생존분석을 위한 통계패키지의 비교 연구
- SAS, SPSS, STATA -

조 미순[1], 김 순귀[2]

요 약

최근 들어 생존분석 기법이 여러 분야에서 관심을 모으고 있을 뿐 아니라 생존자료를 분석하기 위한 여러 패키지들도 개발되어 연구되고 있다. 본고에서는 생존분석의 여러 모형을 간략히 소개하고 생존자료를 분석하기 위하여 널리 사용되고 있는 패키지인 SAS, SPSS, STATA의 기능을 찾아보고 그들의 특징을 비교 조사할 것이다.

주요어 : 층화 로그 순위검정, 와이블 회귀모형, Cox 회귀모형, 가능도비 검정

1. 서론

생존자료는 정해진 한 시점에서 사건이 일어난 시점까지의 기간으로 구성되는데, 이 기간을 생존시간(survival time)이라고 한다. 생존분석이란 연구에 포함된 개체들의 생존시간을 분석하여 집단의 생존경험을 요약하는 방법이다. 생존분석이 다른 분석방법과 다른 점은 연구 도중 여러 가지 이유로 개체의 탈락이 발생하여 절단(censored)이 일어난다는 것이다.

최근 들어 생존분석 기법이 여러 분야에서 관심을 모으면서 생존자료를 분석하기 위한 패키지들이 개발되고 있다. 이 논문에서는 여러 생존 모형을 간단히 소개하였고(송 혜향, 정 갑도, 이 원철(1998), Hosmer, D. W. and Lemeshow, S.(1999) Chap T. Le(1997)), 여러 패키지들 중 널리 사용되고 있는 SAS, SPSS, STATA의 기능을 찾아보고 그 특징을 비교해 보았다(황 영신(1999).

2. 비모수적 방법

2.1 Kaplan-Meier 방법

누적한계추정법(The product limit estimator)이라고 불리는 Kaplan-Meier 추정량은 절단된 관측값을 포함하는 생존자료의 대표적인 분석방법인 생존함수에 대한 추정량이다. i번째 구간에서 생존할 확률의 추정값은

$$p_i = \frac{n_i - d_i}{n_i}$$

이다. 절단 및 사망이 발생한 시간을 순서화하여 $t_{(1)} < t_{(2)} < \cdots < t_{(n)}$ 라고 하면 시간 t에서의

1) (210-702) 강원도 지변동 123번지 강릉대학교 통계학과 석사과정
2) (210-702) 강원도 지변동 123번지 강릉대학교 정보통계학과 교수

Kaplan-Meier 추정량은

$$\hat{S}(t) = \prod_{t_{(i)} < t} \frac{n_i - d_i}{n_i}$$

와 같이 구해진다. 여기에서 d_i는 시간 $t_{(i)}$에서의 사망자수를 나타내고, n_i는 시간 $t_{(i)}$에 위험에 노출된 사람의 수를 나타낸다. Kaplan-Meier 추정량 $\hat{S}(t)$는 시간 $t_{(i)}$ 바로 직전에 생존한 사람이 시간 $t_{(i)}$를 지나는 동안 생존할 조건부 확률의 곱으로 나타난다. Kaplan-Meier 추정량의 분산에 대한 추정은 테일러 전개에 기초한 delta method라고 불리는 방법으로부터 유도된 Greenwood 공식을 이용한다.

$$\text{Greenwood 공식 : } \widehat{Var}(\hat{S}(t)) = [\hat{S}(t)]^2 \sum_{t_{(i)} \le t} \frac{d_i}{n_i(n_i - d_i)}$$

2.2 생명표

생명표(Life table)란 각 환자의 개별적인 생존시간을 고려하지 않고 전체 환자들의 생존시간의 자료를 구간으로 묶어 생존율을 계산한 경우를 말한다. 생존율 계산 원리는 Kaplan-Meier 생존율 계산과 비슷하지만 구간 내에서 절단된 개체들을 어떻게 가정하느냐에 따라 달리 계산될 수 있다.

구간 생존율을 구하는 식은

$$p_i = 1 - \frac{d_i}{n_i - w_i/2}$$

이며, 이 식에서 n_i는 위험에 노출된 개체 수, w_i는 절단된 생존시간을 갖는 개체 수이며 d_i는 구간에서 사망한 개체 수이다. $n_i - w_i/2$를 조정된 n_i(adjusted n_i)라고 부른다.

따라서 누적 생존율은

$$S(t_{(i)}) = S(t_{(i-1)}) \times p_i$$

이다.

2.3 로그 순위 검정법

로그순위 검정법은 K개의 집단의 생존경험을 비교하는데 가장 많이 사용되는 비모수적 검정법이다. 귀무가설은 K개 집단의 생존경험이 동일하다는 것이다.

$$H_0 : S_1(t) = S_2(t) = \cdots = S_K(t), \quad H_1 : \text{not } H_0$$

기대사망자수는

$$\widehat{e_{ki}} = \frac{d_i n_{ki}}{n_i}, \quad k = 1, 2, \cdots, K-1, K$$

이고, 벡터를 이용하여 관측된 사망자 수와 기대 사망자수를 표현하면

$$\underline{d_i}' = (d_{1i}, d_{2i}, \cdots, d_{K-1i}), \quad \underline{\widehat{e_i}}' = (\widehat{e_{1i}}, \widehat{e_{2i}}, \cdots, \widehat{e_{K-1i}})$$

로 나타난다.

검정통계량을 얻기 위해서 d_i의 분산-공분산 행렬의 추정을 필요로 하며, 분산-공분산 행렬

을 $\hat{V}_i$로 놓자. $\hat{V}_i$의 $K-1$개 대각요소는

$$\hat{V}_{kki} = \frac{n_{ki}(n_i - n_{ki})d_i(n_i - d_i)}{n_i^2(n_i - 1)}, \quad k = 1, 2, \cdots, K-1$$

그리고 대각선 아래 요소들은

$$\hat{V}_{kli} = -\frac{n_{ki}n_{li}d_i(n_i - d_i)}{n_i^2(n_i - 1)}, \quad k, l = 1, 2, \cdots, K-1 \quad k \neq l$$

이다. K개 집단의 생존 경험을 비교하기 위한 검정통계량은

$$\chi^2_{LR} = \left[\sum_{i=1}^{m}\left(\underset{-}{d_i} - \underset{-}{\hat{e}_i}\right)\right]\left[\sum_{i=1}^{m}\hat{V}_i\right]^{-1}\left[\sum_{i=1}^{m}\left(\underset{-}{d_i} - \underset{-}{\hat{e}_i}\right)\right]$$

로 주어지며 기대사망자수의 크기가 크다면 χ^2_{LR}은 근사적으로 자유도가 $K-1$인 카이제곱 분포를 따를 것이다. 따라서 $\chi^2(K-1)$을 이용하여 검정하고, p-value는 $\Pr(\chi^2(K-1) \geq \chi^2_{LR})$이다.

로그 순위 검정은 가중값을 1로 두어 초기사망보다는 시간이 흐른 뒤에서의 사망간의 차이에 더 비중을 두게 된다.

2.4 Wilcoxon 검정법

Wilcoxon 검정법(Wilcoxon test)은 독립된 두 집단 이상의 생존 경험을 비교하는데 초기 사망에 큰 비중을 두어 비교하고자 할 때 사용되는 비모수적 검정법이다. 귀무가설은 로그 순위 검정법과 같다. 검정통계량은

$$\chi^2_w = \frac{\left[\sum_{i=1}^{m} n_i(d_{1i} - e_{1i})\right]^2}{\sum_{i=1}^{m} n_i^2 v_{1i}}$$

이며, χ^2_w가 근사적으로 자유도 1인 카이제곱 분포를 따름을 이용하여 검정한다.
비모수적 방법을 행하기 위한 SAS, SPSS, STATA의 명령문을 살펴보면 다음과 같다.

표 1. 비모수적 방법의 명령문

	SAS	SPSS	STATA
Kaplan-Meier	proc lifetest	KM	sts list
생명표	proc lifetest method=LT	survival table	ltable
Log-rank test	proc lifetest/ strata	KM test logrank	sts test
Wilcoxon test	proc lifetest/ strata	KM test breslow	sts test, wilcoxon

3. 모수적 방법

3.1 와이블 회귀모형

i번째 개체의 사망시간을 t_i로 표시하기로 하자. 와이블 분포함수에 대한 가속화고장모형은

$$t = \exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\}\varepsilon^{\sigma}$$

와 같이 주어진다. 여기에서 σ는 척도모수이며, ε은 모수 1을 가지는 지수분포를 따른다. 동일한 분포함수에 대한 비례위험모형 또는 곱의 위험함수는 또한

$$h(t) = \lambda\gamma(\lambda t)^{\gamma-1}\exp\{\beta_1' x_1 + \cdots + \beta_p' x_p\}, \quad t \geq 0 \text{ and } \gamma, \lambda > 0.$$

로 주어진다. 여기에서 γ와 λ는 형상모수와 척도모수를 그리고 $x_1, \cdots, x_p$는 공변량을 각각 나타낸다. 만약 $Y = \ln t$이라 할 때, Y는 극단값분포(extreme-value distribution)를 따른다. 두 모형간에 다음과 같은 관계가 있음을 보일 수 있다.

$$\gamma = 1/\sigma,$$

$$\beta_i = -\beta_i' \times \sigma \text{ for } i = 1, 2, \cdots, p.$$

$$\lambda = \exp\{-\beta_0\}$$

모수적 방법의 모형을 구하기 위한 SAS, SPSS, STATA의 명령문을 살펴보면 다음과 같다.

표 2. 모수적 방법의 명령문

	SAS	STATA
지수회귀모형	proc lifereg/distribution exponential	streg, dist(exponential)
와이블회귀모형	proc lifereg	streg, dist(weibull)

4. 준모수적 방법

4.1 Cox 비례위험 모형

Cox 비례위험함수 회귀모형(proportional hazards regression model)은 생존시간에 대해 어떠한 분포도 가정하지 않고, 위험함수간의 비례관계를 이용하는 모형이다. 위험함수는 시점 t까지 생존한 개체가 그 직후 아주 짧은 시간동안 사망할 순간사망률을 의미하는 것으로 다음과 같다.

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$$

시점 t에서의 Cox 비례위험모형(proportional hazards model)은

$$h_i(t) = h_0(t)\exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})$$

와 같이 나타낸다. 여기에서 x_{ij} ($j = 1, \cdots, k$)는 i번째 개체의 j번째 공변량의 값을, $h_0(t)$는 모든 공변량의 값이 0일 때의 기저위험함수(baseline hazard function)를 나타낸다.

시점 t 에서의 두 개체의 위험함수의 비(hazard ratio)는

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t)\exp(\beta_1 x_{i1}+\cdots+\beta_k x_{ik})}{h_0(t)\exp(\beta_1 x_{j1}+\cdots+\beta_k x_{jk})}$$

$$= \exp\{\beta_1(x_{i1}-x_{j1})+\cdots+\beta_k(x_{ik}-x_{jk})\}$$

로 시간에 관계없이 일정하다. Cox 비례위험모형을 사용하려면 위험함수의 비가 시간에 관계없이 일정한 값을 나타내는 지, 즉 비례성 가정이 성립하는 지를 검토해 보아야 한다.

Cox 비례위험모형의 추정법은 부분 가능도함수(partial likelihood function)를 최대로 하는 β를 추정하는 것이고, 부분 가능도비 검정(partial likelihood ratio test), Wald 검정 그리고 Score 검정법을 이용하여 계수의 유의성을 검정한다.

4.2 동일한 생존시간을 갖는 Cox 모형

앞에서 설명한 Cox 모형은 관측된 생존시간에 동일한 값이 존재하지 않는 경우를 가정하여 설명하였다. 그러나 대부분의 생존자료는 동일한 생존시간(tied survival time)을 갖고 있으므로 동일한 생존시간이 있는 경우에 대하여 살펴볼 필요가 있다. 동일한 생존시간을 갖는 자료를 다루기 위한 방법은 여러 가지가 논의되었으나 여기에서는 Exact partial likelihood, Breslow 근사, Efron 근사에 대해서만 다루기로 하겠다.

Exact partial likelihood 방법은 생존시간을 정확하기 측정하기 어려우므로 시점 t에 d개의 동일한 자료(ties)가 있다는 가정에 기초한다. 따라서 $d!$개의 배열 중 하나에서 자료가 관측된다. Exact partial likelihood는

$$l_p(\beta) = \prod_{i=1}^{m} \frac{e^{x_{(i)}\beta}}{\sum_{j\in R(t_{(i)})} e^{x_j\beta}}$$

의 분모를 $d!$개의 배열을 포함하도록 수정하여 구한다.

Breslow 근사는 Exact partial likelihood 방법보다 쉽게 계산되며. 시점 t에 d개의 동일한 자료(ties)가 있다는 가정에 기초한다. Breslow 근사는 다음의 부분 가능도를 사용한다.

$$l_{p1}(\beta) = \prod_{i=1}^{m} \frac{e^{x_{(i)+}\beta}}{\left[\sum_{j\in R(t_{(i)})} e^{x_j\beta}\right]^{d_i}}$$

여기에서 d_i는 생존시간이 $t_{(i)}$인 개체의 수를 나타내고, $x_{(i)+}$는 d_i개체들의 공변량의 합을 나타낸다.

Efron 근사 역시 Exact partial likelihood 방법보다 쉽게 계산할 수 있으며, 시점 t에 d개의 동일한 자료가 있다는 가정에 기초한다. Efron 근사는

$$l_{p2}(\beta) = \prod_{i=1}^{m} \frac{e^{x_{(i)+}\beta}}{\prod_{k=1}^{d_i}\left[\sum_{j\in R(t_{(i)})} e^{x_j\beta} - \frac{k-1}{d_i}\sum_{j\in D(t_{(i)})} e^{x_j\beta}\right]}$$

를 이용한다. 여기에서 $D(t_{(i)})$는 생존시간이 $t_{(i)}$인 개체들을 나타낸다. Efron 근사는 Breslow 근사와 비교해 볼 때 Exact partial likelihood에 좀 더 가까운 값을 산출하는 것으로 나타난다. $d_i = 1$일 때는 동일한 생존시간이 없는 경우가 될 것이다.

5. 결론

생존자료를 분석하는데 일반적으로 널리 사용되고 있는 패키지인 SAS, SPSS, STATA를 비교해 본 결과 어떤 특정한 패키지가 특별히 좋다고 단정지을 수는 없었다. 왜냐하면 각 패키지 나름대로의 특징이 있었기 때문이다. SAS의 경우에는 비모수적 방법과 관련된 기능에서 다른 패키지에서는 산출할 수 있는 결과를 얻지 못하는 경우가 있었고, SPSS는 모수적 방법과 관련하여 얻을 수 없는 결과가 나타났다. STATA는 앞에서 살펴본 모든 경우에서 결과를 나타내었지만 완벽하다고는 할 수 없어 하나의 패키지만을 이용하여 분석을 할 경우 얻고자 하는 결과를 얻지 못하는 경우가 생길 수 있을 것이다. 따라서 생존자료를 패키지를 이용하여 분석하고자 할 때에는 하나의 패키지만을 이용하여 분석하기보다는 여러 패키지의 장점을 살려 분석하는 것이 바람직할 것이다.

참 고 문 헌

(1) 송 혜향, 정 갑도, 이 원철(1998), *생존분석*, 청문각

(2) 정 병철, 이 재원(1996), *Splus를 이용한 생존분석*, 응용통계 제11권 pp.71-89

(3) 황 영신(1999), 생존분석용 통계패키지의 비교 연구(석사학위논문)

(4) 허 명회, 박 미라(1994) 생존분석, 자유아카데미

(5) Chap T. Le(1997) *Applied Survival Analysis*, John Wiley & Sons Inc., New York

(6) Hosmer, D. W. and Lemeshow, S.(1999) *Applied Survival Analysis*, John Wiley & Sons Inc., New York

(7) Park, Y. S. & S. K. Kim(2003), Comparative Study on Imputation Procedures in Exponential Regression Model with missing values, Journal of the Korean Data & Information Science Society, V14, pp.143-152

(8) SAS/STAT User's Guide V. 8.0(1999)

(9) SPSS User's Guide, V. 10.0(1999)

(10) Stata Corporation(2003) *Getting Started with Stata for Windows*