The Korean Journal of Applied Statistics (2017) DOI: https://doi.org/10.5351/KJAS.2017.30.4.539
30(4), 539–553

Additive hazards models for interval-censored semi-competing risks data with missing intermediate events

Jayoun Kim[a] · Jinheum Kim[b,1]

[a]Research Coordinating Center, Konkuk University Medical Center;
[b]Department of Applied Statistics, University of Suwon

(Received April 18, 2017; Revised June 14, 2017; Accepted June 18, 2017)

Abstract

We propose a multi-state model to analyze semi-competing risks data with interval-censored or missing intermediate events. This model is an extension of the three states of the illness-death model: healthy, disease, and dead. The 'diseased' state can be considered as the intermediate event. Two more states are added into the illness-death model to incorporate the missing events, which are caused by a loss of follow-up before the end of a study. One of them is a state of the lost-to-follow-up (LTF), and the other is an unobservable state that represents an intermediate event experienced after the occurrence of LTF. Given covariates, we employ the Lin and Ying additive hazards model with log-normal frailty and construct a conditional likelihood to estimate transition intensities between states in the multi-state model. A marginalization of the full likelihood is completed using adaptive importance sampling, and the optimal solution of the regression parameters is achieved through an iterative quasi-Newton algorithm. Simulation studies are performed to investigate the finite-sample performance of the proposed estimation method in terms of empirical coverage probability of true regression parameters. Our proposed method is also illustrated with a dataset adapted from Helmer *et al.* (2001).

Keywords: additive hazards model, log-normal frailty, interval-censored or missing intermediate event, multi-state model, semi-competing risks data

1. 서론

MRC UK Cognitive Function and Aging Study (MRC CFAS)는 영국에서 6개 지역을 대상으로 진행된 다기관 연구로 65세 이상 노인들의 인지기능장애와 관련된 요인들을 찾기 위한 목적으로 진행되었다 (MRC CFAS, 1998). 인지기능장애(cognitive impairment; CI)의 유무는 Mini-Mental State Examination (MMSE) 점수로 평가하게 되는데 그 점수가 일정 점수보다 높으면 CI가 있는 것으로 정의하였다. 이 연구에서는 매 1–3년마다 참가자들의 CI의 발생 유무를 추적 관찰하였다. 그런데 연구 도

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2014R1A1A2056869).
[1]Corresponding author: Department of Applied Statistics, University of Suwon, 17 Wauan-Gil, Bongdam-Eup, Hwaseong 18323, Korea. E-mail: jkimdt65@gmail.com

중에 참가자가 사망할 수 있기 때문에 모든 참가자의 CI의 발생 유무는 관측할 수 없었지만 모든 참가자의 사망 유무와 사망 연령은 관측할 수 있었다고 한다. 이처럼 사망으로 인해 CI의 발생이 중도절단(censoring)될 수는 있지만 CI의 발생으로 인해 사망이 중도절단되지 않는 비대칭적인 자료를 준경쟁적위험 자료(semi-competing risks data)라고 한다. CI의 발생처럼 다른 사건의 발생으로 인해 중도절단될 수 있는 사건을 중간사건(intermediate event)이라고 하고, 사망처럼 다른 사건의 발생으로 인해 중도절단되지 않는 사건을 종말사건(terminal event)이라고 한다. 준경쟁적위험 자료에 대한 모형으로 가장 널리 쓰이는 모형은 소위 'illness-death' 모형(IDM)이다 (Andersen 등, 1993). MRC CFAS에서는 준경쟁적위험 자료를 다루는데 있어 두 가지 다른 특징을 갖고 있다. 첫째, 중도 탈락(lost to follow up; LTF)의 발생을 중도절단으로 다루지 않고 CI의 발생과 마찬가지로 중간사건으로 다루었다. 둘째, LTF의 발생으로 인해 CI의 발생이 중도절단될 수 있다고 가정하였다. 이때, CI의 발생 유무를 관찰할 수 없기 때문에 확률 모형을 써서 다루었다.

상술한 것처럼 두 가지 중간사건이 모두 종말사건에 의해 중도절단될 수 있을 뿐만 아니라 두 가지 중간사건 중에서 관심 있는 중간사건, 즉 CI의 발생이 다른 중간사건, 즉 LTF의 발생으로 인해서도 중도절단되는 상황을 고려하는 연구가 활발히 진행되어 왔다. Frydman과 Szarek (2009)는 CI의 발생에 대한 생존함수(survival function)를 비모수적인 방법으로 추정하는 방법을 제안하였고, Siannis 등 (2007)과 Barrett 등 (2011)은 Cox (Cox, 1972)의 비례위험모형(proportional hazards model)을 통해 회귀계수를 추정하는 방법을 제안하였다. 본 논문에서는 Cox의 비례위험모형 대신에 Lin과 Ying (1994)이 제안한 가산위험모형(additive risk model; L&Y 모형)을 써서 회귀계수를 추정하는 방법을 제안하고자 한다. Cox의 모형은 공변량의 효과를 위험률의 비(hazard ratio)로 설명하는 반면에 L&Y 모형은 공변량의 효과를 위험률의 차이(hazard difference)로 설명하는 점이 서로 다르다. 또한 CI의 발생 유무는 주기적으로 관찰되기 때문에 CI의 발생 시점은 정확히 알 수 없고 CI의 발생이 관측되지 않았던 마지막 방문 시점과 CI의 발생이 관측된 최초 시점 사이로 정의한다. 예를 들어 CI가 구간 $(L, R]$에서 발생했을 때 L과 R 사이의 어느 시점에서 발생했는지 알 수 없기 때문에 Barrett 등 (2011)처럼 그 구간의 모든 시점에서 균등하게 발생할 수 있다고 가정할 수 있지만, 즉 CI의 발생 시점을 비조건부 확률(unconditional probability)로 다룰 수 있지만, Lindsey와 Ryan (1998)과 Collett (2015)의 방법에 의하면 구간 $(L, R]$를 몇 개의 부구간(sub-interval)으로 나눌 수 있을 뿐만 아니라 각 부구간에서 CI의 발생 확률도 추정할 수 있기 때문에 본 논문에서는 각 부구간의 조건부 확률(conditional probability)을 가중치로 하는 추정 방법을 제안하고자 한다. 한편 한 개체가 여러 사건, 즉 CI의 발생, LTF의 발생, 사망 등에 노출되어 있기 때문에 본 논문에서는 잠재 변수(latent variable)인 프레일티(frailty)를 써서 사건 간 연관성을 모형에 포함하고자 한다.

본 논문은 다음과 같이 구성되어 있다. 2절에서는 모형을 제안하고 모수를 추정하는 방법을 소개하고자 한다. 3절과 4절에서는 각각 모의실험 자료와 실제 자료를 이용하여 제안한 모형의 소표본 성질을 살펴보고자 한다. 마지막으로 5절에서 본 논문의 한계를 밝히고 향후 연구 과제를 제시하고자 한다.

2. 모형 및 모수 추론

본 논문에서는 Siannis 등 (2007)과 Barrett 등 (2011)처럼 IDM에 LTF가 추가된 모형을 제안하고자 한다 (Figure 2.1). 이 모형은 다섯 가지 상태, 즉 건강한 상태(H), 관심 있는 중간사건의 발생 상태(non-fatal; NF), 종말사건의 발생 상태(fatal; F), LTF의 발생 상태(LTF), 관측 불가능한 중간사건의 발생 상태(NF(LTF))로 구성되어 있다. 상태 H, NF, F, LTF, NF(LTF)를 각각 상태 0, 1, 2, 3, 4라고 표기 하자. Figure 2.1에 표시된 것처럼 상태 간의 전이 중에서 가능한 경우는 일곱 가지, 즉 $0 \to 1$, $0 \to 2$, $0 \to 3$, $1 \to 2$, $3 \to 2$, $3 \to 4$, $4 \to 2$이다. 이 중에서 점선으로 표시된 전이 $3 \to 4$와 $4 \to 2$는

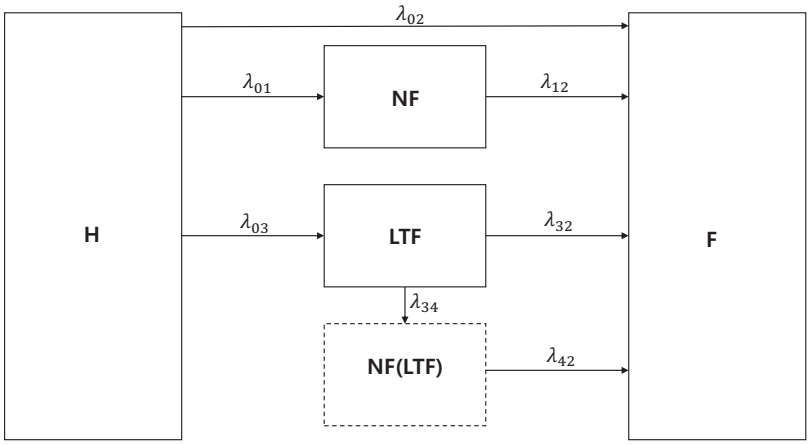

Figure 2.1. A five-state model with lost-to-follow-up (LTF).

관측가능하지 않은 잠재적인 전이에 해당된다.

t는 연구 시작 시점부터의 시간을 나타낸다고 하자. 시점 $t \geq 0$에서 한 개체가 가질 수 있는 상태 S_t는 $S_t \in \{0, 1, 2, 3, 4\}$이다. $\mathcal{A} = \{(r, s) : (r, s) = (0, 1), (0, 2), (0, 3), (1, 2), (3, 2), (3, 4), (4, 2)\}$라고 하자. 시점 t에서 $(r, s) \in \mathcal{A}$에 대하여 상태 r에서 상태 s로의 전이강도(transition intensity) $\lambda_{rs}(t)$를 다음과 같이 정의하고,

$$\lambda_{rs}(t) = \lim_{dt \to 0} \frac{\Pr(S_{t+dt} = s | S_t = r)}{dt}, \quad (r, s) \in \mathcal{A},$$

$(r, s) \notin \mathcal{A}$에 대해서는 $\lambda_{rs}(t) = 0$으로 정의하자. Figure 2.1에서 볼 수 있듯이 LTF 이후에 중간사건의 발생 유무를 알 수 없기 때문에 관측가능한 자료로는 세 가지 전이 $3 \to 2$, $3 \to 4$, $4 \to 2$를 구분할 수 없다 (즉, non-identifiable). 따라서 $\lambda_{34}(t)$와 $\lambda_{42}(t)$가 다음 두 가지 제약 조건을 만족한다고 가정하자 (Siannis 등, 2007; Barrett 등, 2011).

$$\lambda_{02}(t) - \lambda_{01}(t) = r\{\lambda_{32}(t) - \lambda_{34}(t)\}, \quad t \geq 0, \ r > 0, \tag{2.1}$$

$$\lambda_{42}(t) = \lambda_{12}(t), \quad t \geq 0. \tag{2.2}$$

제약 조건 (2.1)은 '상태 H에서 F와 NF로 전이되는 강도의 차이와 상태 LTF에서 F와 NF로 전이되는 강도의 차이가 서로 비례적이다.'는 의미이다. 다시 말해 종말사건으로 전이되는 강도와 중간사건으로 전이되는 강도의 차이가 LTF 발생 이후에 없거나 ($r = 1$인 경우), 줄어들거나 ($r > 1$인 경우), 늘어난다 ($0 < r < 1$인 경우). 제약 조건 (2.2)는 'LTF의 발생 전후에 관계없이 상태 NF에서 F로 전이되는 강도가 같다.'는 의미이다. 본 논문에서는 공변량 $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)'$와 프레일티(혹은 랜덤효과) u가 주어졌을 때 L&Y 모형 (Lin과 Ying, 1994)을 써서 전이강도 $\lambda_{rs}(t)$에 대해 다음과 같은 모형을 가정하고자 한다.

$$\lambda_{rs}(t|\boldsymbol{x}, u) = \gamma \left(\alpha_{rs} \theta_{rs} t^{\theta_{rs}-1} + \boldsymbol{\beta}_{rs}' \boldsymbol{x} \right), \quad (r, s) \in \mathcal{A}. \tag{2.3}$$

단, $\alpha_{rs} (> 0)$과 $\theta_{rs} (> 0)$는 각각 와이블분포의 척도모수와 형상모수이고, $\boldsymbol{\beta}_{rs}$는 회귀계수 벡터이다. $\gamma = \exp(u)$는 로그정규(log-normal) 프레일티이고 u는 $N(0, \sigma^2)$을 따른다고 가정하자. 따라서 본 논문에서는 기저전이강도(baseline transition intensity)를 생존분석에서 가장 널리 쓰이는 분포

인 와이블분포의 위험률로 가정하였으며 프레일티는 전이강도에 대해 승법적으로 영향을 미친다고 가정하였다. 한편 모형 (2.3)에 포함된 모수들 중에서 $\alpha_{34}, \alpha_{42}, \theta_{34}, \theta_{42}, \boldsymbol{\beta}_{34}, \boldsymbol{\beta}_{42}$는 제약 조건 (2.1)과 (2.2)를 만족해야 하므로 추정해야 할 모수 벡터는 $\boldsymbol{\zeta} = (\alpha_{01}, \alpha_{02}, \alpha_{03}, \alpha_{12}, \alpha_{32}, \theta_{01}, \theta_{02}, \theta_{03}, \theta_{12}, \theta_{32}, \boldsymbol{\beta}_{01}, \boldsymbol{\beta}_{02}, \boldsymbol{\beta}_{03}, \boldsymbol{\beta}_{12}, \boldsymbol{\beta}_{32}, \sigma^2)'$이다.

시 구간 $(t_1, t_2]$에서 상태 0, 상태 1, 상태 3, 상태 4를 벗어날 누적전이강도함수(cumulative transition intensity function)는 각각 다음과 같다.

$$\begin{aligned}
H_0(t_1, t_2|\boldsymbol{x}, u) &= \int_{t_1}^{t_2} \{\lambda_{01}(s|\boldsymbol{x}, u) + \lambda_{02}(s|\boldsymbol{x}, u) + \lambda_{03}(s|\boldsymbol{x}, u)\} ds \\
&= \sum_{r=1}^{3} \gamma \left\{ \alpha_{0r}(t_2^{\theta_{0r}} - t_1^{\theta_{0r}}) + (\boldsymbol{\beta}_{0r}'\boldsymbol{x})(t_2 - t_1) \right\}, \\
H_1(t_1, t_2|\boldsymbol{x}, u) &= \int_{t_1}^{t_2} \lambda_{12}(s|\boldsymbol{x}, u) ds = \gamma \left\{ \alpha_{12}(t_2^{\theta_{12}} - t_1^{\theta_{12}}) + (\boldsymbol{\beta}_{12}'\boldsymbol{x})(t_2 - t_1) \right\}, \\
H_3(t_1, t_2|\boldsymbol{x}, u) &= \int_{t_1}^{t_2} \{\lambda_{32}(s|\boldsymbol{x}, u) + \lambda_{34}(s|\boldsymbol{x}, u)\} ds \\
&= \gamma \left\{ \alpha_{32}(t_2^{\theta_{32}} - t_1^{\theta_{32}}) + (\boldsymbol{\beta}_{32}'\boldsymbol{x})(t_2 - t_1) + \alpha_{34}(t_2^{\theta_{34}} - t_1^{\theta_{34}}) + (\boldsymbol{\beta}_{34}'\boldsymbol{x})(t_2 - t_1) \right\}, \\
H_4(t_1, t_2|\boldsymbol{x}, u) &= \int_{t_1}^{t_2} \lambda_{42}(s|\boldsymbol{x}, u) ds = \gamma \left\{ \alpha_{42}(t_2^{\theta_{42}} - t_1^{\theta_{42}}) + (\boldsymbol{\beta}_{42}'\boldsymbol{x})(t_2 - t_1) \right\}.
\end{aligned}$$

단, $\alpha_{34} = r^{-1}(\alpha_{01} - \alpha_{02}) + \alpha_{32}$, $\theta_{34} = \theta_{01} = \theta_{02} = \theta_{32}$, $\boldsymbol{\beta}_{34} = r^{-1}(\boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{02}) + \boldsymbol{\beta}_{32}$, $\alpha_{42} = \alpha_{12}$, $\theta_{42} = \theta_{12}$, $\boldsymbol{\beta}_{42} = \boldsymbol{\beta}_{12}$. 연구 시작 시점부터 연구 종료 시점까지 한 개체가 경험할 수 있는 경로를 나누면 여섯 가지, 즉 경로 1: $0 \to 0$, 경로 2: $0 \to 2$, 경로 3: $0 \to 1$, 경로 4: $0 \to 1 \to 2$, 경로 5: $0 \to 3$, 경로 6: $0 \to 3 \to 2$이다. 경로에 따라 우도함수를 정의하기 위해 몇 가지 기호를 정의하자. 연구 시작 시점부터 관심 있는 중간사건이 발생할 때까지의 시간, LTF가 발생할 때까지의 시간, 종말사건이 발생할 때까지의 시간을 각각 R, L, T라고 하자. 시점 s에서 여전히 상태 0에 있는 개체를 다음과 같이 나타내고,

$$\mathcal{H}_0(s) = \{R \wedge L \wedge T > s\},$$

시점 f에서 LTF가 발생한 개체가 시점 s에서 여전히 상태 3에 있는 개체를 다음과 같이 나타내자.

$$\mathcal{H}_{3,f}(s) = \{L = f, R \wedge T > s, f \leq s\}.$$

개체 i $(i = 1, 2, \ldots, n)$의 관심 있는 중간사건이 관측되지 않은 마지막 방문 시점과 관심 있는 중간사건이 관측된 최초 방문 시점을 각각 a_i, b_i라고 하고, 사망 시점이나 중도절단 시점을 t_i라고 하자. I_{ij}는 개체 i가 경로 j를 따라가면 1, 그렇지 않으면 0의 값을 갖는 지시함수(indicator function)라고 하자. 단, $j = 1, 2, \ldots, 6$. 경로 j를 따라가는 개체들의 집합을 $\mathcal{B}_j = \{i : I_{ij} = 1\}$이라고 하자. 개체 $i \in \mathcal{B}_1 \cup \mathcal{B}_2$인 경우는 시점 t_i 전까지 관심 있는 중간사건이 관측되지 않았기 때문에 $a_i, b_i \geq t_i$이고, 개체 $i \in \mathcal{B}_3 \cup \mathcal{B}_4$인 경우는 관심 있는 중간사건이 시점 a_i와 b_i 사이에서 관측되기 때문에 $a_i < b_i \leq t_i$이고, 개체 $i \in \mathcal{B}_5 \cup \mathcal{B}_6$인 경우는 시점 a_i에서 LTF가 발생했기 때문에 $a_i < t_i$이지만 $b_i < t_i$ 혹은 $b_i \geq t_i$이다. 이때, t_i는 개체 $i \in \mathcal{B}_1 \cup \mathcal{B}_3 \cup \mathcal{B}_5$이면 중도절단 시점이 되고, 개체 $i \in \mathcal{B}_2 \cup \mathcal{B}_4 \cup \mathcal{B}_6$이면 사망 시점이 된다. 따라서 경로 1과 2에 해당하는 개체의 우도함수 Q_1, Q_2는 각각 다음과 같다.

$$\begin{aligned}
Q_{i1}(t_i|\boldsymbol{x}_i, u_i) &= \Pr(R_i \wedge L_i \wedge T_i > t_i | \mathcal{H}_0(0), \boldsymbol{x}_i, u_i) \\
&= \exp\{-H_0(0, t_i|\boldsymbol{x}_i, u_i)\}, \quad i \in \mathcal{B}_1. \qquad (2.4)
\end{aligned}$$

$$Q_{i2}(t_i|\boldsymbol{x}_i,u_i) = \Pr(T=t_i, R\wedge L > t_i|\mathcal{H}_0(0),\boldsymbol{x}_i,u_i)$$
$$= Q_{i1}(t_i|\boldsymbol{x}_i,u_i)\lambda_{02}(t_i|\boldsymbol{x}_i,u_i), \quad i\in\mathcal{B}_2. \tag{2.5}$$

경로 3과 4에 해당하는 개체의 우도함수 Q_3^*, Q_4^*는 각각 다음과 같다.

$$\begin{aligned} Q_{i3}^*(a_i,b_i,t_i|\boldsymbol{x}_i,u_i) &= \Pr\left(R_i\in(a_i,b_i], L_i>t_i, T_i>t_i|\mathcal{H}_0(0),\boldsymbol{x}_i,u_i\right)\\ &= \exp\{-H_0(0,a_i|\boldsymbol{x}_i,u_i)\}\\ &\quad\times\int_{a_i}^{b_i}\exp\{-H_0(a_i,s|\boldsymbol{x}_i,u_i)\}\lambda_{01}(s|\boldsymbol{x}_i,u_i)\exp\{-H_1(s,b_i|\boldsymbol{x}_i,u_i)\}ds\\ &\quad\times\exp\{-H_1(b_i,t_i|\boldsymbol{x}_i,u_i)\}, \quad i\in\mathcal{B}_3. \end{aligned} \tag{2.6}$$

$$\begin{aligned} Q_{i4}^*(a_i,b_i,t_i|\boldsymbol{x}_i,u_i) &= \Pr(R_i\in(a_i,b_i], L_i>t_i, R_i<T_i=t_i|\mathcal{H}_0(0),\boldsymbol{x}_i,u_i)\\ &= Q_3^*(a_i,b_i,t_i|\boldsymbol{x}_i,u_i)\lambda_{12}(t_i|\boldsymbol{x}_i,u_i), \quad i\in\mathcal{B}_4. \end{aligned} \tag{2.7}$$

식 (2.6)과 (2.7)은 집합 $\mathcal{B}_3\cup\mathcal{B}_4$에 속하는 개체 i의 관심 있는 중간사건이 구간 $(a_i,b_i]$의 모든 시점에서 균등하게 발생할 수 있다는 가정 하에서 얻어진 결과이다 (Barrett 등, 2011). 그러나 본 논문에서는 구간 $(a_i,b_i]$를 관심 있는 중간사건이 발생할 수 있는 부구간으로 분할한 후 부구간 별로 가중치를 구해 우도함수를 정의하고자 한다 (Collett, 2015).

- 경로 3이나 4에 속하는 개체들의 중간사건의 발생 구간을 $R_{i'}\in(a_{i'},b_{i'}]$로 하자. 단, $i'\in\mathcal{B}_3\cup\mathcal{B}_4$. 집합 $\mathcal{B}_3\cup\mathcal{B}_4$에 속하는 개체들의 $b_{i'}$ 중에서 가장 작은 $b_{i'}$을 s_1으로 정의하자. 그 다음 s_1보다 크거나 같은 $a_{i'}$을 가진 개체들 중 가장 작은 $b_{i'}$을 s_2로 정의하자. s_m $(m=1,2,\ldots)$보다 크거나 같은 $a_{i'}$을 가진 개체가 없을 때까지 이와 같은 과정을 반복하여 다음과 같은 시점들을 얻었다고 가정하자.

$$0=s_0<s_1<s_2<\cdots<s_l<s_{l+1}=\infty.$$

- 집합 $\mathcal{B}_3\cup\mathcal{B}_4$에 속하는 개체 i'의 시점 s_m $(m=1,\ldots,l)$에서의 가중치 $w_{i'm}$을 다음과 같이 정의하자.

$$w_{i'm} = \frac{d_{i'm}\exp\left\{-H_0(0,s_m|\boldsymbol{x}_{i'},u_{i'})\right\}\lambda_{01}(s_m|\boldsymbol{x}_{i'},u_{i'})}{\sum_{m'=1}^{l}d_{i'm'}\exp\left\{-H_0(0,s_{m'}|\boldsymbol{x}_{i'},u_{i'})\right\}\lambda_{01}(s_{m'}|\boldsymbol{x}_{i'},u_{i'})}. \tag{2.8}$$

단, $d_{i'm}$은 s_m이 구간 $(a_{i'},b_{i'}]$에 포함되는지 유무를 나타내는 지시자, 즉 $d_{i'm}=I(s_m\in(a_{i'},b_{i'}])$이다. 따라서 본 논문에서는 부구간 별 가중치 (2.8)을 써서 경로 3과 4에 해당하는 개체의 우도함수 Q_3, Q_4를 각각 다음과 같이 정의하고자 한다.

$$\begin{aligned} Q_{i3}(a_i,b_i,t_i|\boldsymbol{x}_i,u_i) &= \exp\{-H_0(0,a_i|\boldsymbol{x}_i,u_i)\}\\ &\quad\times\sum_{m=1}^{l}[d_{im}w_{im}\exp\{-H_0(a_i,s_m|\boldsymbol{x}_i,u_i)\}\lambda_{01}(s_m|\boldsymbol{x}_i,u_i)\exp\{-H_1(s_m,b_i|\boldsymbol{x}_i,u_i)\}\\ &\quad\times\exp\{-H_1(s_m,b_i|\boldsymbol{x}_i,u_i)\}]\exp\{-H_1(b_i,t_i|\boldsymbol{x}_i,u_i)\}\\ &= \sum_{m=1}^{l}d_{im}w_{im}\exp\{-H_0(0,s_m|\boldsymbol{x}_i,u_i)\}\lambda_{01}(s_m|\boldsymbol{x}_i,u_i)\exp\{-H_1(s_m,t_i|\boldsymbol{x}_i,u_i)\},\\ &\quad i\in\mathcal{B}_3. \end{aligned} \tag{2.9}$$

$$Q_{i4}(a_i,b_i,t_i|\boldsymbol{x}_i,u_i) = Q_3(a_i,b_i,t_i|\boldsymbol{x}_i,u_i)\lambda_{12}(t_i|\boldsymbol{x}_i,u_i), \quad i\in\mathcal{B}_4. \tag{2.10}$$

마지막으로 경로 5와 6에 해당하는 개체의 우도함수 Q_5, Q_6은 각각 다음과 같다.

$$\begin{aligned} Q_{i5}(a_i, b_i, t_i|\boldsymbol{x}_i, u_i) &= \Pr(R_i \wedge T_i > t_i|\mathcal{H}_{3,a_i}(a_i), \boldsymbol{x}_i, u_i) + \Pr(R_i \in (a_i, t_i], T_i > t_i|\mathcal{H}_{3,a_i}(a_i), \boldsymbol{x}_i, u_i) \\ &= \exp\{-H_0(0, a_i|\boldsymbol{x}_i, u_i)\}\lambda_{03}(a_i|\boldsymbol{x}_i, u_i)\Bigg[\exp\{-H_3(a_i, t_i|\boldsymbol{x}_i, u_i)\} \\ &\quad + \int_{a_i}^{t_i} \exp\{-H_3(a_i, s|\boldsymbol{x}_i, u_i)\}\lambda_{34}(s|\boldsymbol{x}_i, u_i)\exp\{-H_4(s, t_i|\boldsymbol{x}_i, u_i)\}ds\Bigg], \\ &\quad i \in \mathcal{B}_5. \end{aligned} \tag{2.11}$$

$$\begin{aligned} Q_{i6}(a_i, b_i, t_i|\boldsymbol{x}_i, u_i) &= \Pr(R_i > T_i, T_i = t_i|\mathcal{H}_{3,a_i}(a_i), \boldsymbol{x}_i, u_i) \\ &\quad + \Pr(R_i \in (a_i, t_i], R_i < T_i = t_i|\mathcal{H}_{3,a_i}(a_i), \boldsymbol{x}_i, u_i) \\ &= \exp\{-H_1(0, a_i|\boldsymbol{x}_i, u_i)\}\lambda_{03}(a_i|\boldsymbol{x}_i, u_i)\Bigg[\exp\{-H_3(a_i, t_i|\boldsymbol{x}_i, u_i)\}\lambda_{32}(t_i|\boldsymbol{x}_i, u_i) \\ &\quad + \left\{\int_{a_i}^{t_i} \exp\{-H_3(a_i, s|\boldsymbol{x}_i, u_i)\}\lambda_{34}(s|\boldsymbol{x}_i, u_i)\exp\{-H_4(s, t_i|\boldsymbol{x}_i, u_i)\}ds\right\} \\ &\quad \times \lambda_{42}(t_i|\boldsymbol{x}_i, u_i)\Bigg], \quad i \in \mathcal{B}_6. \end{aligned} \tag{2.12}$$

따라서 식 (2.4)–(2.5), (2.9)–(2.11)에 의해 우도함수는 다음과 같이 주어진다.

$$L(\boldsymbol{\zeta}) = \prod_{i=1}^{n}\left\{\prod_{j=1}^{6} Q_{ij}^{I_{ij}}\right\}\phi\left(0, \sigma^2; u_i\right). \tag{2.13}$$

단, $\phi(\cdot)$는 평균이 0이고, 분산이 σ^2인 정규분포의 확률밀도함수이다.

본 논문에서는 SAS NLMIXED procedure를 써서 모수 $\boldsymbol{\zeta}$를 추론하고자 한다. 이를 위해 주변우도(marginal likelihood)를 다음과 같이 정의하고,

$$m(\boldsymbol{\zeta}) = \int \cdots \int L(\boldsymbol{\zeta})du_1 \cdots du_n,$$

함수 $f(\boldsymbol{\zeta}) = -\log m(\boldsymbol{\zeta})$를 가장 작게 하는 $\boldsymbol{\zeta}$를 구하고 ($\hat{\boldsymbol{\zeta}}$), 그 값에서 계산한 헤시안(Hessian) 행렬의 역행렬을 $\hat{\boldsymbol{\zeta}}$의 공분산 행렬의 추정값으로 정의한다. 주변우도를 정의하기 위해서는 프레일티 분포에 대한 적분이 요구되는데 본 논문에서는 Pinheiro와 Bates (1995)가 제안한 조정중요표본추출법(adaptive importance sampling)을 사용하고자 한다. 또한 최적해 $\hat{\boldsymbol{\zeta}}$를 얻기 위해 목적함수 $f(\boldsymbol{\zeta})$의 일차미분치인 경사벡터(gradient vector)와 이차미분치인 헤시안 행렬을 이용하는 반복유사뉴튼(iterative quasi-Newton)방법을 사용하고자 한다.

3. 모의실험

2절에서 제안한 추정량의 소표본 성질을 살펴보기 위해 모의실험을 수행하였다. 모형 (2.3)에서 기저전이강도는 $\theta_{rs} = 1$인 와이블분포를 가정하였으며 프레일티는 $\gamma = \exp(u)$인 로그정규분포를 가정하였다. 단, $u \sim N(0, \sigma^2)$. 공변량 x는 $B(1, 0.5)$를 따른다고 가정하였다. 중도절단 시간은 $C = 181$로 고정하였다. 표본의 크기 n은 200으로 고정하였고 500번 반복 수행하였다. i ($i = 1, 2, \ldots, n$)번째 개체의 모의실험 자료는 다음 절차에 따라 생성하였다.

- 단계0: 관측 주기를 격주로 하여 6개월 동안 중간사건의 발생 유무를 12번, 즉 15, 31,. . ., 166, 181일에 관측하였다. 그러나 실제 관측 시점은 계획된 시점과 다를 수 있기 때문에 평균이 0이고, 표준

Table 3.1. Regression parameters in the different scenarios of the simulation study

Scenario	β_{01}	β_{02}	β_{12}	β_{03}	β_{32}
1	0.004	0.004	0.004	0.006	0.006
2	0.004	0.006	0.006	0.006	0.006
3	0.004	0.004	0.006	0.006	0.006
4	0.004	0.006	0.008	0.006	0.006
5	0.004	0.004	0.002	0.006	0.006
6	0.004	0.006	0.004	0.006	0.006

편차가 5인 정규분포에서 생성한 난수를 계획된 시점에 더하여 중간사건의 실제 관측 시점, 즉 $0 = l_0 < l_{1i} < \cdots < l_{11,i} < l_{12} = 181$로 정의하였다. 공변량 x_i는 $B(1, 0.5)$에서 생성하였으며, 로그 정규 프레일티는 평균이 0이고, 분산이 0.1인 정규분포에서 u_i를 생성하여 $\gamma_i = \exp(u_i)$로 정의하였다. $u_{01i}, u_{02i}, u_{03i}$를 각각 $U(0,1)$에서 생성한 난수라고 하자. R_i, T_i, L_i는 각각 s에 대한 방정식 $\Lambda_{01}(s|x_i, u_i) + \log(1-u_{01i}) = 0$, $\Lambda_{02}(s|x_i, u_i) + \log(1-u_{02i}) = 0$, $\Lambda_{03}(s|x_i, u_i) + \log(1-u_{03i}) = 0$의 해로 정의하자. 단, $\Lambda_{0j}(s|x_i, u_i) = \gamma_i\{(\beta_{0j}x_i)s + \alpha_{0j}s^{\theta_{0j}}\}$, $j = 1, 2, 3$.

- 단계1: 만일 $C \le R_i \wedge T_i \wedge L_i$이면 i번째 개체는 중간사건을 경험하지 않고 시점 C에서 중도절단된 것으로 정의하였고($i \in \mathcal{B}_1$), 만일 $T_i = R_i \wedge T_i \wedge L_i$이면 i번째 개체는 중간사건을 경험하지 않고 T_i에서 사망한 것으로 정의하였다($i \in \mathcal{B}_2$). 그러나 만일 $R_i = R_i \wedge T_i \wedge L_i$이면 단계2로, 만일 $L_i = R_i \wedge T_i \wedge L_i$이면 단계3으로 이동한다.
- 단계2: u_{12i}를 $U(1 - \exp\{\Lambda_{12}(R_i|x_i, u_i)\}, 1)$에서 생성한 난수라고 하자. 단, $\Lambda_{12}(s|x_i, u_i) = \gamma_i\{(\beta_{12}x_i)s + \alpha_{12}s^{\theta_{12}}\}$. T_i를 s에 대한 방정식 $\Lambda_{12}(s|x_i, u_i) + \log(1 - u_{12i}) = 0$의 해로 재정의하자. 만일 $C \le T_i$이면 i번째 개체는 중간사건을 경험하고 시점 C에서 중도절단된 것으로 정의하였고($i \in \mathcal{B}_3$), 그렇지 않으면 i번째 개체는 중간사건을 경험하고 시점 T_i에서 사망한 것으로 정의하였다($i \in \mathcal{B}_4$).
  - 만일 $R_i \in (0, l_{1i})$이면 $a_i = 0$, $b_i = l_{1i}$로 놓았고, 만일 $R_i \in (l_{k-1,i}, l_{ki})$이면 $a_i = l_{k-1,i}$, $b_i = l_{ki}$로 놓았다. 단, $k = 2, 3, \ldots, 11$.
  - 그러나 만일 $R_i \in (l_{11,i}, C)$이면 마지막 관측 시점 이전에 중간사건이 발생하지 않았기 때문에 경로의 유형을 재정의해야 한다. 만일 $C \le T_i$이면 i번째 개체는 중간사건을 경험하지 않고 시점 C에서 중도절단된 것으로 재정의하였고 ($i \in \mathcal{B}_1$), 그렇지 않으면 i번째 개체는 중간사건을 경험하지 않고 T_i에서 사망한 것으로 재정의하였다 ($i \in \mathcal{B}_2$).
- 단계3: u_{32i}와 u_{34i}를 각각 $U(1 - \exp\{\Lambda_{32}(L_i|x_i, u_i)\}, 1)$와 $U(1 - \exp\{\Lambda_{34}(L_i|x_i, u_i)\}, 1)$에서 생성한 난수라고 하자. 단, $\Lambda_{3j}(s|x_i, u_i) = \gamma_i\{(\beta_{3j}x_i)s + \alpha_{3j}s^{\theta_{3j}}\}$, $j = 2, 4$. R_i와 T_i를 각각 s에 대한 방정식 $\Lambda_{32}(s|x_i, u_i) + \log(1 - u_{32i}) = 0$와 $\Lambda_{34}(s|x_i, u_i) + \log(1 - u_{34i}) = 0$의 해로 재정의하자. 만일 $C \le R_i \wedge T_i$이면 i번째 개체는 LTF 후에 중간사건을 경험하지 않고 시점 C에서 중도절단된 것으로 정의하였고 ($i \in \mathcal{B}_5$), 만일 $T_i \le R_i$이면 i번째 개체는 LTF 후에 중간사건을 경험하지 않고 T_i에서 사망한 것으로 정의하였다 ($i \in \mathcal{B}_6$). 그러나 만일 $R_i < T_i$이면 단계4로 이동한다.
- 단계4: u_{42i}를 $U(1 - \exp\{\Lambda_{42}(R_i|x_i, u_i)\}, 1)$에서 생성한 난수라고 하자. 단, $\Lambda_{42}(s|x_i, u_i) = \gamma_i\{(\beta_{42}x_i)s + \alpha_{42}s^{\theta_{42}}\}$. T_i를 s에 대한 방정식 $\Lambda_{42}(s|x_i, u_i) + \log(1 - u_{42i}) = 0$의 해로 재정의하자. 만일 $C \le T_i$이면 i번째 개체는 LTF 후에 중간사건을 경험하고 시점 C에서 중도절단된 것으로 정의하였고($i \in \mathcal{B}_5$), 그렇지 않으면 i번째 개체는 중간사건을 경험하고 시점 T_i에서 사망한 것으로 정의하였다($i \in \mathcal{B}_6$).

Table 3.2. Scale parameters of the weibull distribution in the different death rates of the simulation study

Death rate	α_{01}	α_{02}	α_{12}	α_{03}	α_{32}
Low	0.002	0.001	0.0005	0.004	0.001
Modertae	0.002	0.002	0.0010	0.004	0.002
High	0.002	0.003	0.0015	0.004	0.003

Table 3.3. Average percentage of death event rates in the different scenarios of the simulation study based on 500 data sets of 200 subjects when $x \sim B(1, 0.5)$ and $\theta_{01} = \theta_{02} = \theta_{12} = \theta_{03} = \theta_{32} = 1$

Scenario	Route	Low	Moderate	High
1	2	17	23	29
	4	6	7	8
	6	15	17	19
	2 or 4 or 6	**13**	**16**	**18**
2	2	21	27	32
	4	7	8	8
	6	15	17	18
	2 or 4 or 6	**14**	**17**	**19**
3	2	17	24	29
	4	8	8	8
	6	16	18	19
	2 or 4 or 6	**14**	**17**	**19**
4	2	21	27	32
	4	8	9	9
	6	16	17	19
	2 or 4 or 6	**15**	**18**	**20**
5	2	17	23	29
	4	5	5	6
	6	14	16	18
	2 or 4 or 6	**12**	**15**	**17**
6	2	20	26	32
	4	6	7	7
	6	14	16	18
	2 or 4 or 6	**14**	**16**	**19**

본 모의실험에서는 회귀계수에 대해 여섯 가지 시나리오를 고려하였고 (Table 3.1), 기저전이강도에 대해 세 가지 모형(Low, Moderate, High)을 고려하였다 (Table 3.2). 사망 위험률에 대한 공변량 효과를 시나리오 별로 비교해 보면, 시나리오 1과 2는 중간사건의 발생 유무와 관계없이 일정한 경우이고, 3과 4는 중간사건이 발생한 이후에 증가하는 경우이고, 5와 6은 중간사건이 발생한 이후에 감소하는 경우이다. 또한 중간사건의 발생률과 사망 위험률에 대한 공변량의 효과를 시나리오 별로 비교해 보면, 시나리오 1, 3, 5는 서로 같은 경우이며, 2, 4, 6은 전자보다 후자가 큰 경우이다. Table 3.3은 회귀계수에 대한 시나리오와 사망 위험률의 조합에 따른 모의실험 자료의 평균 사망률을 경로 2, 4, 6 별로 정리한 것이다. Table 3.3을 보면 시나리오에 관계없이 사망 위험률이 'Low'에서 'High'로 변함에 따라 실험적 평균 사망률이 대체로 증가하였다. 2절에서 언급한 것처럼 주변우도함수에 대한 근사는 (즉 프레일티 분포에 대한 적분은) SAS NLMIXED procedure에서 제공하는 adaptive importance sampling 방법을 사용하였고, 최적해는 iterative quasi-Newton 방법을 사용하였으며 최대반복횟수는 500번, 상대경사수렴기준(relative gradient convergence criterion)은 10^{-8}으로 정하였다. Table 3.4는 회귀계수

Table 3.4. Empirical results of the proposed method in terms of the average of the relative bias (R.Bias), the standard errors (SEM), and the coverage probability (CP) based on 500 data sets of 200 subjects when $x \sim B(1, 0.5)$ and $\theta_{01} = \theta_{02} = \theta_{12} = \theta_{03} = \theta_{32} = 1$

			Low				Moderate				High			
Scenario	Param.	True	R.Bias (%)	SD ($\times 10^5$)	SEM ($\times 10^5$)	CP (%)	R.Bias (%)	SD ($\times 10^5$)	SEM ($\times 10^5$)	CP (%)	R.Bias (%)	SD ($\times 10^5$)	SEM ($\times 10^5$)	CP (%)
1	β_{01}	0.004	−8.1	109	116	94.0	−10.3	122	119	90.6	−10.1	115	121	93.2
	β_{02}	0.004	2.7	103	107	94.6	2.5	122	122	94.6	3.1	131	137	96.6
	β_{03}	0.006	−2.8	170	173	96.2	−3.3	178	175	92.8	−5.1	187	176	92.2
	β_{12}	0.004	−10.9	111	112	91.0	−9.7	131	127	92.2	−7.0	142	149	96.2
	β_{32}	0.006	4.5	148	157	95.8	9.8	173	179	96.6	6.3	192	196	96.4
	σ^2	0.100	102.0	16935	12758	90.0	84.9	14554	11999	91.8	65.5	13395	11531	92.6
2	β_{01}	0.004	−10.9	117	118	91.4	−10.5	117	123	92.6	−12.9	123	124	93.0
	β_{02}	0.006	−0.4	114	124	95.6	−0.7	140	139	94.4	0.7	159	153	93.4
	β_{03}	0.006	−2.0	183	176	92.8	−1.9	172	179	96.2	−3.2	185	181	92.4
	β_{12}	0.006	−9.5	143	150	91.0	−6.7	182	167	90.6	−5.6	186	186	94.4
	β_{32}	0.006	6.8	150	154	96.2	7.9	196	177	95.0	9.1	200	194	96.8
	σ^2	0.100	82.2	16444	12239	89.6	75.5	17585	11683	86.4	62.0	16508	11178	87.2
3	β_{01}	0.004	−13.5	103	113	92.4	−13.1	115	118	91.0	−13.2	124	120	89.6
	β_{02}	0.004	1.8	99	105	95.8	1.1	118	120	96.2	0.3	139	135	95.8
	β_{03}	0.006	−4.4	185	170	93.0	−2.4	181	173	93.0	−3.4	178	176	95.0
	β_{12}	0.006	−10.3	146	141	88.4	−7.6	153	158	91.8	−4.8	159	174	95.6
	β_{32}	0.006	7.1	143	155	96.6	6.6	193	177	95.4	7.8	202	196	94.4
	σ^2	0.100	112.8	19191	12698	88.6	90.9	17250	12046	87.6	62.9	15384	11143	86.8
4	β_{01}	0.004	−12.7	117	117	90.6	−13.2	117	120	89.4	−12.3	131	125	89.8
	β_{02}	0.006	1.9	130	125	94.8	0.1	134	137	95.2	3.2	153	155	97.6
	β_{03}	0.006	−5.6	176	174	93.4	−6.7	184	177	90.8	−2.5	191	183	94.2
	β_{12}	0.008	−11.7	167	179	91.0	−6.4	195	200	94.6	−9.5	216	215	90.8
	β_{32}	0.006	8.0	170	156	94.0	8.4	185	175	96.2	8.6	193	195	96.8
	σ^2	0.100	81.2	15998	12026	92.4	88.7	16223	12005	89.2	80.3	15959	11828	92.8
5	β_{01}	0.004	−3.9	110	118	95.8	−6.8	118	119	92.8	−5.2	122	124	94.6
	β_{02}	0.004	2.7	104	108	96.4	−0.1	125	121	95.2	5.0	149	139	95.8
	β_{03}	0.006	0.0	178	174	95.2	−1.6	175	174	96.2	−0.9	167	180	96.4
	β_{12}	0.002	−11.6	81	82	94.2	−12.1	97	100	95.6	−7.4	122	120	95.2
	β_{32}	0.006	2.6	166	157	95.2	5.2	187	177	93.2	4.4	219	198	93.8
	σ^2	0.100	100.3	21007	13223	89.4	100.3	21007	13223	89.4	72.1	15134	11907	90.2
6	β_{01}	0.004	−8.1	129	121	92.4	−7.0	127	125	92.6	−9.6	129	126	92.4
	β_{02}	0.006	−1.9	127	124	94.6	−1.6	134	140	96.0	−1.9	156	152	94.0
	β_{03}	0.006	−0.1	182	179	94.8	0.4	179	183	95.8	−2.7	187	183	94.2
	β_{12}	0.004	−7.8	114	120	93.2	−8.1	124	134	94.0	−6.0	142	155	95.4
	β_{32}	0.006	3.6	157	153	95.8	5.0	171	174	95.6	6.2	180	191	97.0
	σ^2	0.100	77.4	17865	12317	90.4	97.5	19692	12513	86.8	69.4	15080	11906	93.0

에 대한 시나리오와 사망 위험률의 조합에 따른 회귀계수와 프레일티 분포의 분산 모수에 대한 추정량의 상대편향비율(R.Bias), 추정량의 표준편차(SD), 추정량의 표준오차의 평균(SEM), 추정량의 95%신뢰구간 포함률(CP)을 정리한 것이다. 시니리오를 고정했을 때 사망 위험률이 'Low'에서 'High'로 증가함에 따라 SEM이 단조적으로 증가하는 경향을 보였지만 R.Bias와 CP에서는 별다른 경향을 찾아볼 수

Table 3.5. Sensitivity analysis of the proposed method depending on the determination of the ratio r in terms of the average of the relative bias (R.Bias), the standard errors (SEM), and the coverage probability (CP) based on 500 data sets of 200 subjects when $x \sim B(1, 0.5)$ and $\theta_{01} = \theta_{02} = \theta_{12} = \theta_{03} = \theta_{32} = 1$

Param.	True	r											
		0.5			0.75			1.5			2		
		R.Bias (%)	SEM ($\times 10^5$)	CP (%)	R.Bias (%)	SEM ($\times 10^5$)	CP (%)	R.Bias (%)	SEM ($\times 10^5$)	CP (%)	R.Bias (%)	SEM ($\times 10^5$)	CP (%)
β_{01}	0.004	−13.1	118	90.4	−13.5	118	90.2	−13.3	118	91.6	−14.2	118	91.0
β_{02}	0.004	1.4	121	97.2	0.0	120	95.8	1.4	121	96.2	3.5	121	95.4
β_{03}	0.006	−1.3	174	91.4	−1.4	174	92.6	−1.8	174	92.4	−2.6	173	92.0
β_{12}	0.006	−8.2	158	90.8	−9.0	157	91.4	−8.5	158	91.4	−9.6	157	90.6
β_{32}	0.006	7.0	178	95.2	7.4	178	93.6	8.5	179	95.2	7.2	177	95.2
σ^2	0.100	89.8	12307	92.4	100.8	12424	89.2	100.4	12472	86.4	91.7	12207	88.2

Table 3.6. Sensitivity analysis of the proposed method depending on the underlying frailty distribution in terms of the average of the relative bias (R.Bias), the standard errors (SEM), and the coverage probability (CP) based on 500 data sets of 200 subjects when $x \sim B(1, 0.5)$ and $\theta_{01} = \theta_{02} = \theta_{12} = \theta_{03} = \theta_{32} = 1$

Param.	True	$N(0, 0.2)$			$U(-0.775, 0.775)$			$DE(0.316)$			$G(5.483, 0.2)$		
		R.Bias (%)	SEM ($\times 10^5$)	CP (%)	R.Bias (%)	SEM ($\times 10^5$)	CP (%)	R.Bias (%)	SEM ($\times 10^5$)	CP (%)	R.Bias (%)	SEM ($\times 10^5$)	CP (%)
β_{01}	0.004	−14.3	119	86.0	−14.3	119	90.4	−13.5	118	89.4	−15.3	119	88.2
β_{02}	0.004	3.5	123	97.2	3.4	124	95.8	1.8	122	92.8	1.0	123	96.4
β_{03}	0.006	−7.2	175	91.2	−4.3	176	95.0	−6.4	174	92.0	−7.4	174	90.0
β_{12}	0.006	−8.7	160	92.6	−6.0	161	91.2	−6.3	162	93.0	−6.3	161	95.2
β_{32}	0.006	6.8	182	95.2	9.7	184	95.8	6.9	181	96.4	9.6	183	96.6
σ^2	0.200	29.8	13401	83.2	27.8	13288	86.0	21.2	13005	81.2	24.8	13142	83.4

없었다. 그러나 사망 위험률을 고정했을 때 시나리오에 따른 별다른 경향은 없었다. 몇 가지 경우를 제외하고 회귀계수에 대한 CP는 명목값인 0.95에 가까운 것으로 나타났다. 그러나 프레일티의 분산 모수에 대한 CP는 명목값보다 작은 것으로 나타났다.

한편 본 논문에서 제안한 추정량이 제약 조건 식 (2.1)에 포함된 미지의 상수 r의 값과 프레일티 분포의 오명세(mis-specification)에 얼마나 민감한지를 살펴보기 위해 모의실험을 수행하였다. 이를 위해 회귀계수에 대한 시나리오는 'Scenario 3', 사망 위험률은 'Moderate'인 경우로 고정하였다. Table 3.5는 실제 r의 값이 각각 0.5, 0.75, 1.5, 2일 때 r의 값을 1로 놓고 추정했을 때의 추정량의 결과를 정리한 것이다. 처음 두 경우는 종말사건으로 전이되는 강도와 중간사건으로 전이되는 강도의 차이가 LTF 발생 이후에 늘어나는 경우이고, 마지막 두 경우는 반대로 LTF 발생 이후에 줄어드는 경우이다. 실제 r의 값이 1일 때의 결과 (Table 3.4 참조)와 비교해 볼 때 별다른 경향을 찾아볼 수 없었다. 따라서 제안한 추정량은 미지의 상수 r의 값에 대해 로버스트함을 알 수 있었다. Table 3.6은 실제 프레일티 분포가 각각 $U(-0.775, 0.775)$, 이중지수분포 $DE(0.316)$, 감마분포 $G(5.483, 0.2)$를 따를 때 프레일티의 분포를 정규분포로 놓고 모수를 추정했을 때 추정량의 결과를 정리한 것이다. $U(-0.775, 0.775)$와 $DE(0.316)$은 정규분포처럼 대칭적인 반면에 전자는 정규분포보다 꼬리가 짧은 분포이고 후자는 꼬리가 두터운 분포이다. $G(5.483, 0.2)$는 정규분포와 다르게 비대칭적이다. 실제 프레일티의 분포가 정규분포 $N(0, 0.2)$일 때의 결과와 비교해 보면 프레일티의 분산 모수에 대한 CP 값이 명목값보다 작은 경향이 있었지만 회귀계수에 대한 R.Bias와 CP 값은 별다른 경향이 없었다. 따라서 제안한 추정량은 미지의 프레일티 분포에 대해 로버스트함을 알 수 있었다.

Table 4.1. Regression parameter estimates (Est), their standard errors (SE), and p-values (P) according to a choice of r

Covariate	ζ	r											
		0.5			1			1.5			2		
		Est ($\times 10^4$)	SE ($\times 10^4$)	P	Est ($\times 10^4$)	SE ($\times 10^4$)	P	Est ($\times 10^4$)	SE ($\times 10^4$)	P	Est $\times 10^4$	SE ($\times 10^4$)	P
Sex	β_{011}	−7.7	3.3	0.028	−7.7	3.3	0.028	−7.6	3.3	0.028	−7.6	3.3	0.028
	β_{021}	19.5	5.4	0.001	19.6	5.4	0.001	19.5	5.4	0.001	19.5	5.4	0.001
	β_{031}	−5.4	3.7	0.151	−5.4	3.7	0.152	−5.4	3.7	0.151	−5.4	3.7	0.151
	β_{121}	466.4	301.5	0.132	471.1	301.5	0.128	470.4	301.1	0.128	470.7	301.0	0.128
	β_{321}	42.0	89.8	0.644	30.7	90.2	0.736	26.3	90.2	0.772	24.3	90.3	0.789
Certificate	β_{012}	−6.3	3.6	0.088	−6.3	3.6	0.089	−6.3	3.6	0.089	−6.3	3.6	0.089
	β_{022}	1.4	6.0	0.819	1.4	6.0	0.817	1.4	6.0	0.822	1.4	6.0	0.823
	β_{032}	−0.6	4.3	0.887	−0.6	4.3	0.891	−0.6	4.3	0.887	−0.6	4.3	0.887
	β_{122}	−351.1	304.9	0.258	−346.5	305.2	0.265	−346.1	304.9	0.265	−345.5	304.9	0.266
	β_{322}	94.5	119.3	0.434	93.4	120.0	0.442	92.9	120.0	0.445	92.6	120.1	0.446
	σ^2	60.8	65.5	0.360	66.7	68.8	0.339	60.6	65.1	0.359	60.2	64.7	0.360
$-2\log L$		13152			13158			13155			13158		
AIC		13184			13190			13187			13190		

4. 실제 예

Persones Agées Quid (PAQUID) 자료는 치매가 사망에 미치는 영향을 알아보기 위해 프랑스 남서부에 위치한 두 지역(Gironde, Dordogne)에 거주하고 있는 65세 이상의 노인들을 대상으로 1988–1990년까지 2–3년 간격으로 인구사회학적 특성과 정신건강상태 등에 대해 조사한 것이다 (Helmer 등, 2001). 3,675명의 참가자 중에서 연구 기간 동안 832명이 치매로 진단 받았으며(22.6%) 그중에서 639명이 사망하였고(76.8%), 치매로 진단 받지 않은 2,843명 중에서는 2,298명이 사망하였다(80.8%). 그러나 PAQUID 자료를 직접 얻을 수 없어 본 논문에서는 PAQUID 자료에서 1000개를 임의로 추출하여 R 패키지 'SmoothHazard'에 수록한 'Paq1000' 자료를 분석하였다 (Touraine 등, 2014). 이 자료는 치매 진단 유무와 사망 유무, 연구 시작 시 연령, 치매로 진단 받지 않은 마지막 방문 시 연령, 치매로 진단 받은 연령(치매로 진단 받은 경우만 해당), 사망 시간이나 중도절단 시간, 성별, 초등교육 이수 유무 등으로 이루어져 있다. 본 논문에서는 치매 진단 유무와 사망 유무를 각각 중간사건과 종말사건으로 정의하였다. 또한 치매로 진단 받지 않은 마지막 방문 시점부터 4년 이상 추적되지 않았으면 그 시점에서 LTF가 발생한 것으로 정의하였다. 1,000명 중에서 231명이 LTF 되었으며 그중에서 159명이 사망하였다(68.8%). 치매로 진단 받은 186명 중에서 127명이 사망하였고(68.3%), 치매로 진단 받지 않은 583명 중에서 438명이 사망하였다(75.1%). 연구 시작 연령이 같은 개체들은 묶어 군집으로 정의하였으며(군집 갯수 = 33, 군집의 크기 = 1–8개) 군집 내 개체 간의 연관성은 로그정규 프레일티를 가정하였다.

Table 4.1은 제약 조건 (2.1)에 포함된 미지의 상수 r의 값에 따라 회귀계수의 추정값(Est)과 표준오차(SE), 유의확률(P)을 정리한 결과이다. 이때, 공변량으로 성별(sex)과 초등교육 이수 유무(certificate)를 고려하였다. r의 값에 따라 Est와 SE, P의 값이 모두 비슷하였으며 $-2\log L$과 AIC의 값은 $r = 0.5$일 때 가장 작은 것으로 나타났다. 이에 대응하는 결과를 중심으로 살펴보면, 상태 H에서 상태 LTF로 전이되는($0 \to 3$) 강도는 남성보다 여성이 인구 100,000명 당 54명($P = 0.151$), 초등교육을 이수하지 못한 사람보다 이수한 사람이 6명($P = 0.887$) 많았으나 모두 통계적으로 유의하지는 않았다. LTF 이후 사망으로 전이되는($3 \to 2$) 강도는 여성보다 남성이, 초등교육을 이수한 사람보다 이수하지 못한 사람이 각각 인구 100,000명 당 420, 945명 많은 것으로 나타났지만 모두 통계적으로 유의하지 않았다(각각 $P = 0.644, 0.434$). 상태 H에서 상태 NF, 즉 치매로 진단 받은 상태로 전이되는($0 \to 1$)

강도는 남성보다 여성이 많았고(인구 100,000명 당 77명), 초등교육을 이수하지 못한 사람보다 이수한 사람이 많았으며(인구 100,000명 당 63명) 그 차이가 각각 유의수준 5%, 10%에서 통계적으로 유의하였다. 치매로 진단 받은 이후에 사망으로 전이되는($1 \to 2$) 강도는 여성보다 남성이 많았고(인구 100,000명 당 4,664명) 그 차이가 유의수준 5%에서 통계적으로 유의하지 않았으며($P = 0.132$), 초등교육을 이수하지 못한 사람보다 이수한 사람이 많은 것으로 나타났지만(인구 100,000명 당 3,511명) 통계적으로 유의하지 않았다($P = 0.258$). 치매로 진단 받지 않고 사망으로 전이되는($0 \to 2$) 강도는 여성보다 남성이 많았고(인구 100,000명 당 195명) 그 차이가 유의수준 5%에서 통계적으로 매우 유의하였으며, 초등교육을 이수한 사람보다 이수하지 못한 사람이 많은 것으로 나타났으나(인구 100,000명 당 14명) 통계적으로 유의하지 않았다($P = 0.819$). 공통 프레일티의 분산 추정량은 $60.8(\times 10^{-4})$으로 군집 내 개체 간의 연관성은 적은 것으로 나타났다.

5. 맺음말

본 논문에서는 종말사건에 대한 정보는 주어져 있지만 중간사건이 구간중도절단 되었거나 연구 기간 도중에 추적이 끊겨 중간사건의 발생 유무를 모르는 준경쟁적위험 자료에 다중상태모형을 적용하여 상태 간의 전이강도를 추정하였다. 이를 위해 로그정규 프레일티를 랜덤효과로 가진 Lin과 Ying (1994)의 가산위험모형을 가정하고 완전 자료에 기초한 우도를 구축하였다. 조정중요표본추출법을 통해 주변우도를 유도한 후, 반복유사뉴튼 알고리즘을 써서 회귀계수에 대한 최대우도추정량을 얻었다. PAQUID 자료에서 임의로 추출하여 얻은 Paq1000 자료에 제안한 추정 방법을 적용해 본 결과 남성보다 여성이 LTF에 더 많이 노출되어 있었으나 LTF 이후 사망 위험은 여성보다 남성이 더 높은 것으로 나타났다. 건강한 상태에서 치매로 전이되는 위험은 남성보다 여성이 더 많이 노출되어 있었으나, 치매로 진단 받은 이후에는 여성보다 남성의 사망 위험이 더 높은 것으로 나타났다. 이런 경향은 치매로 진단 받지 않은 집단에서도 동일하게 나타났다. 한편, 초등교육을 이수하지 못한 사람보다 이수한 사람이 LTF에 더 많이 노출되어 있었으나 LTF 이후 사망 위험은 초등교육을 이수한 사람보다 이수하지 못한 사람이 더 높은 것으로 나타났다. 건강한 상태에서 치매로 전이되는 위험은 초등교육을 이수하지 못한 사람보다 이수한 사람이 더 많이 노출되어 있었으며, 치매로 진단 받은 이후 사망 위험도 초등교육을 이수하지 못한 사람보다 이수한 사람이 더 높았다. 그러나 치매로 진단 받지 않은 집단에서는 초등교육을 이수한 사람보다 이수하지 못한 사람의 사망 위험이 더 높은 것으로 나타났다. PAQUID 자료의 두 그룹(여성 vs. 남성, 초등학교 이수함 vs. 초등학교 이수 못함) 간 전이강도의 차이에 대한 추정값을 참고하여 모의실험에서 공변량 효과의 크기를 정하였다. 이때, Lin과 Ying (1994, 1995)이 언급한 것처럼 가산위험모형에서는 위험률의 비 대신에 위험률의 차이를 고려하기 때문에 공변량의 효과가 모의실험에서처럼 매우 작은 값을 가질 수도 있다. 제안한 추정량의 효율을 추정량의 상대편향비율과 표준오차, 모수의 95% 신뢰구간 포함률의 측면에서 살펴보기 위해 회귀계수에 대한 여섯 가지 시나리오와 사망 위험률에 대한 세 가지 시나리오의 조합 하에서 소표본 모의실험을 수행하였다. 시나리오를 고정했을 때 사망 위험률이 증가함에 따라 추정량의 표준오차는 단조적으로 증가하는 경향을 보였지만 추정량의 상대편향비율과 모수의 95% 신뢰구간 포함률은 별다른 경향이 없었다. 사망 위험률을 고정했을 때 회귀계수에 대한 시나리오에 관계없이 회귀계수에 대한 95% 신뢰구간 포함률은 명목값에 가까운 것으로 나타났다. 본 논문에서 제안한 추정량은 제약 조건 (2.1)에 포함된 미지의 상수 r의 값과 프레일티 분포의 오명세에 대해 로버스트한 성질을 가지고 있다고 말할 수 있다.

본 논문에서는 다섯 가지 전이강도 모형에 대해 Lin과 Ying (1994)처럼 준모수적 모형을 고려하고자 하였으나 장애모수에 해당되는 기저전이강도에 대한 비모수적 추정이 수월하지 않아서(추정 모수의 개수 $\propto$ 개체수) 생존분석에서 가장 널리 쓰이는 분포인 와이블분포를 가정하였다(추정 모수의 개수 = 10).

그러나 기저전이강도에 대해 특정 분포를 가정하는 것은 연구결과를 임상자료에 적용하는 데 한계점을 가지고 있기 때문에 이를 극복할 수 있는 연구가 필요하다고 생각한다. Lin과 Ying (1994)이 생존자료에서 다룬 것처럼 전이강도 별로 누적기저전이강도(cumulative baseline transition intensity)에 대한 Nelson-Aalen-type의 추정량을 구한 후 프로파일 우도함수에 기초한 모수 추정 방법을 준경쟁적위험 자료로 확장하는 연구나, Joly 등 (1998)과 Leffondré 등 (2013)이 제안한 것처럼 기저전이강도에 대한 스플라인 평활방법을 적용하는 연구가 필요하다고 생각한다. 한편 심사위원의 의견에 따라 PAQUID 자료에 가산위험모형을 적합하는 것이 타당한지를 알아보기 위해 다섯 가지 전이강도 별로 비례위험 가정을 검토하였다. 유의수준 10%에서 성별은 전이 $0 \to 3$를($P = 0.354$) 제외하고 나머지 전이들은 모두($0 \to 1, P = 0.063$; $0 \to 2, P < 0.001$; $1 \to 2, P = 0.093$; $3 \to 2, P = 0.062$) 비례위험 가정을 만족하지 않았지만, 초등교육 이수 여부는 전이 $0 \to 2$를($P < 0.001$) 제외하고 나머지 전이들이 모두($0 \to 1, P = 0.963$; $1 \to 2, P = 0.147$; $0 \to 3, P = 0.754$; $3 \to 2, P = 0.148$) 비례위험 가정을 만족하였다 . 따라서 본 논문처럼 두 공변량을 모두 가산위험모형에 포함하는 것도 바람직하지 않을 뿐만 아니라 두 공변량을 모두 비례위험모형에 포함하는 것도 바람직하지 않다고 생각한다. 따라서 PAQUID 자료에 대한 좀더 좋은 추정량을 얻기 위해서 비례위험모형과 가산위험모형을 동시에 고려하는 확장된 모형에 대한 연구가 필요하다고 생각한다.

References

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*, Springer, New York.

Barrett, J. K., Siannis, F., and Farewell, V. T. (2011). A semi-competing risks model for data with interval-censoring and informative observation: an application to the MRC cognitive function and aging study, *Statistics in Medicine*, **30**, 1–10.

Collett, D. (2015). *Modeling Survival Data in Medical Research* (3rd ed), CRC Press, London.

Cox, D. R. (1972). Regression models and life-tables, *Journal of the Royal Statistical Society, Series B (Methodological)*, **34**, 187–220.

Frydman, H. and Szarek, M. (2009). Nonparametric estimation in a Markov "illness-death" process from interval censored observations with missing intermediate transition status, *Biometrics*, **65**, 143–151.

Helmer, C., Joly, P., Letenneur, L., Commenges, D., and Dartigues, J. F. (2001). Mortality with dementia: results from a French prospective community-based cohort, *American Journal of Epidemiology*, **154**, 642–648.

Joly, P., Commenges, D., and Letenneur, L. (1998). A penalized likelihood approach for arbitrarily censored and truncated data: application to age-specific incidence of dementia, *Biometrics*, **54**, 185–194.

Leffondré, K., Touraine, C., Helmer, C., and Joly, P. (2013). Interval-censored time-to-event and competing risk with death: is the illness-death model more accurate than the Cox model?, *International Journal of Epidemiology*, **42**, 1177-1186.

Lin, D. Y. and Ying, Z. (1994). Semiparametric analysis of the additive risk model, *Biometrika*, **81**, 61–71.

Lin, D. Y. and Ying, Z. (1995). Semiparametric analysis of general additive-multiplicative hazard models for counting processes, *The Annals of Statistics*, **23**, 1712–1734.

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

MRC CFAS (1998). Cognitive function and dementia in six areas of England and Wales: the distribution of MMSE and prevalence of GMS organicity level in the MRC CFA Study. The Medical Research Council Cognitive Function and Ageing Study (MRC CFAS), *Psychological Medicine*, **28**, 319–335.

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model, *Journal of Computational and Graphical Statistics*, **4**, 12–35.

Siannis, F., Farewell, V. T., and Head, J. (2007). A multi-state model for joint modelling of terminal and non-terminal events with application to Whitehall II, *Statistics in Medicine*, **26**, 426–442.

Touraine, C., Joly, P., and Gerds, T. A. (2014). *SmoothHazards: Fitting illness-death model for interval-censored data*, R package version 1.2.3, from: https://www.R-project.org/package=SmoothHazard

결측되었거나 구간중도절단된 중간사건을 가진 준경쟁적위험 자료에 대한 가산위험모형

김자연[a] · 김진흠[b,1]

[a]건국대병원 연구지원센터, [b]수원대학교 응용통계학과

(2017년 4월 18일 접수, 2017년 6월 14일 수정, 2017년 6월 18일 채택)

요 약

본 논문에서는 사망과 같은 종말사건의 발생 유무는 알고 있지만 치매 발병과 같은 중간사건이 구간중도절단 되었거나 연구 기간 도중에 추적이 끊겨 결측된 준경쟁적위험 자료에 대해 다중상태모형을 적용하여 모수를 추정하는 방법을 제안하였다. 이를 위해 본 논문에서는 상태 간의 전이강도는 로그정규 프레일티를 랜덤효과로 가진 Lin과 Ying (1994)의 가산위험모형을 따른다고 가정하였다. 다섯 가지 상태를 가진 다중상태모형에서 가능한 여섯 가지 경로별로 조건부우도를 정의하였고, 주변우도를 구하기 위해 조정중요표본추출법을 적용하였으며 반복유사뉴튼 방법으로 최적해를 구하였다. 소표본 모의실험을 통해 모수의 95% 신뢰구간 포함률이 명목값에 얼마나 가까운지 살펴보았으며, 제안한 모형을 Persones Agées Quid (PAQUID) 자료 (Helmer 등, 2001)에 적용하고 그 결과를 해석하였다.

주요용어: 가산위험모형, 로그정규 프레일티, 결측 되었거나 구간중도절단된 중간사건, 다중상태모형, 준경쟁적위험 자료

이 논문은 2014년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No.NRF-2014R1A1A2056869).

[1]교신저자: (18323) 경기도 화성시 봉담읍 와우안길 17, 수원대학교 응용통계학과.
E-mail: jkimdt65@gmail.com