

Analyzing Clustered and Interval-Censored Data based on the Semiparametric Frailty Model

Jinheum Kim¹ · Youn Nam Kim²

¹Department of Applied Statistics, University of Suwon

²Clinical Trials Center Severance Hospital, Yonsei University Health System

(Received August 1, 2012; Revised August 25, 2012; Accepted September 25, 2012)

Abstract

We propose a semi-parametric model to analyze clustered and interval-censored data; in addition, we plugged-in a gamma frailty to the model to measure the association of members within the same cluster. We propose an estimation procedure based on EM algorithm. Simulation results showed that our estimation procedure may result in unbiased estimates. The standard error is smaller than expected and provides conservative results to estimate the coverage rate; however, this trend gradually disappeared as the number of members in the same cluster increased. In addition, our proposed method was illustrated with data taken from diabetic retinopathy studies to evaluate the effectiveness of laser photocoagulation in delaying or preventing the onset of blindness in individuals with diabetic retinopathy.

Keywords: Cox proportional hazards model, diabetic retinopathy studies, EM algorithm, frailty, interval censoring.

1. 서론

상관된 혹은 군집화된 생존자료(correlated or clustered failure time data)는 의학 및 동물실험연구 등 많은 분야에서 자주 접하게 된다 (Cai와 Prentice, 1995; Kalbfleish와 Prentice, 2002). 때때로 한 개체의 생존시간이 정확하게 관측되기 보다는 그 값이 속하는 구간에 대한 정보만 알려지곤 하는데 이런 형태의 자료를 소위 구간중도절단된 자료(interval-censored data)라고 일컫는다. 이런 자료는 환자들을 연속적으로 모니터링 하는 것이 아니라 주기적으로 관측하는 연구에서 발생하게 된다 (Sun, 2006). 본 논문에서는 군집구간중도절단된 생존자료(clustered and interval-censored failure time data)의 회귀모형과 관련된 통계적 추론을 하고자 한다. 군집구간중도절단된 자료는 군집우중도절단된 자료에 비해 드문 편이지만 Goggins와 Finkelstein (2000)의 AIDS 연구자료와 Goethals 등 (2009)과 Ampe 등 (2012)에 의해 연구가 진행되고 있는 젖소 유선염 연구 자료를 예로 들 수 있다. 전자의 예에서는 환자가 군집이 되고 그 환자의 혈액 및 소변검사는 군집 내 개체가 된다. 후자의 예에서는 군집은 젖소로 정의하고 군집 내 개체는 네 개의 유두로 정의한다.

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.20110010889).

¹Corresponding author: Professor, Department of Applied Statistics, University of Suwon, 17 Wauan-Gil, Bongdam-Eup, Hwaseong 445-743, Korea. E-mail: jkimdt65@gmail.com

군집의 크기가 비정보적(non-informative)일 때(즉, 군집의 크기와 생존시간이 서로 독립일 때), Bellamy 등 (2004)과 Goethals 등 (2009), Ampe 등 (2012)은 군집구간중도절단된 자료에 대한 모형으로 콕스 비례위험 프레일티 모형(Cox proportional hazards frailty model)을 제안하였다. 그들은 전개를 용이하게 하기 위해 이 모형에 포함된 기저위험함수를 비모수적으로 다루지 않고 모수적으로 다루었다. 그들이 적용한 모형은 와이블 분포이며 이와 같이 가정하면 기저위험함수와 관련된 모수의 개수는 무한개에서 유한개로(와이블 분포는 2개임) 줄어든다. 그러나 이 가정을 만족하지 못하면 모형 오설계(model misspecification)로 인해 모수의 추정량이 편향될 수 있다. 위 세 연구의 차이점은 프레일티 분포를 어떤 분포로 가정했느냐에 있다. Bellamy 등 (2004)은 프레일티 분포로 정규분포를 가정한 반면 Goethals 등 (2009)과 Ampe 등 (2012)은 감마분포를 가정하였다. 사실 그들이 가정한 정규분포와 감마분포는 모두 프레일티 분포로 널리 쓰이고 있다. 이에 본 논문에서는 Bellamy 등 (2004)과 Goethals 등 (2009), Ampe 등 (2012)처럼 기저위험함수를 모수적인 모형으로 가정하지 않고 비모수적으로 다루고자 한다. Duchateau와 Janssen (2008)은 군집우중도절단된 자료(clustered and right-censored data)에 콕스 비례위험 프레일티 모형을 가정하고 모수를 추정하였으며 본 논문에서는 Duchateau와 Janssen (2008)의 방법을 군집구간중도절단된 자료로 확장하고자 한다.

2절에서는 콕스 비례위험 프레일티 모형을 제안하고 EM 알고리즘을 써서 모수를 추정하고자 한다. 3절에서는 모의실험을 통해 제안한 추정량의 소표본 성질을 살펴보고자 하며, 4절에서는 제안한 방법을 당뇨병성 망막병증 연구(diabetic retinopathy studies; DRS) 자료에 적용하여 레이저 치료 여부와 질병 발생 연령이 실명에 미치는 영향을 분석하고자 한다. 마지막으로 5절에서는 본 논문의 한계점과 향후 연구에 대해 토의하고자 한다.

2. 모형 설계와 모수 추론

T_{ij} 는 i 번째 군집에 속하는 j 번째 개체의 생존시간이라고 하자. 단, $i = 1, \dots, n$; $j = 1, \dots, n_i$. 그런데 이 생존시간은 직접 관측할 수 없고 단지 관심 있는 이벤트가 발생한 구간 즉, $(L_{ij}, U_{ij}]$ 만 관측할 수 있다고 하자. 이런 자료를 소위 구간중도절단 자료라고 하는데, 만일 $L_{ij} = 0$ 이면 좌중도절단된 자료(right-censored data)를 나타내고, $U_{ij} = \infty$ 이면 우중도절단된 자료(left-censored data)를 나타낸다. 따라서 구간중도절단은 좌중도절단과 우중도절단을 포함하는 중도절단이라고 할 수 있다. 동일한 군집 내에 있는 개체들의 생존시간은 서로 연관되어 있다. 이런 자료를 다루기 위해 대표적으로 주변 모형(marginal models), 프레일티 모형(frailty models), 랜덤효과 모형(random effects models), 동적 모형(dynamic models) 등이 쓰이고 있으며 (Aalen 등, 2008), 본 논문에서는 프레일티 모형을 도입하고자 한다.

u_i 는 군집 간 서로 이질적인 정도를 나타내는 프레일티라고 하자 (단, $i = 1, \dots, n$). 이 프레일티는 관측 가능한 값은 아니며 일반적으로 군집 내 개체들은 프레일티가 같다고 가정한다. 사실 군집 내 개체들은 동일한 환경에 놓여 있거나 유전적인 성향이 유사하기 때문에 이런 가정은 타당하다. 이 프레일티가 주어졌을 때, 동일한 군집 내에 있는 개체들의 생존시간은 서로 독립이라는 가정 아래 생존시간에 대한 콕스 비례위험 프레일티 모형을 다음과 같이 정의하자.

$$\lambda(t|\mathbf{x}_{ij}, u_i) = \lambda_0(t)u_i \exp(\boldsymbol{\beta}'\mathbf{x}_{ij}), \quad (2.1)$$

단, $\mathbf{x}_{ij} = (x_{ij1}, \dots, x_{ijp})'$ 는 T_{ij} 와 연관된 공변량 벡터이고, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ 는 회귀계수벡터이며, $\lambda_0(t)$ 는 임의의 기저위험함수이다. Bellamy 등 (2004)과 Goethals 등 (2009)은 $\lambda_0(t)$ 가 와이블 분포를 따른다고 가정하고 모수적인 방법으로 추론하였다. 비록 와이블 분포가 위험함수로 널리 사용되고 있지만 본 논문에서는 $\lambda_0(t)$ 를 특정한 분포로 국한하지 않고 비모수적인 방법으로 추론하고자 한다. 다시 말

해 Bellamy 등 (2004)과 Goethals 등 (2009)은 모수적인 모형을 가정한 반면 본 논문에서는 준모수적인 모형을 가정하고자 한다.

중도절단과 생존시간이 서로 독립이라고 가정하자. i 번째 군집의 개체 중에서 구간중도절단된 개체들과 우중도절단된 개체들로 이루어진 집합을 각각

$$D_i = \{j|T_{ij} \in (L_{ij}, U_{ij}]\}, \quad R_i = \{j|T_{ij} \in (L_{ij}, \infty)\}$$

라고 하자. \mathbf{x}_{ij} 와 u_i 이 주어졌을 때 T_{ij} 의 생존함수를 $S(t|\mathbf{x}_{ij}, u_i)$ 라고 하자. 따라서 \mathbf{x}_{ij} 와 u_i 이 주어졌을 때, i 번째 군집의 조건부우도함수(conditional likelihood function)는 다음과 같다.

$$L_i(\lambda_0, \boldsymbol{\beta}|\mathbf{x}_{ij}, u_i) = \prod_{j \in R_i} S(L_{ij}|\mathbf{x}_{ij}, u_i) \prod_{j \in D_i} \{S(L_{ij}|\mathbf{x}_{ij}, u_i) - S(U_{ij}|\mathbf{x}_{ij}, u_i)\}. \quad (2.2)$$

모형 (2.1)의 가정 아래 식 (2.2)는 다음과 같이 표현된다.

$$L_i(\lambda_0, \boldsymbol{\beta}|\mathbf{x}_{ij}, u_i) = \prod_{j \in R_i} \exp(-u_i \tilde{L}_{ij}) \prod_{j \in D_i} \left\{ \exp(-u_i \tilde{L}_{ij}) - \exp(-u_i \tilde{U}_{ij}) \right\}, \quad (2.3)$$

단, $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$, $\tilde{L}_{ij} = \Lambda_0(L_{ij}) \exp(\boldsymbol{\beta}'\mathbf{x}_{ij})$, $\tilde{U}_{ij} = \Lambda_0(U_{ij}) \exp(\boldsymbol{\beta}'\mathbf{x}_{ij})$. Kronecker 곱을 이용하면 식 (2.3)의 곱을 간편하게 표현할 수 있다. 길이가 2^{d_i} 인 벡터 \mathbf{a}_i 의 k 번째 원소 a_{ik} 를 다음 Kronecker 곱의 k 번째 원소로 정의하자.

$$\mathbf{a}_i = \left(\exp(-u_i \tilde{L}_{i1}), -\exp(-u_i \tilde{U}_{i1}) \right)' \otimes \cdots \otimes \left(\exp(-u_i \tilde{L}_{id_i}), -\exp(-u_i \tilde{U}_{id_i}) \right)'.$$

따라서 식 (2.3)은 다음과 같이 표현된다.

$$L_i(\lambda_0, \boldsymbol{\beta}|\mathbf{x}_{ij}, u_i) = \exp(-u_i C_i) \sum_{k=1}^{2^{d_i}} a_{ik}, \quad (2.4)$$

단, $C_i = \sum_{j \in R_i} \tilde{L}_{ij}$. 또한 완전자료(즉, u_i 가 관측 가능하다고 가정)에 기초한 i 번째 군집의 완전우도함수(full likelihood function)는 다음과 같이 주어진다.

$$L_{f,i}(\lambda_0, \boldsymbol{\beta}, \theta) = L_i(\lambda_0, \boldsymbol{\beta}|\mathbf{x}_{ij}, u_i)g_\theta(u_i),$$

단, $g_\theta(\cdot)$ 는 프레일티의 확률밀도함수이다. 본 논문에서는 프레일티 분포로 널리 쓰이는 감마분포를 가정하고자 한다. 먼저 감마분포 $G(1/\theta, \theta)(\theta > 0)$ 의 확률밀도함수는 다음과 같다.

$$g_\theta(u) = \frac{u^{\frac{1}{\theta}-1}}{\Gamma(1/\theta)\theta^{\frac{1}{\theta}}} \exp\left(-\frac{u}{\theta}\right).$$

따라서 식 (2.4)는 다음과 같다.

$$L_{f,i}(\lambda_0, \boldsymbol{\beta}, \theta) = \sum_{k=1}^{2^{d_i}} d_{ik} \frac{u_i^{\frac{1}{\theta}-1} \exp\{-u_i(1/\theta + C_i + \log b_{ik})\}}{\Gamma(1/\theta)\theta^{\frac{1}{\theta}}}, \quad (2.5)$$

단, b_{ik} 와 d_{ik} 는 각각 아래 Kronecker 곱으로 이루어진 벡터 \mathbf{b}_i 와 \mathbf{d}_i 의 k 번째 원소를 나타낸다.

$$\mathbf{b}_i = \left(\exp(\tilde{L}_{i1}), \exp(\tilde{U}_{i1}) \right)' \otimes \cdots \otimes \left(\exp(\tilde{L}_{id_i}), \exp(\tilde{U}_{id_i}) \right)'; \quad \mathbf{d}_i = (1, -1)' \otimes \cdots \otimes (1, -1)'.$$

또한 주변우도함수(marginal likelihood function)는 식 (2.5)를 프레일티에 대해 적분하여 얻을 수 있으며 다음과 같이 주어진다.

$$L_{m,i}(\lambda_0, \beta, \theta) = \int_0^\infty L_{f,i}(\lambda_0, \beta, \theta) du_i = \sum_{k=1}^{2^{d_i}} \frac{d_{ik}}{(1/\theta + C_i + \log b_{ik})^{\frac{1}{\theta}} \theta^{\frac{1}{\theta}}}.$$

한편 베イズ 법칙을 써서 프레일티의 사후분포는 다음과 같이 얻을 수 있다.

$$f_{U_i}(u_i|\text{Data}) = \frac{L_{f,i}(\lambda_0, \beta, \theta)}{L_{m,i}(\lambda_0, \beta, \theta)} = \sum_{k=1}^{2^{d_i}} w_{ik} G\left(\frac{1}{\theta}, \frac{1}{(1/\theta + C_i + \log b_{ik})}\right).$$

따라서 사후평균 $E(U_i|\text{Data})$ 과 $E(\log U_i|\text{Data})$ 는 각각 다음과 같다.

$$E(U_i|\text{Data}) = \sum_{k=1}^{2^{d_i}} w_{ik} \frac{1/\theta}{1/\theta + C_i + \log b_{ik}} = u_i^*,$$

$$E(\log U_i|\text{Data}) = \sum_{k=1}^{2^{d_i}} w_{ik} \left\{ \psi\left(\frac{1}{\theta}\right) - \log\left(\frac{1}{\theta} + C_i + \log b_{ik}\right) \right\} = lu_i^*,$$

단, $\psi(x) = d \log(x)/dx = \Gamma'(x)/\Gamma(x)$,

$$w_{ik} = \frac{(-1)^{d_{ik}} / (1/\theta + C_i + \log b_{ik})^{\frac{1}{\theta}}}{\sum_{k=1}^{2^{d_i}} (-1)^{d_{ik}} / (1/\theta + C_i + \log b_{ik})^{\frac{1}{\theta}}}.$$

Goethals 등 (2009)은 주변우도함수

$$L_m(\lambda_0, \beta, \theta) = \prod_{i=1}^n L_{m,i}(\lambda_0, \beta, \theta)$$

에 Newton-Raphson 알고리즘을 적용하여 모수를 추론하는 방법을 제안했는데, 본 논문에서는 완전자료에 기초한 우도함수

$$L_f(\lambda_0, \beta, \theta) = \prod_{i=1}^n L_{f,i}(\lambda_0, \beta, \theta)$$

를 써서 모수를 추론하고자 한다. 완전자료에 기초한 로그우도함수는 다음과 같이 두 개의 항으로 분해되며, 첫 번째 항은 λ_0 와 β 의 함수이고, 두 번째 항은 θ 의 함수이다.

$$l_f(\lambda_0, \beta, \theta) = \sum_{i=1}^n \left\{ \sum_{j \in D_i} \log \left\{ \exp(-u_i \tilde{L}_{ij}) - \exp(-u_i \tilde{U}_{ij}) \right\} - u_i C_i \right\} + \sum_{i=1}^n \left\{ \theta^{-1} \log u_i - \theta^{-1} u_i - \log \Gamma(\theta^{-1}) - \theta^{-1} \log \theta \right\}. \quad (2.6)$$

다시 말해 식 (2.6)의 첫 번째 항에 Sun (2006)의 방법을 적용하여 모수 β 와 $\Lambda_0(t)$ (즉, $S_0(t)$)를 추정하고자 하며, 식 (2.6)의 두 번째 항으로부터 모수 θ 를 추정하고자 한다. 단, $S_0(t) = \exp(-\Lambda_0(t))$. 그런데 첫 번째 항에 포함되어 있는 u_i 는 관측되지 않는 값이기 때문에 u_i 를 u_i^* 로 대체한 후 Sun (2006)의 방법을 적용하고자 하며, 두 번째 항에서는 u_i 와 $\log u_i$ 를 각각 u_i^* 와 lu_i^* 로 대체한 후 Newton-Raphson

방법을 적용하고자 한다. 맨 처음 E-과정을 수행할 때 모수 β 와 기저생존함수 $S_0(\cdot)$ 에 대한 초기값이 필요하다. 이 값들은 $N = \sum_{i=1}^{n_i}$ 개의 모든 개체가 서로 독립이라고 가정하고서 프레이티가 없는 Sun (2006)의 방법을 적용하여 얻은 값으로 사용하고자 한다. 이 때 θ 의 초기값은 보통 1로 놓는다. M-과정에서 모수 β 와 기저생존함수 $S_0(t)$ 의 추정은 Sun (2006)과 같이 Newton-Raphson 방법을 적용하였다. 사실 $S_0(t)$ 에 대한 추정은 비모수적인 방법을 따르는데 이를 위해 먼저 무한차원 모수공간(infinite-dimensional parameter space)를 유한차원 모수공간(finite-dimensional parameter space)으로 축소해야 한다. 모수의 차원을 축소하는 방법은 흔히 주어진 자료의 동등집합(equivalence set)을 만드는 것이다 (Lindsey와 Ryan, 1998). M-과정을 구체적으로 살펴보면 다음과 같다. $D_i \cup R_i$ 에 대응하는 동등집합의 중간값을

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

라고 하고, $S_0(s_q)$ 를 다음과 같이 표현할 수 있다고 가정하자.

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

단, $\alpha = (\alpha_1, \dots, \alpha_q)'$, $\alpha_0 = -\infty$, $\alpha_{m+1} = \infty$. 따라서 $\Lambda_0(t) = \log S_0(t)$ 의 관계를 적용하면,

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

이다. 한편

$$\alpha_{ijq} = I(s_q \in (L_{ij}, U_{ij})), \quad i = 1, \dots, n; \quad j = 1, \dots, n_i; \quad q = 1, \dots, m$$

라고 하자. 따라서 α, β, θ 를 써서 기대로그우도함수 $E\{l_f(\alpha, \beta, \theta)|\text{Data}\}$ 를 표현하면 다음과 같다.

$$\begin{aligned} l_f^*(\alpha, \beta, \theta) &= E\{l_f(\alpha, \beta, \theta)|\text{Data}\} \\ &= \sum_{i=1}^n \left\{ \sum_{j \in D_i} \log \left\{ \sum_{q=1}^{m+1} \alpha_{ijq} (\exp\{-u_i^* a_{q-1} \exp(\beta' x_{ij})\} - \exp\{-u_i^* a_q \exp(\beta' x_{ij})\}) \right\} \right. \\ &\quad \left. - u_i^* \sum_{j \in R_i} \sum_{q=1}^{m+1} I(L_{ij} \in [s_{q-1}, s_q]) a_{q-1} \exp(\beta' x_{ij}) \right\} \\ &\quad + \sum_{i=1}^n \{\theta^{-1} l u_i^* - \theta^{-1} u_i^* - \log \Gamma(\theta^{-1}) - \theta^{-1} \log \theta\} \end{aligned} \tag{2.7}$$

표현을 간단하게 하기 위해

$$\begin{aligned} S^*(t|x_{ij}, u_i^*) &= \exp(-\Lambda_0(t) u_i^* \exp(\beta' x_{ij})), \\ f_{ijq}^* &= S^*(s_q|x_{ij}) \log S^*(s_q|x_{ij}, u_i^*), \\ b_{ijq}^* &= u_i^* \exp(\alpha_q + \beta' x_{ij}), \\ c_{ijq}^* &= \sum_{l=q}^{m+1} (\alpha_{ijl} - \alpha_{ijl+1}) S^*(s_l|x_{ij}, u_i^*), \\ g_{ij}^* &= \sum_{q=1}^{m+1} \alpha_{ijq} \{S^*(s_{q-1}|x_{ij}, u_i^*) - S^*(s_q|x_{ij}, u_i^*)\} \end{aligned}$$

라고 하자. 단, $f_{ij0}^* = f_{ijm+1}^* = 0$, $\alpha_{ijm+2} = 0$. 그러면 β 와 α_q 에 대한 스코어 함수는 다음과 같다.

$$U_\beta^* = \frac{\partial l_f^*}{\partial \beta} = \sum_{i=1}^n \left\{ \sum_{j \in D_i} x_{ij} \frac{\sum_{q=1}^{m+1} \alpha_{ijq} (f_{ijq-1}^* - f_{ijq}^*)}{g_{ij}^*} + u_i \sum_{j \in R_i} x_{ij} \sum_{q=1}^{m+1} I(L_{ij} \in [s_{q-1}, s_q]) a_{q-1} \exp(\beta' x_{ij}) \right\},$$

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

$$U_\theta^* = \frac{\partial l_f^*}{\partial \theta} = \theta^{-2} \left\{ \sum_{i=1}^n (u_i^* - l u_i^*) + n\psi(\theta^{-1}) + n \log \theta - n \right\}.$$

관찰정보행렬(observed information matrix)의 각 블록의 원소는 다음과 같다.

$$I_{11} = - \frac{\partial^2 l_f^*}{\partial \beta \partial \beta'} = \sum_{i=1}^n \left\{ \sum_{j \in D_i} x_{ij} x_{ij}' \left\{ \left(\frac{\sum_{q=1}^{m+1} \alpha_{ijq} (f_{ijq-1}^* - f_{ijq}^*)}{g_{ij}^*} \right)^2 - \frac{\sum_{q=1}^{m+1} \alpha_{ijq} (h_{ijq-1}^* - h_{ijq}^*)}{g_{ij}^*} \right\} + u_i \sum_{j \in R_i} x_{ij} x_{ij}' \sum_{q=1}^{m+1} I(L_{ij} \in [s_{q-1}, s_q]) a_{q-1} \exp(\beta' x_{ij}) \right\},$$

$$I_{12} = I_{21}' = - \frac{\partial^2 l_f^*}{\partial \alpha_q \partial \beta} = - \sum_{i=1}^n \left\{ \sum_{j \in D_i} x_{ij} b_{ijq}^* \left\{ \frac{c_{ijq}^* + \sum_{l=q}^{m+1} (\alpha_{ijl} - \alpha_{ijl+1}) f_{ijl}^*}{g_{ij}^*} - \frac{c_{ijq}^*}{g_{ij}^{*2}} \sum_{l=1}^{m+1} \alpha_{ijl} (f_{ijl-1}^* - f_{ijl}^*) \right\} - \sum_{j \in R_i} x_{ij} b_{ijq}^* I(L_{ij} \in [s_q, \infty)) \right\},$$

$$I_{22} = - \frac{\partial^2 l_f^*}{\partial \alpha_q^2} = - \sum_{i=1}^n \left\{ \sum_{j \in D_i} b_{ijq}^* c_{ijq}^* \left(\frac{1 - b_{ijq}^*}{g_{ij}^*} - \frac{b_{ijq}^* c_{ijq}^*}{g_{ij}^{*2}} \right) - \sum_{j \in R_i} b_{ijq}^* I(L_{ij} \in [s_q, \infty)) \right\},$$

$$I_{22} = - \frac{\partial^2 l_f^*}{\partial \alpha_q \partial \alpha_r} = \sum_{i=1}^n \sum_{j \in D_i} \left(\frac{b_{ijq}^* b_{ijr}^* c_{ijq}^* c_{ijr}^*}{g_{ij}^{*2}} + \frac{b_{ijq}^* b_{ijr}^* c_{ijr}^*}{g_{ij}^*} \right) \quad (q < r),$$

단, $h_{ijq} = f_{ijq}^* (1 + \log S^*(s_q | x_{ij}, u_i^*))$ 이고, $h_{ij0} = h_{ijm+1} = 0$. 또한

$$I_\theta = - \frac{\partial^2 l_f^*}{\partial \theta^2} = \frac{2}{\theta^3} \left\{ \sum_{i=1}^n (u_i^* - l u_i^*) + n\psi(\theta^{-1}) + n \log \theta - \frac{3}{2}n + \frac{1}{2}\theta^{-1}n\psi'(\theta^{-1}) \right\},$$

단, $\psi'(x) = \Gamma''(x)/\Gamma(x) - \psi(x)^2$. 따라서 $(\alpha, \beta)'$ 의 s 번째 해는 다음 식으로부터 얻을 수 있다.

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

단,

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

이고, $I_{11|2} = I_{11} - I_{12}I_{22}^{-1}I_{21}$, $I_{12|2} = I'_{21|2} = -I_{11|2}^{-1}I_{12}I_{22}^{-1}$, $I_{22|1} = I_{22}^{-1} + I_{22}^{-1}I_{21}I_{11|2}^{-1}I_{12}I_{22}^{-1}$. 마찬가지로 θ 의 s 번째 해는 다음 식으로부터 얻을 수 있다.

$$\theta^{(s)} = \theta^{(s-1)} + I_{\theta}^{-1}U_{\theta}^*.$$

다음 조건을 만족할 때까지 위 E-과정과 M-과정을 반복하여 최종 해를 얻는다. $\epsilon > 0$ 에 대해

$$\max \left\{ \left| \alpha_q^{(s)} - \alpha_q^{(s-1)} \right|, \left| \beta_r^{(s)} - \beta_r^{(s-1)} \right|, \left| \theta^{(s)} - \theta^{(s-1)} \right| \right\} < \epsilon.$$

한편 위 수렴조건을 만족한 모수 $(\alpha, \beta)'$ 와 θ 의 해를 각각 $(\alpha^{(s)}, \beta^{(s)})'$ 와 $\theta^{(s)}$ 라고 할 때, 모수 $(\alpha, \beta)'$ 와 θ 의 추정량에 대한 분산은 각각 행렬 $I((\alpha, \beta)')^{-1}|_{(\alpha, \beta)' = (\alpha^{(s)}, \beta^{(s)})'}$ 의 대각원소와 $I_{\theta}^{-1}|_{\theta = \theta^{(s)}}$ 로 추정한다.

3. 모의실험 연구

본 절에서는 2절에서 제안한 추정량의 소표본 성질을 살펴보기 위해 모의실험을 수행하였다. 모수들은 다음과 같이 정의하였다. 기저위험함수는 $\lambda_0(t) = 0.1$ 로 놓았으며, 프레일티 분포의 모수 θ 는 0.5, 1, 1.5로 놓았다. $u_i \sim G(\theta^{-1}, \theta)$, $\theta > 0$ 일 때, T_{i1}, \dots, T_{in_i} 의 조건부생존함수(conditional survival function)는 다음과 같다.

$$S(t_{i1}, \dots, t_{in_i} | x_{ij}, u_i) = \prod_{j=1}^{n_i} S(t_{ij} | x_{ij}, u_i) = \exp \left\{ -u_i \sum_{j=1}^{n_i} \Lambda_0(t_{ij}) \exp(\beta' x_{ij}) \right\}.$$

또한 주변생존함수(marginal survival function)는 다음과 같다.

$$S(t_{i1}, \dots, t_{in_i} | x_{ij}) = \left\{ 1 + \theta \sum_{j=1}^{n_i} \Lambda_0(t_{ij}) \exp(\beta' x_{ij}) \right\}^{-\theta^{-1}}.$$

위 식으로부터 켄달의 타우 τ 는 $\tau = \theta/(\theta + 2)$ 와 같이 표현된다. 따라서 모수 θ 에 대한 위의 값들은 각각 $\tau = 1/5, 1/3, 3/7$ 에 해당한다. 100개 군집을 고려했으며, 군집 내 개체수는 각각 3명, 5명, 7명으로 구성하였다. 0-1 공변량을 하나만 고려하였으며 공변량이 생존시간에 미치는 효과는 0, 0.5, 1로 놓았다. 상술한 모수로부터 군집구간중도절단된 자료를 생성하는 절차는 다음과 같다.

- 단계1: 성공확률이 0.5인 베르누이 분포로부터 x_{ij} 를 생성한다.
- 단계2: 감마분포 $G(\theta^{-1}, \theta)$ 로부터 u_i 를 생성한다.
- 단계3: $U(0, 1)$ 로부터 p 를 생성한다.
- 단계4: 단계1-단계3에서 생성한 x_{ij} 와 u_i, p 에 대해 다음 관계 식을 만족하는 t 를 T_{ij} 로 놓는다.

$$S(t | x_{ij}, u_i) = \exp(-u_i \Lambda_0(t) \exp(\beta x_{ij})) = 1 - p.$$

Table 3.1. Bias, standard deviation(SD), mean of se(SeM), 95% coverage rate(CP) of parameters, β and θ , based on 2,000 replications when $n_i = 3$

π	True		RC	β				θ			
	β	θ		Bias	SD	SeM	CP	Bias	SD	SeM	CP
0		0.5	42.7	-0.008	0.175	0.154	91.5	0.000	0.176	0.065	54.1
		1	48.6	0.000	0.187	0.164	91.6	0.000	0.265	0.124	65.5
		1.5	53.2	-0.004	0.210	0.172	89.2	-0.002	0.361	0.179	66.5
0.5	0.5	0.5	35.5	-0.002	0.167	0.147	91.1	0.007	0.163	0.066	58.1
		1	42.3	-0.005	0.186	0.156	89.9	-0.016	0.246	0.122	67.2
		1.5	47.8	-0.009	0.199	0.165	90.0	-0.018	0.335	0.178	68.2
	1	0.5	29.7	-0.015	0.171	0.146	90.6	-0.016	0.147	0.064	59.9
		1	37.2	-0.029	0.181	0.154	89.8	-0.023	0.233	0.122	69.0
		1.5	43.0	-0.036	0.198	0.162	88.2	-0.049	0.318	0.174	70.8
0.8	0	0.5	39.9	-0.005	0.169	0.150	91.7	0.009	0.168	0.066	56.7
		1	45.9	0.005	0.185	0.158	90.9	-0.002	0.263	0.124	65.0
		1.5	51.0	0.006	0.207	0.167	89.1	0.034	0.373	0.183	67.3
	0.5	0.5	32.9	-0.003	0.163	0.143	91.2	0.005	0.154	0.066	60.6
		1	40.1	-0.003	0.188	0.152	89.0	0.008	0.249	0.125	68.5
		1.5	45.7	-0.003	0.200	0.160	88.4	0.013	0.338	0.181	70.5
1	0.5	27.2	-0.007	0.167	0.142	91.0	-0.005	0.145	0.065	64.6	
	1	34.9	0.002	0.186	0.150	88.5	0.002	0.231	0.124	71.5	
	1.5	41.1	-0.007	0.199	0.157	88.6	0.012	0.324	0.181	73.9	

- 단계5: 모든 개체는 총 12개의 시점 $1, 2, \dots, 12$ 에서 관측된다고 가정하자 (Kim과 Xu, 2002). 성공확률이 π 인 베르누이 분포로부터 서로 독립인 12개의 확률변수 v_{ij} 를 생성하여 이들로 이루어진 확률벡터 $v_i = (v_{i1}, \dots, v_{i12})'$ 를 얻는다. ' $v_{ik} = 0$ '이면 한 개체가 k 번째 시점에 방문하지 않았다는 것을 의미하고, ' $v_{ik} = 1$ '이면 방문했다는 것을 의미한다 (단, $k = 1, \dots, 12$). 예를 들어 $v_i = (1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0)'$ 이면 가능한 증도절단구간은 $(0, 1], (1, 2], (2, 5], (5, 7], (7, 11], (11, \infty)$ 이다.
- 단계6: 단계4에서 얻은 T_{ij} 는 이 값 전후에 있는 가장 가까운 방문시점들 사이에서 증도절단 시킨다. 예를 들어, $t_{ij} = 8.5$ 라고 하면 8.5 이전에 방문한 마지막 시점은 7이고, 8.5 이후에 방문한 최초 시점은 11이므로 T_{ij} 는 구간 $(7, 11]$ 에서 증도절단된 것으로 간주한다.

모의실험에서 π 는 0.5와 0.8를 고려하였다. π 의 값이 클수록 구간증도절단된 자료의 비율이 높아지고(즉, 우증도절단된 비율은 작아지고), 구간증도절단된 자료의 평균 길이는 줄어든다. 모의실험 설계 변수 θ, β, n_i, π 의 모든 조합에 대해 2,000번씩 반복하여 β 와 θ 에 대한 추정량의 편향('bias')과 표준편차('SD'), 표준오차의 평균('SeM'), 95% 실험적 포함률('CP')를 각각 구하였다.

Table 3.1, Table 3.2, Table 3.3은 각각 $n_i = 3, 5, 7$ 일 때의 결과이다. 각 표에서 'RC'는 우증도절단 비율을 의미한다. Table 3.1를 보면 첫째, 우증도절단 비율은 다소 높은 편이며 예상했던 것처럼 π 가 커질수록 그 비율이 감소하였다. 둘째, 'bias'는 나쁘지 않지만 'SD'와 'SeM'의 차이는 예상했던 것보다 큰 편이었다. 이 차이로 인해 'CP'의 값이 명목값보다 작게 나타났다. 셋째, 모수 θ 의 'CP'는 명목값보다 좀 더 작게 나타났다. Table 3.2와 Table 3.3을 보면 n_i 가 커질수록 'RC'는 거의 변하지 않았지만 'bias'와 'SD', 'SeM'은 점차 줄어들었으며 'CP'의 값은 명목값인 95%에 근접하게 나타났다.

Table 3.2. Bias, standard deviation(SD), mean of se(SeM), 95% coverage rate(CP) of parameters, β and θ , based on 2,000 replications when $n_i = 5$

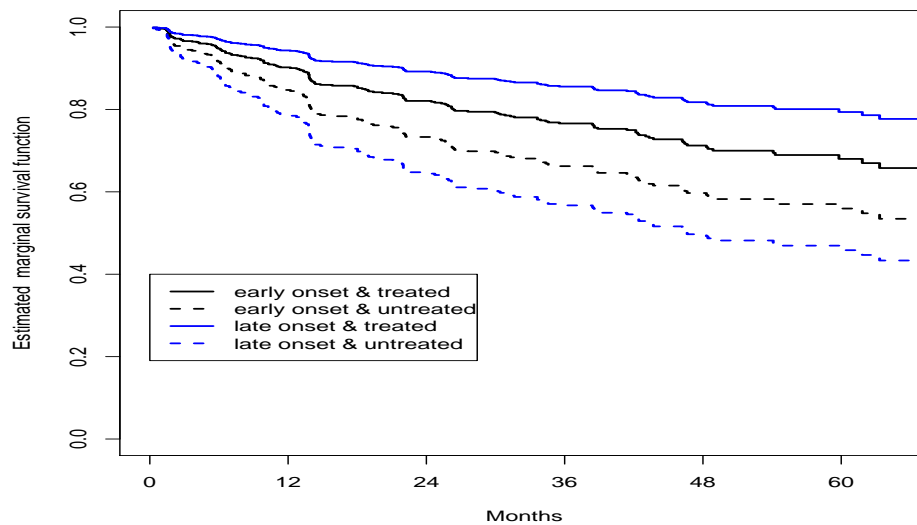
π	True		RC	β				θ			
	β	θ		Bias	SD	SeM	CP	Bias	SD	SeM	CP
0		0.5	42.2	-0.004	0.132	0.119	92.6	-0.005	0.125	0.065	68.9
		1	48.1	0.002	0.142	0.126	91.7	0.003	0.202	0.125	77.7
		1.5	52.6	0.003	0.148	0.132	92.7	-0.007	0.293	0.179	76.9
0.5	0.5	0.5	35.1	-0.002	0.123	0.114	93.1	-0.004	0.117	0.065	73.1
		1	42.0	-0.004	0.138	0.121	91.3	-0.006	0.196	0.124	76.5
		1.5	47.3	-0.005	0.140	0.127	92.6	-0.018	0.266	0.178	80.1
0.5	1	0.5	29.4	-0.010	0.125	0.113	92.1	-0.006	0.114	0.065	74.5
		1	36.7	-0.007	0.134	0.119	91.7	-0.027	0.188	0.121	77.7
		1.5	42.6	-0.014	0.148	0.125	89.6	-0.021	0.264	0.177	79.5
0.8	0	0.5	39.8	-0.001	0.126	0.116	93.3	0.004	0.125	0.066	69.0
		1	46.1	-0.004	0.134	0.123	92.5	0.010	0.203	0.125	78.3
		1.5	50.8	-0.005	0.146	0.129	92.2	-0.017	0.288	0.178	78.2
0.8	0.5	0.5	32.8	-0.001	0.123	0.111	92.4	-0.003	0.115	0.065	73.1
		1	40.0	-0.001	0.136	0.118	90.7	-0.010	0.191	0.123	78.0
		1.5	45.6	-0.001	0.143	0.124	91.9	-0.002	0.277	0.179	79.3
0.8	1	0.5	27.2	-0.001	0.124	0.110	91.9	-0.009	0.109	0.064	75.2
		1	34.9	-0.002	0.130	0.116	91.9	-0.011	0.186	0.123	80.1
		1.5	40.9	-0.012	0.141	0.121	91.3	-0.012	0.261	0.178	81.1

Table 3.3. Bias, standard deviation(SD), mean of se(SeM), 95% coverage rate(CP) of parameters, β and θ , based on 2,000 replications when $n_i = 7$

π	True		RC	β				θ			
	β	θ		Bias	SD	SeM	CP	Bias	SD	SeM	CP
0		0.5	41.9	0.000	0.109	0.100	93.6	-0.006	0.107	0.065	75.4
		1	47.8	-0.004	0.114	0.106	93.4	-0.009	0.185	0.123	80.0
		1.5	52.4	-0.006	0.121	0.112	92.7	-0.011	0.257	0.178	81.8
0.5	0.5	0.5	34.9	-0.002	0.104	0.096	93.0	-0.005	0.102	0.065	78.1
		1	41.8	-0.001	0.110	0.102	93.0	-0.010	0.178	0.123	81.9
		1.5	47.2	0.000	0.122	0.107	91.7	-0.017	0.253	0.178	82.4
0.5	1	0.5	29.2	-0.010	0.104	0.095	92.3	-0.011	0.099	0.064	78.3
		1	36.7	-0.011	0.113	0.101	91.9	-0.013	0.173	0.123	82.0
		1.5	42.5	-0.011	0.119	0.106	92.1	-0.015	0.237	0.178	85.3
0.8	0	0.5	39.7	-0.002	0.106	0.098	92.8	-0.006	0.107	0.065	75.7
		1	45.9	0.003	0.111	0.104	93.5	-0.007	0.178	0.123	81.4
		1.5	50.8	0.000	0.121	0.109	93.4	-0.005	0.259	0.179	81.5
0.8	0.5	0.5	32.9	-0.004	0.100	0.094	93.4	-0.004	0.099	0.065	79.6
		1	40.0	-0.003	0.108	0.099	92.9	-0.005	0.171	0.124	83.8
		1.5	45.5	-0.005	0.118	0.104	91.4	-0.006	0.246	0.179	83.7
0.8	1	0.5	27.3	-0.002	0.102	0.093	92.8	-0.004	0.098	0.065	80.0
		1	35.1	-0.005	0.109	0.098	91.9	-0.003	0.168	0.124	84.3
		1.5	41.0	0.001	0.118	0.103	91.6	-0.011	0.244	0.178	83.6

Table 4.1. Parameter estimates and their standard errors for DRS data

	Type(x_1)		Treatment(x_2)		Interaction(x_3)		θ		
	Est.	SE	Est.	SE	Est.	SE	Est.	SE	
Grouped case									
Proposed	0.40	0.20	-0.52	0.22	-0.96	0.35	0.99	0.09	
Lam et al.	0.41	0.26	-0.51	0.23	-0.99	0.36	0.93	0.31	
Ungrouped case									
Cox model(frailty)	0.40	0.26	-0.51	0.23	-0.99	0.36	0.93	0.24	
Weibull(frailty)	0.42	0.27	-0.53	0.24	-1.03	0.37	1.10	0.37	

**Figure 4.1.** Estimated marginal survival functions

4. 당뇨병성 망막병증 연구 자료에 적용

당뇨병성 망막병증 연구(diabetic retinopathy studies; DRS)는 레이저 광학응고치료법이 실명에 이르는 시기를 늦추는 효과가 있는지를 평가하기 위한 연구이다. 이 연구에서는 당뇨병 환자 중에서 두 눈이 모두 실명에 이를 가능성이 높은 197명을 대상으로 하였다. 두 눈 중에서 랜덤하게 한 눈에는 레이저 치료를 하고 다른 한 눈에는 레이저 치료를 하지 않았다. 치료 전에 시력을 측정하였고 매 4개월에 한 번씩 내원하여 시력을 측정 받도록 하였다. 이 연구에서 생존시간은 시력이 5/200로 떨어지는 데까지 걸린 시간으로 정의하였다. 이 연구에서 군집은 환자이며 군집 내 개체는 두 눈이다. 분석에 포함된 공변량은 당뇨병의 형태(x_1)와 레이저 치료 여부(x_2), 두 변수의 교호작용($x_3 = x_1 \times x_2$)이다. 당뇨병이 20세 이전에 발병했으면 $x_1 = 0$, 20세 이후에 발병하였으면 $x_1 = 1$ 이다. 레이저 치료를 받지 않았으면 $x_2 = 0$, 레이저 치료를 받았으면 $x_2 = 1$ 이다. 한편 Ross와 Moore (1999)는 16개의 구간으로 나누었는데 본 논문에서도 그들의 범주 구분을 따랐다. 그들이 나눈 구간은 (0, 6], (6, 10], (10, 14], ..., (54, 58], (58, 66], (66, 83]이다.

Table 4.1은 네 가지 서로 다른 방법에 따른 모수 추정량과 표준오차를 정리한 것이다. 처음 두 방법은 Ross와 Moore (1999)를 따라 DRS 자료를 그룹화된 자료('Grouped case'로 표시)로 다루었으며 마치

막 두 방법은 우중도절단된 자료('Ungrouped case'로 표시)로 다루었다. 두 번째 방법('Lam *et al.*'로 표시)은 본 논문과 모형은 동일하지만 다중대체법(multiple imputation)을 써서 구간우중도절단된 자료로 변환하여 다룬 점이 본 논문의 방법과 다르다. 마지막 두 방법도 처음 두 방법과 마찬가지로 군집 간 차이를 프레이티로 다루었는데, 세 번째 방법('Cox model'로 표시)은 기저위험함수를 비모수적으로 다룬 반면 네 번째 방법('Weibull'로 표시)은 모수적으로 다루었다. Table 4.1에서 볼 수 있듯이 네 가지 방법에 따른 결과는 거의 유사하다. 다만 제안한 방법('Proposed'로 표시)에서 θ 에 대한 추정량의 표준오차가 나머지 방법보다 작게 나타났다. 'Proposed'의 결과를 중심으로 보면 레이저 치료 여부가 생존 시간에 미치는 영향은 유의하였으며(p -값 = 0.018), 특히 20세 이전에 당뇨병이 발병한 그룹보다 20세 이후에 발병한 그룹에서 더 효과적인 것으로 나타났다(20세 이전 그룹에서 위험률의 비 = 0.59, 20세 이후 그룹에서 위험률의 비 = 0.23). 단, 위험률의 비는 레이저 치료를 받지 않은 그룹에 대한 레이저 치료를 받은 그룹의 위험률의 비이다. 프레이티 효과는 p -값이 < 0.0001 으로 매우 유의하여 환자 간 차이가 있음을 알 수 있었다. Figure 4.1은 당뇨병 형태와 레이저 치료 여부에 따라 나눈 네 그룹의 추정생존함수를 각각 나타낸 것이다. Table 4.1의 결과에서 살펴본 것처럼 당뇨병의 형태에 관계없이 레이저 치료를 받은 그룹의 추정생존함수가 레이저 치료를 받지 않은 그룹보다 높았으며, 레이저 치료에 따른 효과(즉, 레이저 치료 여부에 따른 추정생존함수 차이)는 20세 이전에 당뇨병이 발병한 그룹보다 20세 이후에 발병한 그룹에서 더 큰 것으로 나타났다.

5. 토의

본 논문에서는 군집구간중도절단된 자료를 분석하기 위해 준모수적인 모형을 제안했으며 군집 내 개체들 간의 연관성은 감마 프레이티를 통해 모형에 포함하였다. EM 알고리즘 방법을 써서 모수를 추정하였고 모의실험을 통해 그 추정량의 소표본 성질을 살펴보았다. 모의실험 결과에 따르면 추정량의 편향은 거의 없었지만 추정량의 표준오차가 과소추정되었으며 이로 인해 95% 신뢰구간 포함률이 명목값보다 작게 나타났다. 본 논문에서는 β , θ 및 기저위험함수와 관련된 모수를 추정해야 한다. 그런데 본 논문에서는 기저위험함수를 비모수적으로 추정하고자 하기 때문에 추정해야 할 모수의 개수가 자료의 크기에 따라 변한다. 따라서 기저위험함수의 추정이 관심 모수인 β 와 θ 의 추정에 영향을 미칠 수 있다고 생각된다. 그러나 이런 현상은 군집 내 개체수가 증가함에 따라 점차 나아졌다. 한편 심사위원의 지적에 따라 군집 내 개체수를 고정시킨 후 군집의 개수를 증가시켜 추정량의 성질을 살펴본 결과, 추정량의 bias는 거의 변하지 않았지만 추정량의 SD와 SeM은 작아지는 경향을 보였다. 그러나 CP의 값은 거의 변화가 없었다. 따라서 연관성 모수 θ 의 성질은 군집의 개수 보다는 군집 내 개체수에 더 의존함을 반증하는 것이라고 생각된다. 예상했던 것처럼 π 가 커질수록 우중도절단 되는 비율은 감소하였지만 그 양이 크지 않았으며, 추정량의 bias, SD 및 SeM과 CP는 거의 차이가 없었다. 심사위원의 지적에 따라 π 의 값이 작아짐에 따라 추정량에 미치는 영향을 좀 더 자세히 살펴보기 위해 Table 3.2의 모의실험조건에서 $\pi = 0.2$ 인 경우에 대해 모의실험을 수행하였다. $\pi = 0.5$ 인 경우와 비교해볼 때 bias가 조금 커졌으며 SD와 SeM도 조금 증가하는 경향을 보였다. CP의 값을 보면 β 의 경우는 거의 변화가 없지만 θ 의 경우는 조금 작아지는 경향을 보였다. 각 모의실험조건에서 프로그램 수행에 소요된 시간은 Table 3.1의 경우는 약 260분, Table 3.2와 Table 3.3의 경우는 약 200분이었다. 이는 군집 내 크기가 클수록 수렴속도가 빨라져 오히려 소요시간이 줄어들었기 때문이다.

Table 4.1에 포함된 네 가지 모형 중에서 DRS 자료를 더 잘 적합하는 모형을 선택하는 기준으로 Akaike information criterion(AIC)를 고려할 수 있다. 다만 네 가지 모형 중에서 제안한 모형을 제외한 나머지 모형의 결과는 Lam 등 (2010)에 수록된 결과를 인용했기 때문에 이 세 모형에 대해서는 AIC를 구할 수 없어 Table 4.1에 포함된 네 모형의 적합도를 직접적으로 비교할 수 없었다. 그러나 향

후에 추정량의 표준오차가 과소추정되는 단점을 해결하기 위해 EM 알고리즘과 Lam 등 (2010)에 의해 제안된 다중대체방법을 결합하는 방법을 제안하고자 한다. 비록 프레일티에 대한 추정량의 표준오차가 소표본에서 과소추정되는 단점을 가지고 있지만 제안한 추정방법은 기저위험함수를 비모수적으로 다루었기 때문에 모수적으로 다른 Bellamy 등 (2004), Goethals 등 (2009), Ampe 등 (2012)의 방법보다 임상연구에서 더 실용적이라고 할 수 있다. 또한 제안한 모형이 프레일티 분포에 대한 오설계(misspecification)에 대해 로버스트한지를 추후 연구에서 다루고자 한다.

References

- Aalen, O. O., Borgan, Ø. and Gjessing, H. K. (2008). *Survival and Event History Analysis*, Springer, New York.
- Ampe, B., Goethals, K., Laevens, H. and Duchateau, L. (2012). Investigating clustering in interval-censored udder quarter infection times in dairy cows using a gamma frailty model, *Preventive Veterinary Medicine*, in press.
- Bellamy, S., Li, Y., Ryan, L. M., Lipsitz, S., Canner, M. and Wright, R. (2004). Analysis of clustered and interval censored data from a community-based study in asthma, *Statistics in Medicine*, **23**, 3607–3621.
- Cai, J. and Prentice, R. L. (1995). Estimating equations for hazard ratio parameters based on correlated failure time data, *Biometrics*, **82**, 151–164.
- Duchateau, L. and Janssen, P. (2008). *The Frailty Model*, Springer, New York.
- Goethals, K., Ampe, B., Berkvens, D., Laevens, H., Janssen, P. and Duchateau, L. (2009). Modelling interval-censored, clustered cow udder quarter infection times through the shared gamma frailty model, *Journal of Agricultural, Biological, and Environmental Statistics*, **14**, 1–14.
- Goggins, W. B. and Finkelstein, D. M. (2000). A proportional hazards model for multivariate interval-censored failure time data, *Biometrics*, **56**, 940–943.
- Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, Second Ed., John Wiley, New York.
- Kim, M. Y. and Xu, X. (2002). The analysis of multivariate interval-censored survival data, *Statistics in Medicine*, **21**, 3715–3726.
- Lam, K. F., Xu, Y. and Cheung, T.-L. (2010). A multiple imputation approach for clustered interval-censored survival data, *Statistics in Medicine*, **29**, 680–693.
- Lindsey, J. C. and Ryan, L. M. (1998). Tutorial in biostatistics: Methods for interval-censored data, *Statistics in Medicine*, **17**, 219–238.
- Ross, E. A. and Moore, D. (1999). Modeling clustered, discrete, or grouped time survival data with covariates, *Biometrics*, **55**, 813–819.
- Sun, J. (2006). *The Statistical Analysis of Interval-censored Failure Time Data*, Springer, New York.