응용통계연구 (2010)
23(5), 871–882
DOI: 10.5351/KJAS.2010.23.5.871

종속적인 중도절단을 가진 동물종양 자료의 분석을 위한 모형

김진흠[1] · 김윤남[2]

[1]수원대학교 통계정보학과, [2]연세대학교 보건대학원

(2010년 6월 접수, 2010년 8월 채택)

요 약

동물종양 실험에서는 종양발생 시간이 직접 관찰되지 않고 단지 자연사로 인한 관찰 시점이나 강제적으로 희생시킨 시점 이전에 종양이 발생했는지 유무만을 알 수 있다. 이와 같은 형태의 결측을 가진 자료를 분석하기 위해 3단계(건강→종양발생→사망) 모형이 널리 사용되고 있다. 본 논문에서는 자연사로 인한 사망 시간이 종속적인 중도절단으로 작용하여 사망 시간과 종양발생 시간이 종속될 때, 이를 모형에 반영하기 위해 감마 프레일티 효과를 도입하였다. 모수 추정은 종양발생 시간과 프레일티 효과의 결측을 다루기 위해 EM 알고리즘 방법을 사용하였다. 제안한 추정량의 소표본 성질을 살펴보기 위해 제안한 방법을 Lindsey와 Ryan (1993, 1994)의 방광암 자료에 적용하여 모수를 추정하였으며, 그 추정값을 바탕으로 모의실험을 수행하였다.

주요용어: 3단계 모형, EM 알고리즘, 가우스-라게르 적분, 감마 프레일티 효과, 동물종양 실험, 방광암 자료.

1. 서론

현상태자료(current status data)는 흔히 관찰 연구(observational studies)에서 얻을 수 있다. 전형적 예로 동물종양 실험(tumorigenicity experiment)을 들 수 있는데, 이 실험 연구에서 주요 관심은 공변량이 종양발생에 미치는 효과를 밝히는 데 있다. 그러나 종양발생 시간은 관측할 수 없고 단지 관찰 시점(examination time)에서 종양발생 유무에 대한 정보만을 얻을 수 있다. 따라서 동물종양 실험을 포함하여 현상태자료는 구간중도절단(interval censoring)된 자료의 일종이라고 생각 할 수 있다 (Sun, 2006). 종양발생이 관찰 시점 이전에 있었다면 관찰 시점에서 좌중도절단(left censoring)된 것이고, 그 이후에 있었다면 우중도절단(right censoring)된 것으로 볼 수 있기 때문이다. 구간중도절단된 자료와 현상태자료의 차이점은 현상태자료에는 중도절단 되지 않고 정확하게 관측된 자료가 없다는 데 있다.

현상태자료에 대한 대부분의 연구들은 개체의 공변량이 주어지면 종양발생 시간과 관찰 시점이 서로 독립이라고 가정하고 있다. 만약 관찰 시점을 실험 전에 미리 정한다면 이와 같은 가정은 타당하다고 생각한다. 그러나 동물종양 실험에서는 동물이 자연사(natural death)하는 사건이 발생하거나 강제적으로 희생(sacrifice)시키는 사건이 발생하면 관측이 이루어진다. 따라서 자연사가 발생하면 관측이 이루어지기 때문에 종양발생 시간과 자연사로 인한 사망 시간이 더 이상 서로 독립일 수 없다. 실제로 Lagakos와 Louis (1988)는 종양의 치명도(lethality)를 중심으로 종양발생과 사망의 관계를 다음과 같이 요약하였다. 치명적이지 않은 종양(non-lethal tumor)의 경우는 사망 시간이 종양발생 시간과 독립

이 논문는 2008년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. R01-2008-000-20538-0).

[1]교신저자: (445-743) 경기도 화성시 봉담읍 와우리 산 2-2호, 수원대학교 통계정보학과, 교수.
E-mail: jinhkim@suwon.ac.kr

이고, 급성종양(rapidly lethal tumor)의 경우는 종양발생 직후 바로 사망하기 때문에 사망 시간을 종양발생 시간으로 간주할 수 있디. 따라서 이 두 경우는 우중도절단된 자료의 분석 방법을 적용할 수 있기 때문에 어려움이 없다. 그런데 대부분의 종양들은 중간 종양(intermediate lethal tumor)이기 때문에 사망 시간이 종양발생 시간과 서로 종속될 수 있다. 이와 같은 자료는 건강(health), 종양발생(tumor onset), 사망(death)로 이루어진 3단계 모형(three-state model)으로 분석할 수 있다. Lindsey와 Ryan (1993, 1994)은 관찰 시점의 두 가지 다른 타입(즉, 자연사와 강제적인 희생)과 종양발생 유무를 고려하여 4가지 서로 다른 조합에 대해 우도(likelihood)를 정의하고, EM 알고리즘 (Dempster 등, 1977)을 적용하여 모수를 추정하는 방법을 제안하였다. 이 때 기저위험함수(baseline hazard function)의 차원을 낮추기 위해 조각지수(piecewise exponential) 모형을 가정하였다. 그들이 EM 알고리즘을 도입한 이유는 종양발생 시간이 결측되어 관측된 자료만으로 충분통계량를 추정해야 했기 때문이다. French와 Ibrahim (2002)은 동일한 문제에 대해 베이지안 방법을 적용하였다. Lindsey와 Ryan (1993, 1994), French와 Ibrahim (2002)은 모두 치명도를 종양없이 사망할 위험률과 종양을 가지고 사망할 위험률의 비로 정의하고, 그 값을 기저 치명도(baseline lethality)와 처리 효과에 따른 치명도로 구성하였다.

본 논문에서는 자연사로 인한 관찰 시점이 종속적인 중도절단(informative censoring)으로 작용하는 경우를 다루고자 한다. 종양발생 시간과 사망 시간의 종속 관계를 모형에 포함하기 위해 생존자료분석에서 널리 쓰이는 프레일티(frailty) 효과를 도입하고자 한다 (Huang과 Wolfe, 2002). 2절에서는 감마 프레일티(gamma frailty)를 가진 3단계 모형을 정의하고 우도함수를 유도하고자 한다. 3절에서는 EM 알고리즘을 도입하여 모수를 추정하고자 한다. 4절에서는 Lindsey와 Ryan (1993, 1994)이 분석한 방광암(bladder cancer) 자료에 대해 강제적으로 희생시킨 시점을 Lindsey와 Ryan (1993, 1994)과 다르게 하여 모수를 추정한 후, 그 추정량을 바탕으로 모의실험을 수행하여 추정량의 소표본 성질을 살펴보고자 한다. 5절에서는 제안한 방법의 한계점과 향후 과제에 대해 토론하고자 한다.

2. 감마 프레일티를 가진 3단계 모형

2.1. 모형 구축

동물종양 연구에서 관심 있는 두 사건은 종양발생과 사망이다. 여기서 사망이란 자연사에 의한 사망뿐만 아니라 연구 종료와 같은 강제적 희생에 의한 사망을 의미한다. 두 사건 중에서 먼저 일어난 사건의 발생 시간을 X_i라고 하고, T_i는 사망 시간이라고 하자. 단, $i = 1, \ldots, n$. 한편, T_i는 $T_i = \min(T_{1i}, T_{2i})$로 정의되며, T_{1i}과 T_{2i}는 각각 자연사에 의한 사망 시간과 강제적 희생에 의한 사망시간 이다. 종양 없이 죽은 개체는 $X_i = T_i$이기 때문에 사망 전까지 종양이 발생하지 않았다는 것은 확실하지만 종양발생 시간을 알 수 없다. 종양을 가지고 죽은 개체들도 X_i가 T_i 보다 작다는 것은 확실하지만 종양발생 시간 X_i와 종양발생 후 사망할 때 까지의 시간 $(T_i - X_i)$는 알 수 없다. 전자의 경우는 사망이 종양발생에 대해 우중도절단의 역활을 하고, 후자의 경우는 그 반대로 좌중도절단의 역활을 한다고 말할 수 있다. 만일 $X_i < T_i$이면 $\delta_i = 1$로 정의하고, $X_i = T_i$이면 $\delta_i = 0$으로 정의하자. 또한, 사망의 형태에 따라 자연사이면 $d_i = 1$로 정의하고, 강제적으로 희생되었으면 $d_i = 0$으로 정의하자. 따라서 i번째 개체에 대해 관측 가능한 자료는 $o_i = (t_i, \delta_i, d_i, z_i)$이다. 여기서, z_i는 i번째 개체의 p-차원 공변량 벡터이다.

동물종양 연구에서는 흔히 공변량 벡터 z_i가 주어지면 종양발생 시간과 사망 시간은 서로 독립이라고 가정한다. 그러나 고려하고 있는 공변량에 의해 설명되지 못하는 개체 고유의 랜덤 변량이 있으면 종양발생 시간과 사망 시간은 독립적이지 못하고 종속적일 수 있다. 이와 같은 종속 관계를 모형에 반영하기 위해 본 논문에서는 프레일티 효과를 모형에 포함하고자 한다. 프레일티 r_i는 서로 독립이고 평균이 1,

분산이 ϕ인 감마 분포 $G(\phi^{-1}, \phi)$를 따른다고 가정하자. 단, $\phi > 0$. 공변량 z_i와 감마 프레일티 r_i가 주어졌을 때, 종양발생 시간과 종양발생이전 사망(death without tumor) 시간, 종양발생이후 사망(death with tumor) 시간에 대한 위험함수(hazard function), $\alpha_i(t|z_i, r_i)$, $\tilde{\lambda}_i(t|z_i)$, $\lambda_i(t|x_i, z_i, r_i)$를 각각 Cox 비례위험모형(Cox proportional hazards model)을 이용하여 다음과 같이 정의하자.

$$\alpha_i(t|z_i, r_i) = \lim_{\epsilon \to 0} \Pr\left(t \le X_i < t + \epsilon, \delta_i = 1 | X_i \ge t, z_i, r_i\right) = \alpha_0(t) r_i \exp\left(\beta' z\right), \tag{2.1}$$

$$\tilde{\lambda}_i\left(t|z_i\right) = \lim_{\epsilon \to 0} \Pr\left(t \le X_i < t + \epsilon, \delta_i = 0 | X_i \ge t, z_i\right) = \lambda_0(t) \exp\left(\psi' z_i\right), \tag{2.2}$$

$$\lambda_i\left(t|x_i, z_i, r_i\right) = \lim_{\epsilon \to 0} \Pr\left(t \le T_i < t + \epsilon | X_i = x_i, z_i, r_i, \delta_i = 1\right) = \lambda_0(t) r_i^{\tau} \exp\left\{(\psi + \gamma)' z_i\right\}, \tag{2.3}$$

여기서 $\alpha_0(\cdot)$과 $\lambda_0(\cdot)$는 각각 종양발생 시간과 종양발생이전 사망 시간에 대한 기저위험함수이고, β, ψ, γ는 p-차원 회귀계수벡터이다. τ는 프레일티에 기인한 종양발생 시간과 사망 시간 간의 종속관계를 나타내는 모수이다. 만약 $\tau = 0$이면 공변량이 주어졌을 때 종양발생 시간과 사망 시간은 서로 독립이고, 프레일티는 종양발생 시간에만 영향을 미치는 랜덤 효과에 지나지 않는다. 그러나 $\tau > 0$이면 프레일티가 높을수록 종양발생이후 사망 때까지의 단축되고, $\tau < 0$이면 프레일티가 높을수록 오히려 종양을 가진 상태로 사망 때까지의 시간이 길어진다. Lindsey와 Ryan (1993, 1994), French와 Ibrahim (2002)이 제안한 모형과 위에서 제안한 모형은 두 가지 점에서 다르다고 할 수 있다. 첫째, 공변량이 주어졌을 때 종양발생 시간과 사망 시간의 종속적인 관계를 프레일티 효과로 나타낸 것이며, 둘째, 종양발생으로 인한 치명도가 개체마다 다르다고 가정한 것이다.

한편, Lindsey와 Ryan (1993, 1994), French와 Ibrahim (2002)처럼 $\alpha_0(\cdot)$과 $\lambda_0(\cdot)$에 대해 조각지수 모형을 가정하자. 시구간을 J개 구간으로 나누고, j번째 구간을 $I_j = (s_{j-1}, s_j]\,(j = 1, \ldots, J)$이라고 하자. 단, $s_0 \equiv 0$. 구간 I_j 위에서 각각 $\alpha_0(t) = \alpha_{j0}$, $\lambda_0(t) = \lambda_{j0}$라고 하고, $\alpha = (\alpha_{10}, \ldots, \alpha_{J0})'$, $\lambda = (\lambda_{10}, \ldots, \lambda_{J0})'$라고 놓자.

2.2. 우도

종양발생 유무와 자연사 유무에 따라 4가지 조합이 가능한데 모든 개체는 4가지 형태 중에서 어느 한 그룹에 포함되며 각 그룹에 따라 우도 함수에 기여하는 양이 달라진다. 만약 $(\delta_i, d_i) = (0, 0)$이면 희생되었을 때 아직 종양이 발생하지 않은 그룹(sacrifice with no tumor; SNT)에 속하고, $(\delta_i, d_i) = (0, 1)$이면 자연사 했을 때 아직 종양이 발생하지 않은 그룹(death with no tumor; DNT), $(\delta_i, d_i) = (1, 0)$이면 희생되었을 때 이미 종양이 발생한 그룹(sacrifice with tumor; SWT)에 속하고, $(\delta_i, d_i) = (1, 1)$이면 자연사 했을 때 종양이 발생하지 않은 그룹(death with tumor; DWT)에 속한다. SNT, DNT, SWT, DWT 그룹에 속하는 개체의 우도를 각각 L_{i1}, L_{i2}, L_{i3}, L_{i4}라고 하면 다음과 같이 주어진다. $\theta = (\alpha', \beta, \lambda', \psi, \gamma, \tau, \phi)'$라고 놓자.

$$L_{i1}(\theta; o_i, r_i) = \exp\left[-\int_0^{t_i} \left\{\alpha_i(u|z_i, r_i) + \tilde{\lambda}_i(u|z_i)\right\} du\right],$$

$$L_{i2}(\theta; o_i, r_i) = \tilde{\lambda}_i(t_i|z_i) \exp\left[-\int_0^{t_i} \left\{\alpha_i(u|z_i, r_i) + \tilde{\lambda}_i(u|z_i)\right\} du\right],$$

$$L_{i3}(\theta; o_i, r_i) = \int_0^{t_i} f_X(u|z_i, r_i) \tilde{S}_T(u|z_i) \frac{S_T(t_i|x_i, z_i, r_i)}{S_T(u|x_i, z_i, r_i)} du,$$

$$L_{i4}(\theta; o_i, r_i) = \lambda_i(t_i|x_i, z_i, r_i) \int_0^{t_i} f_X(u|z_i, r_i) \tilde{S}_T(u|z_i) \frac{S_T(t_i|x_i, z_i, r_i)}{S_T(u|x_i, z_i, r_i)} du,$$

여기서 $S_X, \tilde{S}_T, S_T$는 각각 모형 (2.1), (2.2), (2.3)에 대한 생존함수이고, f_X은 모형 (2.1)에 대한 확률밀도함수이다. L_{i3}와 L_{i4}를 살펴보면 종양발생 시간에 대한 적분이 포함되어 있는데, 이는 SWT와 DWT에 속하는 개체의 종양발생시간은 관측할 수 없고 다만 t_i 이전에 발생한 정보 밖에 없기 때문이다. 따라서 관측 가능한 자료 o_i에 기초한 i번째 개체의 조건부 우도(conditional likelihood)는 다음과 같이 주어진다.

$$\text{cLik}(\theta; o_i, r_i) = L_{i1}^{(1-\delta_i)(1-d_i)} L_{i2}^{(1-\delta_i)d_i} L_{i3}^{\delta_i(1-d_i)} L_{i4}^{\delta_i d_i}.$$

만약 $\delta_i = 1$인 개체의 종양발생 시간 x_i가 관측 가능하다면 완전 자료(complete data)를 얻을 수 있고, 이를 $c_i = (x_i, t_i, \delta_i, d_i)$로 나타내자. 완전 자료가 주어지면 L_{i3}와 L_{i4}는 각각 다음과 같이 다시 표현할 수 있다.

$$\begin{aligned}
\tilde{L}_{i3}(\theta; c_i, r_i) &= \alpha_i(x_i|z_i, r_i) \exp\left[-\int_0^{x_i} \left\{ \alpha_i(u|z_i, r_i) + \tilde{\lambda}_i(u|z_i) \right\} du - \int_{x_i}^{t_i} \lambda_i(u|x_i, z_i, r_i) du \right], \\
\tilde{L}_{i4}(\theta; c_i, r_i) &= \alpha_i(x_i|z_i, r_i) \lambda_i(t_i|x_i, z_i, r_i) \\
&\quad \times \exp\left[-\int_0^{x_i} \left\{ \alpha_i(u|z_i, r_i) + \tilde{\lambda}_i(u|z_i) \right\} du - \int_{x_i}^{t_i} \lambda_i(u|x_i, z_i, r_i) du \right].
\end{aligned}$$

따라서 완전 자료 c_i에 기초한 i번째 개체의 조건부 우도와 완전 우도(full likelihood)는 각각 다음과 같이 주어진다.

$$\begin{aligned}
\text{cLik}(\theta; c_i, r_i) &= L_{i1}^{(1-\delta_i)(1-d_i)} L_{i2}^{(1-\delta_i)d_i} \tilde{L}_{i3}^{\delta_i(1-d_i)} \tilde{L}_{i4}^{\delta_i d_i} \\
&= \prod_{j=1}^{J} \left[\alpha_j^{\delta_{ij}} \lambda_j^{d_i d_{ij}} \exp\left[-\left\{ \alpha_j r_i \exp(\beta' z_i) + \lambda_j \exp(\psi' z_i) \right\} T_{ij}^{NT} \right.\right. \\
&\qquad \left.\left. -\lambda_j r_i^{\tau} \exp\{(\psi + \gamma)' z_i\} T_{ij}^{T} \right] \right] \exp\{z_i'(\delta_i \beta + d_i \psi + d_i \delta_i \gamma)\} r_i^{\delta_i + \tau d_i \delta_i}, \\
\text{fLik}(\theta; c_i, r_i) &= \text{cLik}(\theta; c_i, r_i) g(\phi; r_i) \\
&= \prod_{j=1}^{J} \left[\alpha_j^{\delta_{ij}} \lambda_j^{d_i d_{ij}} \exp\left\{ -\lambda_j \exp(\psi' z_i) T_{ij}^{NT} \right\} \right] \exp\{z_i'(\delta_i \beta + d_i \psi + d_i \delta_i \gamma)\} \\
&\quad \times r_i^{\delta_i + \tau d_i \delta_i + \phi^{-1} - 1} \exp\left[-\left\{ \exp(\beta' z_i) \sum_{j=1}^{J} \alpha_j T_{ij}^{NT} + \phi^{-1} \right\} r_i \right] \\
&\quad \times \exp\left[-\exp\{(\psi + \gamma)' z_i\} r_i^{\tau} \sum_{j=1}^{J} \lambda_j T_{ij}^{T} \right] \left\{ \Gamma\left(\phi^{-1}\right) \phi^{\phi^{-1}} \right\}^{-1},
\end{aligned}$$

여기서 $d_{ij} = I(t_i \in I_j)$이고, δ_{ij}는 i번째 개체의 종양이 발생이 구간 I_j에서 있었으면 $\delta_{ij} = 1$로 정의하고, 그렇지 않으면 $\delta_{ij} = 0$으로 정의되는 지시변수(indicator)이다. 또한, T_{ij}^{NT}와 T_{ij}^{T}는 각각 i번째 개체가 구간 I_j에서 종양 없이 지낸 시간과 종양을 가지고 지낸 시간을 나타낸다. g는 감마 분포 $G(\phi^{-1}, \phi)$의 확률밀도함수이다. 따라서 완전 자료에 기초한 로그 완전 우도(log full likelihood)는 다음과 같이 주어진다. $o = (o_1, \ldots, o_n)'$, $c = (c_1, \ldots, c_n)'$, $r = (r_1, \ldots, r_n)'$라고 하면,

$$\begin{aligned}
l_c(\theta; c, r) &= \log\left\{ \prod_{i=1}^{n} \text{fLik}(\theta; c_i, r_i) \right\} \qquad (2.4) \\
&= \sum_{j=1}^{J} \left\{ N_j^T \log \alpha_j + (a_j + b_j) \log \lambda_j - \lambda_j \sum_{i=1}^{n} T_{ij}^{NT} \exp\left(\psi' z_i\right) \right\} + \sum_{i=1}^{n} z_i'(\delta_i \beta + d_i \psi + d_i \delta_i \gamma)
\end{aligned}$$

$$+\sum_{i=1}^{n}\left[\left(\delta_i+\tau d_i\delta_i+\phi^{-1}-1\right)\log r_i-\left\{\exp\left(\beta' z_i\right)\sum_{j=1}^{J}\alpha_j T_{ij}^{NT}+\phi^{-1}\right\}r_i\right. \tag{2.5}$$

$$\left.-\exp\left\{(\psi+\gamma)' z_i\right\}r_i^{\tau}\sum_{j=1}^{J}\lambda_j T_{ij}^{T}\right]-n\log\Gamma\left(\phi^{-1}\right)-n\phi^{-1}\log\phi,$$

여기서 N_j^T는 구간 I_j에서 종양을 가지고 있는 개체수를 나타내고, $a_j=\sum_{i=1}^{n}(1-\delta_i)d_{ij}d_i$와 $b_j=\sum_{i=1}^{n}\delta_i d_{ij}d_i$는 각각 구간 I_j에서 종양 없이 사망한 개체수와 종양을 가지고 사망한 개체수를 나타낸다.

3. EM 알고리즘에 의한 모수 추정

식 (2.5)의 우도 함수에는 두 종류의 결측이 포함되어 있다. 그중 하나는 프레일티로 인해 발생한 것이고, 다른 하나는 종양을 가지고 사망한 개체의 종양발생 시간을 관측할 수 없기 때문에 발생한 것이다. 이와 같이 결측이 포함된 자료의 모수 추정을 위해 흔히 사용되는 EM 알고리즘을 써서 모수를 추정하고자 한다.

프레일티와 관련된 항들을 추정하기 위해 먼저 r_i의 사후 분포를 유도해야 하는데, o_i가 주어졌을 때 r_i의 조건부 분포는 다음과 같이 주어진다.

$$g_{r_i|o_i}(\phi;r_i)=\frac{\mathrm{clik}(\theta;o_i,r_i)g(\phi;r_i)}{\int_0^{\infty}\mathrm{clik}(\theta;o_i,u)g(\phi;u)du}. \tag{3.1}$$

$\delta_i=1$인 개체는 종양발생 시간이 관측되지 않기 때문에 식 (3.1)은 매우 복잡하다. 그래서 본 논문에서는 r_i의 사후 분포를 c_i가 주어졌을 때 r_i의 조건부 분포로 근사하는 방법을 사용하고자 한다. 그러면 r_i의 사후 분포의 확률밀도함수는 다음 식처럼 비례적으로 주어진다.

$$g_{r_i|o_i}(\phi;r_i)$$

$$\propto\ r_i^{\delta_i+\tau d_i\delta_i+\phi^{-1}-1}\exp\left[-\left\{\exp(\beta' z_i)\sum_{j=1}^{J}\alpha_j T_{ij}^{NT}+\phi^{-1}\right\}r_i\right]\exp\left[-\exp\left\{(\psi+\gamma)' z_i\right\}r_i^{\tau}\sum_{j=1}^{J}\lambda_j T_{ij}^{T}\right].$$

따라서 만약 $\tau=1$이면 사후 분포가 다시 감마 분포가 되지만 그 이외의 경우에는 감마 분포가 되지 않는다. 프레일티에 대한 조건부 기댓값을 추정하기 위해 Metropolis-Hastings 알고리즘 (Metropolis 등, 1953; Hastings, 1970) 대신에 Gauss-Laguerre 방법 (Golub와 Welsch, 1969)을 사용하였다. EM 알고리즘의 '최대화 과정(maximization step)'에서 필요한 프레일티의 함수들은 다음과 같다. $\mathrm{E}(r_i|o_i,\theta)$, $\mathrm{E}(\log r_i|o_i,\theta)$, $\mathrm{E}(r_i^{\tau}|o_i,\theta)$, $\mathrm{E}(r_i^{\tau}\log r_i|o_i,\theta)$, $\mathrm{E}\{r_i^{\tau}(\log r_i)^2|o_i,\theta\}$.

한편, 종양발생 시간이 관측 되지 않아 생긴 결측값, $N_j^T, T_{ij}^T, T_{ij}^{NT}$에 대한 추정은 Lindsey와 Ryan (1993, 1994)의 방법을 적용하고자 한다. 이 계산을 위해 Lindsey와 Ryan (1993)의 p.288과 Lindsey와 Ryan (1994)에 p.16에 있는 q 함수를 $q(x_i,t_i)=\tilde{L}_{i4}(\theta;c_i,r_i)$로 바꾸고 난 후, Lindsey와 Ryan (1993, 1994)의 방법을 적용하면 $\mathrm{E}(N_j^T|o,\theta)$, $\mathrm{E}(T_{ij}^{NT}|o_i,\theta)$, $\mathrm{E}(T_{ij}^T|o_i,\theta)$를 각각 추정할 수 있다. 여기서, q 함수는 i번째 개체가 t_i에서 사망하고 x_i에서 종양이 발생할 결합 확률에 해당한다.

모수 θ에 대한 최대우도추정량(maximum likelihood estimate; MLE)은 Newton-Raphson 방법을 써서 구할 수 있는데 이를 구체적으로 설명하면 아래와 같다. θ에 대한 초기값, $\theta^{(0)}$을 가지고, $k=0,1,\ldots$에 대해 다음 식을 통해 $\theta^{(k)}$를 수정해 나간다. $Q(\theta|\theta^{(k)})=\mathrm{E}_{\theta^{(k)}}\{l_c(\theta;c,r)|o\}$로 놓자.

$$\theta^{(k+1)}=\theta^{(k)}-\left\{\frac{\partial^2 Q\left(\theta|\theta^{(k)}\right)}{\partial\theta\partial\theta'}\right\}_{\theta=\theta^{(k)}}^{-1}\left\{\frac{\partial Q\left(\theta|\theta^{(k)}\right)}{\partial\theta}\right\}_{\theta=\theta^{(k)}}.$$

상술한 E-M 단계는 다음 조건을 만족할 때까지 계속적으로 반복된다.

$$\left|\left|\theta^{(k+1)} - \theta^{(k)}\right|\right|_{\infty} < \epsilon,$$

여기서 $||\cdot||_{\infty}$은 최대노름(maximum norm)을 나타내고, ϵ은 임의의 상수이다. Q에 대한 1차 미분 값과 2차 미분 값은 아래와 같이 정리된다. 다만 'E$(\cdot|o,\theta)$'와 같은 표현 대신 o와 θ에 대한 조건을 생략하고 'E$(\cdot)$'와 같이 간략히 나타내고자 한다. 먼저 Q의 1차 미분 값들은 다음과 같다.

$$\begin{aligned}
\frac{\partial Q}{\partial \alpha_j} &= \frac{\mathrm{E}(N_j^T)}{\alpha_j} - \sum_{i=1}^{n} \mathrm{E}\left(T_{ij}^{NT}\right) \exp\left(\beta' z_i\right) \mathrm{E}(r_i), \quad j = 1, \ldots, J; \\
\frac{\partial Q}{\partial \beta} &= \sum_{i=1}^{n} \left\{ \delta_i - \exp\left(\beta' z_i\right) \mathrm{E}(r_i) \sum_{j=1}^{J} \alpha_j \mathrm{E}\left(T_{ij}^{NT}\right) \right\} z_i; \\
\frac{\partial Q}{\partial \lambda_j} &= \frac{(a_j + b_j)}{\lambda_j} - \sum_{i=1}^{n} \left[\mathrm{E}\left(T_{ij}^{NT}\right) \exp\left(\psi' z_i\right) + \mathrm{E}\left(T_{ij}^{T}\right) \exp\left\{(\psi+\gamma)' z_i\right\} \mathrm{E}(r_i^{\tau}) \right], \quad j = 1, \ldots, J; \\
\frac{\partial Q}{\partial \psi} &= \sum_{i=1}^{n} \left[\delta_i - \exp\left(\psi' z_i\right) \sum_{j=1}^{J} \lambda_j \mathrm{E}\left(T_{ij}^{NT}\right) - \exp\left\{(\psi+\gamma)' z_i\right\} \mathrm{E}\left(r_i^{\tau}\right) \sum_{j=1}^{J} \lambda_j \mathrm{E}\left(T_{ij}^{T}\right) \right] z_i; \\
\frac{\partial Q}{\partial \gamma} &= \sum_{i=1}^{n} \left[d_i \delta_i - \exp\{(\psi+\gamma)' z_i\} \mathrm{E}(r_i^{\tau}) \sum_{j=1}^{J} \lambda_j \mathrm{E}\left(T_{ij}^{T}\right) \right] z_i; \\
\frac{\partial Q}{\partial \tau} &= \sum_{i=1}^{n} \left[\delta_i d_i \mathrm{E}(\log r_i) - \exp\left\{(\psi+\gamma)' z_i\right\} \mathrm{E}\left(r_i^{\tau} \log r_i\right) \sum_{j=1}^{J} \lambda_j \mathrm{E}\left(T_{ij}^{T}\right) \right]; \\
\frac{\partial Q}{\partial \phi} &= n\phi^{-2} \psi\left(\phi^{-1}\right) - n\phi^{-2}(1 - \log\phi) - \phi^{-2} \sum_{i=1}^{n} \{\mathrm{E}(\log r_i) - \mathrm{E}(r_i)\},
\end{aligned}$$

여기서 $\psi(x) = d \log \Gamma(x)/dx$는 다이감마(digamma) 함수이다. 또한, Q의 2차 미분 값들은 다음과 같다. 임의의 p-차원 벡터 x에 대해 $x^{\otimes 2} = xx'$은 p-차원 정사각행렬을 나타낸다.

$$\begin{aligned}
\frac{\partial^2 Q}{\partial \alpha_j^2} &= -\frac{\mathrm{E}(N_j^T)}{\alpha_j^2}, \quad j = 1, \ldots, J; \\
\frac{\partial^2 Q}{\partial \beta^2} &= -\sum_{i=1}^{n} \exp\left(\beta' z_i\right) \mathrm{E}(r_i) \sum_{j=1}^{J} \alpha_j \mathrm{E}\left(T_{ij}^{NT}\right) z_i^{\otimes 2}; \\
\frac{\partial^2 Q}{\partial \lambda_j^2} &= -\frac{(a_j + b_j)}{\lambda_j^2}, \quad j = 1, \ldots, J; \\
\frac{\partial^2 Q}{\partial \psi^2} &= -\sum_{i=1}^{n} \left[\exp\left(\psi' z_i\right) \sum_{j=1}^{J} \lambda_j \mathrm{E}\left(T_{ij}^{NT}\right) + \exp\left\{(\psi+\gamma)' z_i\right\} \mathrm{E}\left(r_i^{\tau}\right) \sum_{j=1}^{J} \lambda_j \mathrm{E}\left(T_{ij}^{T}\right) \right] z_i^{\otimes 2}; \\
\frac{\partial^2 Q}{\partial \gamma^2} &= -\sum_{i=1}^{n} \exp\left\{(\psi+\gamma)' z_i\right\} \mathrm{E}\left(r_i^{\tau}\right) \sum_{j=1}^{J} \lambda_j \mathrm{E}\left(T_{ij}^{T}\right) z_i^{\otimes 2}; \\
\frac{\partial^2 Q}{\partial \tau^2} &= -\sum_{i=1}^{n} \exp\left\{(\psi+\gamma)' z_i\right\} \mathrm{E}\left\{r_i^{\tau} (\log r_i)^2\right\} \sum_{j=1}^{J} \lambda_j \mathrm{E}\left(T_{ij}^{T}\right); \\
\frac{\partial^2 Q}{\partial \phi^2} &= -2n\phi^{-3} \psi\left(\phi^{-1}\right) - n\phi^{-4} \psi'\left(\phi^{-1}\right) + n\phi^{-3}(3 - 2\log\phi) + 2\phi^{-3} \sum_{i=1}^{n} \left\{\mathrm{E}\left(\log r_i\right) - \mathrm{E}(r_i)\right\};
\end{aligned}$$

$$\frac{\partial^2 Q}{\partial\beta\partial\alpha_j} = \frac{\partial^2 Q}{\partial\alpha_j\partial\beta} = -\sum_{i=1}^{n} \mathrm{E}\left(T_{ij}^{NT}\right) \exp\left(\beta' z_i\right) \mathrm{E}\left(r_i\right) z_i, \quad j = 1, \ldots, J;$$

$$\frac{\partial^2 Q}{\partial\psi\partial\lambda_j} = \frac{\partial^2 Q}{\partial\lambda_j\partial\psi} = -\sum_{i=1}^{n} \left[\mathrm{E}\left(T_{ij}^{NT}\right) \exp(\psi' z_i) + \mathrm{E}\left(T_{ij}^{T}\right) \exp\left\{(\psi+\gamma)' z_i\right\} \mathrm{E}\left(r_i^{\tau}\right)\right] z_i, \quad j = 1, \ldots, J;$$

$$\frac{\partial^2 Q}{\partial\gamma\partial\lambda_j} = \frac{\partial^2 Q}{\partial\lambda_j\partial\gamma} = -\sum_{i=1}^{n} \mathrm{E}\left(T_{ij}^{T}\right) \exp\left\{(\psi+\gamma)' z_i\right\} \mathrm{E}(r_i^{\tau}) z_i, \quad j = 1, \ldots, J;$$

$$\frac{\partial^2 Q}{\partial\tau\partial\lambda_j} = \frac{\partial^2 Q}{\partial\lambda_j\partial\tau} = -\sum_{i=1}^{n} \mathrm{E}\left(T_{ij}^{T}\right) \exp\left\{(\psi+\gamma)' z_i\right\} \mathrm{E}(r_i^{\tau} \log r_i), \quad j = 1, \ldots, J;$$

$$\frac{\partial^2 Q}{\partial\gamma\partial\psi} = \frac{\partial^2 Q}{\partial\psi\partial\gamma} = -\sum_{i=1}^{n} \exp\left\{(\psi+\gamma)' z_i\right\} \mathrm{E}(r_i^{\tau}) \sum_{j=1}^{J} \lambda_j \mathrm{E}\left(T_{ij}^{T}\right) z_i^{\otimes 2};$$

$$\frac{\partial^2 Q}{\partial\tau\partial\psi} = \frac{\partial^2 Q}{\partial\psi\partial\tau} = \frac{\partial^2 Q}{\partial\tau\partial\gamma} = \frac{\partial^2 Q}{\partial\gamma\partial\tau} = -\sum_{i=1}^{n} \exp\{(\psi+\gamma) z_i\} \mathrm{E}(r_i^{\tau} \log r_i) \sum_{j=1}^{J} \lambda_j \mathrm{E}\left(T_{ij}^{T}\right) z_i.$$

이외 다른 2차 미분 값들은 모두 0이다. 한편, EM 알고리즘을 통해 얻어진 추정량의 분산을 추정하기 위해 음헤시안 행렬(negative Hessian matrix)을 다음과 같이 정의하면,

$$\hat{\Sigma} = -\frac{\partial^2 Q\left(\theta|\theta^{(s)}\right)}{\partial\theta\partial\theta'}\bigg|_{\theta=\theta^{(s)}},$$

행렬 $\hat{\Sigma}$는 3개의 블럭으로 이루어진 대각행렬이 된다. 여기서, $\theta^{(s)}$는 θ의 MLE를 나타낸다. 첫 번째 블럭은 모수 α, β, 두 번째 블럭은 모수 λ, ψ, γ, τ, 세 번째 블럭은 모수 ϕ에 각각 대응한다. 따라서 각 모수의 표준오차는 $\hat{\Sigma}$의 역행렬에서 대응하는 대각원소의 제곱근 값으로 정의된다.

4. 수치 분석

4.1. 방광암 자료 분석

2절에서 제안한 모형을 Lindsey와 Ryan (1993, 1994)의 방광암 자료에 적용하였다. 방광암 자료는 총 671개 개체로 이루어진 동물실험 자료이다. 각 개체는 약물의 투여량에 따라 저용량 그룹(총 개체수 = 387)이나 고용량 그룹(총 개체수 = 284)으로 나누어진다. Lindsey와 Ryan (1993, 1994)은 시구간을 $J = 3$개 구간으로 나누고, 각 구간의 오른족 끝점을 12, 18, 33으로 잡았다. 그러나 본 논문에서는 구간의 개수와 구간의 경계점을 사망 시간에 대한 기저위험함수를 추정하여 정하였다. 그림 4.1은 사망 시간에 대한 누적위험함수의 증가량을 나타낸 것이며, 이 그림으로부터 구간의 개수는 $J = 4$개, 구간의 경계점은 13, 19, 25, 33으로 잡았다. 표 4.1은 각 모수의 추정량과 표준오차, 유의확률 값(p값)을 정리한 것이다. 고용량 그룹이 저용량 그룹보다 종양발생의 가능성이 높았으며($\beta : p$값 < 0.0001), 종양이 사망에 미치는 영향도 고용량 그룹에서 더 높았지만($\gamma : p$값 $= 0.0079$), 투여량이 전체 사망에 미치는 영향은 없는 것으로 나타났다($\psi : p$값 $= 0.6413$). 모수 ϕ의 유의성이 매우 높게 나타나(p값 < 0.0001) 종양발생이 약물의 투여량만으로는 설명되지 못하고 개체마다 고유한 프레일티가 있음을 알 수 있었다. 종양발생과 사망 간의 종속 관계를 나타내는 모수 τ의 유의성도 매우 높게 나타났다(p값< 0.0001).

한편, Lindsey와 Ryan (1993)의 결과에 의하면 저용량 그룹보다 고용량 그룹의 종양발생이 높았으며(p값 < 0.0001), 종양이 사망에 미치는 영향도 고용량 그룹에서 높았지만(p값 $= 0.0020$), 투여량이 전체 사망에 미치는 영향은 크지 않게 나타났다(p값 $= 0.0995$). 종양발생으로 인한 사망률, 즉 치명

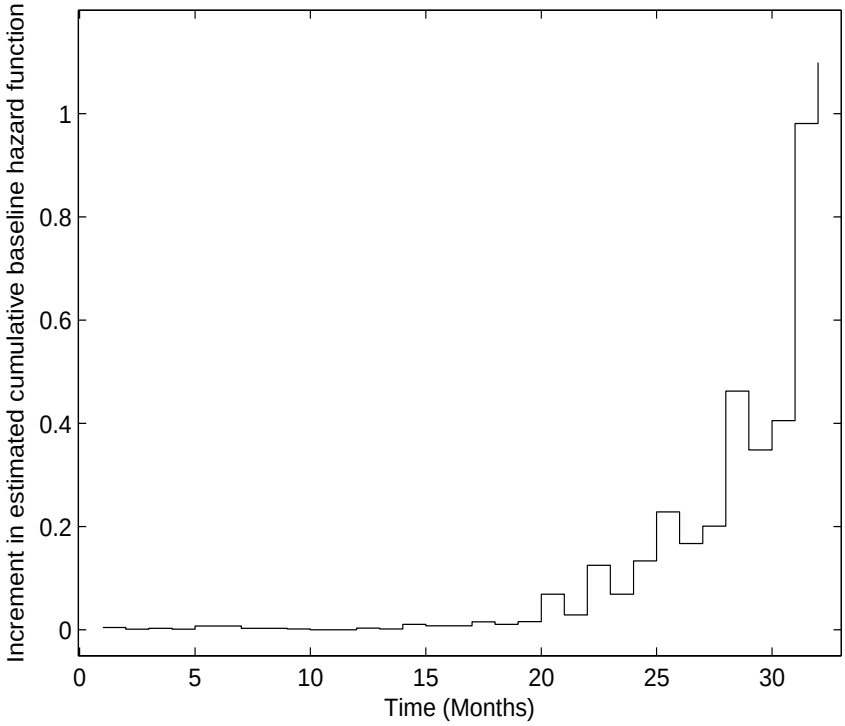

그림 **4.1.** 방광암 자료 (Lindsey와 Ryan, 1993, 1994)에서 사망시간에 대한 누적위험함수의 증가량

표 **4.1.** 방광암 자료 (Lindsey와 Ryan, 1993, 1994)에 대한 모수 추정값, 표준오차 및 유의확률 값

모수	추정값	표준오차	p값
α_{01}	0.0006	0.0002	< 0.0001
α_{02}	0.0019	0.0006	< 0.0001
α_{03}	0.0257	0.0079	< 0.0001
α_{04}	0.0064×10^{-6}	8.3455×10^{-6}	0.9973
λ_{01}	0.0018	0.0005	< 0.0001
λ_{02}	0.0070	0.0016	< 0.0001
λ_{03}	0.0719	0.01310	< 0.0001
λ_{04}	0.2723	0.0462	< 0.0001
β	3.4045	0.3140	< 0.0001
ψ	0.1506	0.3233	0.6413
γ	0.9049	0.3409	0.0079
τ	3.8947	0.4735	< 0.0001
ϕ	0.1252	0.0067	< 0.0001

도는 매우 유의한 것으로 나타났다(p값 < 0.0001). 본 모형의 결과와 Lindsey와 Ryan (1993, 1994)의 결과를 놓고 볼 때, 투여량이 종양발생에 미치는 영향이나 종양발생 유무에 따른 투여량이 사망에 미치는 영향에 대한 결과는 크게 차이가 나지 않았지만 Lindsey와 Ryan 모형으로는 미처 알 수 없었던 종양발생이 투여량에만 의존하는 것이 아니라 개체 고유한 특성에도 의존한다는 점과 종양발생이 사망 시간

과 유의하게 종속되어 있다는 점을 발견할 수 있었다.

또한, 구간의 개수 J를 4 대신 3과 5로 하여 분석하였는데, $J = 3$일 때의 경계점은 $J = 4$일 때의 경계점 중에서 중간에 있는 경계점을 임의로 2개 선택하였고, $J = 5$일 때의 경계점은 $J = 4$일 때의 경계점에 27을 하나 더 추가하였다. 각 모형에 따른 로그 우도값은 다음과 같다. $J = 3$일 때, -1240.42(경계점: $19, 25, 33$), -1315.83(경계점: $13, 19, 33$), -1343.82(경계점: $13, 25, 33$); $J = 5$일 때 -1279.50(경계점: $13, 19, 25, 27, 33$). $J = 4$일 때의 로그 우도값은 -1245.62으로 나타났으며 경계점을 $19, 25, 33$으로 하는 모형을 제외하고는 로그 우도 값이 가장 크게 나와 $J = 4$인 모형이 타당하다고 할 수 있었다.

4.2. 모의실험

개체마다 고유한 프레일티가 존재하고 종양발생과 사망이 종속적일 때 제안한 모형이 얼마나 이를 잘 탐색해내는지를 살펴보기 위해 3.1절에 다룬 방광암 자료 분석 결과를 바탕으로 모의실험을 수행하였다. 저용량 그룹은 130개($z_i = 0, i = 1, \ldots, 130$), 고용량 그룹은 95개($z_i = 1, i = 131, \ldots, 225$) 개체로 구성하였다. $r_i\,(i = 1, \ldots, 225)$는 $G(0.1252^{-1}, 0.1252)$에서 생성하였다. $J = 4$개 각 구간에서 기저위험함수는 다음과 같이 가정하였다. $\alpha_{01} = 0.0006$, $\alpha_{02} = 0.0019$, $\alpha_{03} = 0.0257$, $\alpha_{04} = 0.0001$; $\lambda_{01} = 0.0018$, $\lambda_{02} = 0.0070$, $\lambda_{03} = 0.0719$, $\lambda_{04} = 0.2723$. 또한, $\beta = 3.4045$, $\psi = 0.1506$, $\gamma = 0.9049$, $\tau = 3.8947$. 한편, 각 개체의 강제적으로 희생시킨 시간(t_{i2})은 저용량 그룹에서는 13 ($i = 1, \ldots, 16$), 19 ($i = 17, \ldots, 105$), 33 ($i = 106, \ldots, 130$), 고용량 그룹에서는 13 ($i = 131, \ldots, 145$), 19 ($i = 146, \ldots, 207$), 33 ($i = 208, \ldots, 225$)으로 각각 정하였다. 이와 같은 설계 값을 가지고서 다음 절차에 따라 (t_i, δ_i, d_i)를 결정하였다.

- 단계 1: $t \in I_j$, $j = 1, \ldots, 4$에 대해 $\alpha_0(t) = \alpha_{0j}$을 가진 모형 (2.1)으로 부터 종양발생 시간을 생성한다 (x_{io}).
- 단계 2: $t \in I_j$, $j = 1, \ldots, 4$에 대해 $\lambda_0(t) = \lambda_{0j}$을 가진 모형 (2.2)으로 부터 종양없이 사망한 시간을 생성한다 ($\tilde{t}_{i1}$).
- 단계 3: 만약 $\tilde{t}_{i1} < x_{io}$이면 단계 4로 진행하고, 그렇지 않으면 단계 5로 진행한다..
- 단계 4: 만약 $\tilde{t}_{i1} \le t_{i2}$이면 $t_i = \tilde{t}_{i1}$, $\delta_i = 0$, $d_i = 1$ (DNT)로 정의하고, 그렇지 않으면 $t_i = t_{i2}$, $\delta_i = 0$, $d_i = 0$ (SNT)로 정의한다.
- 단계 5: 종양발생 시간 x_{io}가 주어진 조건 하에서 $t \in I_j$, $j = 1, \ldots, 4$에 대해 $\lambda_0(t) = \lambda_{0j}$을 가진 모형 (2.3)으로부터 종양을 가지고 사망한 시간을 생성한다 (t_{i1}). 만약 $t_{i1} \le t_{i2}$이면 $t_i = t_{i1}$, $\delta_i = 1$, $d_i = 1$ (DWT)로 정의하고, 그렇지 않은 경우에는 x_{io}와 t_{i2}의 대소에 따라 $x_{io} \le t_{i2}$이면 $t_i = t_{i2}$, $\delta_i = 1$, $d_i = 0$ (SWT)로 정의하거나 그렇지 않으면 $t_i = t_{i2}$, $\delta_i = 0$, $d_i = 0$ (SNT)로 정의한다.

그림 4.2는 1,000번 반복하여 얻은 모수 β, ψ, γ, τ, ϕ의 추정값에 대한 박스플롯이다. 전반적으로 추정량의 분포는 대칭적이라도 말할 수 있다. β에 대한 추정값 중에는 참값으로부터 오른쪽으로 벗어난 값들이 많이 관측 되었고 (A열), ψ에 대한 추정값은 참값으로부터 왼쪽으로 벗어난 값이 많이 관측 되었다 (B열). γ에 대한 추정값은 참값을 중심으로 양쪽 방향으로 벗어난 값이 있었고 특히 오른쪽으로 크게 벗어난 값이 더 많았다 (C열). τ에 대한 추정값은 다른 모수에 비해 가장 넓게 분포하였으며 참값에서 벗어난 값도 가장 많았다 (D열). ϕ의 참값은 작아서 추정값의 산포가 작았고 참값에서 크게 벗어난 값이 적었다. 표 4.2는 모수 β, ψ, γ, τ, ϕ의 참값과 1,000번 반복을 통해 얻은 추정값의 평균(Mean), 편향의 평균(Bias), 표준편차(SD), 표준오차의 평균(SEM), 95% 신뢰구간 포함률(CR)을 정리한 것이

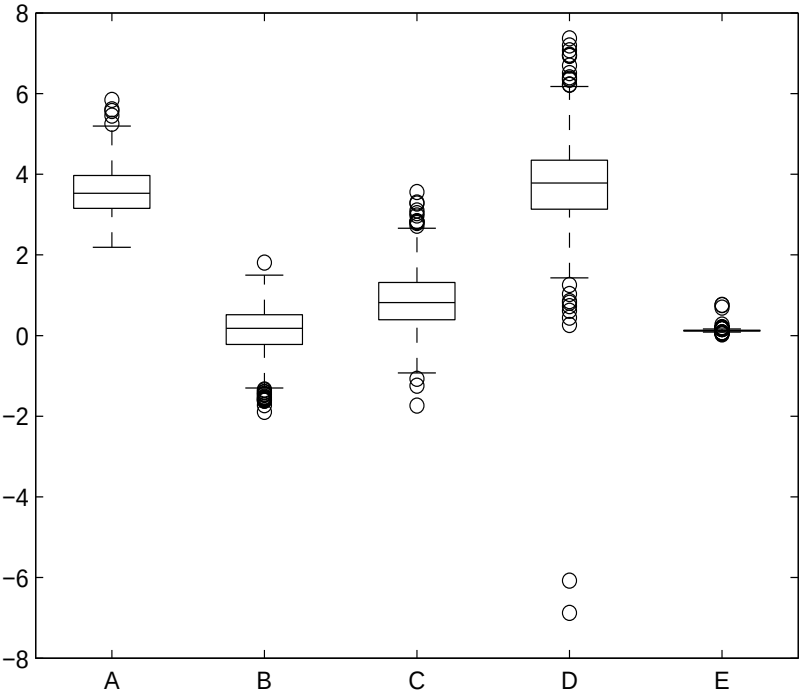

그림 4.2. 1,000번 반복하여 얻은 모수 β, ψ, γ, τ, ϕ의 추정값에 대한 박스플롯 (A, B, C, D, E열은 각각 모수 β, ψ, γ, τ, ϕ에 해당함)

표 4.2. 제안한 모형에 대해 1,000번 반복을 통해 얻은 추정값의 평균(Mean), 편향의 평균(Bias), 표준편차(SD), 표준오차의 평균(SEM), 95% 신뢰구간 포함률(CR)

Parameter	True	Mean	Bias	SE	SEM	CR(%)
α_{01}	0.0006	0.0007	0.0001	0.0005	0.0003	84.4
α_{02}	0.0019	0.0016	−0.0003	0.0013	0.0008	82.5
α_{03}	0.0257	0.0251	−0.0006	0.0152	0.0117	90.8
α_{04}	0.0001	0.0134	0.0133	0.0280	0.0094	69.1
λ_{01}	0.0018	0.0018	0.0002×10^{-1}	0.0008	0.0008	94.3
λ_{02}	0.0070	0.0072	0.0002	0.0026	0.0025	94.3
λ_{03}	0.0719	0.0741	0.0022	0.0218	0.0206	93.6
λ_{04}	0.2723	0.2806	0.0083	0.0870	0.0731	92.1
β	3.4045	3.5897	0.1852	0.6061	0.5335	94.0
ψ	0.1506	0.1334	−0.0172	0.5729	0.5106	94.0
γ	0.9049	0.8646	−0.0402	0.6982	0.5298	88.7
τ	3.8947	3.7618	−0.1328	1.0745	0.6914	84.4
ϕ	0.1252	0.1256	0.0004	0.0380	0.0116	83.5

다. CR을 중심으로 살펴보면 모수 β와 ψ는 각각 94.0%, 94.0%로 명목 값과 다르지 않지만 γ, τ, ϕ는 각각 88.7%, 84.4%, 83.5%로 명목 값보다 낮게 나타났다. 대체적으로 SEM이 SD 보다 작게 나타난 것으로 볼 때 추정량의 분산이 과소추정 된 것으로 생각되며 이로 인해 95% 명목 포함률 보다 낮아진 것으로 생각된다.

5. 고찰

본 논문에서는 자연사로 인한 관찰 시점이 종속적인 중도절단으로 작용하여 생긴 종양발생 시간과 사망 시간의 종속 관계를 모형에 포함하기 위해 감마 프레일티 3단계 모형을 제안하였다. 모수 추정은 종양 발생 시간과 프레일티 효과의 결측을 다루기 위해 EM 알고리즘 방법을 사용하였다. Lindsey와 Ryan (1993, 1994)이 분석한 방광암 자료에 대해 강제적으로 희생시킨 시점을 다르게 하여 모수를 추정한 후, 그 추정량을 바탕으로 모의실험을 수행하여 추정량의 소표본 성질을 살펴보았다. 모의실험 결과에 의하면 전반적으로 추정량의 분포는 대칭적인 형태를 보였지만 모수에 따라 참값으로부터 많이 벗어난 값들이 발견되었다. 특히 τ에 대한 추정값은 다른 모수에 비해 가장 넓게 분포하였으며 참값에서 벗어난 값도 가장 많이 나타났다. 95% 신뢰구간 포함률을 보면 모수 β와 ψ는 명목 값과 거의 같았지만 γ, τ, ϕ들은 명목 값보다 낮게 나타났다. 이와 같이 보수적인 결과가 나온 것은 SEM이 SD 보다 작게 나타난 것으로 볼 때 추정량의 분산이 과소추정 된 데 기인한 것으로 생각된다. 한편, Kim 등 (2010)은 본 논문과 동일한 문제를 다루었으며 그들은 프레일티 분포로 감마 분포 대신 정규 분포를 가정하였다. 두 분포의 모의실험 결과를 종합해 볼 때 종속적인 중도절단을 다루기 위해 제안한 방법은 프레일티 분포에 로버스트하다는 것을 알 수 있었다.

향후에는 추정량의 분산 추정량을 개선하기 위해 EM 알고리즘의 E-단계에서 사용한 Gauss-Laguerre 방법 대신 MCMC 표본 추출법 중에서 Metropolis-Hastings 방법을 적용하고자 한다. 또한 Cox 비례위험모형의 기저위험함수를 추정하기 위해 조각지수 모형을 가정했는데 구간의 개수와 경계점을 정해야 하는 제한점을 가지고 있다. 이를 극복하기 위한 방법으로 두 가지 방법을 고려할 수 있다고 생각된다. 첫째, 모수적인 방법은 조각지수 분포 대신 생존분석에서 널리 쓰이는 와이블 분포를 가정하는 것이다. 와이블 분포의 경우 구간의 개수나 경계점을 결정할 필요는 없지만 이중적분을 해야하기 때문에 컴퓨팅 시간이 길어질 것으로 예상된다. 둘째, 비모수적인 방법으로는 누적기저위험함수를 추정하고 그 추정량으로 프로파일 우도(profile likelihood)를 만들어 나머지 모수를 추정하는 방법을 제안하고자 한다.

참고문헌

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood for incomplete data via the EM algorithms (with discussion), *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

French, J. L. and Ibrahim, J. G. (2002). Bayesian methods for a three-state model for rodent carcinogenicity studies, *Biometrics*, **58**, 906–916.

Golub, G. H. and Welsch, J. H. (1969). Calculation of Gauss quadrature rules, *Mathematics of Computation*, **23**, 221–230.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, **57**, 97–109.

Huang, X. and Wolfe, R. A. (2002). A frailty model for informative censoring, *Biometrics*, **58**, 510–520.

Kim, J., Kim, Y., Nam, C., Choi, E. and Kim, Y. J. (2010). A analysis of tumorigenicity data using a normal frailty effect, *Proceedings for the Spring Conference, 2010, The Korean Statistical Society*, **13**.

Lagakos, S. W. and Louis, T. A. (1988). Use of tumor lethality to interpret tumorigenicity experiments lacking cause-of-death data, *Applied Statistics*, **37**, 169–179.

Lindsey, J. C. and Ryan, L. M. (1993). A three-state multiplicative model for rodent tumorigenicity experiments, *Applied Statistics*, **42**, 283–300.

Lindsey, J. C. and Ryan, L. M. (1994). A comparison of continuous - and discrete time three state models for rodent tumorigenicity experiments, *Environmental Health Perspective Supplements*, **102**, 9–17.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equations of state calculations by fast computing machines, *Journal of Chemical Physics*, **21**, 1087–1091.

Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*, Springer, New York.

Analysis of Tumorigenicity Data with Informative Censoring

Jinheum Kim[1] · Youn Nam Kim[2]

[1]Department of Applied Statistics, University of Suwon
[2]Graduate School of Public Health, Yonsei University

(Received June 2010; accepted August 2010)

Abstract

In animal tumorigenicity data, the occurrence time of tumor is not observed because the existence of a tumor is examined only at either time of natural death or time of sacrifice for the animal. A three-state model (Health-Tumor onset-Death) is widely used to model the incomplete data. In this paper, we employed a frailty effect into the three-state model to incorporate the dependency of death on tumor occurrence when the time of natural death works as an informative censoring against the tumor onset time. For the inference of parameters, then the EM algorithm is considered in order to deal with missing quantities of tumor onset time and random frailty. The proposed method is applied to the bladder tumor data taken from Lindsey and Ryan (1993, 1994) and a simulation study is performed to show the behavior of the proposed estimators.

Keywords: Bladder cancer data, EM algorithm, gamma frailty effect, Gauss-Laguerre method, three-state model, tumorigenicity experiment.

This work was supported by Mid-career Researcher Program through NRF grant funded by the MEST (No. R01-2008-000-20538-0).

[1]Corresponding author: Professor, Department of Applied Statistics, University of Suwon, Gyeonggi-Do 445-743, Korea. E-mail: jinhkim@suwon.ac.kr