship between z and time to event
  - Effect of covariates is multiplicative on t rather than on hazard function
  - Parametric, but providing a good fit if correctly chosen

Refinements of Cox model

- Stratification
  - When the PH assumption is violated for some covariate
  - $\lambda_j(t|z) = \lambda_{0j}(t)\exp\{\beta' z\}, j = 1, \ldots, s$

- Time-dependent covariates
  - eg, BP, cholesterol, size of the tumor, …
  - $\lambda(t|z(t)) = \lambda_0(t)\exp\{\beta' z(t)\}$

BMT data (revisited)

- Platelet recovery를 time-varying 공변량으로 간주한다면,
  - eg, id=1, TDF=2081, DFI=0, TP=13, PI=1:
    - (start,end)=(0,13), p.status=0
    - (start,end)=(13,2081), p.status=1
- 공변량 z_1, z_2를 포함하여 유의한 time-varying 공변량을 전진선택법으로 선택하면, acute GvHD, chronic GvHD, platelet recovery 중에서 platelet recovery 만 선택됨

Making time-varying covariates

```
> head(bmt[1:3,])
  DG   TD  TDF DI RI DFI   TA AI  TC CI TP PI PAGE DAGE PSEX DSEX PCMV DCMV   WT FAB
1  1 2081 2081  0  0   0   67  1 121  1 13  1   26   33    1    0    1    1   98   0
2  1 1602 1602  0  0   0 1602  0 139  1 18  1   21   37    1    1    0    0 1720   0
3  1 1496 1496  0  0   0 1496  0 307  1 12  1   26   35    1    1    1    0  127   0
  H MTX p28 d28
1 1   0  -2   5
2 1   0  -7   9
3 1   0  -2   7
> head(p.bmt)
  id start  end status group p.status
1  1     0   13      0     1        0
2  1    13 2081      0     1        1
3  2     0   18      0     1        0
4  2    18 1602      0     1        1
5  3     0   12      0     1        0
6  3    12 1496      0     1        1
```

Cox PH model (time-varying)

```
> fit.p
Call:
coxph(formula = Surv(start, end, status) ~ factor(group) + p.status,
    data = p.bmt)


                  coef exp(coef) se(coef)     z       p
factor(group)2 -0.497     0.608    0.289 -1.72 0.08600
factor(group)3  0.382     1.465    0.268  1.43 0.15000
p.status       -1.120     0.326    0.329 -3.40 0.00067

Likelihood ratio test=22.9  on 3 df, p=4.29e-05  n= 256, number of events= 83
   (1 observation deleted due to missingness)
> wald
        [,1]
[1,] 11.5813
> p.pvalue
              [,1]
[1,] 0.0006661839
```

Tests of PH assumption

- Use of time-dependent covariate methodology
- To test the proportionality assumption for a fixed-time covariate, z_1, create a time-dependent covariate, $z_2(t) = z_1 \times g(t)$
- In many applications, take $g(t) = \ln t$
- PH model: $\lambda(t|z_1) = \lambda_0(t) \exp\{\beta_1 z_1 + \beta_2 z_2(t)\}$
  - HR=$\frac{\lambda(t|z_1)}{\lambda(t|z_1^*)} = \exp\{\beta_1(z_1 - z_1^*) + \beta_2 g(t)(z_1 - z_1^*)\}$: depend on t if $\beta_2 \neq 0$
  - Test of $H_0: \beta_2 = 0$ is equivalent to test for the proportional hazards assumption
  - Power of detecting non-proportionality: depend on the choice of $g(t)$

BMT data (revisited)

- 공변량들 중에서 MTX만 PH 가정을 만족 못하여 MTX에 대해서는 층화

```
> fit.mtx
Call:
coxph(formula = Surv(TDF, DFI) ~ MTX + tv.mtx, data = bmt)


         coef exp(coef) se(coef)     z       p
MTX     11.65  1.14e+05    1.509  7.72 1.2e-14
tv.mtx -1.95  1.42e-01    0.284 -6.87 6.6e-12

Likelihood ratio test=90.4  on 2 df, p=0  n= 137, number of events= 83
> wald
          [,1]
[1,] 47.12934
> mtx.pvalue
                [,1]
[1,] 6.645335e-12
```

Final model (fixed only)

```
> fit.final
Call:
coxph(formula = Surv(TDF, DFI) ~ factor(DG) + FAB + p28 * d28 +
    strata(MTX), data = bmt)


                coef exp(coef) se(coef)      z      p
factor(DG)2 -1.03391     0.356 0.364650 -2.8354 0.0046
factor(DG)3 -0.33763     0.713 0.367826 -0.9179 0.3600
FAB          0.90843     2.480 0.278913  3.2570 0.0011
p28          0.00547     1.005 0.019968  0.2741 0.7800
d28         -0.00175     0.998 0.018163 -0.0963 0.9200
p28:d28      0.00286     1.003 0.000951  3.0021 0.0027

Likelihood ratio test=31.1  on 6 df, p=2.4e-05  n= 137, number of events= 83
```

콕스비례위험모형

구간중도절된 자료의 콕스비례위험모형

회귀모형

Cox model with IC data

- Data: $\{(l_i, r_i], z_i); i = 1, \ldots, n\}$
- Cox model: $\lambda(t|z) = \lambda_0(t) \exp\{\beta' z\}$
- Unlike right-censored data, estimating β under interval censoring involve estimation of both β and the cumulative baseline hazard function, $\Lambda_0(t) = \int_0^t \lambda_0(s) ds$

ML approach

- Likelihood function: $L = \prod_{i=1}^{n} [S(l_i|z_i) - S(r_i|z_i)]$
  - $S(t|z) = S_0(t)^{\exp\{\beta' z\}}$: Survival function for a subject with covariates z
- Log-likelihood: Assuming that $l_i < r_i, \forall i$,
  - $l(\beta, S_0) = \sum_{i=1}^{n} \log\{[S_0(l_i)^{\exp\{\beta' z\}} - S_0(r_i)^{\exp\{\beta' z\}}]\}$
  - S_0: baseline survival function, $S_0(t) = \exp\{-\Lambda_0(t)\}$
- Focus on estimation of S_0 at the different observation time points, i.e., $0 = t_0 < t_1 < \cdots < t_m < t_{m+1} = \infty$, of the form,
  - $S_0(t) = \prod_{j|t_j \le t} \exp\{e^{\alpha_j}\} = \exp\left\{\sum_{j|t_j \le t} e^{\alpha_j}\right\}$
  - $\alpha = (\alpha_1, \ldots, \alpha_m)$: unknown parameters

ML approach

- $l(\beta, S_0)$ can be rewritten as
  - $l(\beta, \alpha) = \sum_{i=1}^{n} \log\{\sum_{j=1}^{m+1} \alpha_{ij}[\exp\{-a_{j-1}\exp\{\beta' z\}\} - \exp\{-a_j \exp\{\beta' z\}\}]\}$
    - $a_j = \sum_{k=1}^{j} e^{\alpha_k}$
- Use the Newton-Raphson algorithm to determine the MLE of β and α (Finkelstein, BCS, 1986)
- Asymptotic properties (Huang & Wellner, 1997)
  - $\hat{S}_0$: strongly consistent
  - $\hat{\beta}$: asymptotically normal with the usual $\sqrt{n}$-convergence rate and efficient

Breast cosmesis data (revisited)

- intcox package을 사용할 수 있지만 모수 추정만 제공되고 se는 제공 안됨
  - permutation test를 이용
  - 붓스트랩 방법 이용

```
> fit
Call:
intcox(formula = Surv(left, right, type = "interval2") ~ treatment
    data = bcos)


                   coef exp(coef) se(coef)  z  p
treatmentRadChem 0.776      2.17        NA NA NA

Likelihood ratio test=NA  on 1 df, p=NA  n= 94
```

군집생존자료분석
재발사건자료분석
다상태모형
경쟁위험모형

다변량 생존자료분석

다변량 생존자료의 분류

자료유형	순서 존재 여부	
	Yes	No
동일한 유형	재발사건자료 (recurrent event data)	군집생존자료 (clustered survival data)
다른 유형	다상태모형 (multi-state model)	경쟁위험모형 (competing risks model)

군집생존자료분석
재발사건자료분석
다상태모형
경쟁위험모형

다변량 생존자료분석

두 가지 예

- 당뇨병성 환자의 시력 손실 자료
  - 시력 손실을 지연시키기 위한 방법인 laser photocoagulation의 효과를 조사
  - 한쪽 눈에는 레이저 치료를 하고, 다른 한쪽 눈에는 하지 않음
- 발암물질로 인한 쥐의 생존시간 자료
  - 같은 어미에서 태어난 세 마리 새끼를 군집으로 잡고, 그들 중에서 한 마리만 발암물질에 노출시킴

두 가지 예

```
> head(diabetes)
  id      time status trteye treat adult agedx
1  5 46.24967      0      2     1     2    28
2  5 46.27553      0      2     0     2    28
3 14 42.50684      0      1     1     1    12
4 14 31.34145      1      1     0     1    12
5 16 42.30098      0      1     1     1     9
6 16 42.27406      0      1     0     1     9
```

```
> head(rats)
  litter rx time status
1      1  1  101      0
2      1  0   49      1
3      1  0  104      0
4      2  1  104      0
5      2  0  102      0
6      2  0  104      0
```

Conditional vs. Marginal

- Data:
  $\{(t_{ik} = \min(x_{ik}, c_{ik}), \delta_{ik} = I(t_{ik} = x_{ik}), z_{ik}); i = 1, \dots, g, k = 1, \dots, n_i\}$
- Conditional method
  - $\lambda_{ik}(t|z_{ik}, a_i) = \lambda_{0i}(t)\exp(\beta' z_{ik})$
    - λ_{0i}: i번째 군집의 효과
    - z_{ik}가 유일한 binary covariate라면 조건부 방법은 같은 군집 내에 속한 두 그룹을 비교하는 것과 같음. 즉, $\exp(\beta)$는 같은 군집에 속한 두 그룹 간의 사건발생 위험률의 비를 나타냄
  - 군집효과의 추정은, 랜덤효과모형을 이용. 즉, $\lambda_{0i}(t) = \lambda_0(t)\exp(a_i), a_i$ ~Gamma with mean 1 and variance θ,
- Marginal method
  - $\lambda_{ik}(t|z_{ik}) = \lambda_0(t)\exp(\beta' z_{ik})$
    - 군집 내 관측값들이 서로 독립이라고 가정
  - 일치추정량이 되며 (Lin & Wei, JASA, 1989), model misspecificarion은 극복하기 위해 sandwich 분산 추정량을 사용

Diabetic retinopathy data (revisited)

```
> cond.fit
Call:
coxph(formula = Surv(time, status) ~ trt * adult + frailty(id),
    data = drs)

               coef   se(coef) se2   Chisq  DF   p
trt          -0.505 0.225    0.221   5.03  1.0 0.0250
adult         0.397 0.259    0.205   2.35  1.0 0.1300
frailty(id)                          122.55 88.6 0.0098
trt:adult    -0.986 0.362    0.355   7.43  1.0 0.0064

Iterations: 6 outer, 31 Newton-Raphson
     Variance of random effect= 0.926   I-likelihood = -847
Degrees of freedom for terms=  1.0  0.6 88.6  1.0
Likelihood ratio test=218  on 91.1 df, p=1.86e-12  n= 394
```

Diabetic retinopathy data `

```
> marg.fit
Call:
coxph(formula = Surv(time, status) ~ trt * adult + cluster(id),
    data = drs)


             coef exp(coef) se(coef) robust se     z      p
trt        -0.425     0.654    0.218     0.185 -2.30 0.0220
adult       0.341     1.407    0.199     0.196  1.74 0.0810
trt:adult -0.846     0.429    0.351     0.304 -2.79 0.0053

Likelihood ratio test=28.5  on 3 df, p=2.86e-06  n= 394, number of events= 155
```

Rats data (revisited)

```
> rats.cond.fit
Call:
coxph(formula = Surv(time, status) ~ rx + frailty(litter), data = rats)

                 coef  se(coef) se2   Chisq DF   p
rx               0.914 0.323    0.319  8.01  1.0 0.0046
frailty(litter)                       17.69 14.4 0.2400

Iterations: 6 outer, 24 Newton-Raphson
     Variance of random effect= 0.499   I-likelihood = -180.8
Degrees of freedom for terms=  1.0 14.4
Likelihood ratio test=37.6  on 15.4 df, p=0.00124  n= 150
> rats.marg.fit
Call:
coxph(formula = Surv(time, status) ~ rx + cluster(litter), data = rats)


    coef exp(coef) se(coef) robust se    z      p
rx 0.905      2.47    0.318     0.303 2.99 0.0028

Likelihood ratio test=7.98  on 1 df, p=0.00474  n= 150, number of events= 40
```

군집생존자료분석
재발사건자료분석
다상태모형
경쟁위험모형

다변량 생존자료분석

Bladder data

- 방광암을 앓고 있는 환자에 대해, thiothepa 치료의 종양 재발에 미치는 효과를 조사
- 병원을 방문할 때마다 새로운 종양의 발병여부를 조사하고 제거
- 환자마다 재발시점과 재발횟수가 같지 않음

```
> head(bladder2)
  id rx number size start stop event enum
1  1  1      1    3     0    1     0    1
2  2  1      2    1     0    4     0    1
3  3  1      1    1     0    7     0    1
4  4  1      5    1     0   10     0    1
5  5  1      4    1     0    6     1    1
6  5  1      4    1     6   10     0    2
```

Bladder data

```
> bladder
    id rx number size stop event enum
1    1  1      1    3    1     0    1
2    1  1      1    3    1     0    2
3    1  1      1    3    1     0    3
4    1  1      1    3    1     0    4
5    2  1      2    1    4     0    1
6    2  1      2    1    4     0    2
7    2  1      2    1    4     0    3
8    2  1      2    1    4     0    4
9    3  1      1    1    7     0    1
10   3  1      1    1    7     0    2
11   3  1      1    1    7     0    3
12   3  1      1    1    7     0    4
13   4  1      5    1   10     0    1
14   4  1      5    1   10     0    2
15   4  1      5    1   10     0    3
16   4  1      5    1   10     0    4
17   5  1      4    1    6     1    1
18   5  1      4    1   10     0    2
19   5  1      4    1   10     0    3
20   5  1      4    1   10     0    4
```

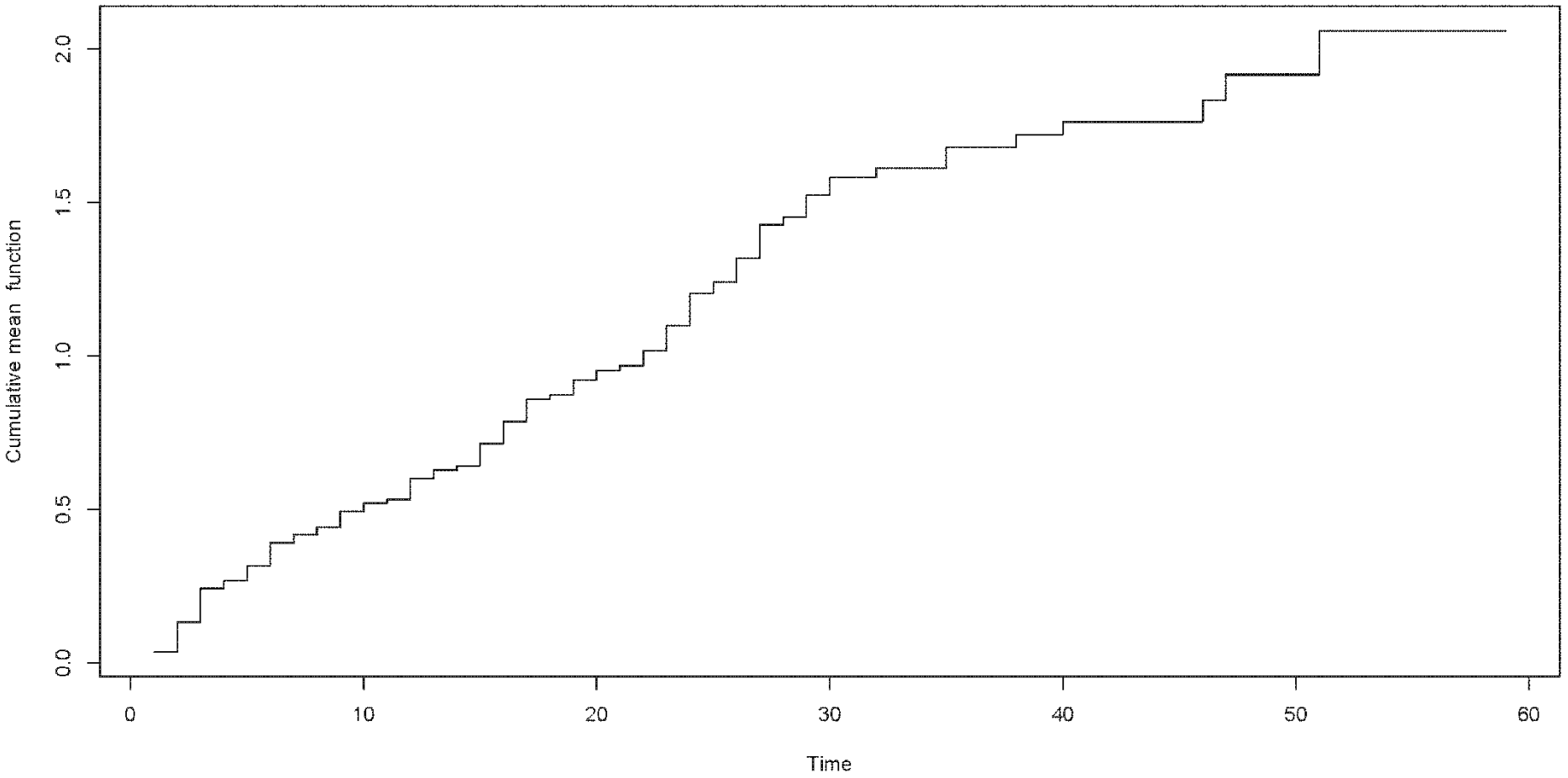
Intensity function vs. Cumulative mean function

- 개체 내 재발사건을 동일한 사건으로 간주
- $t_{i1} < \cdots < t_{im_i} (i = 1, \ldots, n)$: i번째 개체의 재발 사건 시점
- $N_i(t) = \sum_{j=1}^{m_i} I(t_{ij} \leq t)$: i번째 개체의 t 시점까지 발생된 재발사건의 총 횟수
- $H_i(t) = \{N_i(s): 0 \leq s < t\}$: i번째 개체의 t 시점적전까지의 재발사건에 대한 기록
- Intensity function
  - $\lambda(t|H_i(t)) = \lim_{\Delta t \to 0} \frac{P(dN_i(t)=1|H_i(t))}{\Delta t} = \rho(t)$ (w/o covariate)
  - $\lambda(t|H_i(t), z_i) = \rho_i(t) = \rho_0(t) \exp(\beta' z_i)$ (w/ covariate)
- Cumulative mean function
  - $\mu(t) = \int_0^t \rho(s) ds = E\{N_i(t)\}$(w/o covariate)
  - $\mu_i(t|z_i) = \mu_0(t) \exp(\beta' z_i)$ (w/ covariate)
    - $\mu_0(t) = \int \rho_0(s) ds$: baseline cumulative mean function

Non-parametric method

- $t_1 < \cdots < t_H$: distinct event times across all individuals
- NPMLE for $\mu(t)$
  - $\hat{\mu}(t) = \sum_{h:t_h \le t} \frac{\sum_{i=1}^{n} Y_i(t_h) dN_i(t_h)}{\sum_{i=1}^{n} Y_i(t_h)}$
    - $Y_i(t) = I(t \le \tau_i)$, τ_i: termination time, end-of-follow-up time, or censoring time
  - Same as Nelson-Aalen estimator

Bladder data (revisited)



Semi-parametric method

Method	Risk interval	Risk set	Baseline hazard
Andersen & Gill (AG)	Counting process	Unrestricted	Common
Prentice, Williams & Peterson (PWP-CP)	Counting process	Restricted	Event-specific
Prentice, Williams & Peterson (PWP-GP)	Gap time	Restricted	Event-specific
Wei, Lin & Weissfeld (WLW)	Total time	Semi-restricted	Event-specific

Hypothetical example

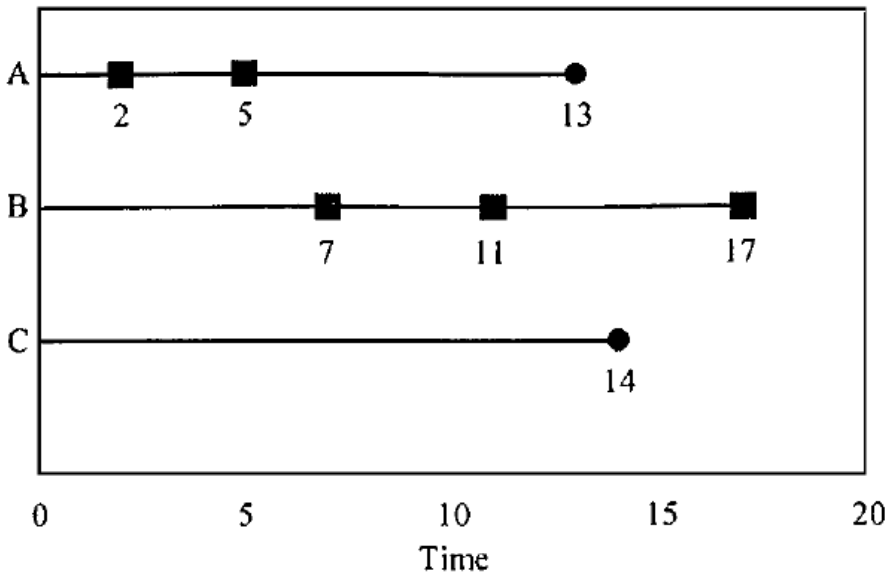


Figure 1. A hypothetical example of three subjects with recurrent events. Observations start at the same time; ■ is the occurrence of an event and ● is censoring. Subject A has two events before being censored; subject B has three events, ending the period of observation with an event; and subject C has no events before being censored

Three types of risk interval

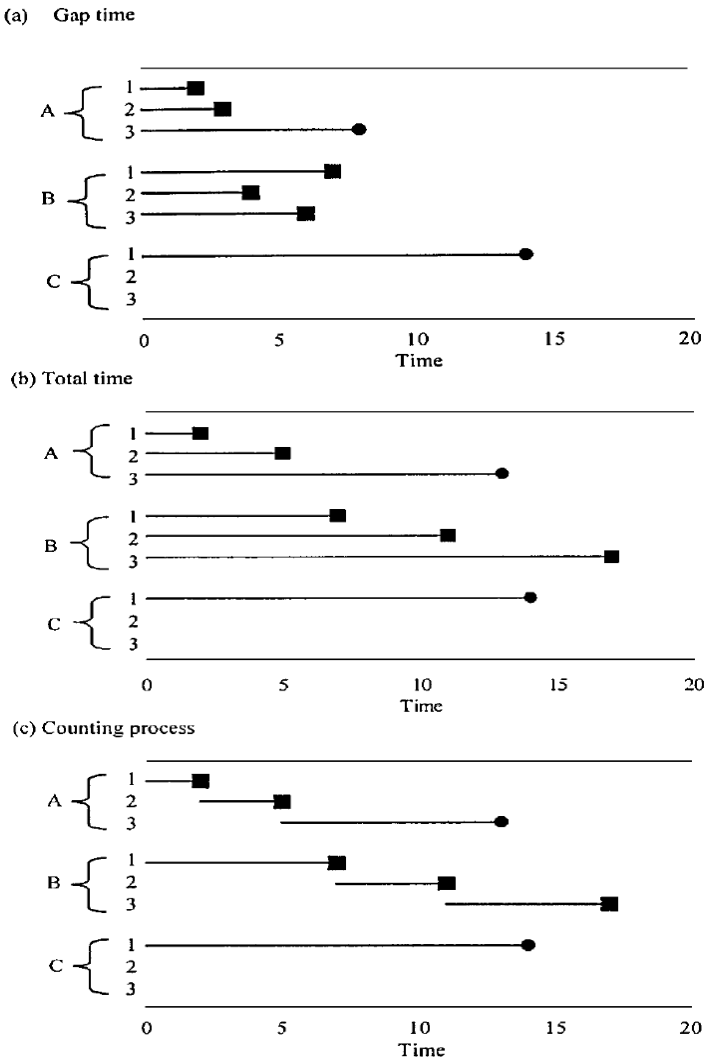


Figure 2. Illustrations of the risk interval formulations: (*a*) gap time; (*b*) total time; (*c*) counting process, using the hypothetical data from Figure 1, where ■ is an event and ● is censoring. Each time to an event or censoring is a separate risk interval, hence subjects A and B have three separate intervals

Partial likelihood

Partial likelihood	Model	Hazard	Example
(a) *Unstratified (unrestricted) models*			
$L(\beta)=\prod_{i=1}^{n}\prod_{k=1}^{K}\left(\frac{e^{\beta' Z_{ik}(X_{ik})}}{\sum_{j=1}^{n}\sum_{l=1}^{K} Y_{jl}(X_{ik})e^{\beta' Z_{jl}(X_{ik})}}\right)^{\delta_{ik}}$	AG	$\lambda_{ik}(t;Z_{ik})=\lambda_0(t)e^{\beta' Z_{ik}(t)}$	$\frac{\lambda_{A1}(2)}{\lambda_{A1}(2)+\lambda_{B1}(2)+\lambda_{C1}(2)}\times\frac{\lambda_{A2}(5)}{\lambda_{A2}(5)+\lambda_{B1}(5)+\lambda_{C1}(5)}$
		$Y_{ik}(t)=I(X_{i,k-1}<t\leqslant X_{ik})$	$\times\frac{\lambda_{B1}(7)}{\lambda_{A3}(7)+\lambda_{B1}(7)+\lambda_{C1}(7)}\times\frac{\lambda_{B2}(11)}{\lambda_{A3}(11)+\lambda_{B2}(11)+\lambda_{C1}(11)}\times\frac{\lambda_{B3}(17)}{\lambda_{B3}(17)}$

Partial likelihood

(b) Stratified models

$$L(\beta) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left(\frac{e^{\beta' Z_{ik}(X_{ik})}}{\sum_{j=1}^{n} Y_{jk}(X_{ik}) e^{\beta' Z_{ik}(X_{ik})}} \right)^{\delta_{ik}}$$

PWP-CP	$\lambda_{ik}(t;Z_{ik}) = \lambda_{0k}(t)e^{\beta' Z_{ik}(t)}$	$\frac{\lambda_{A1}(2)}{\lambda_{A1}(2)+\lambda_{B1}(2)+\lambda_{C1}(2)} \times \frac{\lambda_{A2}(5)}{\lambda_{A2}(5)}$
	$Y_{ik}(t) = I(X_{i,k-1} < t \leqslant X_{ik})$	$\times \frac{\lambda_{B1}(7)}{\lambda_{B1}(7)+\lambda_{C1}(7)} \times \frac{\lambda_{B2}(11)}{\lambda_{B2}(11)} \times \frac{\lambda_{B3}(17)}{\lambda_{B3}(17)}$
PWP-GT	$\lambda_{ik}(t;Z_{ik}) = \lambda_{0k}(t-t_{k-1})e^{\beta' Z_{ik}(t)}$	$\frac{\lambda_{A1}(2)}{\lambda_{A1}(2)+\lambda_{B1}(2)+\lambda_{C1}(2)} \times \frac{\lambda_{A2}(3)}{\lambda_{A2}(3)+\lambda_{B2}(3)}$
	$Z_{jk}(X_{ik})$ replaced by $Z_{ik}(X_{i,k-1}+G_{ik})$	$\times \frac{\lambda_{B1}(7)}{\lambda_{B1}(7)+\lambda_{C1}(7)} \times \frac{\lambda_{B2}(4)}{\lambda_{B2}(4)} \times \frac{\lambda_{B3}(6)}{\lambda_{B3}(6)}$
	$Y_{ik}(t) = I(G_{ik} > t)$	
WLW*	$\lambda_{ik}(t;Z_{ik}) = \lambda_{0k}(t)e^{\beta' Z_{ik}(t)}$	$\frac{\lambda_{A1}(2)}{\lambda_{A1}(2)+\lambda_{B1}(2)+\lambda_{C1}(2)} \times \frac{\lambda_{A2}(5)}{\lambda_{A2}(5)+\lambda_{B2}(5)+\lambda_{C1}(5)}$
	$Y_{ik}(t) = I(X_{ik} \geqslant t)$	$\times \frac{\lambda_{B1}(7)}{\lambda_{B1}(7)+\lambda_{C1}(7)} \times \frac{\lambda_{B2}(11)}{\lambda_{B2}(11)+\lambda_{C1}(11)} \times \frac{\lambda_{B3}(17)}{\lambda_{B3}(17)}$

Bladder data (revisit)

```
> AG.fit
Call:
coxph(formula = Surv(start, stop, event) ~ rx + number + size,
    data = bladder2)


           coef exp(coef) se(coef)      z      p
rx      -0.4647     0.628   0.1997 -2.327 0.0200
number   0.1750     1.191   0.0471  3.717 0.0002
size    -0.0437     0.957   0.0691 -0.632 0.5300

Likelihood ratio test=17.5  on 3 df, p=0.000553  n= 178, number of events= 112

> PWP.CP.fit
Call:
coxph(formula = Surv(start, stop, event) ~ rx + number + size +
    strata(enum), data = bladder2)


            coef exp(coef) se(coef)      z     p
rx      -0.33349     0.716   0.2162 -1.543 0.120
number   0.11962     1.127   0.0533  2.243 0.025
size    -0.00849     0.992   0.0728 -0.117 0.910

Likelihood ratio test=6.51  on 3 df, p=0.0893  n= 178, number of events= 112
```

Bladder data (revisit)

```
> PWP.GP.fit
Call:
coxph(formula = Surv(stop - start, event) ~ rx + number + size +
    strata(enum), data = bladder2)


            coef exp(coef) se(coef)      z      p
rx      -0.27900     0.757   0.2073 -1.346 0.1800
number   0.15805     1.171   0.0519  3.043 0.0023
size     0.00742     1.007   0.0700  0.106 0.9200

Likelihood ratio test=9.33  on 3 df, p=0.0252  n= 178, number of events= 112
> WLW.fit
Call:
coxph(formula = Surv(stop, event) ~ rx + number + size + strata(enum),
    data = bladder)


           coef exp(coef) se(coef)     z       p
rx      -0.5848     0.557   0.2011 -2.91 3.6e-03
number   0.2103     1.234   0.0468  4.50 6.9e-06
size    -0.0516     0.950   0.0697 -0.74 4.6e-01

Likelihood ratio test=25.3  on 3 df, p=1.36e-05  n= 340, number of events= 112
```

군집생존자료분석
재발사건자료분석
다상태모형
경쟁위험모형

다변량 생존자료분석

Multi-state model

- 두 가지 상태(0:alive, 1:dead)를 가진 생존자료
  - 1: absorbing state (흡수)
- K개의 상태, $\tilde{K} = \{1, \ldots, K\}$, 를 가진 다상태 모형
- Illness-death model: 3 states (0: healthy, 1: diseased, 2:dead)

Transition intensity vs. transition probability

- $P_{hj}(s,t) = P(X(t) = j | X(s) = h, F_s-)$: transition probability
- $\alpha_{hj}(t) = \lim_{\Delta t \to 0} \frac{P_{hj}(t,t+\Delta t)}{\Delta t}$: transition intensity
  - eg, in illness-death model, $\alpha_{01}(t), \alpha_{02}(t), \alpha_{12}(t)$
  - eg, in competing risks model, $\alpha_{01}(t), \ldots, \alpha_{0K}(t)$
- Relation
  - eg, in illness-death model, $P_{00}(s,t) = \exp\{-\int(\alpha_{01}(s) + \alpha_{02}(s))ds\}$
  - eg, in competing risks model, $P_{00}(s,t) = \exp\{-\int -\sum_{k=1}^{K} \alpha_{0k}(s)ds\}$

Estimation of transition probabilities in finite-state Markov processes

- $\alpha_{hji}(t|z_i) = \alpha_{hj0}(t)\exp(\beta'_{hj}z_i)$: transition intensity from state h to state j for the individual i
- Using $\hat{\beta}$ and $\hat{A}_{hj0}(t,\hat{\beta})$, obtain $\hat{A}_{hj}(t|z) = \hat{A}_{hj0}(t,\hat{\beta})\exp(\hat{\beta}'z)$ for $h \neq j$ and $\hat{A}_{hh}(t|z) = -\sum_{j\neq h}\hat{A}_{hj}(t|z)$, and then $P_{hj}(s,t|z)$ can be estimated

BMT data (revisit)

처리그룹	Patients	Platelet recovery 0→1	Relapse or death 0→2	After plate recovry, relapse or death 1→2
ALL	38	35	3	12
AML low risk	54	48	6	9
AML high risk	45	38	7	18
Total	137	121	16	67

BMT (revisited)

```
> p.fit
Call:
coxph(formula = Surv(TP, PI) ~ factor(DG) + p28 * d28 + FAB,
    data = bmt)

                coef exp(coef) se(coef)      z      p
factor(DG)2  0.35769     1.430 0.249368  1.434 0.150
factor(DG)3  0.15916     1.173 0.298347  0.533 0.590
p28          0.01623     1.016 0.016091  1.009 0.310
d28         -0.01447     0.986 0.013804 -1.049 0.290
FAB         -0.11727     0.889 0.236910 -0.495 0.620
p28:d28     -0.00168     0.998 0.000886 -1.901 0.057

Likelihood ratio test=8.83  on 6 df, p=0.184  n= 137, number of events= 120
> d.fit
Call:
coxph(formula = Surv(TDF, status) ~ factor(DG) + p28 * d28 +
    FAB, data = bmt)

                coef exp(coef) se(coef)      z      p
factor(DG)2  0.74966     2.116  0.75997  0.986 0.3200
factor(DG)3  0.74242     2.101  0.91818  0.809 0.4200
p28         -0.14357     0.866  0.05168 -2.778 0.0055
d28          0.13241     1.142  0.04442  2.981 0.0029
FAB         -0.05788     0.944  0.67271 -0.086 0.9300
p28:d28      0.00505     1.005  0.00184  2.743 0.0061

Likelihood ratio test=16.6  on 6 df, p=0.0107  n= 137, number of events= 16
> pd.fit
Call:
coxph(formula = Surv(TP, TDF, pdstatus) ~ factor(DG) + p28 *
    d28 + FAB, data = bmt)

                coef exp(coef) se(coef)     z       p
factor(DG)2 -1.71593     0.180  0.42548 -4.03 5.5e-05
factor(DG)3 -0.75516     0.470  0.40750 -1.85 6.4e-02
p28          0.03860     1.039  0.02178  1.77 7.6e-02
d28         -0.02927     0.971  0.02050 -1.43 1.5e-01
FAB          1.21197     3.360  0.32223  3.76 1.7e-04
p28:d28      0.00268     1.003  0.00123  2.17 3.0e-02

Likelihood ratio test=37.8  on 6 df, p=1.22e-06  n= 120, number of events= 67
   (17 observations deleted due to missingness)
```

군집생존자료분석
재발사건자료분석
다상태모형
경쟁위험모형

다변량 생존자료분석

Competing risks model

- Multi-state model의 일종, absorbing state가 여러 개 있음
- 한 사건의 발생은 다른 사건의 발생을 중도절단시킴
  - eg, 주관심이 암의 재발일 때 어떤 환자들이 사망한 경우, 사망사건의 발생은 암 재발에 대해서는 중도절단 사건이 됨
  - $K = 2$ 일 때, 한 사건이 다른 사건으로의 전이를 허용하지만 그 역은 성립하지 않는 경쟁위험모형을 준경쟁위험모형(semi-competing risks model)이라고 함

Cause-specific cumulative incidence function

- $\left(T = \min\{T_j : j = 1, \ldots, K\}, \epsilon = argmin_j\{T_j\}\right)$
  - T_j: j번째 원인에 의한 사건발생시간
- $\lambda_j(t) = \lim_{h \to 0} \frac{P(t \le T < t+h, \epsilon = j | T \ge t)}{h}, j = 1, \ldots, K$: 원인별 위험함수
  - $\lambda(t) = \sum_{j=1}^{K} \lambda_j(t)$: t시점에서 어떤 원인에 의해서든 실패할 확률
- $F_j(t) = P(T \le t, \epsilon = j) = \int_0^t S(s-)\lambda_j(s)ds = \int_0^t \lambda_j(s) \exp(-\int_0^s \lambda(u)du)\, ds$: cause-specific cumulative incidence function(CIF), 즉 모든 다른 실패 원인이 존재할 때, t시점까지 원인 j에 의해 실패할 확률
  - $S(t) = P(T > t) = \exp\left\{\int_0^t \lambda(s)ds\right\}$: overall survival function

Estimated CIF

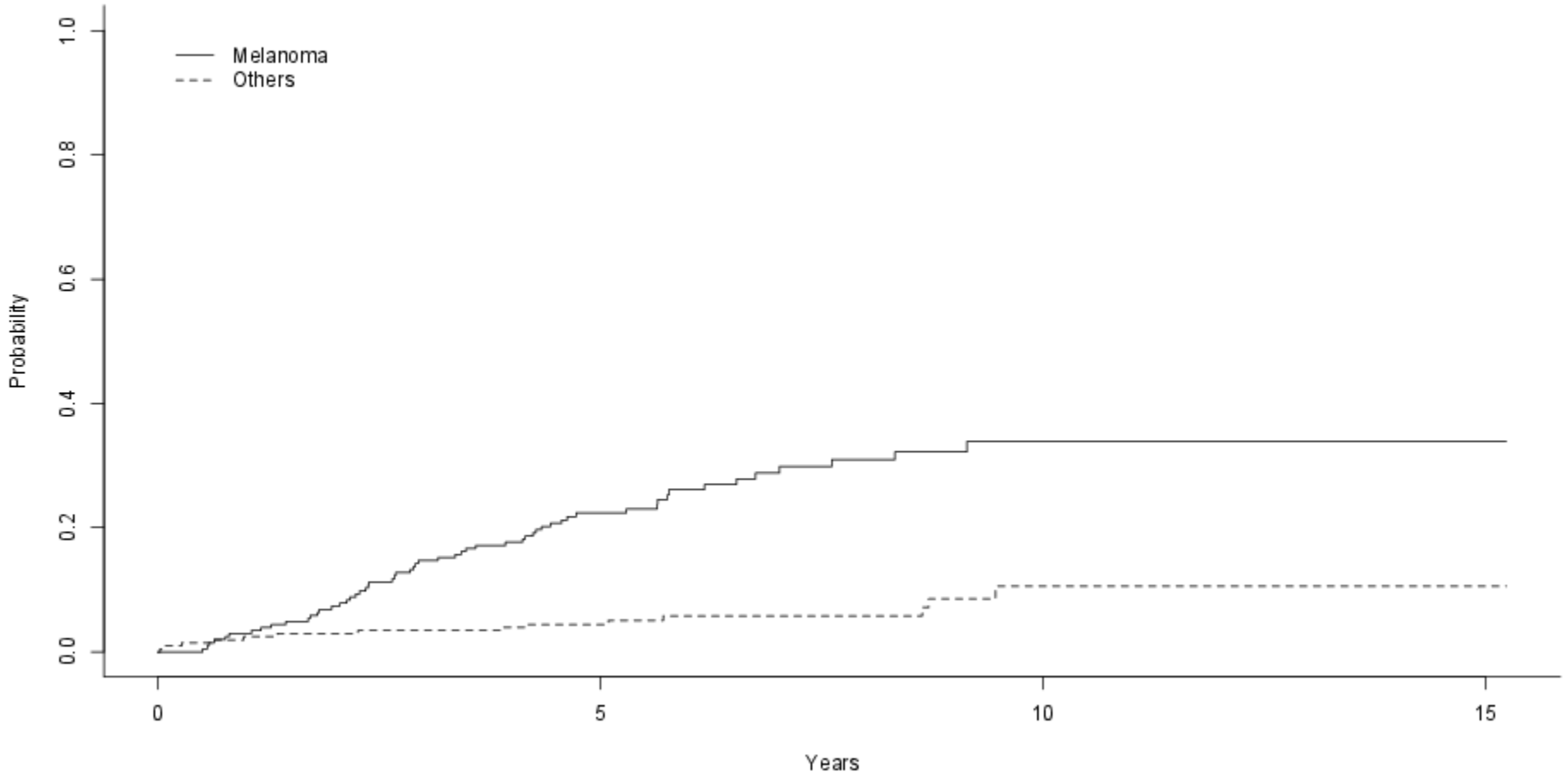
- $t_1 < \cdots < t_r$: unique ordered uncensored time points
- d_{ij}: # of events of type j that occur at time t_j
- n_i: # at risk at t_i
- $\hat{F}_j(t) = \sum_{m:t_m \leq t} \frac{d_{ij}}{n_m} \hat{S}(t_m -)$: estimated CIF
  - $\hat{S}$: PL estimator of the probability of being free of any event by time t
  - $\hat{F}_1(t) \leq 1 - \hat{S}_1(t)$ ('1': event of interest, '2': set of competing risks event), i.e., at risk of being overestimated

Melanoma data

- 피부암 환자 256명을 대상으로 사망원인을 조사
  - 피부암이 원인: 57(status=1), 중도절단: 134(status=2), 다른 원인: 14(status=3),

```
> head(melanoma)
  no status days ulc thick sex
1 789      3   10   1   676   1
2  13      3   30   0    65   1
3  97      2   35   0   134   1
4  16      3   99   0   290   0
5  21      1  185   1  1208   1
6 469      1  204   1   484   1
```

Cause-specific incidence function



Cause-specific hazard function approach

- Cox model for CSH
  - $\lambda_j(t|z_i) = \lambda_{0j}(t)\exp\left(\beta_j' z_i\right), j = 1, \ldots, K$
- Use the partial likelihood principle
  - $t_{1j} < \cdots < t_{n_j,j}$: n_j times of type j failures
  - z_{ij}: covariate for the individual that fails at t_{ij}
  - $L(\beta_1, \ldots, \beta_K) = \prod_{j=1}^{K} \prod_{i=1}^{n_j} \frac{\exp\left\{\beta_j' z_{ij}\right\}}{\sum_{l \in R(t_{ij})} \exp\left\{\beta_j' z_l\right\}}$

Melanoma data (revisited)

```
> m.ftype.fit
Call:
coxph(formula = Surv(days, m.ftype) ~ ulc + thick + sex, data = melanoma)


         coef exp(coef) se(coef)    z       p
ulc   1.16681      3.21 0.311461 3.75 0.00018
thick 0.00113      1.00 0.000379 2.99 0.00280
sex   0.45949      1.58 0.266758 1.72 0.08500

Likelihood ratio test=39.4  on 3 df, p=1.44e-08  n= 205, number of events= 57
> o.ftype.fit
Call:
coxph(formula = Surv(days, o.ftype) ~ ulc + thick + sex, data = melanoma)


          coef exp(coef) se(coef)     z    p
ulc   0.189846      1.21 0.588614 0.323 0.75
thick 0.000904      1.00 0.000844 1.071 0.28
sex   0.521108      1.68 0.543930 0.958 0.34

Likelihood ratio test=2.91  on 3 df, p=0.405  n= 205, number of events= 14
```

Sub-distribution function approach

- j번째 CIF 는 j번째 원인별 위험함수뿐만 아니라 j번째 원인을 제외한 다른 원인별 위험함수의 합에도 의존
- $F_j(t|z) = P(T \le t, \epsilon = j|z)$: sub-distribution function
  - $\lambda_j^s(t|z) = \lim_{\Delta t \to 0} \frac{P(t \le T_j < t+\Delta t | (T \ge t) \cup (T \le t \cap \epsilon \ne j), z)}{\Delta t}$: sub-distribution hazard function
  - 다른 유형으로 이미 사망한 사람도 위험 그룹에 포함됨
  - $F_j(t|z) = 1 - \exp\left\{-\int_0^t \lambda_j^s(s)ds\right\}$
- Parameter estimation: under Cox PH model,
  - Censoring complete data: 중도절단시점을 사전에 앎
  - Inverse probability censoring weight approach: $w_i(t) = I(C_i \ge T_i \wedge t)\frac{\hat{G}_c(t)}{\hat{G}_c(T_i \wedge t)}$, $\hat{G}_c$: PL estimator of G_c

Melanoma data (revisited)

```
> m.crr
convergence:  TRUE
coefficients:
  melanoma.ulc melanoma.thick   melanoma.sex
     1.1360000      0.0009428      0.4187000
standard errors:
[1] 0.3041000 0.0003809 0.2743000
two-sided p-values:
  melanoma.ulc melanoma.thick   melanoma.sex
       0.00019        0.01300        0.13000
> o.crr
convergence:  TRUE
coefficients:
  melanoma.ulc melanoma.thick   melanoma.sex
    -0.0044030      0.0006069      0.4257000
standard errors:
[1] 0.5953000 0.0007646 0.5509000
two-sided p-values:
  melanoma.ulc melanoma.thick   melanoma.sex
          0.99           0.43           0.44
```

References

- Kalbfleish, JD and Prentice, RL. (2002). The Statistical Analysis of Failure Time Data, Second Edition, Wiley.
- Moeschberger, ML and Klein, JP. (2003). Survival Analysis: Techniques for Censored and Truncated Data, Springer.
- Andersen, PK, Borgan, O, Gill, RD, and Keiding, N. (1999). Statistical Models Based on Counting Processes, Springer.
- Sun, J. (2006). The Statistical Analysis of Interval-censored Failure Time Data, Springer.
- Chen, D-G, Sun, J, and Peace, KE. (2012). Interval-censored Time-to-event Data: Methods and Applications, CRC.

THANK YOU!!!