응용통계연구 제 18권 1호, 2005년, pp. 67-81

동측치가 많은 FRAILTY 모형의 분석 *

김용대[1] 박진경[2]

요 약

프레일티모형에 대한 기존의 추론방법은 동측치가 많은 경우에 그 성능이 떨어진다. 그 이유는 사용된 경험적 우도함수가 동측치가 많은 자료에는 적합하지 않기 때문이다. 본 논문에서는 동측치가 많은 프레일티모형에서의 새로운 추론방법을 제안한다. 이항형태의 경험적우도함수를 바탕으로 베이지안 부스트랩을 사용하여 모수의 사후분포를 구한다. 제안된 방법의 장점은 기존에 제안된 주변최대우도추정량에 비하여 계산이 수월하고 안정적인 결과를 제공하는데 있다. 이를 실증적으로 비교하기 위하여 제안된 방법을 주변최대우도추정량과 가상실험을 통하여 비교한다.

주요용어: 프레일티모형, 포아송형태의 경험적우도함수, 이항형태의 경험적우도함수, 베이지안 부스트랩

1. 서론

프레일티(frailty) 모형이란 랜덤항이 들어간 Cox(1972)의 비례위험모형으로, 프레일티는 같은 그룹 내의 개체들이 공유하는 랜덤항을 의미한다. 프레일티 모형은 Vaupel 등(1979)에 의해 처음 소개되었고, 주로 cluster design이나 paired match design에서 주로 적용된다. 각 그룹 내에서는 개체간에 종속성이 존재하는 반면, 그룹간의 변이성은 서로 다르다. 또한, 프레일티의 값이 주어졌을 때, 관측된 생존자료들은 서로 독립이라고 가정한다. 생존자료에서 프레일티가 존재할 경우, 이에 대한 중요성이나 프레일티를 무시할 경우 발생할 문제에 대해 논의된 바 있다(Heckman 과 Singer (1984), Struchers 와 Kalfleisch (1986), Schumacher 등 (1987)).

프레일티모형에 대한 통계적 추론방법은 많은 연구자들에 의하여 연구가 되어졌는데, Clayton(1991)이 베이지안 방법을, McGilchrist 와 Aisbett(1991)이 벌점우도함수 접근방법(penalized likelihood approach)을, Klein(1992) 과 Nielsen 등(1992)이 EM알고리즘을 이용한 주변우도함수 접근방법(marginal likelhiood approach)을, 그리고 Ha 등 (2001)이 계층적우도함수 접근방법(hierachical likelihood approach)을 개발하였다. 이러한 방법들의 문제점으로는, 동측치가 많은 자료에서는 성능이 매우 떨어진다는 것이다. 그 이유는, 위의

* 이 논문은 2002년도 한국학술진흥재단의 지원에 의하여 연구되었음 (KRF-2002-003-C00029)

1) (151-742) 서울시 관악구 신림동 산56-1 서울대학교 통계학과, 조교수
E-mail: ydkim@stats.snu.ac.kr

2) (151-181) 서울시 관악구 봉천7동 산4-8 서울대 연구공원 내, 국제백신연구소(IVI),
Associate Scientific Analyst,
E-mail: jkpark@ivi.int

열거한 방법들이 기저위험함수(baseline hazard function)을 이산화한 포아송형태의 경험적 우도함수 (empirical likelihood)를 사용하기 때문이다. 프레일티가 없는 모형에서 포아송형태의 경험적 우도함수는 Breslow(1974)의 부분우도함수와 같게 되며, 이 우도함수가 동측치가 많은 경우에 성능이 떨어진다는 것은 잘 알려져 있다.

본 논문에서는 동측치를 많이 포함하는 프레일티 모형에서의 새로운 추론방법을 개발한다. Hedecker 등 (2000)은 그룹화된 생존자료에서 Prentice 와 Gloeckler (1978)가 제안한 이항형태의 경험적우도함수를 기반으로 하여 프레일티모형의 추론방법을 제안하였다. 이 방법에서는 수치적으로 랜덤항을 적분하여 주변우도함수를 구하고 이를 최대로 하는 최대주변우도추정량을 사용하였다. 하지만, 이 방법은 계산상의 어려움으로 인하여 그룹의 수가 많은 경우에는 사용하기가 곤란하다.

이러한 계산상의 문제를 해결하기 위하여, 본 논문에서 우리는 베이지안 부스트랩방법을 제안한다. 베이지안 부스트랩은 Rubin (1981)에 의하여 처음 제안되었으며, 그 후 Lo (1993)에 의하여 중도절단자료로 확장되었고, 최근에 Kim 과 Lee (2003)에 의하여 비례위험모형에 적용되었다. 특히, Kim 과 Lee (2003)는 가상실험을 통하여 이항형태의 우도함수를 기반으로 하는 베이지안 부스트랩의 성능이 동측치가 많은 경우에 잘 작동함을 보였다. 본 논문에서는 Kim 과 Lee (2003)가 제안한 베이지안 부스트랩방법을 프레일티 모형으로 확장한다.

본 논문은 다음과 같이 구성된다. 2장에서는 프레일티모형을 소개하고 포아송형태의 우도함수를 기반으로하는 추론방법에 대한 간단한 개요와 왜 이 방법이 동측치가 많은 자료에서는 잘 작동하지 않는 이유에 대하여 설명한다. 3장에서는 본 논문에서 제안하는 베이지안 부스트랩방법을 소개하고, 사후분포를 구할 수 있는 MCMC 알고리즘을 제안한다. 4장에서는 제안된 방법과 Hedecker 등 (2000)이 제안한 최대주변우도추정량과 가상실험을 통하여 비교하고, 5장은 예제를 통하여 자료의 형태와 방법에 따라 결과가 어떻게 달라질 수 있는지를 살펴본다. 마지막 장은 토의 및 결론이다.

2. Frailty 모형

$(T_{ij}, \delta_{ij}, z_{ij}), (i = 1, \ldots, n,\ j = 1, \ldots, n_i)$를 i번째 개체 j번째 자료라 하자. 여기서, $T_{ij} = \min\{X_{ij}, C_{ij}\}$, $\delta_{ij} = I(X_{ij} \leq C_{ij})$이고 X_{ij} 와 C_{ij}는 각각 생존시간과 중도절단시간이다. 그리고, z_{ij}는 위험요인을 나타내는 변량이다. 프레일티모형에서는, 프레일티항 $\epsilon_1, \ldots, \epsilon_n$이 주어졌을 때, X_{ij}는 서로 독립이며, 그 위험함수는 다음의 모형을 따른다.

$$a(t|Z_{ij}, \epsilon_i) = \exp(\beta z_{ij} + \epsilon_i) a_0(t)$$

여기서, $a_0(t)$를 기저위험함수이다. 프레일티항 ϵ_i들은 독립이고 평균이 0, 분산이 σ^2인 정규분포를 따른다고 가정한다.

프레일티항 $\underline{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$들이 주어진 상태에서 모수 β 와 a_0의 우도함수는

$$\begin{aligned} \mathrm{L}(a_0, \beta|\underline{\epsilon}) &= \prod_{i=1}^{n}\prod_{j=1}^{n_i}\left[\exp(\beta^T Z_{ij} + \epsilon_i)a_0(T_{ij})\right]^{I(\delta_{ij}=1)} \\ &\quad \times \exp\left(-\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^{T_{ij}} \exp(\beta^T Z_{ij} + \epsilon_i)a_0(s)ds\right) \end{aligned} \tag{2.1}$$

이다.

관측시간 T_{ij} 중 중도절단이 되지 않은 서로 다른 생존시간의 개수를 q_n, 그리고 $t_1 < \cdots < t_{q_n}$을 서로 다른 생존시간이라 하자. 누적기저위험함수 A_0를 $A_0(t) = \int_0^t a_0(s)ds$로 정의한 후, A_0가 $t_1, \ldots, t_{q_n}$에서만 증가하는 이산형 함수로 가정하고 $a_0(t)$를 $\Delta A_0(t)$로 근사시키면, 우도함수 (2.1)은

$$\begin{aligned} \mathrm{L}^P(A_0, \beta|\underline{\epsilon}) &= \prod_{i=1}^{n}\prod_{j=1}^{n_i}\left[\exp(\beta^T Z_{ij} + \epsilon_i)\Delta A_0(T_{ij})\right]^{I(\delta_{ij}=1)} \\ &\quad \times \exp\left(-\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_0^{T_{ij}} \exp(\beta^T Z_{ij} + \epsilon_i)dA_0(s)\right) \end{aligned} \tag{2.2}$$

이 되고, 이 우도함수를 포아송형태의 경험적우도함수라고 한다. 이 우도함수에 대한 자세한 내용은 Kim 과 Lee (2003)와 그 안에 있는 참고문헌들을 참조바란다.

프레일티모형의 추론에 대한 기존의 대부분의 연구가 포아송형태의 경험적우도함수를 기반으로 하여 제안되었다. 하지만, 포아송형태의 경험적우도함수는 동측치가 많은 경우에 잘 작동하지 않는데, 그 이유는, 누적기저위험함수가 이산형인 경우에 (2.2)가 실제 우도함수가 되지 않는다는 것이다. 동측치가 많은 경우에는 실제 분포가 이산형일 경우가 많고, 따라서, 우도함수가 되지 않은 함수를 이용한 추론은 나쁜 결과를 제공할 수 있다. 실제로, 프레일티항이 없는 경우에 (2.2)는 Breslow (1974)의 부분우도함수와 같아지며, 동측치가 많은 경우에 잘 작동하지 않는다는 것이 잘 알려져 있다.

3. Frailty 모형에 대한 베이지안 부스트랩

베이지안 부스트랩방법은 Rubin (1981)에 의해서 처음 제안되었고, Lo (1993)에 의해서 중도 절단 자료(right censored data)에 대해서 확장되었다. 베이지안 부스트랩은 다음과 같이 세가지 관점에서 설명된다. 첫째로는, Efron (1979)의 부스트랩방법을 확장한 개념이고, 둘째로는, 사전분포의 정보가 사라지면서 베이지안 부스트랩의 사후분포가 완전한 베이지안방법의 사후분포의 극한으로 설명되는 관점이다(Kim 과 Lee, 2003). 마지막으로는, 베이지안 부스트랩의 사후분포가 사전분포와 경험적우도함수(empirical likelihood)의 곱으로 설명된다는 관점으로, 본 논문에서는 세번째 관점을 이용한다.

Kim 과 Lee (2003)는 비례위험모형에 대해서 베이지안 부스트랩방법을 적용하였다. Kim 과 Lee (2003)는 Cox (1972)의 부분우도함수 형태인 포아송형태의 경험적우도함수와,

Prentice 와 Gloeckler (1978)가 제안한 이항형태의 경험적우도함수 두가지 형태에 대해서 베이지안 부스트랩방법을 적용하였다. Kim 과 Lee (2003)는 비교 연구를 통해서 이 둘의 수행 효과에 대해서 논의 한 바 있는데, 이들은 그룹화 된 자료에서는 이항형태의 경험적 우도함수를 적용한 베이지안 부스트랩이 다른 방법보다 좋다고 제안하였다.

이 장에서는 Kim 과 Lee (2003)가 제안한 방법 중, 이항형태의 경험적우도함수를 이용한 방법을 확장하여, 프레일티모형을 적합하고자 한다.

3.1. 이항 형태의 경험적우도함수

이항형태의 경험적우도함수는, 포아송형태의 경험적우도함수에 사용했던 방법인 누적위험함수의 이산화과정 대신에, 생존함수를 이산화 하여 얻어진다. 즉, 식 (2.1)의 우도함수를 생존함수를 사용하여 다시 쓰면

$$\mathrm{L}(a_0, \beta|\underline{\epsilon}) = \prod_{i=1}^{n}\prod_{j=1}^{n_i}\left[f(T_{ij}|Z_{ij},\epsilon_i)^{I(\delta_{ij}=1)}S(T_{ij}|Z_{ij},\epsilon_i)^{I(\delta_{ij}=0)}\right] \tag{3.1}$$

이 된다. 여기서, f는 S의 확률밀도함수이다. 생존함수 $S(t)$가 $t_1,\ldots,t_{q_n}$에서 증가하는 이산형이라 가정하고, $f(t)$를 $-\Delta S(t)$로 근사한 후, $S(t)$를 누적위험함수를 이용하여 $\prod_{s\le t}(1-\Delta A_0(s))$로 치환하면, 다음의 이항형태의 경험적우도함수를 얻는다.

$$\begin{aligned} L^B(A,\beta|\underline{\epsilon}) = \; & \prod_{t\in T_n}\left[\prod_{(i,j)\in D(t)} 1-\{1-\Delta A_0(t)\}^{\exp(\beta z_{ij}+\epsilon_i)}\right] \\ & \times\{1-\Delta A_0(t)\}^{\sum_{(i,j)\in R(t)-D(t)}\exp(\beta z_{ij}+\epsilon_i)}. \end{aligned} \tag{3.2}$$

여기서, $T_n=\{t_1,\ldots,t_{q_n}\}$, $R(t)=\{(i,j): T_{ij}\ge t\}$, $D(t)=\{(i,j): T_{ij}=t, \delta_{ij}=1\}$ 이다.

3.2. 사전분포

이항형태의 경험적우도함수를 이용한 베이지안 부스트랩을 위해서는 모수에 대한 사전분포가 필요하다. 모수로는 회귀계수 β, 누적기저위험함수 $\Delta A_0(t_1),\ldots,\Delta A_0(t_{q_n})$, 그리고 프레일티항의 분산인 σ^2가 있다. 회귀계수 β에는 무정보사전분포인 균등분포를 사용하고, σ^2에 대해서는, 다음과 같이 모수가 α와 κ 인 역-감마분포(inverse-Gamma distribution)를 이용하였다.

$$\begin{aligned} \pi(\sigma^{-2}) &\sim \mathrm{G}(\alpha,\kappa) \\ &\propto (\sigma^{-2})^{\alpha-1}\exp(-\sigma^{-2}/\kappa). \end{aligned} \tag{3.3}$$

$\Delta A_0(t), t\in T_n$의 사전분포로는 서로 독립이고,

$$\pi\{\Delta A_0(t)\} = \Delta A_0(t)^{-1}\{1-\Delta A_0(t)\}^{-1+\eta} \tag{3.4}$$

라 정하였다. Kim 과 Lee (2003)에 의하여 프레일티항이 없는 경우에, 식 (3.4)사전분포에서 $\eta=0$이면, 베이지안 부스트랩의 사후분포가 완전베이지안 방법의 사후분포의 극한으

로 정의됨이 보여졌다. 프레일티항이 있는 경우에도 같은 결과를 얻을 수 있다. 하지만, η가 0 인 경우, 사후분포가 proper하지 않는 경우가 많이 발생하고, 이러한 문제를 해결하기 위하여 작지만 양의 상수인 η를 이용하였다.

3.3. Frailty 모형의 사후분포와 MCMC 알고리즘

설정된 모수들을 이용하여 사후분포를 나타내면 다음과 같다.

$$\prod_{t\in T_n} \pi(\Delta A_0, \beta, \underline{\epsilon}, \sigma^2|\text{data}) \propto$$
$$\prod_{t\in T_n}\left(\prod_{(i,j)\in D(t)}\left[1-\{1-\Delta A_0(t)\}^{\exp(\beta z_{ij}+\epsilon_i)}\right]\right)$$
$$\times\{1-\Delta A_0(t)\}^{\sum_{(i,j)\in R(t)-D(t)}\exp(\beta z_{ij}+\epsilon_i)}\Delta A_0(t)^{-1}\{1-\Delta A_0(t)\}^{-1+\eta}$$
$$\times(\sigma^{-2})^{\alpha-1}\exp(-\sigma^{-2}/\kappa)\times(\sigma^{-2})^{n/2}\exp\left(-\frac{\sum_{i=1}^n\epsilon_i^2}{2\sigma^2}\right). \tag{3.5}$$

MCMC 알고리즘을 개발하기 위해서 변수변환과 Laud 등 (1998)과 Damien 등 (1999)이 제안한 방법인 보조변수를 이용한 Gibbs 샘플링을 이용한다. 우선, 시점 t에서의 기저 위험함수 $\Delta A_0(t)$를 $v_t = -\log\{1-\Delta A_0(t)\}$와 같이 변환한다.
또한, (3.6)의 변수간의 관계를 이용하여

$$\frac{f(v|y,w)}{f(w,y|v)} \propto f(v), \tag{3.6}$$

변환된 변수 v_t에 대해 보조변수 y_t 와 $\underline{w}$를 다음과 같이 설정한다. 우선, y_t 에 대해서는

$$f(y_t|v_t) \propto \{1-\exp(-v_t)\}\exp(-v_t y_t) \tag{3.7}$$

와 같이 기하분포(geometric distribution)를 이용한다. 여기서, $t\in T_n$ 이고, $y_t = 0, 1\ldots,$ 이다. $\underline{w} = \{w_{ij}|(i,j)\in D(t), t\in T_n\}$ 에 대해서는

$$f(w_{ij}|v_t) \propto v_t\exp(\beta z_{ij}+\epsilon_i)\exp\{-\exp(\beta z_{ij}+\epsilon_i)v_t w_{ij}\}$$
$$\times[1-\exp\{-\exp(\beta z_{ij}+\epsilon_i)v_t\}]^{-1} I_{(0,1)}(w_{ij}) \tag{3.8}$$

와 같이 절단된 지수분포(truncated exponential distribution)를 이용한다. (3.7)은 관측된 생존시간에 대해서만 적용되고, (3.8)은 관측된 생존시간에서의 개체들에 대해서만 적용된다.

하나의 변환된 변수와 두개의 보조변수들을 이용하여 만들어진 사후분포는 다음과 같다.

$$\pi(\underline{v},\beta,\sigma^2,\underline{\epsilon},\underline{w},\underline{y}|\text{data}) \propto$$
$$\prod_{t\in T_n} \exp\left\{-v_t \sum_{(i,j)\in R(t)-D(t)} \exp(\beta z_{ij}+\epsilon_i)\right\} \exp\{-v_t(\eta+y_t)\}$$
$$\times \exp\left\{-\sum_{(i,j)\in D(t)} v_t w_{ij} \exp(\beta z_{ij}+\epsilon_i)\right\} \exp(-\sigma^{-2}/\kappa) \exp\left(-\frac{\sum_{i=1}^n \epsilon_i^2}{2\sigma^2}\right)$$
$$\times (\sigma^{-2})^{\alpha+n/2-1} {v_t}^{k_t} \left\{\prod_{(i,j)\in D(t)} \exp(\beta z_{ij}+\epsilon_i)\right\}. \tag{3.9}$$

여기서, $\underline{v}=(v_1,\ldots,v_{q_n})$, $\underline{y}=(y_1,\ldots,y_{q_n})$ 이다.

Gibbs 샘플링을 알고리즘을 위한 각 모수들의 조건부분포들은 다음과 같이 얻을 수 있다. 첫째로, v_t의 조건부분포는

$$\pi(v_t|\beta,\sigma^2,\epsilon_i,w_{ij},\underline{y},\text{data})$$
$$\propto \quad {v_t}^{k_t} \exp\left[-v_t\left\{\sum_{(i,j)\in R(t)} \exp(\beta z_{ij}+\epsilon_i)+y_t+\eta-\sum_{(i,j)\in D(t)}(1-w_{ij})\exp(\beta z_{ij}+\epsilon_i)\right\}\right]$$

으로, 모수가 k_t+1 과

$$\left\{\sum_{(i,j)\in R(t)} \exp(\beta z_{ij}+\epsilon_i)+y_t+\eta-\sum_{(i,j)\in D(t)}(1-w_{ij})\exp(\beta z_{ij}+\epsilon_i)\right\}^{-1}$$

인 감마분포를 따른다. 여기서, k_t 는 시점 t 에서 발생한 사건의 개체 수이다. 보조변수 y_t와 w_{ij}에 대해서는 각각, (3.7) 와 (3.8)를 이용하여 구한다. β 의 조건부분포는 (3.10)과 같이 표준 분포를 따르지 않아, Metropolis-Hastings의 알고리즘을 이용하여 그 값을 생성한다.

$$\pi(\beta|\underline{v},\sigma^2,\underline{\epsilon},\underline{w},\underline{y},\text{data})$$
$$\propto \prod_{t\in T_n} \exp\left[-v_t\left\{\sum_{(i,j)\in R(t)} \exp(\beta z_{ij}+\epsilon_i)-\sum_{(i,j)\in D(t)}(1-w_{ij})\exp(\beta z_{ij}+\epsilon_i)\right\}\right]$$
$$\times \left\{\prod_{(i,j)\in D(t)} \exp(\beta z_{ij}+\epsilon_i)\right\}. \tag{3.10}$$

프레일티의 분산인 σ^2의 조건부분포는 모수가 $(\alpha+n/2)$ 과 $\left(\kappa^{-1}+\sum_{i=1}^n {\epsilon_i}^2/2\right)^{-1}$ 인 감마분포를 따른다.

$\underline{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ 의 조건부분포는

$$\begin{aligned}
&\pi(\underline{\epsilon}|\beta, \underline{v}, \sigma^2, \underline{w}, \underline{y}, \text{data}) \\
&\propto \prod_{t \in T_n} \exp\left[-v_t \left\{ \sum_{(i,j)\in R(t)} \exp(\beta z_{ij} + \epsilon_i) - \sum_{(i,j)\in D(t)} (1 - w_{ij}) \exp(\beta z_{ij} + \epsilon_i) \right\}\right] \\
&\qquad \times \prod_{(i,j)\in D(t)} \exp(\epsilon_i) \exp\left(-\frac{\sum_i \epsilon_i^2}{2\sigma^2}\right) \qquad (3.11)
\end{aligned}$$

와 같이 나타나는데, 이를 관측된 생존시간 t, $t \in T_n$ 에 대해서 표현하는 대신, n개의 부 그룹에 대해서 표현하고, 새로운 보조변수 $\underline{u} = (u_1, \ldots, u_n)$ 를 이용하여 다음과 같이 표현할 수 있다. 즉,

$$\begin{aligned}
&\pi(\underline{\epsilon}|\beta, v_t, \sigma^2, \underline{w}, y_t, \text{data}) \\
&\quad \propto \prod_{i=1}^{n} \exp\{-\exp(\epsilon_i) M]\} \exp\left\{-\frac{\epsilon_i^2 - 2\sigma^2(k_{T_i}\epsilon_i)}{2\sigma^2}\right\} \\
&\quad \propto \prod_{i=1}^{n} \exp(-u_i) I\left[u_i > \exp(\epsilon_i) M\right] \exp\left\{-\frac{(\epsilon_i - \sigma^2 k_{T_i})^2}{2\sigma^2}\right\}, \qquad (3.12)
\end{aligned}$$

여기서, $M = \sum_{t \in T_n} v_t \left[\sum_{j=1}^{n_i} \{\exp(\beta z_{ij}) I(T_{ij} \geq t) - (1 - w_{ij}) \exp(\beta z_{ij}) I(T_{ij} = t, \delta_{ij} = 1)\}\right]$ 이다. (3.12)을 이용하여 $\underline{u}$와 $\underline{\epsilon}$는 각각 절단된 지수분포와 절단된 정규분포(truncated normal distribution)를 이용하여 샘플링할 수 있다.

4. 모의실험

Hedeker 등 (2000)의 방법과 3 장에서 소개된 베이지안 부스트랩방법을 그룹 수에 따라 비교하기 위해서, 그룹의 수를 3개인 경우($S_d = 3$)와 10개인 경우($S_d = 10$)에 대해서 비교해보도록 한다. 그룹의 수가 3개인 경우에는 시점이 $0.3, 0.5, 1.0$ 으로 발생 확률을 동일하게 하였고, 그룹의 수가 10개인 경우에는 시점을 1부터 10까지 발생 확률을 p^i ($p = 0.9$, $i = 1, \ldots, 10$) 에 비례하게끔 설정하였다. 프레일티가 없는 경우($\sigma^2 = 0.0$), 작은 경우($\sigma^2 = 0.1$), 큰 경우($\sigma^2 = 1.0$)일 때를 비교하였고, 프레일티가 존재할 경우 paired match design 을 이용하여 자료를 생성하였다. 그룹은 변량 Z을 이용하여 0과 1로 표시하였다. 생존자료에서 중도 절단된 정도에 따라 비교하기 위해서는 중도 절단 확률(p_c)을 중도절단이 없는 경우($p_c = 0.0$), 중도절단이 조금 있는 경우($p_c = 0.2$), 중도절단이 많이 발생한 경우($p_c = 0.5$) 를 고려하였다. 위 조건에 대해 각 그룹의 표본의 크기를 $n = 30, 50, 100$으로 설정하여 비교하였다.

표 6.1 부터 표 6.5 까지는 모의 실험 결과의 일부를 95% 신뢰구간에 대해 참값을 포함한 확률과 신뢰구간의 평균 길이를 나타낸 것이다. 표에서 'Bayesian'은 베이지안부스트랩 방법을 이용하여 신뢰구간을 계산할 때에, 꼬리부분을 같게 하여 계산하는 사분위수방법을 이용한 것이고, 'App.Bayesian'은 분산을 계산하여 정규분포로 근사시켜 계산한 방법이다. 이 둘을 비교하여 보면 근사시킨 베이지안 부스트랩방법이 사분위수 방법보다 참값을 더 많이 포함하고 신뢰구간의 평균 길이도 더 짧아 안정적인 결과를 보였다. 한편, Hedeker 등 (2000)의 방법은 그룹화된 자료에 대해서는 전반적으로 안정적인 결과보였다. 중도 절단이 없고 회귀계수의 참값이 0일 때에는 자료의 수에 관계 없이 베이지안 부스트랩방법은 Hedeker 등 (2000)의 방법 만큼 좋은 결과를 보였다. 회귀계수의 참값이 1인 경우의 베이지안 부스트랩방법은 사분위수 방법으로 신뢰구간을 나타냈을 때, 프레일티가 없는 경우에 결과가 조금 안좋게 나타났지만, 근사적인 방법을 사용했을 경우에는 결과가 좋게 나타났다. 중도 절단율이 클 경우, 즉, $p_c = 0.5$인 경우, 회귀계수의 참값이 1이고 자료의 수가 작을 때 Hedeker 등의 방법과 베이지안 부스트랩방법에서 종종 수렴하지 못하는 못하는 경우가 발생했지만, 수렴하지 못하는 경우를 제외하고는 결과가 좋았다. 전반적으로, Hedeker 등 (2000)의 방법은 결과가 좋지만, 베이지안 부스트랩방법보다 신뢰구간의 평균길이가 길고, 베이지안 부스트랩의 경우에서는, 대부분의 결과가 좋게 나타났지만, 실제 프레일티가 없는데, 프레일티가 있는 것처럼 가정하고 베이지안 부스트랩방법을 적용 시켰을 때에는 회귀 계수의 참값을 포함하는 경우가 95%에 미치지 못하는 못하지만, Hedeker 등 (2000)의 방법보다 신뢰구간의 평균길이가 짧았다.

베이지안부스트랩의 또 다른 장점은 계산속도가 Hedecker 등 (2000)의 방법에 비하여 매우 빠르다는 것이다. 계산상의 장점을 비교하기 위하여 그룹자료와 연속자료가 혼합된 모형을 고려하였다. 예를들어, 환자의 질병 정도에 따라 의사들의 진료 방법이 달라진다고 하자. 즉, 의사의 진료 형태가 정기 검진, 의사의 진료에 대한 정보가 없는 경우, 다음 번의 진료를 바로 전 검진 때 알려주는 경우, 혹은, 환자가 자신의 상태에 따라 진료를 요청하는 경우도 발생할 수 있다 (Grüger 등,1991). 이러한 경우에는 생존자료가 혼합형태가 된다. 이러한 혼합형태의 생존시간에서 랜덤항이 포함된 경우, Hedeker 등 (2000)의 방법은 계산이 불가능하게 된다. 실제로, 표본의 수가 60이고, $\sigma^2 = 1.0$, 중도 절단율이 0.0이고 그룹의 수(S_d)가 3인 자료를 혼합형태의 생존자료로 생성하여, 자료를 한번 분석시 소요되는 시간을 비교하였는데, Hedeker 등 (2000)의 방법은 약 40분이 소요되는 반면, 베이지안 부스트랩방법은 15초 안팍으로 소요되었다. 물론, 사용된 소프트웨어가 다르다는 고려해야할 점이 있지만, 계산 시간에서 많은 차이를 보였고, 혼합 형태의 생존자료에서 랜덤항을 고려할 때의 Hedeker 등 (2000)의 방법에 대한 모의 실험은 불가능한 상태임을 보여주고 있다. 표 6.4 은 혼합모형에서 베이지안부스트랩의 모의실험결과이다. 자료의 반은 그룹화된 자료이고 나머지 반은 연속형자료를 이용하여 가상실험을 하였다. 표본수가 작은 경우에, 성능이 많이 떨어지지만, 표본수가 증가하면서 점차 개선됨을 알 수 있다. 반면에, 혼합모형에서의 Hedeker 등 (2000)의 방법은 계산시간이 가중으로 인하여 할 수가 없었다.

표 5.1: 여러방법을 이용한 β 의 추정값과 신뢰구간

Method	Point Estimate	95% C.I
Breslow	-1.5092	(-2.3119, -0.7064)
Efron	-1.5721	(-2.3804, -0.7638)
Exact	-1.5979	(-2.4242, -0.7715)
Prentice	-1.6587	(-2.4869, -0.8304)
Hedeker	-1.6586	(-2.5401, -0.7771)
Bayesian	-2.1139	(-3.2894, -1.1156)
App. Bayesian	-2.1139	(-3.2102, -1.0176)

5. 예제

이 장에서는 예제를 통하여 앞서 언급된 방법들의 계수 추정값을 비교해보고자 한다. 사용된 자료는 42명의 급성백혈병 환자들을 대상으로, 각각 21명씩 치료그룹과 비교그룹으로 나뉘어, 치료그룹의 환자에게는 6-mercaptopurine(6-MP)를, 비교그룹은 위약을 처방하여 비교연구한 자료이다 (Klein 과 Moeschberger, 1997). 이 실험은 환자의 상태에 따라 대응짝(matched pair)으로 설계되었으며, 환자의 병이 완치되었거나 연구종료시까지 매달 추적조사되었다. 관측치 중 12개의 자료는 중도절단되었고 17개의 생존시간이 관측되었다. 관측된 생존자료 중, 7개가 서로 다르게 관측되었고, 8개가 두개의 관측치를, 세개와 네개를 갖는 관측치가 각각 한개씩 관측되었다.

표 5.1는 여러가지 방법을 이용하여 얻어진 점추정치와 95% 신뢰구간을 나타내었다. 첫 블럭은 랜덤항을 고려하지 않는 방법들로, SAS를 이용하여 결과를 나타내었다. 'Prentice'는 Prentice와 Gloeckler가 제안한 방법이다. 두번째 블럭은 랜덤항을 고려한 방법이다.

프레일티가 없는 방법에 대한 결과를 보면, 동측치 효과를 보정할 수록 추정값이 작아짐을 알 수 있다. 프레일티가 있는 방법을 사용한 경우에는 Bayesian 방법으로 추정치가 Hedecker 등의 방법에 비하여 작음을 알 수 있다. 이에 대한 하나의 가능한 설명은, 프레일티항의 효과가 매우 강하여, 쓰여지는 방법에 따라 다른 결과를 준다는 것이다. 하지만, 이러한 설명도 확실한 것은 아니며, 좀더 깊은 연구가 필요하다. 본 자료의 분석을 통하여 한가지 명확하게 밝혀진 것은 동측치의 효과를 보정해주는 것이 프레일티항이 존재하는 경우 매우 중요하다는 것이다.

6. 토의 및 결론

생존 모형에서 추론에 사용되는 우도함수는 포아송형태의 경험적우도함수와 이항형태의 경험적우도함수로 나뉜다. 포아송 형태의 경험적우도함수로 추론하는 방법은 Breslow (1974)의 부분우도함수와 같게 된다. 하지만 포아송형태의 경험적우도함수 접근방법은 자

료의 형태가 연속형인 경우에는 설명을 잘 하지만, 동위 자료의 수가 많은 경우에는 편의가 발생한다. 이항형태의 경험적우도함수로 추론하는 경우, 일반 선형 모형의 틀에서 이항형태의 반응변수로 나타내고, 로그-로그 연결함수를 이용하여 모수를 설명할 수 있고, 누적기저위험함수값을 추정할 수 있는 장점이 있다. 프레일티가 없는 경우의 이항형태의 경험적우도함수 방법(Prentice 와 Gloeckler, 1978)은 SAS GENMOD 프로시저를 이용할 수 있고, 프레일티가 포함된 경우(Hedeker 등,2000)는 SAS NLMIXED 프로시저를 이용하여 자료를 적합시킬 수 있다. 따라서 동위 자료가 많은 경우에는 Prentice 와 Gloeckler (1978)의 방법이나 Hedeker 등 (2000)의 방법을 이용한 모형적합이 좋다.

하지만, 자료가 그룹자료와 연속자료가 섞여 있는 경우에는, Hedeker 등 (2000)의 방법은 계산량의 폭증으로 실제 문제에 쓰여지기가 어렵다. 또한, 랜덤항이 많아지는 경우 Hedeker 등 (2000)의 방법은 계산시간과 정확성의 문제가 알려져 있다. 이러한 문제를 해결하고자 본 논문에서 베이지안 부스트랩방법을 제안하였다. 랜덤항이 많아지는 경우 베이지안 부트스랩방법은 단지 추정할 모수의 수가 늘어나는 것으로 간주하여 계산상의 문제에 큰 영향을 끼치지 않는다. 프레일티의 분포가 정규분포가 아닌 다른 분포를 따를 경우에는 사후분포의 형태가 매우 복잡해져서 계산상의 어려움이 있다. 모의실험을 통해서 제안된 베이지안 부스트랩방법은 그룹자료에서는 Hedeker 등 (2000)의 방법과 비슷한 성능을 나타냈으며, 혼합자료에서도 계산량이 많지 않았다. 단, 혼합자료에서 성능이 표본수가 작은 경우 떨어지는 경향이 있었다. 좀 더 심도있는 비교를 위해 설명변수나 랜덤항의 수를 달리하여 회귀계수 및 프레일티의 분산과 같은 산포 모수를 비교하거나, Gibbs 샘플링의 수를 달리하는 등의 여러가지 차원에서의 비교를 고려할 수 있고, 작은 표본수에서 베이지안부스트랩의 성능을 개선하는 문제는 아주 중요한 문제라 사료되며 이들은 향후 연구과제로 남겨놓는다.

참고문헌

Breslow N.E. (1974). Covariance analysis of censored survival data, *Biometrics,* **30**, 89–99.

Clayton, G.D. (1991). A monte carlo method for Bayesian inference in frailty models, *Biometrics,* **47**, 467–485.

Cox D.R. (1972). Models and life tables (with discussion), *Journal of the Royal Statistical Society B,* **34**, 187–220.

Damien P., Wakefield J., and Walker S.(1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables, *Journal of the Royal Statistical Society B,* **61**, 331–344.

Efron B. (1979). Bootstrap methods: another look at the jackknife, *Annals of Statistics,* **7**, 1–26.

Grüger J., Kay R., and Schumacher M. (1991). The validity of inferences based in incomplete observations in disease state models, *Biometrics,* **47**, 595–605.

Ha I.D., Lee Y., and Song J.K. (2001). Heirachical likelihood approach for frailty models, *Biometrika,* **88**, 233–243.

Heckman J.J. and Singer(1984). A method for minimizing the impact of distributional assumptions in economic models of duration data, *Econometrica,* **52**, 271-320.

Hedeker, D. and Siddiqui, O. and Hu, F.B. (2000). Random-effects regression analysis of correlated grouped-time survival data, *Statistical Methods in Medical Research,* **9**,161-179.

Kim Y and Lee J (2003). Bayesian bootstrap for proportional hazard model, *Annals of Statistics,* **31**, 1905–1922.

Klein J.P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm, *Biometrics,* **48**, 795–806.

Klein J.P. and Moeschberger M.L. (1997). *Survival Analysis : Techniques for Censored and Truncated Data*, Springer-Verlag.

Laud, P.W. and Damien, P. and Smith, A.F.M.(1998). Bayesian nonparametric and covariate analysis of failure time data, chapter in practical nonparametric and semiparametric Bayesian statistics, Springer-Verlag, 213–225.

Lo A.Y. (1993). A Bayesian bootstrap for censored data, *Annals of Statistics,* **21**, 100-123.

McGlichrist C.A. and Aisbett C.W.(1991). Regresion with frailty in survival analysis, *Biometrics,* **47**, 461-466.

Nielsen G.G., Gill D.R., Andersen P.K., and Sorensen I.A.T. (1992). A counting process approach to maximum likelihood estimation in frailty models, *Scandinavian Journal of Statistics*, **19**, 25-44.

Prentice R.L. and Gloeckler L.A.(1978). Regression analysis of grouped survival data with application to breast cancer data, *Biometrics,* **34**, 57–67.

Rubin D.B. (1981). The Bayesian bootstrap, *Annals of Statistics,* **9**, 130-134.

Schumacher, M. and Olschewski, M. and Schmoor, C. (1987). The impact of heterogeneity on the comparison of survival times, *Statistics in Medicine,* **6** 773–784.

Struther C.A. and Kalbfleisch J.D.(1986). Misspecified proportional hazards models, *Biometrika,* **73**, 363-369.

Vaupel, J.W. and Manton, K.G. and Stallard, E.(1979). The impact of heterogeneity in individual frailty on the dynamics of mortality, *Demography,* **16**, 439–454.

[2003년 12월 접수, 2004년 9월 채택]

표 6.1: Coverage of coefficients ($p_c = 0.0$)

			$S_d = 3$			$S_d = 10$		
β	n	σ^2	Hedeker	Bayesian	App. Bayesian	Hedeker	Bayesian	App. Bayesian
0	30	0	0.948	0.941	0.948	0.945	0.909	0.913
		0.1	0.965	0.940	0.947	0.957	0.930	0.936
		1	0.953	0.962	0.969	0.958	0.960	0.970
	50	0	0.949	0.945	0.946	0.935	0.936	0.938
		0.1	0.954	0.952	0.959	0.957	0.936	0.942
		1	0.978	0.972	0.972	0.970	0.953	0.954
	100	0	0.945	0.947	0.948	0.948	0.933	0.935
		0.1	0.958	0.948	0.949	0.959	0.955	0.955
		1	0.949	0.965	0.968	0.945	0.961	0.964
1	30	0	0.960	0.912	0.936	0.944	0.919	0.937
		0.1	0.974	0.926	0.951	0.965	0.948	0.952
		1	0.971	0.955*	0.967*	0.958	0.925	0.939
	50	0	0.956	0.932	0.952	0.949	0.932	0.937
		0.1	0.963	0.949	0.961	0.957	0.948	0.955
		1	0.974	0.950	0.952	0.946	0.949	0.953
	100	0	0.953	0.929	0.949	0.952	0.935	0.938
		0.1	0.950	0.933	0.956	0.945	0.944	0.947
		1	0.954	0.938	0.943	0.947	0.938	0.942

표 6.2: Average Length of Confidence Intervals ($p_c = 0.0$)

			$S_d = 3$			$S_d = 10$		
β	n	σ^2	Hedeker	Bayesian	App. Bayesian	Hedeker	Bayesian	App. Bayesian
0	30	0	1.3685	1.2993	1.2972	1.1552	1.1418	1.1399
		0.1	1.4061	1.3111	1.3083	1.1784	1.1537	1.1512
		1	1.5714	1.3840	1.3807	1.3186	1.2305	1.2284
	50	0	1.0224	0.9967	0.9956	0.8574	0.8563	0.8557
		0.1	1.0351	1.0001	0.9987	0.8740	0.8678	0.8670
		1	1.1612	1.0641	1.0630	0.9876	0.9414	0.9406
	100	0	0.7047	0.6971	0.6964	0.5876	0.5928	0.5924
		0.1	0.7147	0.7013	0.7008	0.5994	0.6016	0.6015
		1	0.8037	0.7538	0.7534	0.6867	0.6637	0.6632
1	30	0	1.5147	1.4925	1.4851	1.3201	1.3391	1.3364
		0.1	1.5566	1.5077	1.5016	1.3496	1.4559	1.4546
		1	1.9815	1.6519	1.6385	1.5461	0.9923	0.9910
	50	0	1.0824	1.0901	1.0873	0.9685	0.9898	0.9878
		0.1	1.1174	1.1017	1.0989	0.9840	1.0135	1.0115
		1	1.4152	1.2472	1.2448	1.1499	1.1106	1.1103
	100	0	0.7226	0.7368	0.7355	0.6561	0.6763	0.6760
		0.1	0.7516	1.1408	1.1340	0.6715	0.6956	0.6955
		1	0.9443	0.8822	0.8820	0.7895	0.7716	0.7714

표 6.3: Coverage of coefficients ($p_c = 0.5$)

			$S_d = 3$			$S_d = 10$		
β	n	σ^2	Hedeker	Bayesian	App. Bayesian	Hedeker	Bayesian	App. Bayesian
0	30	0	0.969	0.928	0.950	0.945	0.908	0.929
		0.1	0.973	0.941*	0.958*	0.952	0.916	0.938
		1	0.976	0.945*	0.976*	0.959	0.917	0.939
	50	0	0.963	0.936	0.947	0.946	0.913	0.926
		0.1	0.956	0.958	0.965	0.947	0.933	0.937
		1	0.976	0.954	0.968	0.960	0.926	0.938
	100	0	0.942	0.939	0.947	0.929	0.943	0.946
		0.1	0.964	0.942	0.946	0.957	0.933	0.938
		1	0.946	0.965	0.970	0.951	0.945	0.950
1	30	0	0.957*	0.938*	0.965*	0.969	0.895	0.944
		0.1	0.974*	0.952*	0.965*	0.978	0.886	0.943
		1	0.951*	0.943*	0.953*	0.962	0.886	0.929
	50	0	0.964*	0.907*	0.961*	0.955	0.922	0.945
		0.1	0.974	0.908*	0.955*	0.957	0.910	0.945
		1	0.952	0.956*	0.976*	0.953	0.919	0.948
	100	0	0.960	0.924	0.948	0.946	0.941	0.944
		0.1	0.965	0.933	0.956	0.953	0.942	0.956
		1	0.940	0.945	0.952	0.943	0.954	0.961

표 6.4: Average Length of Confidence Intervals ($p_c = 0.5$)

			$S_d = 3$			$S_d = 10$		
β	n	σ^2	Hedeker	Bayesian	App. Bayesian	Hedeker	Bayesian	App. Bayesian
0	30	0	3.3543	1.8829	1.8732	1.8308	1.7620	1.7560
		0.1	1.9213	1.9585	1.9450	1.7607	1.7937	1.7858
		1	2.0183	2.1499	2.1085	1.9545	1.9993	1.9918
	50	0	1.3509	1.3753	1.3708	1.2174	1.2780	1.2756
		0.1	1.4179	1.3779	1.3737	1.2858	1.3072	1.3043
		1	1.7368	1.4978	1.4905	1.4093	1.4198	1.4152
	100	0	0.9126	0.9184	0.9173	0.8534	0.8512	0.8501
		0.1	0.9435	0.9309	0.9294	0.8814	0.8602	0.8591
		1	1.1615	1.0139	1.0113	0.9862	0.9379	0.9361
1	30	0	4.1972	2.3839	1.8655	2.5066	2.3276	2.3279
		0.1	3.5235	2.3779	1.7707	2.1795	2.4242	2.4322
		1	5.3922	2.3554	1.3888	3.0671	2.8401	2.8575
	50	0	2.2163	1.9165	1.9082	1.4213	1.5221	1.5195
		0.1	1.7836	1.9813	1.9713	1.4079	1.5730	1.5690
		1	2.8858	2.2536	2.2376	1.6701	1.8232	1.8225
	100	0	1.0329	1.1019	1.0961	0.9561	0.9917	0.9897
		0.1	1.0581	1.1408	1.1340	1.0188	1.1586	1.1575
		1	1.6383	1.3339	1.3301	1.0856	1.1676	1.1664

표 6.5: Coverage of coefficients for MIXTURE data

			$p_c = 0.0$		$p_c = 0.5$	
β	n	σ^2	Bayesian	App. Bayesian	Bayesian	App. Bayesian
0	30	0	0.884	0.893	0.892	0.912
		0.1	0.874	0.882	0.897	0.911
		1	0.874	0.882	0.898	0.914
	50	0	0.902	0.906	0.916	0.925
		0.1	0.908	0.913	0.923	0.934
		1	0.868	0.873	0.927	0.934
	100	0	0.941	0.945	0.938	0.936
		0.1	0.927	0.931	0.935	0.936
		1	0.893	0.896	0.944	0.949
	200	0	0.924	0.929	0.938	0.942
		0.1	0.944	0.943	0.937	0.941
		1	0.921	0.921	0.938	0.944
1	30	0	0.916	0.930	0.893	0.927
		0.1	0.913	0.928	0.896	0.935
		1	0.892	0.894	0.890	0.919
	50	0	0.926	0.939	0.919	0.942
		0.1	0.918	0.929	0.922	0.937
		1	0.889	0.900	0.931	0.943
	100	0	0.926	0.930	0.925	0.933
		0.1	0.943	0.954	0.934	0.948
		1	0.911	0.912	0.926	0.931
	200	0	0.927	0.929	0.947	0.951
		0.1	0.931	0.938	0.926	0.936
		1	0.930	0.932	0.957	0.960

Analysis of the Frailty Model with Many Ties*

Yongdai Kim [1] Jin-Kyung Park [2]

ABSTRACT

Most of the previously proposed methods for the frailty model do not work well when there are many tied observations. This is partly because the empirical likelihood used is not suitable for tied observations. In this paper, we propose a new method for the frailty model with many ties. The proposed method obtains the posterior distribution of the parameters using the binomial form empirical likelihood and Bayesian bootstrap. The proposed method yields stable results and is computationally fast. To compare the proposed method with the maximum marginal likelihood approach, we do simulations.

Keywords: Frailty model, Poisson form empirical likelihood, Binomial form empirical likelihood, Bayesian bootstrap.

* This research is supported by Korean Research Foundation Grant (KRF-2002-003-C00029)

1) Assistant Professor, Dept. of Statistics, Seoul National University, San 56-1, Silim-dong, Kwanak-gu, Seoul, Korea, 151-742
E-mail: ydkim@stats.snu.ac.kr

2) Associate Scientific Analyst, International Vaccine Institute, SNU Research Park, San 4-8, Bongcheon-7 dong, Kwanak-gu, Seoul, Korea, 151-818
E-mail: jkpark@ivi.int