생존 분석과 EM알고리즘

2019.07.31.

박 은 령

eunlyeong0530@gmail.com

9.0 Intro

- 생존자료(Survival data)

  : 각 개인이 관심 결과변수(사망 또는 질병의 재발 등)에 도달할 때까지 걸린 시간과 관련된 자료를 의미.

- **Random variable: 기간(time), 사건발생(event)**

  ex) 특정 질병으로의 사망 ⇒ time: 질병발생일 ~ 사망일 / event: 사망여부

  폐암 수술 후 특정 질병 발생 ⇒ time: 폐암수술일 ~ 질병발생일 / event: 질병발생여부

- 주요 특성

  ① 환자가 종료점에 도달할 때까지 걸린 시간(length of time)

  ② 중도절단(censored) : 환자가 실제 종료점에 도달한 시점을 모르는 경우

  ※ event가 발생하지 않으면, censored.

  time: censored time 이용 (study end time, f/u loss time…)

9.1 Life tables and Hazard rates

- **생명표(life table):** 일정한 간격마다 구간생존율을 구하는 방법. 각 개인별로 종료점에 도달하는 시간은 알 수 없지만 일정구간 내에 종료점이 발생한 일정구간 내(ex. 1year) 종료점이 발생한 N수는 알 수 있는 경우에 사용. → *일반적으로 Kaplan-Meier Estimate 사용*

✓ 생존확률(Survival Probability): 기준시점 이후의 모든 시점에 대해 각 환자가 종료점에 도달하지 않을 수 있는 누적 확률, S(t)

$$P(t) = \frac{t \text{ 시점의 생존자 수}}{t \text{ 시점의 대상자수}}$$

$$S(t) = 1 - F(t) = \Pr(T > t) = S(t-1) \times P(t)$$

$$\widehat{S(t)} = \prod_{i;t_i<t} \frac{n_i - y_i}{n_i}$$

t	$S(t)$
1	$S(1) = P(T > 1)$
2	$S(2) = P(T > 2)$
3	$S(3) = P(T > 3)$
.	.
.	.
.	.

Theoretical $S(t)$:

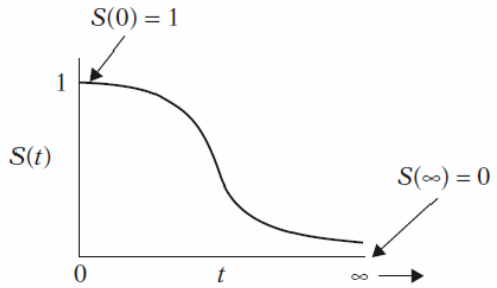


Table 9.1 *Insurance company life table; at each age, n = number of policy holders, y = number of deaths, $\hat{h}$ = hazard rate y/n, $\hat{S}$ = survival probability estimate (9.6).*

Age	n	y	$\hat{h}$	$\hat{S}$	Age	n	y	$\hat{h}$	$\hat{S}$
30	116	0	.000	1.000	60	231	1	.004	.889
31	44	0	.000	1.000	61	245	5	.020	.871
32	95	0	.000	1.000	62	196	5	.026	.849
33	97	0	.000	1.000	63	180	4	.022	.830
34	120	0	.000	1.000	64	170	2	.012	.820
35	71	1	.014	.986	65	114	0	.000	.820
36	125	0	.000	.986	66	185	5	.027	.798
37	122	0	.000	.986	67	127	2	.016	.785
38	82	0	.000	.986	68	127	5	.039	.755
39	113	0	.000	.986	69	158	2	.013	.745
40	79	0	.000	.986	70	100	3	.030	.723
41	90	0	.000	.986	71	155	4	.026	.704
42	154	0	.000	.986	72	92	1	.011	.696
43	103	0	.000	.986	73	90	1	.011	.689
44	144	0	.000	.986	74	110	2	.018	.676
45	192	2	.010	.976	75	122	5	.041	.648
46	153	1	.007	.969	76	138	8	.058	.611
47	179	1	.006	.964	77	46	0	.000	.611
48	210	0	.000	.964	78	75	4	.053	.578
49	259	2	.008	.956	79	69	6	.087	.528
50	225	2	.009	.948	80	95	4	.042	.506
51	346	1	.003	.945	81	124	6	.048	.481
52	370	2	.005	.940	82	67	7	.104	.431
53	568	4	.007	.933	83	112	12	.107	.385
54	1081	8	.007	.927	84	113	8	.071	.358
55	1042	2	.002	.925	85	116	12	.103	.321
56	1094	10	.009	.916	86	124	17	.137	.277
57	597	4	.007	.910	87	110	21	.191	.224
58	359	1	.003	.908	88	63	9	.143	.192
59	312	5	.016	.893	89	79	10	.127	.168

9.2 Censored Data and the Kaplan–Meier Estimate

- Data from a randomized clinical trial run by the Northern California Oncology Group comparing two treatments for head and neck cancer:
  - Chemotherapy (Arm A)
  - Chemotherapy + Radiation (Arm B)
- Survival time in days
- Kaplan–Meier (Nonparametric Survival Function Estimation)
  : 사건이 발생하는 시점마다 구간생존율을 구하여
  이들의 누적으로 누적생존율을 추정하는 방식.

Table 9.2 *Censored survival times in days, from two arms of the NCOG study of head/neck cancer.*

Arm A: Chemotherapy

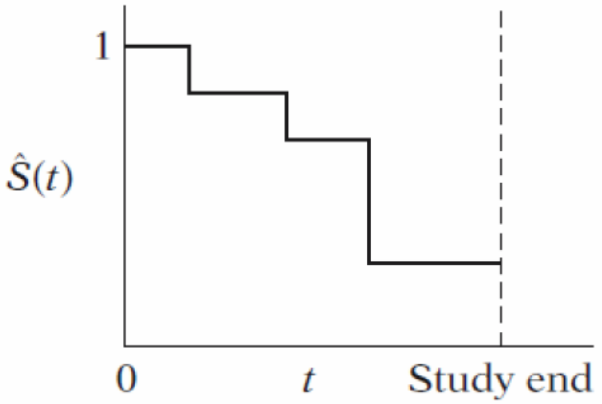
7	34	42	63	64	74+	83	84	91
108	112	129	133	133	139	140	140	146
149	154	157	160	160	165	173	176	185+
218	225	241	248	273	277	279+	297	319+
405	417	420	440	523	523+	583	594	1101
1116+	1146	1226+	1349+	1412+	1417			

Arm B: Chemotherapy+Radiation

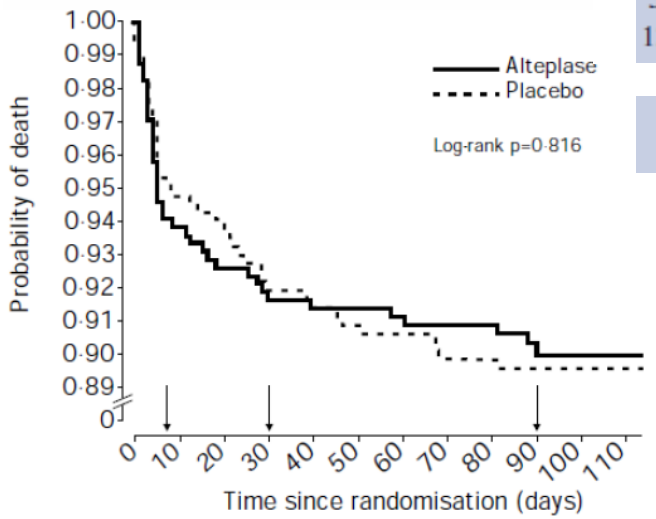
37	84	92	94	110	112	119	127	130
133	140	146	155	159	169+	173	179	194
195	209	249	281	319	339	432	469	519
528+	547+	613+	633	725	759+	817	1092+	1245+
1331+	1557	1642+	1771+	1776	1897+	2023+	2146+	2297+

+ Lost to f/u

$\hat{S}(t)$ in practice:



(ECASS II Study (1998))



9.2 Censored Data and the Kaplan–Meier Estimate

- Ex. WHAS Study

Table 2.1 Survival Times and Vital Status (Censor=1 for deaths) for Five Subjects from the WHAS Study

Subject	Time	Censor
1	6	1
2	44	1
3	21	0
4	14	1
5	62	1

Table 2.2 Estimated Survival Function Computed from the Survival Times for the Four Subjects from the WHAS100 Study and One Hypothetical Subject Shown in Table 2.1

Interval	Conditional probability	$\hat{S}(t)$
$0 \le t < 6$	1.0	1.0
$6 \le t < 14$	$1.0 \times (4/5) = 0.8$	0.8
$14 \le t < 21$	$1.0 \times (4/5) \times (3/4) = 0.6$	0.6
$21 \le t < 44$	$1.0 \times (4/5) \times (3/4) \times (3/3) = 0.6$	0.6
$44 \le t < 62$	$1.0 \times (4/5) \times (3/4) \times (3/3) \times (1/2) = 0.3$	0.3
$t \ge 62$	$1.0 \times (4/5) \times (3/4) \times (3/3) \times (1/2) \times (0/1) = 0$	0.0

Reference: David W. Hosmer and Stanley Lemeshow. 2008. Applied Survival Analysis: Regression Modeling of Time to Event Data (2nd ed.). John Wiley & Sons, Inc., New York, NY, USA.

- Kaplan–Meier Estimate

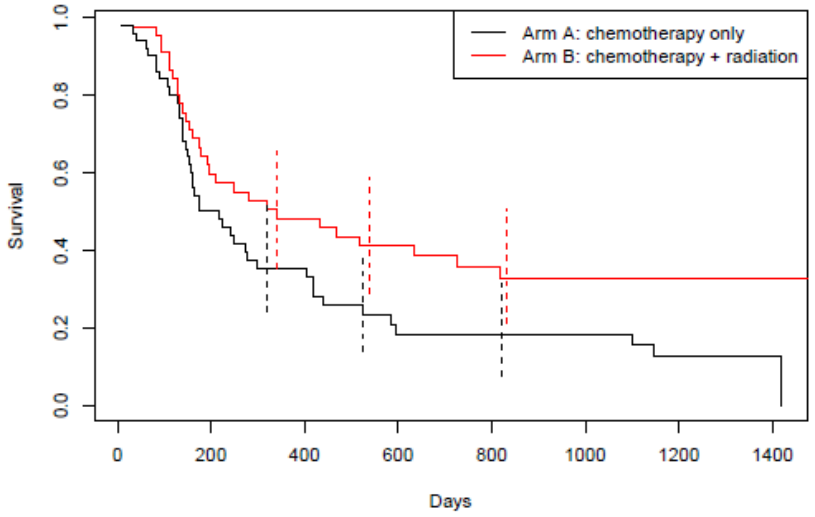


Figure 9.1 NCOG Kaplan–Meier survival curves; lower Arm_A (chemotherapy only); upper Arm_B (chemotherapy+radiation). Vertical lines indicate approximate 95% confidence intervals.

$$\hat{S}_{(j)} = \prod_{k \le j} \left(\frac{n-k}{n-k+1} \right)^{d_{(k)}}$$

$$\mathrm{sd}\left(\hat{S}_{(j)}\right) = \hat{S}_{(j)} \left[\sum_{k \le j} \frac{y_k}{n_k (n_k - y_k)} \right]^{1/2}$$: Greenwood formula

9.2 Censored Data and the Kaplan–Meier Estimate

- Parametric approach 는 곡선의 정확도를 크게 향상시킨다

$$y_k \overset{\text{ind}}{\sim} \text{Bi}(n_k, h_k), \tag{9.19}$$

and that the logits $\lambda_k = \log\{h_k/(1-h_k)\}$ satisfy some sort of regression equation

$$\lambda = X\alpha, \tag{9.20}$$

- The parametric hazard-rate estimates were instead based on a cubic-linear spline.
- The logistic regression maximum likelihood estimate $\hat{\alpha}$ produced hazard rate curves.

$$\hat{h}_k = 1 \Big/ \left(1 + e^{-x_k'\hat{\alpha}}\right)$$

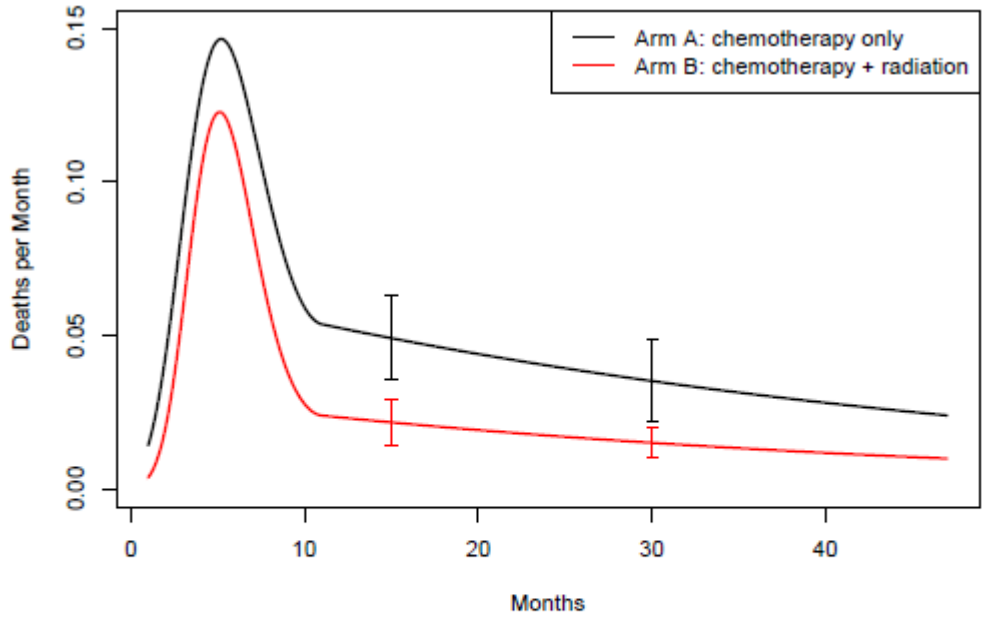


Figure 9.2 Parametric hazard rate estimates for the NCOG study. Arm_A, black curve, has about 2.5 times higher hazard than Arm_B for all times more than a year after treatment. Standard errors shown at 15 and 30 months.

9.3 The Log-Rank Test

- Provides overall comparison of KM curves.
- Uses observed versus expected counts over categories of outcomes, where categories are defined by ordered failure times for entire set of data.
- **즉, 집단간 생존시간에 유의한 차이가 없다는 귀무가설을 검정하기 위한 비모수적인 검정**
- **한 요인에 대해 생존곡선 상의 모든 시점에서 발생하는 사건들을 서로 비교하는 방법**
- 단, 두개 이상의 서로 독립적인 관심요인들이 종료점이 발생할 때까지의 시간에 어떠한 영향을 미치는지 비교하고자 할 때는 사용할 수 없음 → 9.4 비례위험모형 사용

H_0 : There is no overall difference between the two survival curves.

H_1 : Not H_0

$$\text{Log-rank statistic} = \frac{(O_2 - E_2)^2}{\text{Var}(O_2 - E_2)} \sim \chi^2 \text{ with 1 } df \text{ under } H_0$$

where $O_2 - E_2$ = summed observed minus expected score for group 2,

$$\text{Var}(O_i - E_i) = \sum_j \frac{n_{1j} n_{2j} (m_{1j} + m_{2j})(n_{1j} + n_{2j} - m_{1j} - m_{2j})}{(n_{1j} + n_{2j})^2 (n_{1j} + n_{2j} - 1)}, \quad i = 1, 2$$

9.4 The Proportional Hazards Model

- **콕스의 비례위험모형**: hazard에 대해 설명변수들(요인들)의 독립적인 효과를 검정. Semi-parametric Model

  ($\because h_0(t)$: unspecified function)

- Similar to logistic regression, but Cox regression assesses relationship between survival time and covariates .

$$h(t, \mathbf{X}) = h_0(t)\, e^{\sum_{i=1}^{p} \beta_i X_i}$$

where $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ explanatory/predictor variables

$h_0(t) \times e^{\sum_{i=1}^{p} \beta_i X_i}$

Baseline hazard	Exponential
Involves t but not X's	Involves X's but not t (X's are time-independent)

$X_1 = X_2 = \cdots = X_p = 0$

$$h(t, \mathbf{X}) = h_0(t)\, e^{\sum_{i=1}^{p} \beta_i X_i} = h_0(t)\, e^0 = h_0(t)$$

Baseline hazard

Table 9.6 *Pediatric cancer data, first 20 of 1620 children.* `Sex` *1 = male, 2 = female;* `race` *1 = white, 2 = nonwhite;* `age` *in years;* `entry` *= calendar date of entry in days since July 1, 2001;* `far` *= home distance from treatment center in miles;* `t` *= survival time in days;* `d` *= 1 if death observed, 0 if not.*

sex	race	age	entry	far	t	d
1	1	2.50	710	108	325	0
2	1	10.00	1866	38	1451	0
2	2	18.17	2531	100	221	0
2	1	3.92	2210	100	2158	0
1	1	11.83	875	78	760	0
2	1	11.17	1419	0	168	0
2	1	5.17	1264	28	2976	0
2	1	10.58	670	120	1833	0
1	1	1.17	1518	73	131	0
2	1	6.83	2101	104	2405	0
1	1	13.92	1239	0	969	0
1	1	5.17	518	117	1894	0
1	1	2.50	1849	99	193	1
1	1	.83	2758	38	1756	0
2	1	15.50	2004	12	682	0
1	1	17.83	986	65	1835	0
2	1	3.25	1443	58	2993	0
1	1	10.75	2807	42	1616	0
1	2	18.08	1229	23	1302	0
2	2	5.83	2727	23	174	1

9.4 The Proportional Hazards Model

$$\widehat{HR} = \frac{\hat{h}(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})} = \frac{\cancel{\hat{h}_0(t)}\, e^{\sum_{i=1}^{p} \hat{\beta}_i X_i^*}}{\cancel{\hat{h}_0(t)}\, e^{\sum_{i=1}^{p} \hat{\beta}_i X_i}} = e^{\sum_{i=1}^{p} \hat{\beta}_i (X_i^* - X_i)}$$

where $\mathbf{X}^* = (X_1^*, X_2^*, \cdots, X_p^*)$ and $\mathbf{X} = (X_1, X_2, \cdots, X_p)$ denote the set of X's for two individuals

- 해석
  - Age is a mildly significant factor, with older children doing better (i.e., the estimated regression coefficient is negative).
  - However, the dramatic effects are date of entry and far. Individuals who entered the study later survived longer – Perhaps the treatment protocol was being improved – while children living farther away from the treatment center did worse.

Table 9.7 *Proportional hazards analysis of pediatric cancer data (`age`, `entry` and `far` standardized). `Age` significantly negative, older children doing better; `entry` very significantly negative, showing hazard rate declining with calendar date of entry; `far` very significantly positive, indicating worse results for children living farther away from the treatment center. Last two columns show limits of approximate 95% confidence intervals for* $\exp(\beta)$.

	β	sd	z-value	p-value	$\exp(\beta)$	Lower	Upper
`sex`	−.023	.160	−.142	.887	.98	.71	1.34
`race`	.282	.169	1.669	.095	1.33	.95	1.85
`age`	−.235	.088	−2.664	.008	.79	.67	.94
`entry`	−.460	.079	−5.855	.000	.63	.54	.74
`far`	.296	.072	4.117	.000	1.34	1.17	1.55

9.4 The Proportional Hazards Model

참고1) 비례위험가정(proportional hazard assumption)

- 콕스의 비례위험가정은 전체 연구기간 동안 상대위험이 일정하게 유지된다는 가정 (=비교하고자 하는 집단들 간의 위험(hazard)은 서로 일정하게 비례(proportional)한다는 가정)이 필요

$$\frac{h(t, \mathbf{X}^*)}{h(t, \mathbf{X})} = \theta$$

- 그림을 이용하는 방법: 로그-로그 그림[x축: log(시간), y축: log(log(생존확률))]에서 선들이 대략적으로 평행한지 확인.
  이때, 교차하게 되면 비례위험 가정을 위배.
- 모형에 공변량과 log(시간)의 상호작용 항을 포함시켜 분석한 후, 이 항이 유의하지 않는지를 확인.
- 쉔펠드 잔차에 기초한 전반적인 카이제곱 검정(global chi-squared test based on Schoenfeld residuals)

- 위 가정을 위배하는 경우, 위험함수의 형태에 대해 특정한 확률분포를 가정하는 모형들인 지수모형(exponential model), 웨이블 모형(Weibull model) 또는 곰페르츠 모형(Gompertz model) 등을 사용할 수 있다.

참고2) 경쟁위험(competing risk model)

- 서로 다른 여러 개의 결과변수가 관심모수일때, 이 중 일부가 발생함으로 인해 다른 결과변수들이 발생하지 않는 경우

9.5 Missing Data and the EM Algorithm

- Imputation of missing data
  - 단일대체: 명백한 모형을 근거로 한 대체방법
    - ✓ 다변량 정규분포를 가정 : 평균대체 , 회귀대체 , 확률적 회귀 대체
    - ✓ 여러가지 분포를 가정 : 순차회귀 다중대체
    - ✓ 함축적인 모형을 근거로 한 대체방법 : 핫덱대체 , 콜드덱대체
  - 다중대체 (Multiple imputation)
  - EM 대체
- EM 알고리즘(Expectation–Maximization algorithm)

  : 우도함수를 최대화 시키는 모수를 찾는 방법,

  즉, MLE를 찾는 방법 중 하나.

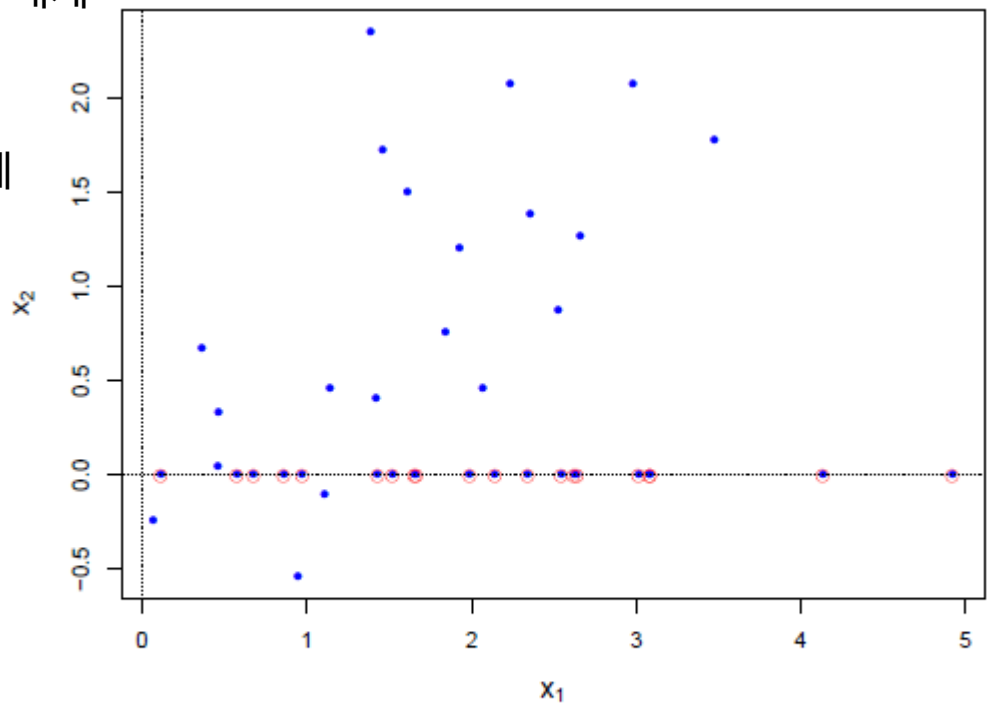


Figure 9.3 Forty points from a bivariate normal distribution, the last 20 with x_2 missing (circled).