생존분석 (Survival Analysis)

연세의과대학 의학통계학과 김 동 기

1. 생존분석의 기초

1.1 생존시간

가. 생존시간(survival time): 어떤 사건(event : 사망)이 발생할 때까지의 시간(time)

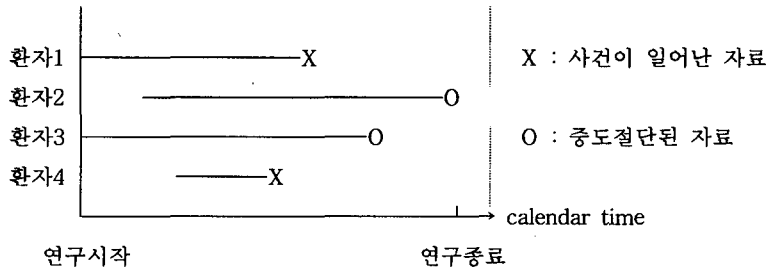
From	To
진단	사망
완치	재발

나. 생존분석(Survival analysis): 생존시간과 관련된 제반 통계학적 분석

다. 생존시간의 특징
- 시간변수가 항상 양의(positive) 수이다.
- 매우 치우친 분포이다(highly skewed distribution)
- 중도절단된 자료(censored data)를 포함하고 있다.

1.2 중도절단(censoring)

가. 중도절단 자료의 예



나. 중도절단의 이유
- loss to follow up
- drop out
- termination of the study
- death from unrelated cause

다. 중도절단의 형태

- Right censoring: $T \in [R, \infty)$
- Left censoring: $T \in (0, L]$
- Interval censoring: $T \in [L, R]$

1.3 생존함수

가. 확률함수

- 생존시간 임의변수: T
- 확률밀도함수: f(t)
- 분포함수: $F(t)=Pr(T \leq t)$

나. 생존함수

- $S(t)=Pr(T \geq t)$
- t시점까지 사망하지 않고 생존할 확률

다. 위험함수(hazard function)

- $h(t)=f(t)/S(t)$
- t시점까지 생존한 사람이 t시점 바로 직후에서 순간적으로 사망할 조건확률
- hazard rate, failure rate, force of mortality
- $S(t)=\exp(-H(t))$, H(t)= 누적위험함수

라. 위험함수의 형태

모형	위험함수의 형태
Exponential	일정
Weibull	단조감소 혹은 단조증가
Lognormal	증가하다가 감소

1.4 일반적인 통계분석방법과 생존분석의 비교

	일반적 방법	생존분석
자료의 요약	기술통계량 - 히스토그램, 돗수분포표 - 평균,중위수,분산,사분위수범위	기술통계량 - life table 작성 - survival curve의 작성
k개 집단의 평균 비교	모수적 방법 - t-검정, 분산분석 비모수적 방법 -윌콕슨검정,크루스칼-왈리스검정	모수적 방법 - 우도비검정(likelihood ratio test) 비모수적 방법 - 로그순위검정(log-rank test) - 일반화된 윌콕슨검정
회귀분석	다중회귀분석 로지스틱회귀분석	모수적 회귀모형 Cox'의 비례위험회귀모형

2. 생존함수의 추정

2.1 Life table method

- 시간을 (k+1)개의 구간 $I_j=[a_{j-1}, a_j), j=1,2,\ldots,k+1$ 로 나눈다.
- N_j=시간 a_{j-1}에서 아직 살아 있는("at risk") 사람의 수
- $D_j=I_j$ 구간에서 죽은 사람의 수
- $W_j=I_j$에서 중도 절단된("withdrawals") 사람의 수
- $q_j=\frac{D_j}{N_j-W_j/2}$, $p_j=1-q_j$
- Probability of survival: $S_j=p_1\cdot p_2\cdot\cdots\cdot p_j, j=1,2,\ldots,k+1$
- variance of survival(Greenwood)

$$var(S_j)=S_j^2\sum_{i=1}^{j}\frac{q_j}{p_j(N_j-0.5W_j)}$$

Ex: Life Tables

Interval in years (I_j)	No. at risk (N_j)	No. of withdrawals (W_j)	No. of deaths (D_j)	N_j'	q_j	p_j	Estimated survival P_j
[0,1)	374	0	90	374	0.241	0.759	0.959
[1,2)	284	0	76	284	0.268	0.732	0.556
[2,3)	208	0	51	208	0.245	0.755	0.420
[3,4)	157	12	25	151	0.164	0.834	0.350
[4,5)	120	5	20	117.5	0.170	0.830	0.291
[5,6)	95	9	7	90.5	0.077	0.923	0.268
[6,7)	79	9	4	74.5	0.054	0.946	0.254
[7,8)	66	3	1	64.5	0.016	0.984	0.250
[8,9)	62	5	3	59.5	0.05	0.950	0.237
[9,10)	54	5	2	51.5	0.039	0.961	0.228
[10,∞)	47	0	47	47.0	1.000	0.000	0.000

2.2 Nonparametric method

가. 중도절단된 자료가 없는 경우

- empirical survival function: $S(t)=\frac{No.\ of\ obs.\geq t}{n}$, $t\geq 0$

나. 중도절단된 자료가 있는 경우

- $t_1, t_2, \ldots, t_k$: 사망이 일어나는 시간

- 생존함수의 추정

$$S(t) = \prod_{j:t_j<t} \frac{n_j - d_j}{n_j}$$

- 생존함수의 추정치의 variance

$$var(S(t)) = S(t)^2 \sum_{j:t_j<t} \frac{d_j}{n_j(n_j - d_j)}$$

3. Kaplan-Meier방법에의한 생존함수의 비교

가. 두집단의 생존함수가 차이가 있는 가를 검정하기 위한 가설

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

나. 검정통계량

- 우도비검정(Likelihood ratio test)
- 로그-순위 검정(log-rank test)
- 윌콕슨검정(Wilcoxon test)

Data: Lung Cancer Survival Data

t	x_1	x_2	x_3	t	x_1	x_2	x_3
Standard, Squamous				Test, Squamous			
411	70	64	5	999	90	54	12
126	60	63	9	231*	50	52	8
118	70	65	11	991	70	50	7
92	40	69	10	1	20	65	21
8	40	63	58	201	80	52	28
25*	70	48	9	44	60	70	13
11	70	48	11	15	50	40	13
Standard, Small				Test, Small			
54	80	63	4	103*	70	36	22
153	60	63	14	2	40	44	36
16	30	53	4	20	30	54	9
56	80	43	12	51	30	59	87
21	40	55	2				
287	60	66	25				
10	40	67	23				
Standard, Adeno				Test, Adeno			
8	20	61	19	18	40	69	5
12	50	63	4	90	60	50	22
				84	80	62	4
Standard, Large				Test, Large			
177	50	66	16	164	70	68	15
12	40	68	12	19	30	39	4
200	80	41	12	43	60	49	11
250	70	53	8	340	80	64	10
100	60	37	13	231	70	67	18

x_1=performance status

x_2=age

x_3=months from diagnosis to entry into study

x_4=1 if tumor type is squamous, 0 otherwise

x_5=1 if tumor type is small, 0 otherwise

x_6=1 if tumor type is adeno, 0 otherwise

x_7=1 if treatment is test, 0 if it is standard

<SAS program>

```
data lung;
      infile 'a:lung1.txt';
      input days status perform age mondiag squam small adeno treat @@
run;

proc lifetest data=lung plots=(s);
      time days*status(0);
      strata treat;
run;
```

<SAS output>

Product-Limit Survival Estimates

TREAT = 0

DAYS	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.000	1.0000	0	0	0	21
8.000	.	.	.	1	20
8.000	0.9048	0.0952	0.0641	2	19
10.000	0.8571	0.1429	0.0764	3	18
11.000	0.8095	0.1905	0.0857	4	17
12.000	.	.	.	5	16
12.000	0.7143	0.2857	0.0986	6	15
16.000	0.6667	0.3333	0.1029	7	14
21.000	0.6190	0.3810	0.1060	8	13
25.000*	.	.	.	8	12
54.000	0.5675	0.4325	0.1090	9	11
56.000	0.5159	0.4841	0.1106	10	10
92.000	0.4643	0.5357	0.1109	11	9
100.000	0.4127	0.5873	0.1099	12	8
118.000	0.3611	0.6389	0.1076	13	7
126.000	0.3095	0.6905	0.1039	14	6
153.000	0.2579	0.7421	0.0985	15	5
177.000	0.2063	0.7937	0.0913	16	4
200.000	0.1548	0.8452	0.0818	17	3
250.000	0.1032	0.8968	0.0689	18	2
287.000	0.0516	0.9484	0.0502	19	1
411.000	0	1.0000	0	20	0

* Censored Observation

Summary Statistics for Time Variable DAYS

Quantile	Point Estimate	95% Confidence Interval [Lower, Upper)	
75%	177.000	92.000	250.000
50%	92.000	16.000	153.000
25%	12.000	10.000	56.000

Mean 109.079 Standard Error 25.096

The LIFETEST Procedure

Product-Limit Survival Estimates

TREAT = 1

DAYS	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.000	1.0000	0	0	0	19
1.000	0.9474	0.0526	0.0512	1	18
2.000	0.8947	0.1053	0.0704	2	17
15.000	0.8421	0.1579	0.0837	3	16
18.000	0.7895	0.2105	0.0935	4	15
19.000	0.7368	0.2632	0.1010	5	14
20.000	0.6842	0.3158	0.1066	6	13
43.000	0.6316	0.3684	0.1107	7	12
44.000	0.5789	0.4211	0.1133	8	11
51.000	0.5263	0.4737	0.1145	9	10
84.000	0.4737	0.5263	0.1145	10	9
90.000	0.4211	0.5789	0.1133	11	8
103.000*	.	.	.	11	7
164.000	0.3609	0.6391	0.1119	12	6
201.000	0.3008	0.6992	0.1082	13	5
231.000	0.2406	0.7594	0.1019	14	4
231.000*	.	.	.	14	3
340.000	0.1604	0.8396	0.0944	15	2
991.000	0.0802	0.9198	0.0738	16	1
999.000	0	1.0000	0	17	0

* Censored Observation

Summary Statistics for Time Variable DAYS

Quantile	Point Estimate	95% Confidence Interval [Lower, Upper)	
75%	231.000	84.000	991.000
50%	84.000	20.000	231.000
25%	19.000	15.000	84.000

Mean 243.085 Standard Error 86.694

Summary of the Number of Censored and Uncensored Values

TREAT	Total	Failed	Censored	%Censored
0	21	20	1	4.7619
1	19	17	2	10.5263
Total	40	37	3	7.5000

The LIFETEST Procedure

Testing Homogeneity of Survival Curves over Strata
Time Variable DAYS

Rank Statistics

TREAT	Log-Rank	Wilcoxon
0	3.2286	41.000
1	-3.2286	-41.000

Covariance Matrix for the Log-Rank Statistics

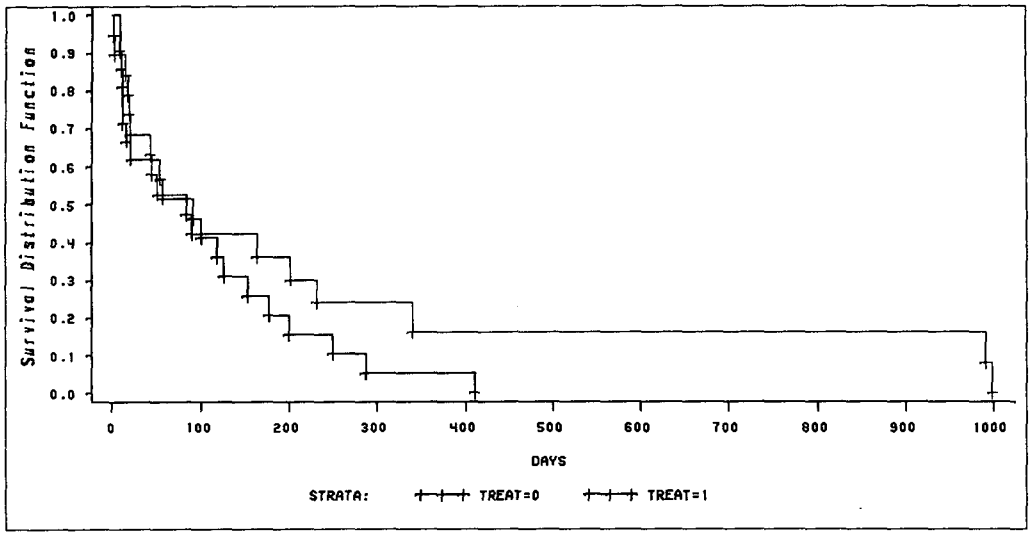
TREAT	0	1
0	8.53645	-8.53645
1	-8.53645	8.53645

Covariance Matrix for the Wilcoxon Statistics

TREAT	0	1
0	5261.19	-5261.19
1	-5261.19	5261.19

Test of Equality over Strata

Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	1.2211	1	0.2691
Wilcoxon	0.3195	1	0.5719
-2Log(LR)	4.3999	1	0.0359



4. 중도절단된 자료에서의 회귀분석적 방법

4.1 회귀분석의 적용

두 집단 혹은 그 이상의 집단에 따라서 생존확률이 차이가 있는지는 log-rank test등으로 비교할 수 있다. 경우에 따라서 연구대상이 이질적인 모집단으로 추출되어 제반 이질적인 특성을 고려하여 집단 간의 생존확률을 비교할 필요가 있다. 이 때 제반 특성변수를 통제하는 방법이 회귀분석이다.

4.2 모수적 회귀분석

가. 특징

생존함수, 위험함수에 특정한 분포를 가정한다.

나. 모수적 모형의 종류

- Exponential model
- Weibull model
- Extreme value model
- Normal model
- Log-normal model
- Gamma model
- Log-gamma model
- Log-logistic model

다. SAS에서의 절차

proc lifereg;

라. an example

<SAS program>

```
proc lifereg data=lung;
     model days*status(0)=perform age mondiag squam small adeno treat
     /d=exponential;
run;
```

<SAS output>

```
                                The SAS System        11:02 Monday, October 6, 1997  13

                               Lifereg  Procedure

Data Set          =WORK.LUNG
Dependent Variable=Log(DAYS)
Censoring Variable=STATUS
Censoring Value(s)=     0
Noncensored Values=    37  Right Censored Values=      3
Left Censored Values=   0  Interval Censored Values=   0


Log Likelihood for EXPONENT -56.92056385

                                The SAS System        11:02 Monday, October 6, 1997  14

                               Lifereg  Procedure

      Variable  DF   Estimate  Std Err ChiSquare  Pr>Chi Label/Value

      INTERCPT   1 0.83295227 1.370156  0.369573  0.5432 Intercept
      PERFORM    1 0.05400238 0.010812  24.94587  0.0001
      AGE        1 0.00903528 0.019666  0.211073  0.6459
      MONDIAG    1 0.00339933 0.011675  0.084781  0.7709
      SQUAM      1 0.36261291 0.444564  0.665301  0.4147
      SMALL      1 -0.1270625 0.486347  0.068256  0.7939
      ADENO      1 -0.8689621 0.586136  2.197886  0.1382
      TREAT      1 0.26974154 0.388209  0.482797  0.4872
      SCALE      0          1         0                    Extreme value scale parameter
            Lagrange Multiplier ChiSquare for Scale 1.255606 Pr>Chi is 0.2625.
```

<SAS program>

```
proc lifereg data=lung;
     model days*status(0)=perform age mondiag squam small adeno treat
     /d=weibull;
run;
```

<SAS output>

```
                                    The SAS System        11:02 Monday, October 6, 1997  15

                                   Lifereg  Procedure

Data Set          =WORK.LUNG
Dependent Variable=Log(DAYS)
Censoring Variable=STATUS
Censoring Value(s)=      0
Noncensored Values=     37  Right Censored Values=      3
Left Censored Values=    0  Interval Censored Values=   0


Log Likelihood for WEIBULL -56.41988629

                                    The SAS System        11:02 Monday, October 6, 1997  16

                                   Lifereg  Procedure

      Variable  DF   Estimate  Std Err ChiSquare  Pr>Chi Label/Value

      INTERCPT   1 0.82902474 1.215583  0.465121  0.4952 Intercept
      PERFORM    1 0.05379252 0.009586 31.48849  0.0001
      AGE        1 0.00972409 0.017518  0.308124 0.5788
      MONDIAG    1 0.00411114 0.010421  0.155646 0.6932
      SQUAM      1 0.39951807 0.394528  1.025458 0.3112
      SMALL      1 -0.1316886 0.425171  0.095934 0.7568
      ADENO      1 -0.8809291 0.513423  2.943954 0.0862
      TREAT      1 0.25698121 0.346468  0.550144 0.4583
      SCALE      1 0.87276657 0.115169                   Extreme value scale parameter
```

4.3 준모수적 회귀분석모형

가. 비례위험모형(porportional hazard model)

- multiplicative effect
- $h(t|x) = h_0(t)\exp(x\beta)$

  $h_0(t)$: baseline hazard function

  x: covariate

  β: regression parameter

나. 준모수적 비례위험모형(semi-parametric proportional hazard model)

$h_0(t)$: nonparametric hazard function

다. 비례위험 회귀모형에서 회귀계수의 해석

$$\frac{h(t, X_k = x+1)}{h(t, X_k = x)} = \frac{h_0(t)\exp(b_1X_1 + \cdots + b_k(x+1) + \cdots + b_pX_p)}{h_0(t)\exp(b_1X_1 + \cdots + b_k(x) + \cdots + b_pX_p)} = \exp(b_k)$$

다른 독립변수가 일정할때 X_k가 1 단위 증가하면 사망할 상대위험도는

$\exp(b_k)$ 증가한다.

라. an example

<SAS program>

```
proc phreg data=lung;
     model days*status(0)=perform age mondiag squam small adeno treat;
run;
```

<SAS output>

The PHREG Procedure

Data Set: WORK.LUNG
Dependent Variable: DAYS
Censoring Variable: STATUS
Censoring Value(s): 0
Ties Handling: BRESLOW

Summary of the Number of
Event and Censored Values

Total	Event	Censored	Percent Censored
40	37	3	7.50

Testing Global Null Hypothesis: BETA=0

Criterion	Without Covariates	With Covariates	Model Chi-Square
-2 LOG L	204.801	175.776	29.026 with 7 DF (p=0.0001)
Score	.	.	30.138 with 7 DF (p=0.0001)
Wald	.	.	25.664 with 7 DF (p=0.0006)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Risk Ratio
PERFORM	1	-0.058459	0.01368	18.25079	0.0001	0.943
AGE	1	-0.013052	0.02058	0.40221	0.5260	0.987
MONDIAG	1	0.000762	0.01182	0.00415	0.9486	1.001
SQUAM	1	-0.367046	0.48477	0.57328	0.4490	0.693
SMALL	1	-0.007721	0.50675	0.0002321	0.9878	0.992
ADENO	1	1.112940	0.63306	3.09069	0.0787	3.043
TREAT	1	-0.379709	0.40580	0.87554	0.3494	0.684