The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.2.331
27(2), 331–343

Modeling Clustered Interval-Censored Failure Time Data with Informative Cluster Size

Jinheum Kim[a,1] · Youn Nam Kim[b]

[a]Department of Applied Statistics, University of Suwon
[b]Clinical Trials Center Severance Hospital, Yonsei University Health System

(Received January 3, 2014; Revised February 26, 2014; Accepted March 14, 2014)

Abstract

We propose two estimating procedures to analyze clustered interval-censored data with an informative cluster size based on a marginal model and investigate their asymptotic properties. One is an extension of Cong *et al.* (2007) to interval-censored data and the other uses the within-cluster resampling method proposed by Hoffman *et al.* (2001). Simulation results imply that the proposed estimators have a better performance in terms of bias and coverage rate of true value than an estimator with no adjustment of informative cluster size when the cluster size is related with survival time. Finally, they are applied to lymphatic filariasis data adopted from Williamson *et al.* (2008).

Keywords: Informative cluster size, interval censoring, marginal model, weighted estimating equation, within-cluster resampling.

1. 서론

많은 생존자료는 한 개체에 대해 관심 있는 한 가지 이벤트가 발생할 때까지의 생존시간을 관측하지만 때로는 두 가지 이상의 이벤트가 발생할 때까지의 생존시간을 관측하곤 한다. 또한 유전역학 연구와 같이 가족 단위로 하는 추적 연구에서는 가족 내에 있는 임의 한 개체보다 가족 모두의 생존시간을 동시에 관측한다. 이와 같은 생존자료를 상관 생존자료(correlated failure time data) 혹은 군집 생존자료(clustered failure time data)라고 한다. 이런 자료는 의학 및 동물실험 연구 등 많은 분야에서 자주 접하게 된다 (Cai와 Prentice, 1995; Kalbfleish와 Prentice, 2002). 군집 생존자료에서는 동일한 개체에서 관측된 생존시간들이 서로 종속되어 있으며, 또한 같은 군집 내에 있는 개체들의 생존시간이 서로 종속되어 있기 때문에 모든 개체의 생존시간이 서로 독립이라는 가정 아래서 얻는 추론 결과는 편향될 수밖에 없다. 한편 주기적으로 관측하는 연구에서는 개체의 생존시간이 정확하게 관측되기보다 그 값이 속한 구간에 대한 정보만 관측되곤 하는데 이런 자료를 구간중도절단된 자료(interval-censored data)라고 한다. 본 논문에서는 군집 구간중도절단된 자료(clustered interval-censored data)의 회귀모형에서

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (No.2011-0010889).
[1]Corresponding author: Department of Applied Statistics, University of Suwon, 17 Wauan-gil, Bongdam-eup, Hwaseong, Gyeonggi 445-743, Korea. E-mail: jkimdt65@gmail.com

통계적 추론 문제를 다루고자 한다. 군집 자료를 분석할 때 일반적으로 같은 군집 내에 있는 개체들의 개수와 생존시간은 서로 무관하다고 가정한다. 그러나 다음 예에서 볼 수 있듯이 군집의 크기가 생존시간에 대해 어떤 정보를 제공할 수도 있는데 군집 내에 있는 개체들 간의 생존시간이 군집의 크기와 상관되어 있을 때 군집의 크기가 생존시간에 영향력이 있다라고 말한다. 살충제가 새끼들의 생존시간에 미치는 영향에 대한 연구에서 보면 같은 어미에서 난 새끼의 수가 적을수록(즉, 군집의 크기가 작을수록) 생존시간이 짧은 경향이 있는데 이는 살충제에 쉽게 영향을 받는 어미일수록 결점을 지닌 새끼들을 낳거나 태아재흡수(fetal resorption)로 인해 새끼들의 수가 줄어들기 때문이다 (Hoffman 등, 2001; Cong 등, 2007). 치아 연구에서도 만성 치주염을 앓고 있는 사람일수록 어금니의 수가 적은데 어금니의 수가 적을수록(즉, 군집의 크기가 작을수록) 다른 어금니들의 생존시간이 짧아지는 경향이 있다 (McGuire와 Nunn, 1996; Cong 등, 2007).

생존시간이 군집의 크기와 상관되어 있을 때 군집화 된 생존자료를 다루는 방법은 크게 두 가지로 나눌 수 있다. 한 방법은 주변모형(marginal model)을 가정하고 (Huster 등, 1989; Wei 등, 1989; Lee 등, 1992) 군집의 크기를 가중값으로 처리하는 방법이고, 다른 방법은 Hoffman 등 (2001)이 제안한 군집 내 재추출(within-cluster resampling; WCR) 방법을 사용하는 것이다. 전자는 군집의 크기를 직접적으로 추정 방정식에 반영하는 방법이며 이미 잘 알려진 것처럼 주변모형은 군집 내에 있는 개체들 간의 종속성을 고려하지 않기 때문에 일반화 추정 방법(generalized estimating equation; GEE)을 써서 로버스트한 방법으로 분산을 추정한다. 반면에 후자는 군집의 크기를 모수 추정에 간접적으로 반영하는 방법이다. 다시 말해 WCR 방법은 군집의 크기가 생존시간에 미치는 영향을 제거하기 위해 군집별로 랜덤하게 하나의 표본만을 뽑아 재구성하는 것이다. 그러면 이 자료는 더 이상 군집화 된 자료가 아니고 서로 독립인 생존자료에 불과하기 때문에 이미 잘 알려진 방법을 써서 모수를 추정할 수 있다. WCR 방법은 반복적으로 재추출하기 때문에 컴퓨팅 시간이 많이 필요하지만 군집 내 자료들 간의 종속성과 생존시간과 상관된 군집의 크기를 동시에 해결할 수 있는 장점이 있다. Cong 등 (2007)은 군집 우중도절단된 자료(clustered right-censored data)에서 콕스 비례위험모형(Cox proportional hazards model)을 가정하고 생존시간과 상관된 군집의 크기를 모수 추정에 반영하기 위해 군집의 크기의 역수를 가중값으로 하는 가중 추정 방법을 제안하였다. Zhang과 Sun (2010)은 군집 구간중도절단된 자료에서 와이블 분포 모형을 가정하고 군집의 크기의 역수를 가중값으로 하는 추정 방법을 제안하였다. Zhang과 Sun (2010)의 방법은 Williamson 등 (2008)이 군집 우중도절단된 생존자료에 대해 제안한 방법을 군집 구간중도절단된 자료로 확장한 것이다. 본 논문에서는 Cong 등 (2007)의 아이디어를 군집 구간중도절단된 자료로 확장하고자 한다. 한편 Cong 등 (2007)과 Williamson 등 (2008)은 군집 우중도절단된 자료에 WCR 방법을 적용하여 모수를 추정하였는데, 전자는 콕스 비례위험모형을 가정한 반면에 후자는 와이블 분포 모형을 가정하였다. Zhang과 Sun (2010)은 군집 구간중도절단된 자료에 WCR 방법을 적용하여 모수를 추정하였는데 이 때 그들은 기저위험함수(baseline hazard function)가 와이블 분포를 따른다고 가정하였다. 또한 본 논문에서는 Cong 등 (2007)처럼 콕스비례위험모형을 가정하고 군집 구간중도절단된 자료에 WCR 방법을 적용하여 모수를 추정하는 방법을 제안하고자 한다.

상술한 것처럼 2절에서는 가중 추정 방법과 군집 내 재추출 방법을 제안하고 그 추정량의 점근적 성질을 살펴보고자 한다. 3절에서는 모의실험을 통해 제안한 추정량의 소표본 성질을 살펴보고, 4절에서는 제안한 방법을 림프성 사상충(lymphatic filariasis; LF) 자료 (Williamson 등, 2008; Zhang과 Sun, 2010)에 적용하고자 한다. 5절에서는 연구 결과를 요약하고 기존 연구 결과와 비교하고자 한다.

2. 모수 추론

첨자 $i = 1, \ldots, n$는 군집을 나타내고, 첨자 $j = 1, \ldots, n_i$는 i번째 군집 내에 있는 개체를 나타낸다.

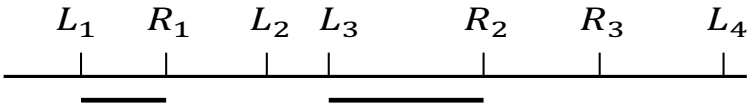

Figure 2.1. Example of equivalence set for interval-censored data

T_{ij}는 i번째 군집 내에 있는 j번째 개체의 생존시간이다. 서로 다른 군집에 있는 개체들의 생존시간은 서로 독립이지만, 같은 군집 내에 있는 개체들의 생존시간은 서로 종속될 수 있다. $\boldsymbol{x}_{ij}$는 T_{ij}와 상관된 p-차원 공변량 벡터이다. 이 때 T_{ij}는 다음과 같은 위험함수를 가진다고 가정하자 (Huster 등, 1989; Wei 등, 1989; Lee 등, 1992).

$$\lambda(t|\boldsymbol{x}_{ij}) = \lambda_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij}). \tag{2.1}$$

단, $\lambda_0(t)$는 미지의 공통 기저위험함수이고, $\boldsymbol{\beta}$는 p-차원 회귀계수 벡터이다. 본 논문에서는 i번째 군집의 T_{ij}들이 n_i에 의존한다고 가정한다. 다시 말해 같은 군집 내에 있는 개체들의 생존시간이 해당 군집의 크기에 의존한다고 가정한다. 한편 T_{ij}는 직접 관찰할 수 없고 단지 T_{ij}가 속한 구간만 관측할 수 있다고 한다. 즉, $(L_{ij}, R_{ij}]$. 이런 유형의 자료를 구간중도절단된 자료라고 한다. 특히 $L_{ij} = R_{ij}$이면 생존시간이 정확히 관측된 경우이고, $L_{ij} = 0$이면 좌중도절단된(left-censored data) 경우, $R_{ij} = \infty$이면 우중도절단된 경우이다. 따라서 관측된 자료는 다음과 같다.

$$\{((L_{ij}, R_{ij}], \boldsymbol{x}_{ij}) : \ i = 1, \ldots, n; \ j = 1, \ldots, n_i\}.$$

공변량 $\boldsymbol{x}_{ij}$가 주어졌을 때, T_{ij}는 L_{ij} 및 R_{ij}와 서로 독립이라고 가정한다.

2.1. 가중 추정 방법

i번째 군집이 우도 함수(likelihood function)에 미치는 기여도 즉, 가중값을 ω_i라고 하면, i번째 군집의 우도 함수는 다음과 같이 주어진다.

$$L_i(\lambda_0, \boldsymbol{\beta}|\boldsymbol{x}_{ij}) = \prod_{j\in D_i}\{S(L_{ij}|\boldsymbol{x}_{ij}) - S(R_{ij}|\boldsymbol{x}_{ij})\}^{\omega_i} \prod_{j\in R_i} S(L_{ij}|\boldsymbol{x}_{ij})^{\omega_i}.$$

단, $S(t|\boldsymbol{x}_{ij}) = \exp\{-\int_0^t \lambda(s|\boldsymbol{x}_{ij})ds\}$, D_i는 i번째 군집 내에 있는 개체 중에서 구간중도절단된 개체로 이루어진 집합이고, R_i는 우중도절단된 개체로 이루어진 집합이다. 따라서 군집 전체의 우도 함수는 다음과 같다.

$$\begin{aligned} L_f(\lambda_0, \boldsymbol{\beta}) &= \prod_{i=1}^{n} L_i(\lambda_0, \boldsymbol{\beta}|\boldsymbol{x}_{ij}) \\ &= \prod_{i=1}^{n}\left\{\prod_{j\in D_i}\left[\exp\left\{-\Lambda_0(L_{ij})\exp\left(\boldsymbol{\beta}'\boldsymbol{x}_{ij}\right)\right\} - \exp\left\{-\Lambda_0(R_{ij})\exp\left(\boldsymbol{\beta}'\boldsymbol{x}_{ij}\right)\right\}\right]^{\omega_i}\right. \\ &\quad \left.\times \prod_{j\in R_i}\exp\left\{-\omega_i\Lambda_0(L_{ij})\exp\left(\boldsymbol{\beta}'\boldsymbol{x}_{ij}\right)\right\}\right\}. \end{aligned}$$

단, $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$. 관찰된 자료 집합 $\mathcal{D} = \{(L_{ij}, R_{ij}] \,|\, i = 1, \ldots, n; \ j = 1, \ldots, n_i\}$에 대응하는 동등집합(equivalence set)의 중간값을

$$0 = s_0 < s_1 < \cdots < s_m < s_{m+1} = \infty$$

라고 하자 (Lindsey와 Ryan, 1998). 여기서 동등집합이란 모든 $\{L_{ij} : i = 1, \ldots, n;\ j = 1, \ldots, n_i\}$와 $\{R_{ij} : i = 1, \ldots, n;\ j = 1, \ldots, n_i\}$를 Figure 2.1과 같이 크기 순으로 한 줄에 나타냈을 때, L 다음에 곧바로 R이 나오는 구간 $[L, R]$ $(L \leq R)$들로 이루어진 집합을 의미한다. Figure 2.1과 같은 예에서 동등집합은 $\{(L_1, R_1], (L_3, R_2]\}$와 같이 정의된다. 또한 기저위험함수가 각 시점 s_q $(q = 1, \ldots, m)$에서 조각 지수분포(piecewise exponential distribution)를 따른다고 가정하면, $S_0(t) = \exp\{-\Lambda_0(t)\}$를 다음과 같이 표현할 수 있다.

$$S_0(s_q) = \prod_{k=0}^{q} \exp\{-\exp(\alpha_k)\} = \exp\left\{-\sum_{k=0}^{q} \exp(\alpha_k)\right\}, \quad q = 0, \ldots, m+1.$$

단, $\alpha_0 = -\infty$, $\alpha_{m+1} = \infty$. 따라서

$$\Lambda_0(s_q) = \sum_{k=0}^{q} \exp(\alpha_k) = a_q, \quad q = 0, \ldots, m+1$$

이다. 한편

$$I_{ijq} = I(s_q \in (L_{ij}, R_{ij}]), \quad i = 1, \ldots, n\ ;\ j = 1, \ldots, n_i;\ q = 1, \ldots, m$$

라고 하자. $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)', \boldsymbol{\beta}$를 써서 로그 우도 함수를 표현하면 다음과 같다.

$$\begin{aligned} l_f(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \log L_f(\lambda_0, \boldsymbol{\beta}) \\ &= \sum_{i=1}^{n} \omega_i \left\{ \sum_{j \in D_i} \log \left\{ \sum_{q=1}^{m+1} I_{ijq} \left[\exp\left\{-a_{q-1} \exp\left(\boldsymbol{\beta}' \boldsymbol{x}_{ij}\right)\right\} - \exp\left\{-a_q \exp\left(\boldsymbol{\beta}' \boldsymbol{x}_{ij}\right)\right\} \right] \right\} \right. \\ &\quad \left. - \sum_{j \in R_i} \sum_{q=1}^{m+1} I(L_{ij} \in [s_{q-1}, s_q)) a_{q-1} \exp\left(\boldsymbol{\beta}' \boldsymbol{x}_{ij}\right) \right\}. \end{aligned}$$

가중 스코어 벡터(weighted score vector)와 가중 관찰 정보행렬(weighted observed information matrix)를 간략히 표현하기 위해

$$\begin{aligned} f_{ijq} &= S(s_q|\boldsymbol{x}_{ij}) \log S(s_q|\boldsymbol{x}_{ij}), \quad q = 1, \ldots, m, \\ b_{ijq} &= \exp\left(\alpha_q + \boldsymbol{\beta}' \boldsymbol{x}_{ij}\right), \quad q = 1, \ldots, m, \\ c_{ijq} &= \sum_{l=q}^{m+1} (I_{ijl} - I_{ijl+1}) S(s_l|\boldsymbol{x}_{ij}), \quad q = 1, \ldots, m, \\ g_{ij} &= \sum_{q=1}^{m+1} I_{ijq} \{S(s_{q-1}|\boldsymbol{x}_{ij}) - S(s_q|\boldsymbol{x}_{ij})\} \end{aligned}$$

라고 하자. 단, $f_{ij0} = f_{ijm+1} = 0$, $I_{ijm+2} = 0$. 그러면 $\boldsymbol{\beta}$와 α_q에 대한 가중 스코어 함수는 각각 다음과 같다.

$$\begin{aligned} U_{\boldsymbol{\beta}} &= \frac{\partial l_f}{\partial \boldsymbol{\beta}} \\ &= \sum_{i=1}^{n} \omega_i \left\{ \sum_{j \in D_i} \boldsymbol{x}_{ij} \frac{\sum_{q=1}^{m+1} I_{ijq}(f_{ijq-1} - f_{ijq})}{g_{ij}} + \sum_{j \in R_i} \boldsymbol{x}_{ij} \sum_{q=1}^{m+1} I\left(L_{ij} \in [s_{q-1}, s_q)\right) a_{q-1} \exp\left(\boldsymbol{\beta}' \boldsymbol{x}_{ij}\right) \right\} \\ &= \sum_{i=1}^{n} \omega_i U_{\boldsymbol{\beta}}^{i}, \end{aligned}$$

$$U_{\boldsymbol{\alpha}} = \left(\frac{\partial l_f}{\partial \alpha_q}\right) = \sum_{i=1}^{n} \omega_i U^i_{\boldsymbol{\alpha}}.$$

단,

$$\frac{\partial l_f}{\partial \alpha_q} = \sum_{i=1}^{n} \omega_i \left\{ \sum_{j\in D_i} \frac{b_{ijq} c_{ijq}}{g_{ij}} - \sum_{j\in R_i} b_{ijq} I(L_{ij} \in [s_q, \infty)) \right\} = \sum_{i=1}^{n} \omega_i U^i_{\alpha_q}, \quad q = 1, \ldots, m,$$

$$U^i_{\boldsymbol{\alpha}} = \left(U^i_{\alpha_1}, \ldots, U^i_{\alpha_m}\right)'.$$

따라서 $\boldsymbol{\beta}$와 $\boldsymbol{\alpha}$에 대한 최대우도추정량은 다음 가중 스코어 방정식으로부터 얻을 수 있다.

$$U_{\boldsymbol{\beta}} = 0, \qquad U_{\boldsymbol{\alpha}} = 0.$$

한편 가중 관찰 정보행렬

$$I(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix}$$

의 각 블럭의 원소는 다음과 같다.

$$\begin{aligned} I_{11} &= -\frac{\partial^2 l_f}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \\ &= \sum_{i=1}^{n} \omega_i \left\{ \sum_{j\in D_i} \boldsymbol{x}_{ij} \boldsymbol{x}'_{ij} \left\{ \left(\frac{\sum_{q=1}^{m+1} I_{ijq}(f_{ijq-1} - f_{ijq})}{g_{ij}} \right)^2 - \frac{\sum_{q=1}^{m+1} I_{ijq}(h_{ijq-1} - h_{ijq})}{g_{ij}} \right\} \right. \\ &\quad \left. + \sum_{j\in R_i} \boldsymbol{x}_{ij} \boldsymbol{x}'_{ij} \sum_{q=1}^{m+1} I(L_{ij} \in [s_{q-1}, s_q)) a_{q-1} \exp(\boldsymbol{\beta}' \boldsymbol{x}_{ij}) \right\}, \\ I_{12} &= I'_{21} = -\frac{\partial^2 l_f}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\beta}}, \\ I_{22} &= -\frac{\partial^2 l_f}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}'}. \end{aligned}$$

단, $h_{ij0} = h_{ijm+1} = 0$이고, $q = 1, \ldots, m$에 대해, $h_{ijq} = f_{ijq}(1 + \log S(s_q|\boldsymbol{x}_{ij}))$,

$$\begin{aligned} (I_{12})_{\cdot q} &= -\frac{\partial^2 l_f}{\partial \alpha_q \partial \boldsymbol{\beta}} \\ &= -\sum_{i=1}^{n} \omega_i \left\{ \sum_{j\in D_i} \boldsymbol{x}_{ij} b_{ijq} \left\{ \frac{c_{ijq} + \sum_{l=q}^{m+1} (I_{ijl} - I_{ijl+1}) f_{ijl}}{g_{ij}} - \frac{c_{ijq}}{g^2_{ij}} \sum_{l=1}^{m+1} I_{ijl}(f_{ijl-1} - f_{ijl}) \right\} \right. \\ &\quad \left. - \sum_{j\in R_i} \boldsymbol{x}_{ij} b_{ijq} I(L_{ij} \in [s_q, \infty)) \right\}, \\ (I_{22})_{qq} &= -\frac{\partial^2 l_f}{\partial \alpha_q^2} \\ &= -\sum_{i=1}^{n} \omega_i \left\{ \sum_{j\in D_i} b_{ijq} c_{ijq} \left(\frac{1 - b_{ijq}}{g_{ij}} - \frac{b_{ijq} c_{ijq}}{g^2_{ij}} \right) - \sum_{j\in R_i} b_{ijq} I(L_{ij} \in [s_q, \infty)) \right\}, \\ (I_{22})_{qr} &= -\frac{\partial^2 l_f}{\partial \alpha_q \partial \alpha_r} = \sum_{i=1}^{n} \omega_i \sum_{j\in D_i} \left(\frac{b_{ijq} b_{ijr} c_{ijq} c_{ijr}}{g^2_{ij}} + \frac{b_{ijq} b_{ijr} c_{ijr}}{g_{ij}} \right) (q < r). \end{aligned}$$

따라서 $(\boldsymbol{\beta}', \boldsymbol{\alpha}')'$의 $s\,(=1,2,\ldots)$번째 해 $(\boldsymbol{\beta}^{(s)'}, \boldsymbol{\alpha}^{(s)'})'$는 Newton-Raphson 알고리즘을 써서 다음과 같이 얻을 수 있다.

$$\begin{pmatrix} \boldsymbol{\beta}^{(s)} \\ \boldsymbol{\alpha}^{(s)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}^{(s-1)} \\ \boldsymbol{\alpha}^{(s-1)} \end{pmatrix} + I^{-1} \begin{pmatrix} U_{\boldsymbol{\beta}} \\ U_{\boldsymbol{\alpha}} \end{pmatrix},$$

단, $(\boldsymbol{\beta}^{(0)'}, \boldsymbol{\alpha}^{(0)'})'$는 $(\boldsymbol{\beta}', \boldsymbol{\alpha}')'$의 초기값을 나타내고,

$$I^{-1} = \begin{pmatrix} I_{11|2}^{-1} & I_{12|2} \\ I_{21|1} & I_{22|1} \end{pmatrix},$$

$I_{11|2} = I_{11} - I_{12} I_{22}^{-1} I_{21}$, $I_{12|2} = I'_{21|2} = -I_{11|2}^{-1} I_{12} I_{22}^{-1}$, $I_{22|1} = I_{22}^{-1} + I_{22}^{-1} I_{21} I_{11|2}^{-1} I_{12} I_{22}^{-1}$. $(\boldsymbol{\beta}', \boldsymbol{\alpha}')'$의 최대우도추정량을 $(\hat{\boldsymbol{\beta}}'_{wee}, \hat{\boldsymbol{\alpha}}'_{wee})'$라고 하자. 한편 모형 (2.1)을 만족하면 Huang과 Wellner (1997)에 의해 적절한 조건 하에서 $(\hat{\boldsymbol{\beta}}'_{wee}, \hat{\boldsymbol{\alpha}}'_{wee})'$는 일치성과 점근정규성을 만족한다. 특히 $n \to \infty$에 따라

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{wee} - \boldsymbol{\beta}\right) \longrightarrow N\left(0, \Sigma^{-1}\right)$$

이고, 분산-공분산 행렬 Σ^{-1}는 샌드위치 추정량(sandwich estimator)에 의해 다음과 같이 추정된다.

$$\hat{\Sigma}_{wee}^{-1} = I_{11|2}\left(\hat{\boldsymbol{\beta}}_{wee}, \hat{\boldsymbol{\alpha}}_{wee}\right)^{-1} V\left(\hat{\boldsymbol{\beta}}_{wee}, \hat{\boldsymbol{\alpha}}_{wee}\right) I_{11|2}\left(\hat{\boldsymbol{\beta}}_{wee}, \hat{\boldsymbol{\alpha}}_{wee}\right)^{-1},$$

단, $V(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \omega_i^2 U_{\boldsymbol{\beta}}^i U_{\boldsymbol{\beta}}^{i\,\prime}$.

2.2. 군집 내 재추출 방법

n개의 군집에서 각각 개체 하나를 랜덤하게 복원추출하여 n개의 개체로 이루어진 재추출 자료집합을 얻는다. 다시 말해 i번째 군집 $\{((L_{ij}, R_{ij}], \boldsymbol{x}_{ij}) : \ j = 1, \ldots, n_i\}$에서 추출한 랜덤 표본을 $((L_i^b, R_i^b], \boldsymbol{x}_i^b)$라고 하면, b번째로 재추출된 자료 집합은 다음과 같다. 단, $b = 1, \ldots, B$.

$$\mathcal{D}^b = \left\{ \left(\left(L_i^b, R_i^b\right], \boldsymbol{x}_i^b \right) : \ i = 1, \ldots, n \right\}.$$

그런데 $\mathcal{D}^b$는 서로 독립인 개체로 이루어져 있기 때문에 2.1절에서 제안한 방법을 $\mathcal{D}^b$에 적용하여 모수를 추정할 수 있다. 다만 이 때 모든 i에 대해 $\omega_i = 1$로 놓는다. $\mathcal{D}^b$에 기초한 $(\boldsymbol{\beta}', \boldsymbol{\alpha}')'$의 최대우도추정량을 $(\hat{\boldsymbol{\beta}}_{wee}^{b'}, \hat{\boldsymbol{\alpha}}_{wee}^{b'})'$라고 하자. 이와 동일하게 B번 반복하여 $(\boldsymbol{\beta}', \boldsymbol{\alpha}')'$의 최대우도추정량 $\{(\hat{\boldsymbol{\beta}}_{wee}^{b'}, \hat{\boldsymbol{\alpha}}_{wee}^{b'})' : \ b = 1, \ldots, B\}$을 얻는다. 한편 $\boldsymbol{\beta}$와 $\boldsymbol{\alpha}$에 대한 군집 내 재추출 추정량을 각각 다음과 같이 정의한다.

$$\hat{\boldsymbol{\beta}}_{wcr} = \frac{1}{B}\sum_{b=1}^{B} \hat{\boldsymbol{\beta}}_{wee}^{b}, \qquad \hat{\boldsymbol{\alpha}}_{wcr} = \frac{1}{B}\sum_{b=1}^{B} \hat{\boldsymbol{\alpha}}_{wee}^{b}.$$

적절한 조건 하에서 $n \to \infty$에 따라 $\hat{\boldsymbol{\beta}}_{wee}^{b}$은 정규분포로 수렴한다 (Huang과 Wellner, 1997). 또한 Hoffman 등 (2001)에 의해 적절한 조건 하에서 $n \to \infty$에 따라 $\hat{\boldsymbol{\beta}}_{wcr}$은

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{wcr} - \boldsymbol{\beta}\right) \longrightarrow N(0, \Sigma)$$

을 만족하고, Σ에 대한 일치 추정량 $\hat{\Sigma}_{wcr}$은 다음과 같이 주어진다.

$$\hat{\Sigma}_{wcr} = \frac{1}{B}\sum_{b=1}^{B} I_{11|2}\left(\hat{\boldsymbol{\beta}}_{wee}^{b}, \hat{\boldsymbol{\alpha}}_{wee}^{b}\right)^{-1} - \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\boldsymbol{\beta}}_{wee}^{b} - \hat{\boldsymbol{\beta}}_{wcr}\right)\left(\hat{\boldsymbol{\beta}}_{wee}^{b} - \hat{\boldsymbol{\beta}}_{wcr}\right)'.$$

3. 모의실험

본 절에서는 모의실험을 통해 2.1절과 2.2절에서 제안한 추정 방법(WEE, WCR)과 군집의 크기가 생존시간에 종속적일 때 이를 무시한 추정 방법(UWE)을 서로 비교하고자 한다. 단, UWE는 모든 군집에 대해 $w_i = 1$을 가정하고 추정한 WEE 추정량을 의미한다.

생존시간 T_{ij}는 아래 모형으로부터 생성하였다.

$$\lambda(t|x_{ij}, w_i) = \lambda_0(t) w_i \exp(\beta_1 x_{i1} + \beta_2 x_{i2}). \tag{3.1}$$

프레일티 w_i는 군집 내에 있는 개체 간의 상관 정도를 나타내는 모수 $\phi \in (0, 1)$를 가진 양의 안정분포(positive stable distribution)에서 생성하였으며 (Chambers 등, 1976), 군집 고유 특성(cluster-specific) 공변량 x_{i1}과 개체 고유 특성(subject-specific) 공변량 x_{i2}는 각각 성공율이 0.5인 이항분포 $B(1, 0.5)$와 균등분포 $U(-0.5, 0.5)$에서 생성하였다. 이 때, $\beta_1 = \beta_2 = 1, \lambda_0(t) = 0.3$으로 놓았다. 한편 (3.1)를 양의 안정분포에 대해 적분하면 Huster 등 (1989), Wei 등 (1989), Lee 등 (1992)이 제안한 주변모형을 얻을 수 있다. 이 주변모형의 누적위험함수와 회귀계수는 각각 $\Lambda_0^*(t) = t^\phi \Lambda_0(t)$, $\beta_1^* = \phi\beta_1$, $\beta_2^* = \phi\beta_2$로 주어진다. 군집의 크기 n_i는 생존시간이 군집의 크기에 의존하는 경우(informative case)와 의존하지 않는 경우(non-informative case)에 따라 다르게 생성하였다. 'informative case'의 경우에는 w_i가 생존시간의 표본 중앙값보다 작으면 이항분포 $B(7, 0.75)$에서, 그렇지 않으면 이항분포 $B(7, 0, 25)$에서 생성하였고, 'non-informative case'의 경우에는 균등분포 $U(0, 1)$에서 발생된 난수값이 0.5보다 작으면 이항분포 $B(7, 0.75)$에서, 그렇지 않으면 이항분포 $B(7, 0, 25)$에서 생성하였으며, 이 때 0 또는 7은 제외하였다. 따라서 n_i가 가질 수 있는 값은 $1, \ldots, 6$이다. 관찰시점은 0.1부터 6.6까지 0.5 간격으로 14개를 잡았다. 즉, $0.1, 0.6, 1.1, \ldots, 6.6$. 이 중에서 처음 7개의 관찰시점에 대해서는 방문하지 않을 확률을 0.3으로 하였고, 나머지 관찰시점에 대해서는 방문하지 않을 확률을 0.5로 하였다. 방문한 시점이 결정되면 구간중도절단된 시점은 T_{ij}가 속하는 인접한 두 시점으로 잡았다. 이 때 $T_{ij} \leq 0.1$이면 0.1시점에서 좌중도절단된 것으로 간주하였고, $T_{ij} \geq 6.6$이면 6.6시점에서 우중도절단된 것으로 간주하였다. 가중값은 $w_i = 1/n_i$로 잡았으며, 군집의 개수는 $n = 150, 300$으로 잡았다. 500개의 데이터 셋을 생성하였으며 재추출은 $B = 1{,}000$번 하였다.

Table 3.1과 Table 3.2는 각각 회귀계수 β_1과 β_2에 대한 모의실험결과이다. $\phi = 0.2$, 0.5, 0.8에 대응하는 β_1, β_2의 참값은 각각 0.2, 0.5, 0.8이다. 표에서 'Bias'는 추정량의 추정값에서 참값을 뺀 값의 평균, 'SD'는 추정량의 표준편차, 'SEM'은 추정량의 표준오차의 평균, 'CP'는 참값에 대한 95% 포함률을 각각 나타낸다. 먼저 Table 3.1을 살펴보면 예상했던 대로 'non-informative case'의 경우에는 세 가지 추정 방법이 별다른 차이를 보이지 않지만 'informative case'의 경우에는 제안한 추정 방법인 'WEE'와 'WCR'이 'UWE'보다 우수하였다. 다시 말해 n에 관계 없이 'Bias'가 눈에 띄게 작았을 뿐만 아니라 'CP'도 명목값인 0.95에 훨씬 더 가까웠다. 그러나 제안한 두 추정량은 'Bias'나 'CP'로 볼 때 매우 유사한 경향을 보였다. ϕ의 값이 1에 가까울수록 즉, 군집 내 개체들이 서로 독립에 가까울수록 추정량의 산포가 줄어드는 경향을 보였는데 이는 주변모형의 가정으로부터 모수를 추정했기 때문이다. 한편 β_2의 추정에 대한 결과인 Table 3.2는 Table 3.1과 유사한 경향을 보였다.

4. 실제 예

림프사상충은 모기로 인해 전염된 기생충병이다. 기생충의 전염을 억제하기 위해 흔히 DEC, 6mg/kg과 ALB, 400mg을 함께 사용하여 치료한다. LF 연구에서는 DEC와 ALB를 병행하는 방법('DEC + ALB')과 DEC만 사용하는 방법('DEC only')이 성충집(adult worm nest)을 박멸하는 데 미치는 효과

Table 3.1. Simulation results for β_1 with true values of 0.2, 0.5, and 0.8 corresponding to ϕ =0.2, 0.5, and 0.8, respectively

True	WEE				UWE				WCR			
	Bias	SD	SEM	CP	Bias	SD	SEM	CP	Bias	SD	SEM	CP
					$n = 150$, non-informative case							
0.2	−1.92E-02	0.190	0.189	0.948	−1.18E-02	0.210	0.209	0.948	−1.83E-02	0.191	0.189	0.946
0.5	1.55E-02	0.158	0.160	0.948	1.21E-02	0.168	0.168	0.934	2.01E-02	0.160	0.160	0.946
0.8	2.96E-03	0.137	0.134	0.942	1.93E-03	0.131	0.128	0.932	1.38E-02	0.140	0.133	0.940
					$n = 300$, non-informative case							
0.2	−5.61E-03	0.135	0.133	0.942	−6.87E-03	0.153	0.147	0.930	−5.10E-03	0.136	0.133	0.940
0.5	−5.89E-03	0.119	0.112	0.930	−5.67E-03	0.124	0.118	0.948	−3.80E-03	0.120	0.112	0.930
0.8	5.23E-03	0.097	0.094	0.956	3.12E-03	0.095	0.090	0.936	1.04E-02	0.098	0.095	0.952
					$n = 150$, informative case							
0.2	−6.87.E-03	0.184	0.187	0.952	7.10.E-02	0.213	0.214	0.944	−5.76.E-03	0.185	0.187	0.952
0.5	1.55.E-02	0.153	0.152	0.938	1.21.E-01	0.161	0.157	0.848	2.12.E-02	0.155	0.151	0.932
0.8	8.05.E-03	0.137	0.127	0.924	6.51.E-02	0.125	0.118	0.906	2.25.E-02	0.140	0.125	0.908
					$n = 300$, informative case							
0.2	4.79.E-03	0.132	0.131	0.944	8.01.E-02	0.158	0.151	0.898	5.42.E-03	0.132	0.131	0.944
0.5	−1.32.E-03	0.107	0.107	0.948	9.98.E-02	0.108	0.111	0.866	1.42.E-03	0.108	0.107	0.940
0.8	4.06.E-03	0.086	0.087	0.950	6.06.E-02	0.077	0.082	0.904	1.08.E-02	0.087	0.087	0.950

Table 3.2. Simulation results for β_2 with true values of 0.2, 0.5, and 0.8 corresponding to $\phi = 0.2$, 0.5, and 0.8, respectively

True	WEE				UWE				WCR			
	Bias	SD	SEM	CP	Bias	SD	SEM	CP	Bias	SD	SEM	CP
					$n = 150$, non-informative case							
0.2	7.50E-03	0.228	0.221	0.936	7.18E-03	0.185	0.184	0.948	8.68E-03	0.232	0.223	0.930
0.5	−1.88E-02	0.214	0.203	0.944	−3.38E-03	0.178	0.170	0.942	−1.40E-02	0.219	0.207	0.926
0.8	8.43E-03	0.220	0.198	0.926	5.15E-03	0.182	0.167	0.908	2.00E-02	0.225	0.202	0.904
					$n = 300$, non-informative case							
0.2	−4.06E-04	0.159	0.156	0.948	1.42E-03	0.130	0.129	0.956	4.56E-04	0.160	0.161	0.950
0.5	8.89E-05	0.149	0.146	0.944	1.37E-03	0.120	0.121	0.946	2.56E-03	0.150	0.152	0.954
0.8	2.61E-03	0.146	0.141	0.938	4.55E-03	0.122	0.118	0.926	8.21E-03	0.148	0.148	0.930
					$n = 150$, informative case							
0.2	−1.63.E-02	0.199	0.200	0.952	6.89.E-02	0.196	0.199	0.940	−1.45.E-02	0.202	0.203	0.946
0.5	−7.11.E-04	0.195	0.184	0.940	9.43.E-02	0.189	0.178	0.912	6.28.E-03	0.199	0.183	0.926
0.8	1.34.E-02	0.187	0.188	0.942	7.21.E-02	0.174	0.173	0.930	2.91.E-02	0.193	0.186	0.918
					$n = 300$, informative case							
0.2	−4.82.E-03	0.137	0.142	0.970	7.71.E-02	0.139	0.141	0.910	−4.59.E-03	0.138	0.147	0.970
0.5	4.12.E-03	0.128	0.130	0.952	1.00.E-01	0.124	0.127	0.868	7.45.E-03	0.130	0.136	0.954
0.8	−5.57.E-03	0.129	0.128	0.952	5.14.E-02	0.120	0.119	0.916	1.80.E-03	0.130	0.135	0.950

를 서로 비교하고자 한다 (Williamson 등, 2008; Zhang과 Sun, 2010). LF를 가진 남자 47명 중에서 랜덤하게 뽑은 22명에게는 전자의 치료법을 적용하고 나머지 25명에게는 후자의 치료법을 적용한 후 주기적으로 남자의 음낭 부근을 초음파로 검사하여 성충집이 완전히 제거될 때까지의 시간(단위: 일)을 측

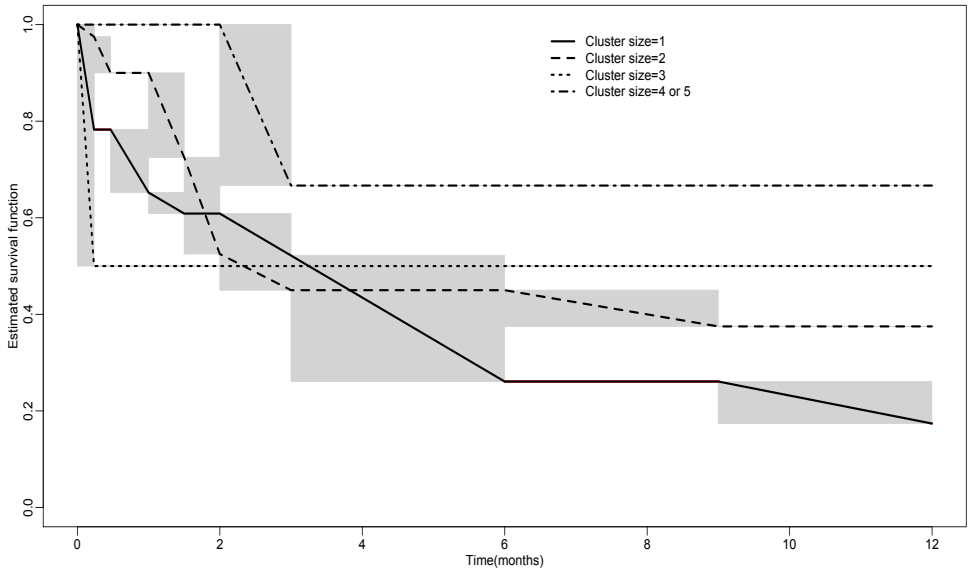

Figure 4.1. Plot of estimated baseline survival functions by the cluster size

Table 4.1. Parameter estimates(est), their standard errors(SE), 95% confidence intervals(95%CI), and p-values for their significance tests(P)

Covariate	WEE				UWE				WCR				
	est	SE	95%CI	P	est	SE	95%CI	P	est	SE	95%CI		P
Group	0.591	0.372	−0.138 1.320	0.112	0.861	0.445	−0.011 1.732	0.053	0.591	0.373	−0.140	1.321	0.113
									(0.594)	(0.372)	(−0.136)	(1.324)	(0.111)
Age	0.005	0.020	−0.035 0.044	0.815	0.003	0.024	−0.045 0.050	0.913	0.005	0.049	−0.091	0.100	0.923
									(0.005)	(0.037)	(−0.068)	(0.077)	(0.900)

정하였다. 초음파 검사는 7, 14, 30, 45, 60, 90, 180, 270, 360일에 있었다. LF 연구에서 남자는 군집에 해당되고($n = 47$), 성충집은 군집 내 개체에 해당되는데 성충집의 개수는 사람에 따라 1개에서 5개까지 가지고 있었고 전체 성충집의 개수는 78개 이었다. 연구기간 동안에 성충집이 박멸된 비율을 보면, 성충집이 1개인 군집들은 81.8%, 2개인 군집들은 62.5%, 3개인 군집들은 50.0%, 4개 혹은 5개인 군집들은 33.3%가 박멸되었다. 따라서 성충집이 많을수록 치료효과가 떨어지는 경향을 보였다. 다시 말해 성충집의 개수와 성충집이 완전히 박멸될 때까지 걸리는 시간이 서로 독립이기보다 종속되는 경향을 보였다. Figure 4.1은 성충집의 개수에 따라 생존함수를 추정한 것인데 대략 180일까지는 뚜렸한 경향이 없었지만 그 이후부터는 성충집을 많이 가지고 있는 사람일수록 성충집이 완전히 박멸될 때까지 시간이 많이 소요되었다. 실재로 연구의 종료 시점에서 생존함수를 비교해보면 예상했던 대로 성충집의 개수가 적을수록 박멸된 비율이 높게 나타났다. Figure 4.1에서 사각형 상자는 동등집합에 해당하는 구간의 양끝 값에서는 추정 생존함수 값이 유일하게 정의되지만 구간 내에서는 그렇지 않다는 것을 나타낸다.

Zhang과 Sun (2010)은 기저위험함수를 와이블 분포로 가정하고 주변모형의 모수를 추정하였는데 Figure 4.2에서 볼 수 있듯이 치료방법에 따라 구분하지 않은 경우(왼쪽 그림)과 구분한 경우(오른쪽 그림)에서 모두 기저생존함수(baseline survival fumction)가 와이블 분포에서 많이 벗어난 것처럼 보인다. 따라서 본 논문에서는 기저생존함수에 대해 특정 분포를 가정하지 않았으며 공변량으로는 Williamson 등 (2008)과 Zhang과 Sun (2010)처럼 군집과 상관된 공변량인 치료방법('Group')과 나이('Age')을 고

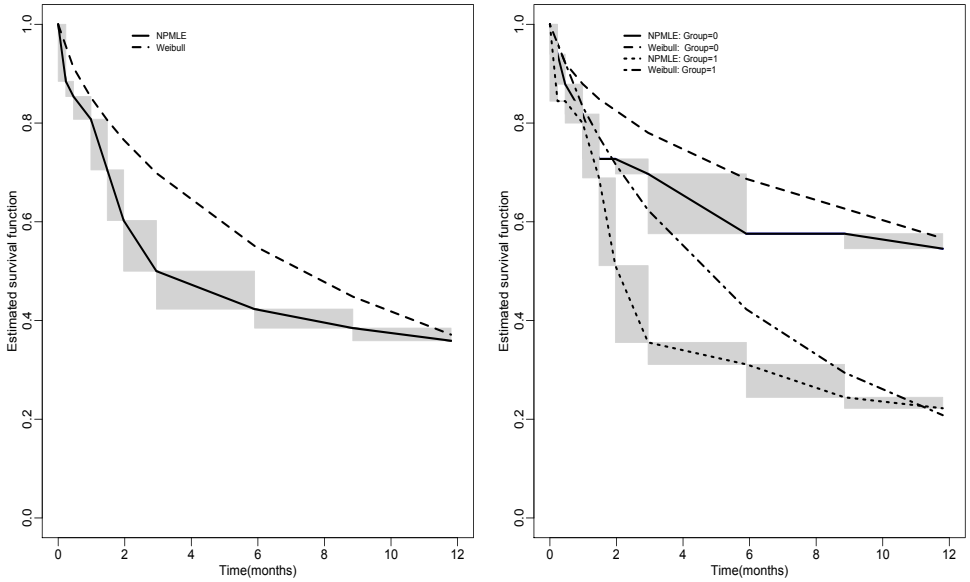

Figure 4.2. Plots of estimated baseline survival functions by assumed model(left) and by assumed model and treatment(right)

Table 4.2. Estimated hazard rate and(est.h) estimated cumulative hazard rates(est.ch) at specified time points, 7, 14, 30, 45, 60, 90, 180, and 360.

Days	est.h				est.ch			
	WEE	UWE	WCR		WEE	UWE	WCR	
			B = 1,000	B = 2,000			B = 1,000	B = 2,000
7	0.105	0.070	0.105	0.105	0.105	0.070	0.105	0.105
14	0.020	0.020	0.020	0.020	0.125	0.090	0.125	0.125
30	0.063	0.031	0.063	0.063	0.188	0.121	0.188	0.188
45	0.094	0.077	0.094	0.094	0.282	0.198	0.281	0.282
60	0.098	0.092	0.097	0.098	0.380	0.290	0.378	0.379
90	0.120	0.116	0.120	0.120	0.500	0.405	0.498	0.499
180	0.222	0.107	0.222	0.223	0.722	0.512	0.721	0.722
270	0.068	0.062	0.070	0.068	0.790	0.574	0.790	0.790
360	0.104	0.046	0.105	0.104	0.893	0.620	0.895	0.894

려하였다. 단, 'DEC only'이면 Group = 1이고 'DEC + ALB'이면 Group = 0. Table 4.1는 LF 자료에 세 가지 추정 방법을 적용하여 얻은 결과이며, 특히 'WCR' 추정량은 1,000번(2,000번) 재추출하여 얻은 결과이다. 'Group(치료방법) 효과를 보면 'UWE'는 다소 유의한 것 같지만(P값 = 0.053), 본 논문에서 제안한 두 방법은 P값이 각각 0.112('WEE'), 0.111(0.111)('WCR')로 유의수준 0.05에서 유의하지 않다. 'Group'의 추정량의 값이 양수이기 때문에 표준 치료방법(DEC + ALB)보다 오히려 새로운 치료방법(DEC only)이 오히려 성충집이 완전히 박별될 때까지의 시간을 단축시키는 효과가 있지만 두 치료방법에 따라 성충집이 완전히 박별될 때까지 걸리는 시간이 다르지 않다고 할 수 있다. 'AGE(나이) 효과를 보면 세 가지 방법의 P값이 각각 0.815('WEE'), 0.913('UWE'), 0.923(0.900)('WCR')로 유의수준 0.05에서 모두 유의하지 않았다. 한편 기저위험률과 누적기저위험률의 추정량은 Table 4.2와 같으며 예상했던 대로 'WEE' 방법과 'WCR' 방법은 서로 유사하며 'UWE' 방법은 'WEE' 방법과

'WCR' 방법보다 과소 추정하는 경향이 있었다. 'WCR' 방법에서 'Group' 효과와 기저위험률은 재추출 회수에 따라 차이가 없었디. 그러나 'Age' 효과는 표준오차의 추정량이 재추출 횟수에 따라 약간 차이가 있어 95% 신뢰구간과 'Age' 효과의 유의성(P값)이 서로 달랐지만 그 차이는 크지 않았다.

5. 맺음말

본 논문에서는 군집 구간중도절단된 자료에서 생존시간이 군집의 크기에 의존할 때 주변모형으로부터 가중 추정 방법(WEE)과 군집 내 재추출 방법(WCR)을 제안하고 그 추정량의 점근적 성질을 살펴보았다. 또한 모의실험을 통해, 생존시간이 군집의 크기와 무관한 경우(non-informative case)에는 세 추정 방법이 서로 유사하였지만(WEE $\approx$ UWE $\approx$ WCR), 생존시간이 군집의 크기에 종속된 경우(informative case)에는 'Bias'와 'CP' 측면에서 볼 때 제안한 두 추정 방법이 이 종속 관계를 무시한 방법보다 우수한 것으로 나타났다(WEE $\approx$ WCR $>$ UWE). LF 자료는 성충집의 개수에 따라 기저생존함수가 다르고 성충집의 개수가 많을수록 성충집이 완전히 박멸될 때까지 걸리는 시간이 길어지는 것으로 나타났다. 한편 두 치료방법(DEC + ALB, DEC only)이 서로 유의하게 다르지 않았으며 나이 효과는 매우 유의하지 않은 것으로 나타났다. LF 자료를 분석한 기존 연구 결과와 비교해보면 본 논문에서 제안한 두 추정 방법의 결과는 Zhang과 Sun (2010)의 결과보다 Williamson 등 (2008)의 결과와 유사하다고 할 수 있다. 실제로 Williamson 등 (2008)의 결과에 따르면 $\hat{\beta}_1 = 0.585$(P값 $= 0.089$), $\hat{\beta}_2 = 0.005$(P값 $= 0.800$)으로 치료 방법의 효과와 나이 효과가 모두 유의하지 않았다. 반면에, Zhang과 Sun(2010)의 결과에 따르면 $\hat{\beta}_1 = -0.267$(P값 $= 0.277$), $\hat{\beta}_2 = -0.088$(P값 < 0.001)으로 치료 방법의 효과는 유의하지 않지만 나이 효과는 매우 유의한 것으로 나왔으며, 더불어 회귀계수의 부호가 Williamson 등 (2008)과 제안한 두 추정 방법의 부호와 서로 정반대로 나왔다. 그 이유가 명확하지는 않지만 Zhang과 Sun (2010)은 기저생존함수를 와이블분포로 가정했는데 Figure 4.2에서 살펴본 것처럼 그 가정이 타당하지 않기 때문이라고 생각된다.

References

Cai, J. and Prentice, R. L. (1995). Estimating equations for hazard ratio parameters based on correlated failure time data, *Biometrics*, **82**, 151–164.

Chambers, J. M., Mallows, C. L. and Stuck, B. W. (1976). A method for simulating stable random variables, *Journal of the American Statistical Association*, **71**, 340–344.

Cong, X. J., Yin, G. and Shen, Y. (2007). Marginal analysis of correlated failure time data with informative cluster sizes, *Biometrics*, **63**, 663–672.

Hoffman, E. B., Sen, P. K. and Weinberg, C. R. (2001). Within cluster resampling, *Biometrika*, **88**, 1121–1134.

Huster, W. J., Brookmeyer, R. and Self, S. G. (1989). Modelling paired survival data with covariates, *Biometrics*, **45**, 145–156.

Huang, J. and Wellner, J. A. (1997). Interval censored survival data: a review of recent progress, *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*,. D. Lin and T. Fleming(eds). Springer-Verlag, New Jersey.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, Second Ed., John Wiley, New York.

Lee, E. W., Wei, L. J. and Amato, D. A. (1992). Cox-type regression analysis for large number of small groups of correlated failure time observations, In *Survival Analysis: State of Arts*, J.P. Klein and P.M. Goel(eds), Kluwer Academic Publishers, Dordrecht.

Lindsey, J. C. and Ryan, L. M. (1998). Tutorial in biostatistics: Methods for interval-censored data, *Statistics in Medicine*, **17**, 219–238.

McGuire, M. K. and Nunn, M. E. (1996). Prognosis versus actual outcome III: The effectiveness of clinical parameters in accurately predicting tooth survival, *Journal of Periodontology*, **67**, 666–674.

Wei, L. J., Lin, D. Y. and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions, *Journal of the American Statistical Association*, **84**, 1065–1073.

Williamson, J., Kim, H. Y., Manathuga, A. and Addiss, D. G. (2008). Modeling survival data with informative cluster size, *Statistics in Medicine*, **27**, 543–555.

Zhang, X. and Sun, J. (2010). Regression analysis of clustered interval-censored failure time data with informative cluster size, *Computational Statistics and Data Analysis*, **54**, 1817–1823.

군집의 크기가 생존시간에 영향을 미치는 군집 구간중도절단된 자료에 대한 준모수적 모형

김진흠[a,1] · 김윤남[b]

[a]수원대학교 통계정보학과, [b]세브란스병원 임상시험센터

(2014년 1월 3일 접수, 2014년 2월 26일 수정, 2014년 3월 14일 채택)

요 약

본 논문에서는 군집 구간중도절단된 자료에서 생존시간이 군집의 크기에 의존할 때 주변모형으로부터 가중 추정 방법과 군집 내 재추출 방법을 써서 모수를 추정하고 그 추정량의 점근적 성질을 살펴보았다. 모의실험을 통해 추정량의 편향의 크기와 신뢰구간의 포함율 측면에서 볼 때 제안한 두 추정 방법이 생존시간과 군집의 크기 간의 종속 관계를 무시한 방법보다 우수한 것으로 나타났다. 제안한 추정 방법을 림프성 사상충 자료에 적용한 결과에 따르면 서로 다른 두 치료방법이 유의하게 다르지 않았으며 나이 효과도 매우 유의하지 않은 것으로 나타났다.

주요용어: 군집의 크기, 구간중도절단, 주변모형, 가중 추정 방정식, 군집 내 재추출.

이 논문은 2011년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.2011-0010889).

[1]교신저자: (445-743) 경기도 화성시 봉담읍 와우안길 17, 수원대학교 통계정보학과.
E-mail: jkimdt65@gmail.com