

임의중도절단된 이표본자료에서 누적위험함수의 비에 대한 동시신뢰대

송 명 언¹⁾, 이 원 기²⁾, 박 희 주³⁾, 송 재 기⁴⁾

임의중도절단된 이표본의 수명자료에서 누적위험함수의 비는 시간의 변화에 따른 두 집단의 위험률을 서로 비교할 수 있는 수단이 된다. 본 논문에서는 먼저 두 집단의 누적위험함수의 비에 대한 추정량을 제안하고, 제안된 추정량의 점근성을 유도하였다. 또한 누적위험함수의 비에 대한 여러 가지 동시신뢰대들을 제안하였으며, 실제 자료에 적용하여 보았다.

1. 서 론

처리군(treatment group)과 대조군(control group)으로 이루어진 임의중도절단(random censoring)된 이표본(two-sample)의 수명자료에서 두 집단의 위험률(hazard rate)을 서로 비교하는 것은 신뢰성분야, 의학분야, 그리고 보건학분야에서 연구의 주요 대상이며, 이때 두 집단의 누적위험함수(cumulative hazard function)의 비는 이들 연구에서 중요한 척도가 된다.

특히 이표본 비례위험모형(two-sample proportional hazards model)이 적절한 경우 두 집단의 누적위험함수의 비는 시간에 영향을 받지 않는 상수(constant)가 되며 이를 상대위험(relative risk)이라 하고, 지금까지 많은 연구가 되어왔다. 특히 Begun과 Reid(1983)는 상대위험에 대한 비모수적 추정량을 제안하였으며, Andersen(1983)은 위험비(hazard ratio)에 대한 추정량들을 제안하고, 이 추정량들을 이용하여 상대위험에 대한 검정을 하였다. Dabrowska, Doksum과 Song(1989)은 두 집단의 누적위험함수들을 이용하여 상대위험의 증가와 감소를 측정하는 함수들을 제시하고, 이들 함수에 대한 동시신뢰대를 구하였다.

최근에 Lin, Fleming과 Wei(1994)는 일반적인 비례위험모형하에서, Song, Jeong과 Song(1997)은 가법위험모형하에서 주어진 공변량(covariate)에 대한 생존함수의 동시신뢰대들을 각각 제안하였고, Parzen, Wei와 Ying(1997)은 특별한 모형을 가정하지 않은 이표본에서 두 집단의 생존함수의 차에 대한 동시신뢰대를 구하였다. 따라서 어떤 특별한 모형을 가정하지 않은 이표본에서, 시간의 변화에 따라 두 집단의 누적위험함수의 비가 어떠한 형태로 변화하는지 조사하여 두 집단을 서로 비교하는 것 또한 의미가 있다.

본 논문에서는 먼저 두 집단의 누적위험함수의 비에 대한 추정량을 제안하고, 제안된 추정량의 점근성을 유도하고자 한다. 또한 누적위험함수의 비에 대한 여러 가지 동시신

- 1) (701-702) 대구광역시 북구 산격동 경북대학교 자연과학대학 통계학과.
- 2) (701-702) 대구광역시 북구 산격동 경북대학교 자연과학대학 통계학과.
- 3) (712-701) 경북 경산시 하양읍 부호리 경일대학교 공과대학 컴퓨터공학과 교수.
- 4) (701-702) 대구광역시 북구 산격동 경북대학교 자연과학대학 통계학과 교수.

퇴대들을 제안하고, 이들 신뢰대들을 이용하여 두 집단의 위험률을 서로 비교하고자 한다.

2. 누적위험함수의 비에 대한 추정

먼저 표본크기 n_1 과 n_2 ($n = n_1 + n_2$)를 갖는 이표본 임의중도절단된 수명자료를 고려하면, 생존시간 T_{ij} ($i=1, 2; j=1, 2, \dots, n_i$)는 연속분포함수 F_i 로부터 임의 추출된 확률표본이며, 중도절단시간 C_{ij} ($i=1, 2; j=1, 2, \dots, n_i$)는 연속분포함수 G_i 로부터 임의 추출된 확률표본이라 하고, T_{ij} 와 C_{ij} 는 서로 독립이라 가정하자. 이때 실제 우리가 관측할 수 있는 자료는 $\{X_{ij}, \delta_{ij}\}$ 이며, 여기서

$$X_{ij} = \min(T_{ij}, C_{ij}), \quad \delta_{ij} = I(T_{ij} \leq C_{ij}) \quad (i=1, 2; j=1, 2, \dots, n_i)$$

이다.

또한 셈과정(counting process) $N_i(t)$ ($i=1, 2$)는 그룹 i 에서 시간 t 까지 사망한 관찰치의 수로서

$$N_i(t) = \sum_{j=1}^{n_i} I(X_{ij} \leq t, \delta_{ij} = 1)$$

이며, 위험집합(risk set) $Y_i(t)$ ($i=1, 2$)는 그룹 i 에서 생존시간이 t 보다 크거나 같은 관찰치의 수라고 하면, 승법강도모형(multiplicative intensity model)하에서 $N_i(t)$ 는

$$M_i(t) = N_i(t) - \int_0^t Y_i(s) d\Lambda_i(s)$$

으로 유일하게 표현되어 질 수 있다. 여기서 $M_i(t)$ 는 제곱적분가능 마팅게일(square integrable martingale)이며, Λ_i 는 그룹 i 에서의 누적위험함수이다. 그리고 만약 N_1 과 N_2 가 동시에 변화하지 않는다면 M_1 과 M_2 는 서로 직교(orthogonal)한다.

이제 두 그룹간의 누적위험함수의 비 $\theta(t)$ 는 다음과 같이 정의되며,

$$\theta(t) = \Lambda_2(t) / \Lambda_1(t)$$

각 그룹에서의 누적위험함수에 대한 Nelson-Aalen 추정량(Nelson(1972), Aalen(1978))을 이용하여 추정량 $\hat{\theta}(t)$ 를

$$\hat{\theta}(t) = \hat{\Lambda}_2(t) / \hat{\Lambda}_1(t)$$

과 같이 제안 할 수 있다. 여기서 그룹 i 에서의 누적위험함수에 대한 Nelson-Aalen 추정량 $\hat{\Lambda}_i(t)$ 는

$$\hat{\Lambda}_i(t) = \int_0^t \frac{J_i(s)}{Y_i(s)} dN_i(s), \quad J_i(s) = I(Y_i(s) > 0)$$

이다. 그러면 $\hat{\theta}(t) - \theta(t)$ 는

$$\begin{aligned} \hat{\theta}(t) - \theta(t) &= \frac{1}{\widehat{\Lambda}_1(t)\Lambda_1(t)} [\Lambda_1(t)\widehat{\Lambda}_2(t) - \widehat{\Lambda}_1(t)\Lambda_2(t)] \\ &= \frac{1}{\widehat{\Lambda}_1(t)} [\widehat{\Lambda}_2(t) - \Lambda_2(t)] - \frac{\Lambda_2(t)}{\widehat{\Lambda}_1(t)\Lambda_1(t)} [\widehat{\Lambda}_1(t) - \Lambda_1(t)] \\ &= \frac{1}{\widehat{\Lambda}_1(t)} \int_0^t \frac{J_2(s)}{Y_2(s)} dM_2(s) - \frac{\theta(t)}{\widehat{\Lambda}_1(t)} \int_0^t \frac{J_1(s)}{Y_1(s)} dM_1(s) \end{aligned}$$

과 같이 표현되어 질 수 있으므로, 정칙조건(regularity condition)하에서 $n \rightarrow \infty$ 일 때 Lenglart's 부등식에 의해

$$\sup_{0 \leq t < \tau} |\hat{\theta}(t) - \theta(t)| \xrightarrow{p} 0$$

이 만족한다. 여기서 $\tau = \inf\{t \geq 0; H_1(t) = 1 \text{ or } H_2(t) = 1\}$ 이며 H_i 는 관찰치 X_{ij} 의 분포 함수이다.

그리고 $Z_n(t) = \sqrt{n}(\hat{\theta}(t) - \theta(t))$ 라 두고, 만약 $n \rightarrow \infty$ 일 때

$$\sup_{0 \leq t < \tau} |Y_i(t)/n - y_i(t)| \xrightarrow{p} 0$$

을 만족하는 함수 $y_i(\cdot) \in (0, 1)$ ($i=1, 2$)들이 존재한다면, 마팅계일 중심극한정리에 의해 과정 $Z_n(\cdot)$ 는 점근적으로 평균 0, 공분산함수가

$$Cov(U(u), U(t)) = \frac{1}{\Lambda_1(u)\Lambda_1(t)} \int_0^{u \wedge t} \frac{1}{y_2(s)} d\Lambda_2(s) + \frac{\theta(u)\theta(t)}{\Lambda_1(u)\Lambda_1(t)} \int_0^{u \wedge t} \frac{1}{y_1(s)} d\Lambda_1(s)$$

인 가우스 과정(Gaussian process) $U(\cdot)$ 으로 분포 수렴한다.

이제 위의 $Z_n(\cdot)$ 의 점근분포를 이용하여 누적위험함수의 비 $\theta(t)$ 에 대한 동시신뢰대를 구하고자 한다. 그러나 $Z_n(\cdot)$ 의 점근분포는 공분산함수의 형태로부터 알 수 있듯이 독립증분구조(independent increment structure)를 갖고 있지 않기 때문에, 브라운 다리과정(Brownian Bridge process)으로 변환되어 질 수 없으므로, 두 그룹간의 누적위험함수의 비 $\theta(t)$ 에 대한 동시신뢰대를 직접적으로 구할 수 없다. 그래서 과정 $Z_n(\cdot)$ 의 점근분포를 간단한 시뮬레이션 과정(simulated process) (Lin 등(1994), Song 등(1997), Parzen 등(1997) 등 참조)을 이용하여 구하고자 한다.

먼저 임의의 i 와 t 에 대해

$$E(M_i(t)) = 0, \quad Var(M_i(t)) = E(N_i(t))$$

(Fleming과 Harrington(1991))이므로, $Z_n(t)$ 에서 $M_i(t)$ ($i=1, 2$)를 평균과 분산이 같은

$$\sum_{j=1}^{n_i} I(X_{ij} \leq t, \delta_{ij} = 1) G_{ij} \text{으로 대치하고, } \Lambda_i \text{를 강일치 추정량인 } \widehat{\Lambda}_i \text{으로 대치하면. 다음과}$$

같은 시뮬레이션 과정 $\hat{Z}_n(\cdot)$ 를 얻을 수 있으며, 여기서 $\{G_{ij}; i=1, 2, j=1, \dots, n_i\}$ 는 표준정규모집단으로부터의 임의표본이다.

$$\hat{Z}_n(t) = \frac{\sqrt{n}}{\widehat{\Lambda}_1(t)} \sum_{j=1}^{n_2} \frac{J_2(X_{2j})}{Y_2(X_{2j})} I(X_{2j} \leq t) \delta_{2j} G_{2j} - \frac{\sqrt{n}\hat{\theta}(t)}{\widehat{\Lambda}_1(t)} \sum_{j=1}^{n_1} \frac{J_1(X_{1j})}{Y_1(X_{1j})} I(X_{1j} \leq t) \delta_{1j} G_{1j}$$

이때 관찰치 $\{X_{ij}, \delta_{ij}\}$ 가 주어진 경우, $\hat{Z}_n(\cdot)$ 에서 확률변수는 $\{G_{ij}\}$ 뿐이므로, $\hat{Z}_n(\cdot)$ 은 서로 독립인 표준정규확률변수들의 합이 된다. 그러므로 시물레이션 과정 $\hat{Z}_n(\cdot)$ 은 엄격(tight)하게 되며, Lindeberg-Feller의 중심극한정리에 의해 관찰치 $\{X_{ij}, \delta_{ij}\}$ 가 주어진 경우 $\hat{Z}_n(\cdot)$ 의 조건부 분포는 $Z_n(\cdot)$ 과 같은 점근분포를 가지게 된다. 또한 조건부 공분산함수는

$$\begin{aligned} & Cov(\hat{Z}_n(u), \hat{Z}_n(t) | \{X_{ij}, \delta_{ij}\} \ (i=1,2 ; j=1,2, \dots, n_i)) \\ &= \frac{n}{\hat{\Lambda}_1(u)\hat{\Lambda}_1(t)} \int_0^{u \wedge t} \frac{J_2(s)}{Y_2^2(s)} dN_2(s) + \frac{n\hat{\theta}(u)\hat{\theta}(t)}{\hat{\Lambda}_1(u)\hat{\Lambda}_1(t)} \int_0^{u \wedge t} \frac{J_1(s)}{Y_1^2(s)} dN_1(s) \end{aligned}$$

이 되며, 대수의 법칙(law of large number)에 의해 $Cov(U(u), U(t))$ 으로 확률 수렴하고, 조건부 분산은

$$\hat{\sigma}^2(t) = \frac{n}{\hat{\Lambda}_1^2(t)} \int_0^t \frac{J_2(s)}{Y_2^2(s)} dN_2(s) + \frac{n\hat{\theta}^2(t)}{\hat{\Lambda}_1^2(t)} \int_0^t \frac{J_1(s)}{Y_1^2(s)} dN_1(s)$$

과 같다. 이제 위의 사실들을 이용하여, 관찰치 $\{X_{ij}, \delta_{ij}\}$ 가 주어진 경우 표준정규모집단의 임의표본 $\{G_{ij}\}$ 를 반복해서 생성함으로써, 시물레이션 과정 $Z_n(\cdot)$ 의 점근분포를 구할 수 있다.

3. 동시신뢰대의 구성

임의의 시간구간 $[t_1, t_2]$ ($0 \leq t_1 \leq t_2 < \tau$)에서 두 그룹간의 누적위험함수의 비 $\theta(t)$ 에 대한 $100(1-\alpha)\%$ 동시신뢰대를 구하기 위해서 아래와 같은 과정 $B(\cdot)$ 를 고려하자.

$$B(t) = \sqrt{n}g(t) [\phi(\hat{\theta}(t)) - \phi(\theta(t))]$$

여기서 $\phi(\cdot)$ 은 미분 ϕ' 이 연속이며 시간구간 $[t_1, t_2]$ 에서 0이 아닌 값을 가지는 알려진 변환함수이며, $g(\cdot)$ 는 시간구간 $[t_1, t_2]$ 에서 비음(non-negative)의 유계함수(bounded function)로 일양적으로 확률 수렴하는 가중함수(weight function)이다. 그러면 델타방법(delta-method)에 의해, $B(\cdot)$ 는 $\tilde{B}(t) = g(t)\phi'(\theta(t))Z_n(t)$ 와 같은 점근분포를 가지게 된다. 또한 $\tilde{B}(\cdot)$ 의 점근분포는 2절의 사실들에 의해, $\hat{B}(t) = g(t)\phi'(\hat{\theta}(t))\hat{Z}_n(t)$ 의 분포로 근사되어 질 수 있다.

이제 누적위험함수의 비 $\theta(t)$ 의 동시신뢰대를 구하기 위해서는 $\sup_{t_1 \leq t \leq t_2} \{B(t)\}$ 의 점근분포에서의 $100(1-\alpha)$ 백분위수 값인 임계치(critical point) q_α 값이 필요하며, 위의 사실들에 의해 q_α 는 $\Pr\{\max_{t_1 \leq X_{ij} \leq t_2} |\hat{B}(X_{ij})| > q_\alpha\} = \alpha$ 을 만족하는 값으로 근사되어 질 수 있으므로, $\{X_{ij}, \delta_{ij}\}$ 가 주어진 경우 표준정규모집단의 임의표본 $\{G_{ij}\}$ 를 반복해서 생성함으로써 q_α 값을 근사적으로 구할 수 있다. 따라서 시간구간 $[t_1, t_2]$ 에서 $\phi(\theta(t))$ 에 대한

100(1- α)% 동시신뢰대는

$$\phi(\hat{\theta}(t)) \pm \frac{1}{\sqrt{n}} q_\alpha / g(t)$$

이 되며, 변환함수 ϕ 와 가중함수 g 에 따라 여러 가지 종류의 $\theta(t)$ 에 대한 100(1- α)% 동시신뢰대를 구할 수 있다.

먼저 ϕ 와 g 가

$$\phi(x) = x, \quad g(t) = \frac{1}{\hat{\alpha}(t)}$$

이면, 두 집단의 누적위험함수의 비 $\theta(t)$ 에 대한 100(1- α)% 동시신뢰대는

$$\hat{\theta}(t) \pm \frac{q_\alpha}{\sqrt{n}} \hat{\alpha}(t)$$

이 되며, 이는 Nair(1984)의 등정도(equal precision) 형태의 신뢰대이다. 한편 $\theta(t)$ 는 항상 비음의 값을 가지지만 위의 신뢰대는 그 구조상 음의 값을 가질 수도 있다. 그래서 Bie, Borgan과 Liestøl(1987)이 누적위험함수의 신뢰구간과 동시신뢰대를 구할 때 사용한 log변환과 arcsin변환을 적용하고자 하며, 각 변환에 대한 ϕ 와 g 는

$$\phi(x) = \log x, \quad g(t) = \frac{\hat{\theta}(t)}{\hat{\alpha}(t)}$$

$$\phi(x) = \arcsin(\exp(-x/2)), \quad g(t) = \frac{2\sqrt{\exp(\hat{\theta}(t))-1}}{\hat{\sigma}(t)}$$

이다. 그러면 log변환된 $\theta(t)$ 에 대한 100(1- α)% 동시신뢰대는

$$\hat{\theta}(t) \cdot \exp\left[\pm \frac{q_\alpha}{\sqrt{n}} \cdot \frac{\hat{\sigma}(t)}{\hat{\theta}(t)}\right]$$

이 되며, arcsin변환된 신뢰대는

$$-2 \log\{\sin(\min[\pi/2, a+b])\} \leq \theta(t) \leq -2 \log\{\sin(\max[0, a-b])\}$$

이 된다. 여기서 a 와 b 는

$$a = \arcsin[\exp\{-\hat{\theta}(t)/2\}], \quad b = \frac{q_\alpha}{2\sqrt{n}} \hat{\sigma}(t) [\exp\{\hat{\theta}(t)\}-1]^{-1/2}$$

이다.

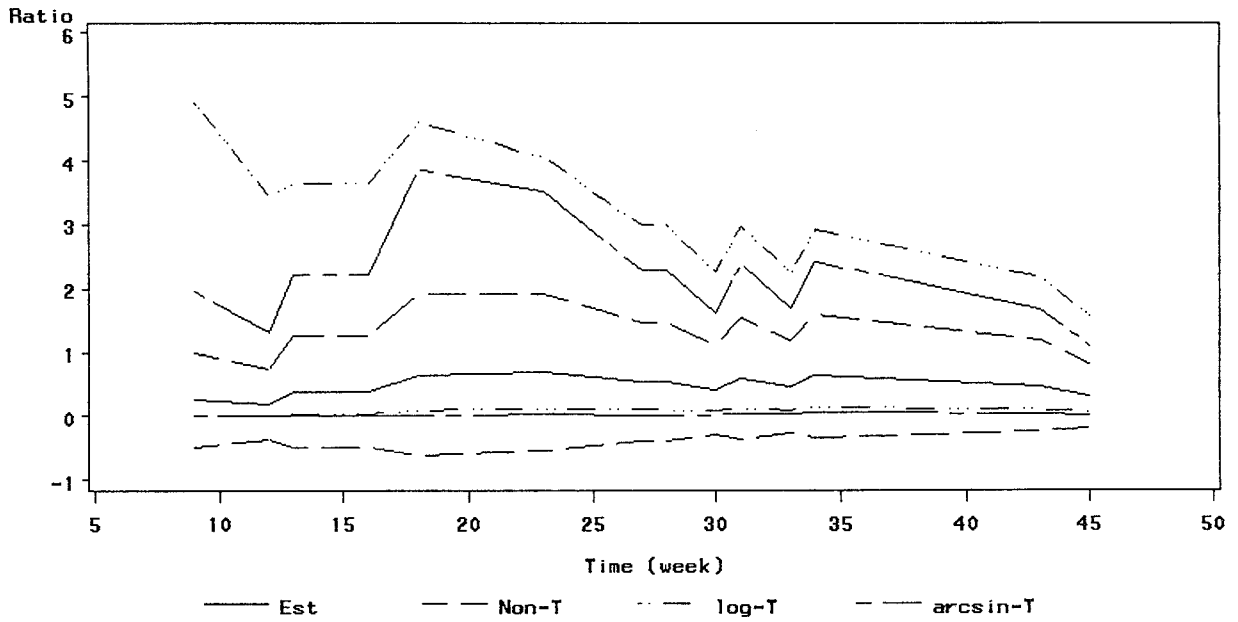
그래서 이제 두 그룹간의 누적위험함수 비 $\theta(t)$ 에 대한 100(1- α)% 동시신뢰대들이 시간의 흐름에 따라 어떻게 변화하는지 조사함으로써 두 집단을 서로 비교할 수 있다. 또한 만약 이표본 비례위험모형이 타당한 경우 $\theta(t)$ 는 상수인 상대위험이 되며, 전구간에서 $\theta(t)$ 에 대한 100(1- α)% 동시신뢰대들에 포함되게 될 것이다. 따라서 이를 이용하면 신뢰대들의 구성은 이표본 비례위험모형을 검정하는 데에도 적용되어 질 수 있다.

4. 예 제

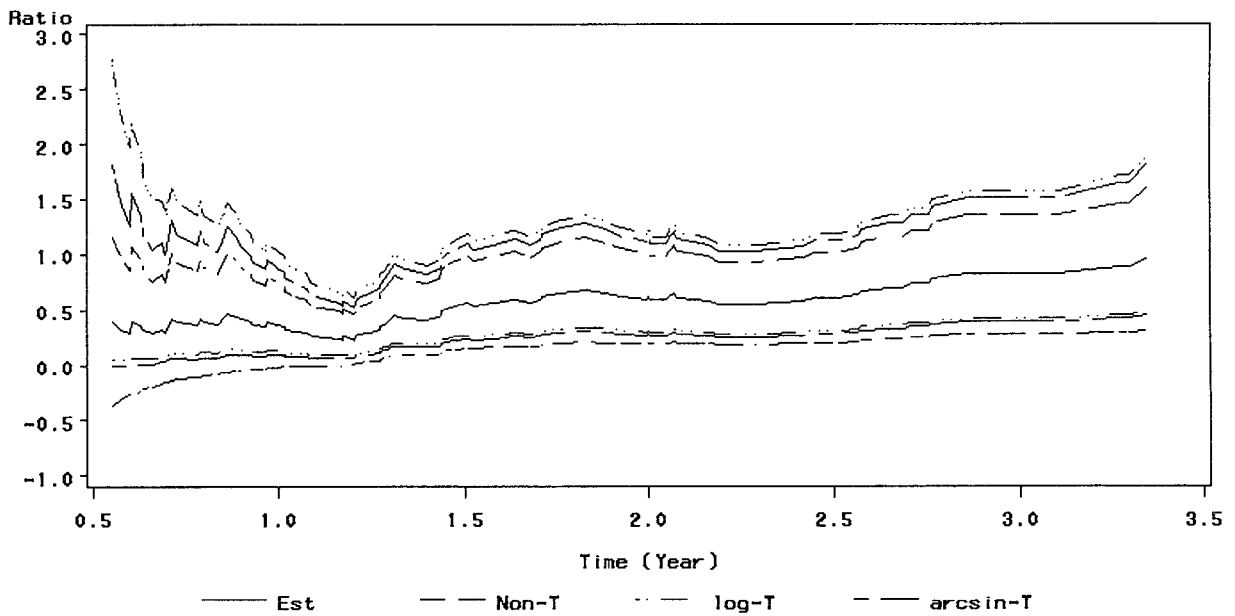
이제 두 가지 실제 자료들을 이용하여 두 집단의 누적위험함수의 비 $\theta(t)$ 에 대한 제안된 95% 동시신뢰대들을 구하여 두 집단을 서로 비교하고자 한다. 각 경우에서 q_a 값은 표준정규모집단의 임의표본 $\{G_{ij}; i=1,2, j=1, \dots, n_i\}$ 를 생성하여 얻은 $\hat{B}(\cdot)$ 을 10000번 반복하여 구하였다.

첫 번째 예제는 급성 골수성 백혈병(acute myelogenous leukemia, AML)환자에 관한 자료로 화학요법의 사용에 대한 효율성을 연구하기 위한 것이다(Embury 등(1977)). 이 연구에서 총 23명의 환자 중 대조집단은 12명으로 화학요법을 사용하지 않고, 치료집단은 11명으로 화학요법을 사용하여 생존시간을 주 단위로 조사하였으며, 비례위험모형이 타당하다고 알려져 있다. 이 자료를 이용하여 대조집단과 치료집단의 누적위험함수의 비 $\theta(t)$ 에 대한 추정치 $\hat{\theta}(t)$ 와 95% 동시신뢰대들을 구한 결과를 그림 1에 나타내었다. 그 결과 추정치는 0.18에서 0.70사이에 있으므로 화학요법의 사용이 더 효율적인 것 같으나, 동시신뢰대들의 상한들이 1보다 더 크게 나타나서 확실히 유의하다고 할 수 없다. 그러나 이 결과는 자료의 수가 너무 적기 때문인 것으로 생각된다. 한편 상대위험의 강일치 추정치는 0.31로 전구간에 모든 동시신뢰대들에 포함되므로, 이 사실에 의해서도 비례위험모형이 적절하다는 것을 알 수 있다.

두 번째 예제는 Ying, Jung과 Wei(1994)에서 분석된 자료로서 폐암환자들의 생존시간을 측정된 것이다. 이들 환자들에 대한 대표적인 치료방법은 etoposide와 cisplatin의 조합으로 투여하는 것으로, 최적의 조합순서는 아직 알려져 있지 않다. 그래서 이 연구에서 총 121명의 폐암환자를 대상으로, 임의로 59명을 추출하여 etoposide을 투여한 후 cisplatin을 투여하고(A방법), 나머지 62명은 반대의 순서로 치료하였다. 특히 이 자료는 비례위험모형이 적합하지 않은 것으로 알려져 있다. 그림 2는 이 자료를 이용한 결과이다. 이를 살펴보면 누적위험함수의 비에 대한 추정치 $\hat{\theta}(t)$ 는 0.25에서 0.95사이에 있으므로 B치료방법이 A치료방법보다 더 효과적인 것처럼 보이며, 또한 신뢰대들을 보면 1년에서 1.5년 사이에서는 유의하게 B치료방법이 A치료방법보다 효과적임을 알 수 있으며, 2년과 2.5년 사이에서도 어느 정도 유의함을 알 수 있다. 그리고 모든 동시신뢰대들은 상대위험의 강일치추정치인 0.95를 포함하지 않는 구간이 있으므로, 이 사실에 의해서도 비례위험모형이 적절하지 않음을 확인할 수 있다.



<그림 1> AML환자들의 누적위험함수의 비에 대한 추정치 $\hat{\theta}(t)$ 와 95% 동시신뢰대



<그림 2> 폐암환자들의 누적위험함수의 비에 대한 추정치 $\hat{\theta}(t)$ 와 95% 동시신뢰대

참 고 문 헌

- [1] Aalen, O.O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics*, Vol. 6, 701-726.
- [2] Andersen, P.K. (1983). Comparing survival distributions via hazard ratio estimates. *Scandinavian Journal of Statistics*, Vol. 10, 77-85.
- [3] Andersen, P.K., Borgan, Ø., Gill, R.D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.
- [4] Begun, J.M. and Reid, N. (1983). Estimating the relative risk with censored data. *Journal of the American Statistical Association*, Vol. 78, 337-341.
- [5] Bie, O., Borgan, Ø., and Liestøl, K. (1987). Confidence intervals and confidence bands for the cumulative hazard rate function and their small sample properties. *Scandinavian Journal of Statistics*, Vol. 14, 221-233.
- [6] Embury, S.H., Elias, L., Heller, P.H., Hood, C.E., Greenberg, P.L., and Schrier, S.L. (1977). Remission maintenance therapy in acute myelogenous leukemia. *Western Journal of Medicine*, Vol. 126, 267-272.
- [7] Dabrowska, D.M., Doksum, K.A., and Song, J.K. (1989). Graphical comparison of cumulative hazards for two populations. *Biometrika*, Vol. 76, 763-773.
- [8] Fleming, T.R. and Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. John Willey & Sons, New York.
- [9] Nair, V.N. (1984). Confidence bands for survival functions with censored data : A comparative study. *Technometrics*, Vol. 26, 265-275.
- [10] Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, Vol. 14, 945-965.
- [11] Lin, D.Y., Fleming, T.R. and Wei, L.J. (1994). Confidence bands for survival curves under the proportional hazards model. *Biometrika*, Vol. 81, 73-81.
- [12] Parzen, M.I., Wei, L.J., and Ying, Z. (1997). Simultaneous confidence intervals for the difference of two survival functions. *Scandinavian Journal of Statistics*, Vol. 24, 309-314.
- [13] Song, M.U., Jeong, D.M., and Song, J.K. (1997). Confidence Bands for survival curve under the additive risk model. *Journal of the Korean Statistical Society*, Vol. 26, 429-443.
- [14] Ying, Z., Jung, S.H., and Wei, L.J. (1994). Survival analysis with median regression model. *Journal of the American Statistical Association*, Vol. 90, 178-184.