응용통계연구 (2012)
25(2), 279–291
DOI: http://dx.doi.org/10.5351/KJAS.2012.25.2.279

결측이 있는 이산형 공변량에 대한 Cox비례위험모형의 패턴-혼합 모델

육태미[1] · 송주원[2]

[1]고려대학교 통계학과, [2]고려대학교 통계학과

(2011년 9월 30일 접수, 2012년 3월 12일 수정, 2012년 3월 27일 채택)

요 약

공변량에 결측이 발생한 Cox 비례위험 모형을 적합할 때, 결측이 발생하는 개체를 모두 제거한 후 분석을 실시한다면 정보 손실에 의해 비효율적이고 결측의 발생 메커니즘이 완전 임의 결측(missing completely at random; MCAR)이 아니라면 모수의 추정값에 편향이 발생할 수 있다. Cox 비례위험 회귀모형의 공변량에 결측이 있는 경우 적용할 수 있는 여러 가지 방법들이 제안되어져 왔으나 이 분석들은 선택모델(selection model)에 기반하고 있다. 본 연구에서는 Little (1993)이 제안한 패턴-혼합 모델(pattern-mixture model)을 사용하여 Cox 비례위험 회귀모형에서 생존시간과 결측 메커니즘의 결합분포를 모델화 하고, 여러 가지 제약에 근거한 생존 분석의 결과를 비교하였다. 모의실험을 통해서 패턴-혼합 모델의 제약(restrictions)에 따른 모수 추정의 민감도를 확인하였고 결측을 무시한 채 분석한 결과 및 선택모형에 근거한 분석결과와 비교하였다. 패턴-혼합 모델의 제약에 따라 공변량의 결측으로 인한 모수 추정의 민감성 정도를 쥐백혈병 자료 예제를 통해 설명하였다.

주요용어: Cox 비례위험 모형, 패턴-혼합 모델, 민감도 분석.

1. 서론

v개의 변수 $Z = (Z_1, \ldots, Z_v)$에 대해 n개의 개체에 대한 측정값으로 이루어진 $(n \times v)$ 자료 행렬 $\mathbb{Z} = \{z_{ij}\}$에 대한 분석 시 변수 Z가 결측을 포함하지 않고 모두 관측이 된 상태라면 일반적인 통계분석을 시행할 수 있지만, 결측을 포함하고 있는 상태라면 추정된 모수에 편향이 발생할 수 있다. 이런 경우 결측된 개체를 모두 제거하여 완전 관측 상태(complete-case; CC)로 흔히 분석하는데 이 방법은 간단하지만 비효율적이고 추정량에 편향이 발생하지 않으려면 결측이 완전히 임의로 발생하는 완전 임의 결측(missing completely at random; MCAR)이라는 강한 가정을 필요로 한다 (Little과 Rubin, 2002). 결측의 생성 원인은 결측을 포함한 변수의 결측이 완전히 임의로 발생 되었는지, 관찰된 변수들의 응답값과 연관되어 있는지 혹은 결측값 발생여부가 결측값 자체와 연관되어 있는지에 따라 각각 완전 임의 결측(MCAR), 임의 결측(missing at random; MAR), 그리고 비임의 결측(not missing at random; NMAR)으로 불린다.

Cox 비례위험 모형에서 결측이 발생하는 경우는 생존시간의 중도절단(censoring)과 공변량에 결측이 있는 두 가지의 경우가 가능하다. 공변량에 결측이 없으나 중도절단에 의해 불완전한 생존 자료의 분석을 위해 부분가능도(partial likelihood) 함수를 최대화 시키는 모수추정량을 찾는 방법이 제안되어져 왔다 (Cox, 1972, 1975). 공변량에 결측이 존재한다면 Cox의 비례위험 회귀모형의 모수 추정은 중도절단

[2]교신저자: (136-701) 서울시 성북구 안암동, 고려대학교 통계학과, 부교수. E-mail: jsong@korea.ac.kr

상태뿐 아니라 결측 자료 메커니즘(missing data mechanism)도 함께 고려하여 분석되어야 한다. 이 결측자료에 대한 모형은 Cox 비례위험 모형과 결측 자료 메커니즘에 대한 분포의 결합분포로 표현되는데 현재까지 진행되어온 대부분의 연구에서는 변수들과 결측 메커니즘의 결합 분포를 선택 모델(selection model) (Hogan과 Laird, 1997)로 가정하여 분석하는 방법들이 제안되어 왔다. 결측 자료 메커니즘이 MCAR일 때는 근사 부분가능도(approximate partial likelihood)함수를 최대화 시키는 모수 추정량을 찾는 방법으로 분석할 수 있으며 (Lin과 Ying, 1993), 이 방법은 제한된 공분산 행렬을 가지고 있는 간단하고 일치성 있는 추정량을 구조화 하고 부분가능도 함수를 최대화 시키는 모수 추정량보다 더 효율적이다. 결측자료 메커니즘이 임의결측(MAR)일 때는 비모수적 가능도(non-parametric likelihood; NPL)함수를 최대화 시켜 모수를 추정할 수 있고 (Chen과 Little, 1999), 이 모수 추정량은 관측 상태인 개체만 가지고 분석하는 방법보다 더 효율적인 것으로 나타났다. 또한 결측 자료 메커니즘이 비임의 결측(NMAR)일 때는 공변량의 모수적 분포와 결측 메커니즘을 특성화 시킨 후, 몬테카를로 EM(monte carlo EM) 알고리즘을 통해서 모수들을 추정하는 방법이 제안되었다 (Herring 등, 2004).

기존의 연구들은 모두 선택 모델을 가정한 연구이며 Cox 비례위험 회귀모형의 공변량에 결측이 있을 때 패턴-혼합 모델(pattern-mixture model) (Little, 1993)을 이용한 연구는 찾기 힘들다. 본 논문에서는 패턴-혼합 모델을 이용하여 Cox 비례위험 모형을 위한 생존 자료와 결측 메커니즘의 결합분포를 모델화하고, 이 모델을 바탕으로 회귀계수의 추정값을 계산하였다. 또한 다양한 제약을 고려하여 공변량의 결측이 모수의 추정에 미치는 영향을 살펴보았다.

2절에서는 패턴-혼합 모델의 적합을 소개하고 이를 적합하기 위한 몇 가지 제약들과 가정에 대해서 논의하고 3절에서는 Cox 비례위험 회귀모형의 공변량에 결측이 있는 경우 패턴-혼합 모델을 적용하는 분석방법을 제안한다. 4절에서는 모의실험을 통하여 패턴-혼합 모델의 Cox 비례위험 회귀모형 적용시 민감도 분석의 타당성에 대해 확인하고, 5절에서는 실제 생존 자료에 대하여 패턴-혼합 모델을 적용한 결과를 설명한다. 6절에서 결과 요약 및 토의를 통해 결론을 맺는다.

2. 패턴-혼합 모델

자료행렬 $\mathbb{Z} = \{z_{ij}\}$ $(i = 1, \ldots, n,\ j = 1, \ldots, v)$에 결측이 존재하는 경우, 결측 지시자(missing indicator) m_{ij}는

$$m_{ij} = \begin{cases} 1, & \text{만약 } z_{ij}\text{가 결측이면,} \\ 0, & \text{만약 } z_{ij}\text{가 관측이면} \end{cases}$$

로 정의되며, 결측지시행렬은 $M = \{m_{ij}\}$으로 표현할 수 있다. 결측이 발생한 자료의 분석은 자료 $\mathbb{Z}$와 결측지시행렬 M의 결합분포에 근거하여야 하는데 이 결합분포를 모델화하는 방법에는 선택 모델과 패턴-혼합 모델이 있다 (Little과 Rubin, 2002). 선택 모델은 Z와 M의 결합 분포를 Z의 주변 분포와 Z가 주어졌을 때 M의 조건부 분포로 나누어 모델화 하는 것으로

$$P(Z, M|\theta, \psi) = P(Z|\theta)P(M|Z, \psi)$$

으로 표현되고, 이 때 (θ, ψ)는 각각 Z의 주변 분포와 Z가 주어졌을 때 M의 조건부 분포의 모수를 나타낸다. 선택모델의 경우 $P(M|Z, \psi)$에 대하여 MCAR, MAR, 또는 NMAR과 같은 가정을 부여하여 분석을 실시할 수 있으며 지금까지 제안된 결측을 포함한 자료 분석 방법의 대부분이 선택모델을 가정하고 있다. 결측 자료 메커니즘이 MCAR이거나 MAR이고 모수 θ와 ψ가 별개(distinct)이면 선택모델에서 Z가 주어졌을 때 M의 조건부 분포는 무시할 수 있고(ignorable missing data mechanism이라 부름) θ의 모수추정은 Z의 주변 분포에만 근거하여 진행할 수 있게 된다.

반면, 패턴-혼합 모델에서는 Z와 M의 결합 분포가

$$P(Z, M|\phi, \pi) = P(M|\pi)P(Z|M, \phi)$$

으로 표현되며, 이때 (π, ϕ)는 각각 M의 주변 분포와 M이 주어졌을 때 Z의 조건부 분포의 알려지지 않은 모수를 나타낸다. 이 모델은 결측 메커니즘이 MCAR이 아닌 자료에 대한 유연성 있는 분석이 가능하도록 한다. 패턴이란 자료에서 결측이 발생되어 있는 모양을 나타내는 것으로 개체의 결측 지시자 $(m_{i1}, m_{i2}, \ldots, m_{iv})$의 형태에 의해서 결정된다. 동일한 결측 지시자를 갖는 개체는 같은 패턴에 속하게 되고 v개의 변수들에 대해서 이론적으로는 완전관측상태의 패턴인 P_0와 불완전관측상태의 패턴들인 $P_1, \ldots, P_{2^v-1}$가 존재하지만 실제 자료에는 일반적으로 $T \leq 2^v - 1$인 $T + 1$개의 패턴, 즉 $P_0, P_1, \ldots, P_T$가 존재한다. $P_t\,(0 \leq t \leq T)$에 속하는 개체의 개수를 $n_t\,(\sum_{t=0}^{T} n_t = n)$라 하고 P_t에 포함된 i번째 $(i = 1, \ldots, n)$ 개체에 대하여 관찰된 변수들을 $z_{obs,i}^{(t)}$로, 관찰되지 않은 변수들을 $z_{mis,i}^{(t)}$로 표현하자. 또한 r_i는 i번째 개체의 패턴이 r번째 패턴을 갖는지 여부를 나타낼 때, i번째 관찰값이 r번째 패턴에 속할 확률을 $p\,(r_i = r) = \pi_r$이라 하자. 이때 r_i의 분포는 다항분포가 되며, r_i가 주어졌을 때 z_i의 분포는 z_i의 관측된 부분과 관측 값이 주어졌을 때 z_i의 결측된 부분의 조건부 부분으로 나누어진다. 즉,

$$p\left(z_i|r_i = r, \phi^{(r)}\right) = p\left(z_{obs,i}^{(r)}|r_i = r, \phi_{obs,i}^{(r)}\right) p\left(z_{mis,i}^{(r)}|r_i = r, z_{obs,i}, \phi_{mis,i}^{(r)}\right)$$

이고, 이 때 $\phi^{(r)}$는 r번째 패턴에서의 관심 모수이며 관측된 부분의 모수 $\phi_{obs,i}^{(r)}$와 관측되지 않은 부분의 모수 $\phi_{mis,i}^{(r)}$는 $\phi^{(r)}$의 함수이다. 결측을 포함한 자료의 경우 관찰자료의 가능도 함수(observed likelihood function)는

$$L(\pi, \phi|Z, M) = L\left(\pi, \phi|Z_{obs}, M\right) = \prod_{r=0}^{T} \left\{ \pi_r^{n_r} \prod_{i \in P_r} P\left(z_{obs,i}^{(r)}|r_i = r, \phi_{obs,r}^{(r)}\right) \right\}$$

으로 표현되고 이것을 최대화시키는 모수 $\phi^{(r)}$의 값을 추정할 수 있다.

패턴에 따라 모수를 다르게 간주하고 각 패턴 중 관측된 자료에 근거해 추정할 수 있는 모수와 관측되지 않아 추정할 수 없는 모수를 구분하는 점에서 패턴-혼합 모델이 선택 모델(selection model)보다 좀 더 정직한 모형이라 할 수 있지만, 추정할 수 없는 모수에 대한 식별성(identifiability)의 문제점을 가지고 있다. 패턴-혼합 모델의 식별성 문제를 해결하는 일반적인 방법은 일종의 제약(restriction)을 적용하여, 불완전 패턴의 추정 할 수 없는 모수들이 관측된 패턴의 모수들(또는 모수들의 함수)과 같다고 가정하는 것이다. 패턴-혼합 모델은 제약이 포함되어도 기본적으로 각 패턴 별 모형은 동일한 모형에 근거하므로 관측 자료에 근거한 모델 적합을 더 복잡하게 하지 않는 장점을 지닌다 (Wang과 Daniels, 2011).

가장 흔히 고려되는 제약은 '완전히 관측된 자료-결측 변수(complete-case missing-variable; CCMV) 제약'으로, 완전 관측인 패턴 P_0상의 모수와 모든 불완전하게 관측된 패턴들에서의 식별할 수 있는 모수들이 같다고 가정하고 분석하는 것이다. 즉,

$$\phi_{mis,r}^{(r)} = \phi_{mis,r}^{(0)}$$

으로 가정한다. 패턴-혼합 모형의 CCMV 제약은 완전 관측된 패턴상의 분포를 불완전 패턴상의 분포와 동일하다고 가정 한다는 관점으로 보면 선택 모델에서 결측 메커니즘이 MCAR일 때와 대응되는 관계라 볼 수 있다. CCMV 제약은 개체의 대부분이 완전 관측상태의 패턴인 P_0에 속하고 오직 작은 비율로만 불완전 관측인 패턴들에 개체가 속할 때 사용하는 것이 합당하고 자료가 단조 형태가 아닐 때에도

부분적으로 확장 가능하다 (Thijs 등, 2002). CCMV 제약은 간단하지만 완전 관측 상태의 패턴 P_0의 표본 크기가 작으면 불안정하게 추정되어 문제가 생길 수 있다.

두 번째 고려할 수 있는 제약은 CCMV 제약을 보완하여 관측된 자료가 주어져 있을 때 관측되지 않은 자료의 조건부 분포가 완전 관측인 패턴 P_0와 불완전 관측인 패턴들 모두가 포함되어 있는 전체 패턴의 관찰된 부분에서 계산되어지는 분포와 동일하다고 가정하는 '이용 가능한 자료-결측 변수(available-case missing value, ACMV) 제약'이다. 이는 패턴들을 결합하여 $s\,(s \leq t)$개의 부분집합으로 분리한 후, 부분집합 안에서는 응답된 자료만을 사용하고 각 부분집합들끼리 모수들의 동일화를 가정하는 것이다. $\gamma^{(r)}$을 포화된 패턴-혼합(saturated pattern-mixture) 모델에서 패턴 r을 위한 i번째 개체의 분포 내의 모수라 하자. 그러면 $\gamma^{(r)}$가 S집합의 패턴과 동일화 된다는 것은

$$\gamma^{(r)} = \frac{\sum\limits_{s \in S} \pi_s \gamma^{(s)}}{\sum\limits_{s \in S} \pi_s}$$

을 나타내며 여기서 π_s는 개체의 패턴이 패턴 s를 갖는 확률을 뜻한다.

패턴-혼합 모델을 위한 CCMV, ACMV 등의 식별성 제약들은 결측에 의해 야기되는 모수 추정의 불확실성을 다루는 민감도(sensitivity)와 연결 될 수 있다. 패턴-혼합 모델의 일부 제약들은 선택 모델 하에서의 MCAR, MAR 또는 NMAR과 관련 되어질 수 있으며, 이 경우 선택 모델 하에서의 분석 결과와 연결 되어 질 수 있다 (Molenberghs 등, 1998; Wang과 Daniels, 2011).

3. 결측이 있는 이산형 공변량에 대한 Cox 비례위험 모형의 패턴-혼합 모델 적용

v개의 공변량, 생존 시간 t, 그리고 중도 절단 여부를 나타내는 δ를 포함한 생존 자료의 경우, 모든 변수가 완전 관측된 개체 i는 결측 지시자가 $m_i = (0, 0, \ldots, 0)$으로 표현 되고 불완전 관측된 개체는 0과 1의 다양한 조합으로 구성된 결측 지시자를 갖게 된다. 또한 t와 δ는 항상 관측이 된 상태를 가정한다. 결측 지시자의 형태에 따라 서로 다른 패턴이 결정되고, 결측 지시자 $(0, 0, \ldots, 0)$를 갖는 개체는 P_0의 패턴에, 그 외의 결측 지시자를 갖는 개체는 각각 $P_1, \ldots, P_T$의 패턴에 속하게 된다. 이 패턴들을 기준으로 패턴-혼합 모델에 여러 가지 제약을 설정할 수 있다.

일반적인 형태의 생존 자료에서 선택 모델의 결측 메커니즘이 MCAR이라 가정하는 것은 완전 관측된 패턴 P_0상의 분포가 불완전 패턴 상의 분포와 동일하다는 것으로, 패턴-혼합 모델의 CCMV 제약과 같아질 수 있다. 즉, 결측 메커니즘이 MCAR인 자료라면 패턴-혼합 모델에 CCMV 제약을 주어서 분석하는 것이 가능하게 된다. 그림 3.1의 경우 패턴-혼합 모델의 CCMV 제약은 패턴 P_0의 개체들에 근거하여 t, δ 그리고 $Z_1, \ldots, Z_{v-1}$가 주어졌을 때 Z_v의 조건부 분포에 대한 모형을 설정하여 모수를 추정하고 이 모수를 사용하여 패턴 P_1의 결측된 부분을 대체하고, t, δ 그리고 $Z_1, \ldots, Z_{v-2}, Z_v$가 주어졌을 때 Z_{v-1}의 조건부 분포에 대한 모형을 설정하여 패턴 P_2의 결측된 부분을 대체한다. 결측된 공변량이 두 개 이상인 경우에는 공변량들의 결합 분포를 일차원 조건부 분포들의 곱으로 특성화 시킬 수 있다 (Herring과 Ibrahim, 2001). 다른 불완전 패턴들에 대해서도 유사하게 패턴 P_0상의 개체의 관측된 값으로 모형을 설정하여 결측된 부분을 대체한다. 이때 Cox 비례위험 회귀모형의 공변량이 연속형이라면 관측된 자료가 주어져 있을 때 관측되지 않은 자료의 조건부 분포로 선형 회귀모형 등을 설정 할 수 있고, 만약 공변량이 이산형이라면 로지스틱 회귀모형을 설정하여 예측 확률을 설정 할 수 있다. 본 연구에서는 패턴-혼합 모델 적용 방법은 관측된 자료를 이용하여 제약에 따라 불완전 패턴의 결측된 부분을 예측확률을 통해서 대체하였으나 예측값에 오차 부분을 포함하여서 확률적 대체도 가능하다.

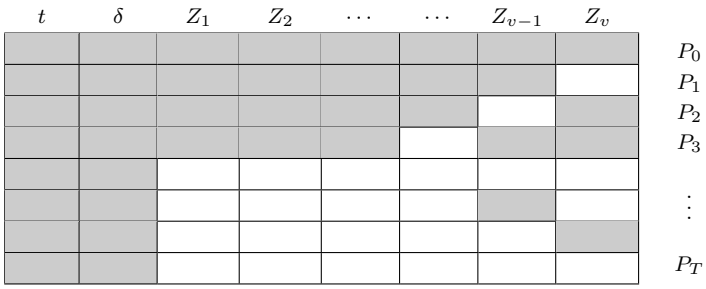

그림 3.1. 공변량을 포함한 생존 자료의 패턴(회색으로 표시된 칸들은 관측된 자료를 의미하고 흰색으로 표시된 칸들은 결측된 자료를 의미함)

단조 형태의 자료의 경우에 선택 모델의 결측 메커니즘이 MAR이라고 가정하는 것은 결측의 원인이 관측된 자료들에 의존하고 있음을 나타내고, 이는 관측된 자료가 주어져 있을 때 관측되지 않은 자료의 조건부 분포가 전체패턴의 부분군과 같다는 관점에서 ACMV 제약과 동일해 질 수 있다 (Molenberghs 등, 1998).

단조 형태의 자료에서 패턴-혼합 모델의 ACMV 제약은 전체 패턴의 부분군을 어떻게 나누느냐에 따라서 다른 결과를 제공할 수 있다. 예를 들어, 2개의 변수 Z_1, Z_2가 있는 단조 형태의 자료에서 개체들은 결측 지시자 $(0,0),(0,1),(1,1)$ 중 하나를 갖고 각각 패턴 P_0, P_1, P_2에 속하게 되며 혼합 패턴을 이용하는 패턴-혼합 모델의 ACMV 제약은 다음과 같은 단계로 진행되어 질 수 있다. 먼저, 완전 관측된 패턴 P_0상의 분포와 패턴 P_1상의 분포가 같다는 CCMV 제약 하에서 패턴 P_1의 모형을 설정하고 이 모형들의 모수를 사용하여 패턴 P_1의 결측된 부분을 대체한다. 다음으로 P_0과 대체된 P_1을 혼합하여 하나의 패턴으로 간주하고, P_0과 P_1패턴의 자료를 더한 P_{01}상의 분포와 패턴 P_2의 분포가 동일하다는 가정하에 패턴 P_2를 위한 모형을 설정하여 결측된 부분을 대체한다. 3개 이상의 변수를 포함한 단조 형태의 자료에서도 이와 유사한 방법으로 확장 시킬 수 있다. 결측된 부분이 모두 대체된 후에는 공변량에 결측이 없는 형태의 대체된 자료가 생성되므로, 결측이 없는 완전한 자료의 형태로 Cox 비례위험 모형을 적합할 수 있다. 그러나 단조형태가 아닌 일반적인 형태의 자료에서는 패턴-혼합 모델의 ACMV 제약은 결측 메커니즘이 MAR이라 가정하는 것과는 달라진다 (Wang과 Daniels, 2011).

4. 모의실험

Cox 비례위험 회귀 모형에서의 패턴-혼합모델 적용을 통한 민감도 분석의 효과를 확인하기 위하여 다음과 같은 모의실험을 시행하였다. 생존시간 t_i는 모수 $\lambda_i = \exp(z_i'\beta)$인 지수분포에서 생성하고 중도절단의 분포는 절단이 발생하지 않은 경우와 30% 절단이 발생한 경우의 두 가지 상태를 고려하였다. Cox 비례위험 회귀모형의 공변량들은 이산형 변수 2개가 존재하는 경우를 고려하였고, 첫 번째 공변량은 확률 0.5로 0 또는 1의 값을 갖고 두 공변량들은 서로 오즈비가 9라고 가정 하였다. 표본의 크기는 200개로 정하고 실험을 1000번 반복 시행하였다. 첫 번째 모의실험은 Z_1의 모든 개체가 관측되고 Z_2의 50%가 결측인 단조형태를 가정하였고, MCAR과 MAR 2가지 결측자료 메커니즘을 고려하였다. 결측 메커니즘이 MCAR일 때는 완전 임의로 결측을 생성하였고 MAR일 때는 추적(follow-up) 시간이 긴 쪽의 50%를 결측시켰다. Cox 비례위험 회귀모형의 실제 계수 값은 $(0,0),(1,0),(1,1),(1,-1),(2,-2)$로 변화시켜가며 비교하였다. 두 번째 모의실험은 두 개의 공변량에 모두 결측이 있는 경우로써 아래와 같

은 확률적 선택 과정에 의해 MAR 가정 하에서 결측을 생성하였다.

$$\text{logit}\,[p(m_1 = 1)|t, \delta, z_1, z_2] = -\frac{3t}{2},$$
$$\text{logit}\,[p(m_2 = 1)|t, \delta, z_1, m_1, z_2] = -t - (1 - m_1)z_1.$$

Cox 비례 위험 회귀 모형의 실제 계수 값은 $(0,0), (0,1), (0,2), (1,1), (2,2), (-1,1), (-2,2)$로 변화시켜 가며 비교하였다. 두 가지 모의실험을 통해서 완전히 응답된 자료 만에 근거한 분석인 '완전자료에 근거한 분석(complete-case; CC) 방법'과 완전히 응답된 자료에 패턴-혼합 모델을 적용한 방법의 Cox 비례 위험 회귀모형의 회귀계수를 결측이 발생하기 전 완전하게 응답된 자료의 회귀계수와 비교하였다.

첫 번째 모의실험의 결과는 표 4.1에 나타난다. 표에서의 완전 자료 열은 공변량에 결측이 생성되기 전의 모의실험 자료의 모수 추정값을 나타내고 각 행은 결측 메커니즘이 MCAR과 MAR일 때의 실제 추정된 계수를 나타낸다. 표 4.1(a)는 중도절단이 발생하지 않는 경우의 모수추정 결과로서 Z_1은 모두 관측 되어 있고 Z_2에 50% 결측이 있는 자료에 대한 추정이기 때문에 결측 자료 메커니즘이 MCAR일 때의 Z_1에 대응하는 회귀 계수 β_1의 추정값은 대부분의 경우 완전 자료의 추정값과 CC 방법과 패턴-혼합 방법에서 모두 비슷하게 추정 되었고 Z_2에 대응하는 회귀 계수 β_2의 추정값도 CC 방법과 패턴-혼합 방법 모두 완전 자료의 추정값과 유사 하였다. 그러나 CC 방법에서의 추정된 모수의 표준오차가 결측된 관찰값의 제외로 인해서 커짐을 볼 수 있다. 또한 계수 실제값의 절대값이 2 이상이 되는 경우의 패턴-혼합 방법의 계수 추정치는 실제값보다 과소추정되는 경향을 관찰 할 수 있다. 한편 결측 자료 메커니즘이 MAR일 때는 CC 방법은 대부분의 경우에서 심각하게 편향된 추정치를 제공해 주고 있는 반면에 패턴-혼합 방법은 결측 메커니즘이 MCAR 가정 하에서의 결과와 비슷한 추정치를 보여주고 있다. 이는 추적 시간의 긴 쪽에서 50%를 결측시킨 MAR에서 CC 방법은 추적 시간이 짧은 자료만으로 모수를 추정하므로 모수추정의 정확도가 떨어지고 패턴-혼합 방법은 이용 가능한 정보를 모두 사용하여 모수를 추정하기 때문에 생기는 결과로 CC 방법에서 오는 단점을 보완하게 된다. 그리고 결측 자료 메커니즘이 MCAR일 때와 마찬가지로 실제 계수의 절대값이 2 이상으로 커지면 패턴-혼합 방법의 제약에 사용된 대체를 위한 공변량의 로짓 모형에서의 생존시간과의 모형이 실제 Cox 비례위험 모형과 차이에서 오는 영향을 심하게 받기 때문에 계수 추정치는 과소추정되고 있음을 볼 수 있다. 표 4.1(b)는 30% 중도절단을 고려한 경우로서 중도절단의 분포가 생존시간이나 다른 공변량들의 분포와 독립인 경우이기 때문에 모수 추정은 중도절단이 없는 경우와 유사한 결과를 볼 수 있으나 중도절단으로 인한 정보 손실에 의해 중도절단이 없는 경우보다 추정의 정확성이 떨어진다.

표 4.2는 두 번째 모의실험 자료에 대하여 단일대체를 실시한 결과로써 자료의 패턴은 결측 지시자가 $(0,0)$을 갖는 경우, $(0,1)$을 갖는 경우, $(1,0)$을 갖는 경우, $(1,1)$을 갖는 경우로 나눠지고 패턴-혼합 방법은 CCMV 제약과 두 가지 ACMV 제약에 의해 분석 되었다. 두 공변량 중 한 개의 공변량만 결측인 비율은 15.4%∼22.3%에 해당하고 두 공변량이 모두 결측인 비율은 9.9%∼19.9%로 나타난다. ACMV_1 제약은 P_0의 모수에 근거하여 각각 패턴 P_1, P_2의 결측값을 대체하고 이들 P_0, P_1, P_2를 합하여 혼합패턴 P_{012}를 만들어서 P_3를 대체한 방법이다. ACMV_2 제약은 먼저 P_0의 모수에 근거하여 패턴 P_1의 결측값을 대체하고 이들 P_0, P_1을 통합하여 혼합패턴 P_{01}을 생성 한 후, 이를 이용하여 다시 모수를 추정하고 이를 이용해 패턴 P_2의 결측값을 대체하여 P_0, P_1, P_2를 합한 혼합패턴 P_{012}를 분석에 이용한 방법이다. 이 실험의 결측 자료 메커니즘은 Z_2의 결측 원인이 Z_1의 값에 의존하고 있는 MAR이다. 실제 계수값 중 적어도 하나에 0을 포함하고 있는 경우에서는 제약을 준 패턴-혼합 방법이 CC 방법보다 결측 생성 전 자료의 모수 추정값에 조금 더 근접한 값을 제공해 주고 있는데, 이는 실제 계수값이 0이라 함은 본래 공변량이 Cox 비례위험 회귀모형에 영향력을 주지 못하는 변수임을 뜻하는 것이지만 CC 방법에서는 유의하지 않는 변수의 완전히 관측된 일부분만을 사용하여 모수를 추정하기 때문에 패

표 4.1. 완전 관측된 첫 번째 공변량과 결측을 포함한 두 번째 공변량에 대한 패턴-혼합 모형 하에서의 결측값 대체 후 Cox 비례위험 회귀 모형의 회귀계수 추정치와 CC 방법 하에서의 회귀계수 추정치의 비교

			완전 자료		CC 방법		패턴-혼합 방법 (CCMV 제약)	
		TP[1)]	β_1	β_2	β_1	β_2	β_1	β_2
(a)	MCAR	$(0,0)$	-0.002[2)]	-0.002	0.002	-0.012	0.004	-0.010
			(0.174)[3)]	(0.160)	(0.252)	(0.230)	(0.175)	(0.162)
		$(1,0)$	1.015	0.001	1.019	-0.004	1.013	0.011
			(0.191)	(0.160)	(0.277)	(0.231)	(0.192)	(0.162)
		$(1,1)$	1.006	1.007	1.025	1.020	0.971	1.002
			(0.192)	(0.173)	(0.279)	(0.251)	(0.195)	(0.178)
		$(1,-1)$	1.005	-1.013	1.014	-1.032	0.928	-0.959
			(0.187)	(0.172)	(0.270)	(0.249)	(0.182)	(0.170)
		$(2,-2)$	2.005	-2.006	2.016	-2.020	1.725	-1.755
			(0.218)	(0.203)	(0.315)	(0.295)	(0.202)	(0.191)
	MAR	$(0,0)$	-0.002	-0.002	-0.003	0.002	0.032	-0.014
			(0.174)	(0.160)	(0.251)	(0.229)	(0.172)	(0.169)
		$(1,0)$	1.015	0.001	0.165	-0.015	1.020	-0.091
			(0.191)	(0.160)	(0.298)	(0.223)	(0.187)	(0.167)
		$(1,1)$	1.006	1.007	0.116	0.207	0.986	0.845
			(0.192)	(0.173)	(0.335)	(0.244)	(0.199)	(0.188)
		$(1,-1)$	1.005	-1.013	0.249	-0.254	0.759	-0.961
			(0.187)	(0.172)	(0.265)	(0.227)	(0.174)	(0.173)
		$(2,-2)$	2.005	-2.006	0.756	-0.769	1.524	-2.147
			(0.218)	(0.203)	(0.287)	(0.239)	(0.192)	(0.205)
(b)	MCAR	$(0,0)$	-0.010	-0.002	-0.007	0.000	-0.009	-0.001
			(0.203)	(0.186)	(0.293)	(0.268)	(0.205)	(0.189)
		$(1,0)$	1.004	0.015	1.009	0.005	1.005	0.015
			(0.212)	(0.175)	(0.306)	(0.252)	(0.214)	(0.177)
		$(1,1)$	0.994	1.003	1.007	1.024	0.965	1.008
			(0.233)	(0.196)	(0.337)	(0.283)	(0.236)	(0.201)
		$(1,-1)$	1.003	-1.012	1.009	-1.015	0.930	-0.942
			(0.216)	(0.194)	(0.311)	(0.279)	(0.212)	(0.192)
		$(2,-2)$	2.018	-2.015	2.039	-2.030	1.748	-1.745
			(0.246)	(0.223)	(0.356)	(0.323)	(0.233)	(0.211)
	MAR	$(0,0)$	-0.010	-0.002	-0.006	-0.004	0.007	0.011
			(0.203)	(0.186)	(0.266)	(0.243)	(0.201)	(0.192)
		$(1,0)$	1.004	0.015	0.136	0.004	0.987	-0.047
			(0.212)	(0.175)	(0.308)	(0.230)	(0.208)	(0.180)
		$(1,1)$	0.994	1.003	0.148	0.210	1.012	0.743
			(0.233)	(0.196)	(0.351)	(0.254)	(0.240)	(0.208)
		$(1,-1)$	1.003	-1.012	0.258	-0.263	0.769	-0.934
			(0.216)	(0.194)	(0.279)	(0.239)	(0.204)	(0.193)
		$(2,-2)$	2.018	-2.015	0.767	-0.768	1.544	-2.051
			(0.246)	(0.223)	(0.299)	(0.248)	(0.223)	(0.220)

(a) 중도절단이 없는 경우, (b) 30%중도절단을 고려한 경우;
1) 실제 회귀계수값; 2) 1000번 반복시행한 회귀계수 추정치들의 평균값; 3) 괄호는 1000번 반복 시행한 회귀계수의 추정량들의 표준오차 평균값

표 **4.2.** 첫 번째 공변량 Z_1과 두 번째 공변량 Z_2 모두에 결측이 있을 경우 패턴-혼합 모형 하에서의 결측값 대체 후 Cox 비례위험 회귀 모형의 회귀계수 추정치와 CC 방법 하에서의 회귀계수 추정치의 비교

							패턴-혼합 방법					
			완전 자료		CC 방법		CCMV 제약		ACMV_1 제약		ACMV_2 제약	
	TP[1)]	MP[2)]	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
(a)	$(0,0)$	(15.4, 15.4, 9.9)	-0.007[3)]	0.006	0.086	0.006	0.039	-0.010	0.026	-0.011	0.000	0.000
			(0.160)[4)]	(0.175)	(0.213)	(0.236)	(0.161)	(0.177)	(0.161)	(0.176)	(0.161)	(0.176)
	$(0,1)$	(19.5, 17.5, 14.5)	-0.002	1.015	0.073	1.228	0.044	1.027	0.028	1.022	-0.001	1.035
			(0.160)	(0.191)	(0.244)	(0.287)	(0.163)	(0.195)	(0.162)	(0.194)	(0.161)	(0.193)
	$(0,2)$	(21.2, 18.1, 17.7)	-0.008	2.038	0.046	2.520	0.002	2.079	-0.008	2.065	-0.032	2.079
			(0.160)	(0.240)	(0.268)	(0.414)	(0.164)	(0.247)	(0.163)	(0.245)	(0.162)	(0.244)
	$(1,1)$	(21.0, 18.0, 16.9)	0.999	1.023	1.269	1.248	1.024	1.026	1.012	1.028	0.979	1.040
			(0.173)	(0.192)	(0.288)	(0.303)	(0.178)	(0.198)	(0.177)	(0.197)	(0.175)	(0.196)
	$(2,2)$	(22.3, 18.3, 19.9)	2.018	2.019	2.339	2.459	1.787	1.903	1.781	1.910	1.744	1.925
			(0.215)	(0.234)	(0.423)	(0.424)	(0.213)	(0.245)	(0.212)	(0.244)	(0.209)	(0.244)
	$(-1,1)$	(16.5, 16.1, 11.2)	-1.010	1.009	-1.082	1.172	-0.891	0.922	-0.911	0.919	-0.940	0.933
			(0.172)	(0.187)	(0.249)	(0.267)	(0.173)	(0.188)	(0.173)	(0.187)	(0.172)	(0.186)
	$(-2,2)$	(16.9, 16.1, 12.3)	-2.008	2.013	-2.283	2.361	-1.549	1.590	-1.573	1.591	-1.590	1.592
			(0.204)	(0.218)	(0.325)	(0.340)	(0.190)	(0.205)	(0.190)	(0.204)	(0.189)	(0.203)
(b)	$(0,0)$	(15.4, 15.4, 9.9)	-0.008	0.008	0.104	0.000	0.044	-0.008	0.025	-0.007	-0.006	0.005
			(0.186)	(0.203)	(0.260)	(0.288)	(0.188)	(0.205)	(0.188)	(0.205)	(0.188)	(0.204)
	$(0,1)$	(19.5, 17.5, 14.5)	-0.003	1.017	0.077	1.241	0.048	1.029	0.029	1.028	-0.003	1.040
			(0.181)	(0.221)	(0.287)	(0.345)	(0.184)	(0.226)	(0.183)	(0.225)	(0.182)	(0.224)
	$(0,2)$	(21.2, 18.0, 17.9)	-0.003	2.023	0.049	2.577	0.011	2.088	-0.001	2.072	-0.021	2.078
			(0.181)	(0.275)	(0.320)	(0.487)	(0.186)	(0.286)	(0.185)	(0.283)	(0.183)	(0.281)
	$(1,1)$	(21.0, 18.0, 16.8)	1.007	1.005	1.305	1.252	1.043	1.000	1.034	1.001	0.988	1.015
			(0.198)	(0.236)	(0.345)	(0.395)	(0.205)	(0.244)	(0.204)	(0.243)	(0.201)	(0.242)
	$(2,2)$	(22.2, 18.3, 19.8)	2.020	2.048	2.321	2.490	1.780	1.906	1.777	1.906	1.750	1.921
			(0.234)	(0.286)	(0.464)	(0.537)	(0.231)	(0.296)	(0.229)	(0.296)	(0.227)	(0.295)
	$(-1,1)$	(16.5, 16.1, 11.4)	-1.016	1.013	-1.103	1.210	-0.882	0.917	-0.906	0.921	-0.946	0.934
			(0.200)	(0.223)	(0.306)	(0.343)	(0.201)	(0.225)	(0.202)	(0.224)	(0.201)	(0.223)
	$(-2,2)$	(16.9, 16.1, 12.5)	-2.018	2.010	-2.368	2.478	-1.545	1.586	-1.567	1.578	-1.608	1.586
			(0.228)	(0.253)	(0.383)	(0.426)	(0.215)	(0.242)	(0.215)	(0.240)	(0.215)	(0.239)

(a) 중도절단이 없는 경우, (b) 30%중도절단을 고려한 경우; 1) 실제 회귀계수값; 2) 결측 비율의 평균(첫 번째 공변량만 결측, 두 번째 공변량만 결측, 두 공변량 모두 결측); 3) 1000번 반복시행한 회귀계수 추정치들의 평균값; 4) 괄호는 1000번 반복시행한 회귀계수 추정량들의 표준오차 평균값

턴-혼합 방법보다 모수 추정에 이용하는 정보가 부족하여 생기는 결과로 추측된다. 그리고 CCMV 제약에서의 모수 추정값은 CC 방법의 추정값보다 ACMV 제약의 추정값에 더 근접한 것을 볼 수 있는데 이는 자료가 MAR인 결측 매커니즘으로 인해 CC 방법에서 제외된 관찰값이 CCMV 제약의 모수 추정에 영향을 주었기 때문이다. 또한 관찰된 자료의 정보를 다르게 사용하는 ACMV_1 제약과 ACMV_2 제약은 서로 유사한 결과를 보여주고 있음을 볼 수 있다.

자료가 일반적인 형태이기 때문에 ACMV 가정을 한 패턴-혼합 방법들이 결측 자료 메커니즘 MAR에 대응한다고 볼 수 없고 그로 인해 실제 계수의 변화에 따라 일정한 결과를 제공하지 않은 것으로 보여진다. 하지만 패턴-혼합 방법에서의 계수 추정값은 실제 계수의 절대값이 커질수록 추정의 정확도가 떨어지며 모수를 과소추정하고 있었고 이는 첫 번째 모의실험에서 확인한 것과 같은 결과이다. CC 방법 또한 결측 메커니즘이 MAR에서는 좋은 결과를 나타내지는 않았다. 30% 중도절단이 고려되는 경우에서는 첫 번째 모의실험의 결과와 동일하게 중도절단이 없는 경우보다 추정의 정확성이 떨어진다.

결과적으로, 두 개의 이산형 공변량을 포함하는 Cox 비례위험 회귀 모형에서 단조 형태를 따르는 생존자료는 결측 메커니즘이 MCAR일 때는 CC 방법과 패턴-혼합 방법이 비슷한 성능을 보여주었고 결측 메커니즘이 MAR일 때 CC 방법은 실제 계수값이 $(0, 0)$일 때를 제외한 대부분 상황에서 정확한 모수 추정에 실패하였고 패턴-혼합 방법이 CC 방법보다 실제 계수 추정값에 유사함을 확인할 수 있었다. 또한, 패턴-혼합 모델을 사용하여 변수와 결측 자료 메커니즘의 결합 분포를 모델화 하는 것은 여러 가지 제약에 따라 다른 추정 결과를 제공해주기 때문에 결측 자료로 인한 모수 추정의 민감도 분석을 가능하게 한다. 그러나 Cox 비례위험 모형을 위한 두 개의 이산형 공변량의 실제 모수의 절대값이 2 이상인 경우에서는 패턴-혼합 방법을 이용한 결측값의 대체는 결과의 신뢰성을 떨어뜨릴 수 있으며, 전체 자료의 크기에 비해 공변량의 결측이 차지하는 비율이 높다면 모수의 추정에 있어서 잘못된 결과를 제공할 수 있다. 모의실험에 의하면 Cox 비례위험 모형의 공변량이 결측을 포함한 자료에 대한 패턴-혼합 모델의 적용은 결측 메커니즘이 MCAR일 때와 일부 MAR인 상황에서 결측된 개체를 모두 제거하고 완전관측 상태의 개체로만 분석하는 방법보다 패턴-혼합 방법을 적용하여 분석한 결과가 더 나을 수 있음을 보여주고 가정에 따른 모수 추정의 민감도를 평가할 수 있다.

5. 예제

Kalbfleisch와 Prentice (1980)의 쥐 백혈병 자료에 대하여 패턴-혼합 모형의 분석을 시행하였다. 이 자료는 Fred Hutchinson Cancer Center의 Dr. Robert Nowinski의 연구실에서 진행된 실험 자료로써 유전요인과 바이러스 요인이 백혈병에 걸린 쥐의 생존기간에 미치는 영향에 관한 연구를 목적으로 하고 있다. 자료에는 204마리의 쥐의 생존시간과 중도절단 여부를 포함한 9개 변수가 기록되어 있다. 본 연구에서는 두 개의 공변량 Gpd-1 형질(Z_1)과 바이러스 레벨(Z_2)을 고려한다. 각 공변량은 상태에 따라 0 또는 1의 값을 갖는다. 공변량의 패턴은 두 공변량이 모두 관측인 P_0, Z_2에만 결측을 가진 P_1, Z_1에만 결측을 가진 P_2 그리고 두 공변량이 모두 결측인 P_3로 총 4개의 패턴으로 나눌 수 있다. 각각의 패턴에 100, 1, 75, 28개의 개체가 존재하였으며 패턴-혼합 모형을 설정하고 CCMV 제약과 ACMV_1 제약을 적용하였다.

표 5.1은 CC 방법과 NPL 방법, 그리고 ACMV_1 제약을 적용한 패턴-혼합 방법을 사용하여 Cox 회귀 모델의 회귀계수를 추정한 결과이다. NPL 방법은 Chen과 Little (1999)이 제안한 모수 추정 방법으로, 결측 메커니즘이 MAR이라는 가정 하에서 비모수적 가능도 함수를 최대화시키는 모수를 추정하는 선택 모델에 근거한 방법이며 공변량이 범주형인 자료에서 좋은 추정치를 제공해준다고 알려져 있다. 패턴 P_1에 포함된 개체가 1개이기 때문에 앞에서 제시한 ACMV_1 제약과 ACMV_2 제약이 유사한 결과를

표 **5.1.** 쥐 백혈병 자료의 CC 방법, NPL 방법과 패턴-혼합 방법의 Cox 비례위험 회귀모형의 회귀계수 비교

	CC 방법		NPL 방법		CCMV 제약		ACMV_1 제약	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
추정값	1.483	0.732	1.496	1.007	1.551	0.834	1.400	1.011
추정값의 분산	0.301	0.297	0.250	0.180	0.119	0.126	0.113	0.131
p-값	0.007*	0.179	<.001*[1)]	<.001*	<.001*	0.019*	<.001*	0.005*

1)∗: 귀무가설 "계수 값이 0이다"에 대해 유의수준 5%에서 유의함

보여주므로 ACMV_1 제약을 적용한 패턴-혼합 방법만을 시행하였다. Cox 회귀분석에서 불완전한 관측 값들을 포함했을 때 Gpd-1형질과 바이러스 레벨이 백혈병에 의한 죽음에 유의한 연관을 보인다고 할 수 있다. 추적 시간이 짧은 개체에 대해서 결측이 많이 발생하였다고 알려진 이 자료의 성질 때문에 CC 방법의 모수 추정값은 편향이 있을 것이고 모수 β_2의 추정은 p-value가 0.179로 유의하지 않은 것을 볼 수 있다. 또한 CC 방법의 모수 추정값은 NPL 방법과 패턴-혼합 모형의 ACMV_1 제약과 CCMV 제약의 경우보다 분산이 큰 것을 확인할 수 있다. 패턴-혼합 방법의 모수 추정값은 NPL 방법보다 다소 작은 분산을 가지고 있음을 볼 수 있고, ACMV_1 제약과 CCMV 제약의 모수 추정값은 다소 차이가 있으나 각각 p-value가 0.005와 0.019로 나타나 두 추정 값 모두 유의수준 5% 하에서 0과 유의하게 다르다. 쥐 백혈병 자료의 패턴은 일반적인 형태로 단조 형태가 아니지만, 패턴 P_1에 속하는 개체가 1개로 거의 단조 형태를 보이는 자료라고 말할 수 있기 때문에 ACMV_1 가정이나 CCMV 가정을 준 패턴-혼합 모델에서의 분석이 CC 방법보다 합당 할 것으로 고려된다. 그러나 이 분석에서는 Gpd-1형질과 바이러스 레벨 이외의 다른 요소들을 고려하지 않았기 때문에 백혈병에 의한 죽음과 두 요소간의 관계에 대한 결과들이 확정적이라고 말할 수는 없다.

6. 토의

패턴-혼합 모델에서 추정할 수 없는 모수가 존재하는 문제를 해결하기 위한 하나의 대안으로 모델에 CCMV 또는 ACMV 등의 제약을 주는 방법이 있으며, 패턴-혼합 모델은 모델의 오설정(misspecification)에 매우 민감한 것으로 알려져 있다 (Demirtas, 2005). 모의실험을 통해서 Cox 비례위험모형의 제약을 준 패턴-혼합 모델의 적용은 CC 방법과 완전 자료의 모수 추정치의 중간정도의 추정량을 제공하고 있음을 확인하였고 패턴-혼합 모델의 여러 가지 제약에 따라 모수 추정량에 차이가 있음을 보았다. 이는 결측에 의한 모수값의 불확실성을 제시하는 것으로 추정량의 민감도를 의미하며 결측이 있는 공변량에 대한 Cox 비례위험 모형에 대해 CCMV 제약과 여러 가지 ACMV 제약을 설정한 패턴-혼합 모델의 적용은 모수 추정의 민감도 분석을 가능하게 한다.

모의실험의 결과에 따르면 단조 형태의 Cox 비례위험 회귀 모형의 생존 자료에서 결측 메커니즘이 MCAR일 때와 본 연구에서 고려한 추적(follow-up)시간이 긴 쪽의 일부가 결측이 된 MAR 가정에서는 패턴-혼합 방법이 CC(complete- Case) 방법의 회귀계수 추정 값보다 더 나은 성능을 보여주었다. 그리고 비단조 형태인 일반적인 자료의 경우에는 패턴-혼합 방법이 CC 방법보다 더 좋은 결과를 제공해 줄 수 있음을 보았다. 이와 같은 결과는 CCMV 제약 하에서의 패턴-혼합 방법에서도 발견되었는데 이는 본 연구에서 고려한 Cox 비례위험 모형에 대한 CCMV 제약 하에서의 모수의 추정 방법에 기인한 것으로 보인다. 일반적인 CCMV 제약 하에서는 완전한 자료의 모수에 근거하여 전체 자료의 모수를 추정하는 데 반하여 본 연구에서는 완전하게 측정된 자료들로부터 추정된 모수를 사용하여 공변량의 결측값을 대체한 후 대체된 자료를 사용하여 Cox 비례위험 모형의 모수를 추정하는 방법을 고려하였다. 대체를 실시할 때 완전하게 응답된 자료만에 근거한 모수를 사용하지만 대체는 연관된 변수들을 포함하므

로 MAR 가정 하에서 실시되어 대체된 자료에 근거한 Cox 비례위험 모형의 모수의 추정치는 CC 방법에 근거한 모수의 추정치보다 나은 결과를 보이는 것으로 나타났다.

한편, 결측 자료 메커니즘이 MCAR이나 MAR일 때, 실제 계수의 절대값이 커질수록 패턴-혼합 방법에서의 모수 추정이 실패하는 경향이 나타났는데 이는 대체를 위해 사용된 공변량과 생존시간의 모형이 실제 Cox 비례위험 모형과 차이가 있기 때문에 일어나는 현상으로 추측된다. 즉, 실제 자료 분석에서 대체 모형을 선택할 때는 결측이 발생한 변수의 형태에 따라 모형을 선택하는데 이산형 공변량에 결측이 발생하였으므로 대체 모형으로 로짓모형을 선택하였고 이 때 연관변수로 생존시점을 포함하고 이를 일반적으로 사용하는 선형 회귀계수를 이용해 추정하였다. 하지만, Cox 비례위험 모형에서 생존시간과 공변량의 연관관계는 비선형이므로 대체모형의 선형 연관성 가정은 실제와 차이가 발생하게 되며 이에 따라 대체모형의 예측에 편향이 발생하는 것으로 파악되었다. 이와 같은 문제점은 본 연구에서 고려한 대체모형보다 더 적절한 모형설정으로 해결될 수 있을 것으로 기대한다. 모의실험의 결과에는 제시하지 않았지만 패턴-혼합모델의 다중 대체방법 적용 역시 단순 대체와 유사한 결과를 제공해 주었다.

본 연구에서 사용한 Cox 비례위험 모형의 모수추정에 패턴-혼합 모델을 이용하는 것은 특정한 제약 하에서 공변량의 결측된 부분을 대체한다는 의미로 볼 때 베이지안 기법인 MCMC(markov chain monte carlo) 기법을 사용하여 확장하는 것도 가능할 것이다. 또한 결측 메커니즘이 MCAR일 때와 일부 MAR인 상황에서 결측된 개체를 모두 제거하고 완전 관측 상태의 개체로만 분석하는 방법보다 패턴-혼합 방법을 적용하여 분석한 결과가 더 나을 수 있음을 보여주고 가정에 따른 추정의 민감도를 평가하였다.

본 연구의 모의실험에서는 다양한 가정 하에서의 패턴-혼합 모형의 추정 결과를 CC 방법의 결과와 비교하였다. CC 방법은 선택모형 하에서 결측자료 메커니즘을 MCAR로 가정한 분석의 결과와 유사한데 결측자료 메커니즘을 MAR 또는 NMAR이라 가정한 후 선택모형을 사용한 추정 결과와 비교를 진행하는 것도 선택모형과 패턴-혼합 모형의 유사성 및 차이점을 파악하는 데 도움이 될 것으로 기대된다.

본 연구의 모의실험은 MCAR과 MAR 가정 하에서 생성된 자료에 대하여 실시하였다. 결측 자료에 대하여 일반적으로 위 두 가정 하에서 분석이 실시되기 때문에 패턴-혼합 모형 하에서 흔히 사용되는 제약들이 MCAR이나 MAR 가정과 유사하게 제안되어 왔으며 이들 가정 하에서의 패턴-혼합 모형의 민감도 분석 결과를 살펴보고 이를 선택모형의 결과와 유사하게 해석하려는 시도였다. NMAR 가정 하에서 생성된 자료에 대하여 다양한 제약을 고려하고 이 제약들 하에서의 민감도 분석 결과를 비교하는 연구가 추후에 진행되면 민감도 분석 및 결과 해석에 중요한 의미를 지닐 것이다.

참고문헌

Chen, H. Y. and Little, R. J. A. (1999). Proportional hazards regression with missing covariates, *Journal of the American Statistical Association*, **94**, 896–908.

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

Cox, D. R. (1975). Partial likelihood, *Biometrika*, **62**, 269–279.

Demirtas, H. (2005). Multiple imputation under Bayesianly smoothed pattern-mixture models for non-ignorable drop-out, *Statistics in Medicine*, **24**, 2345–2363.

Herring, A. H. and Ibrahim, J. G. (2001). Likelihood-based methods for missing covariates in the Cox proportional hazards model, *Journal of the American Statistical Association*, **96**, 292–302.

Herring, A. H., Ibrahim, J. G. and Lipsitz, S. R. (2004). Non- ignorable missing covariate data in survival analysis: A case-study of an International Breast Cancer Study Group trial, *Journal of the Royal Statistical Society*, **53**, 293–310.

Hogan, J. W. and Laird, N. M. (1997). Model-based approaches to analysing incomplete longitudinal and failure time data, *Statistics in Medicine*, **16**, 259–272.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*, Wiley, New York.

Lin, D. Y. and Ying, Z. (1993). Cox regression with incomplete covariate measurements, *Journal of the American Statistical Association*, **88**, 1341–1349.

Little, R. J. A. (1993). Pattern-Mixture models for multivariate incomplete data, *Journal of the American Statistical Association*, **88**, 125–134.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis With Missing Data*, Wiley, New York.

Molenberghs, G., Michiels, B., Kenward, M. G. and Diggle, P. J. (1998). Monotone missing data and pattern-mixture models, *Statistica Neerlandica*, **52**, 153–161.

Thijs, H., Molenberghs, G., Michiels, B., Verbeke, G. and Curran, D. (2002). Strategies to fit pattern-mixture models, *Biostatistics*, **3**, 245–265.

Wang, C. and Daniels, M. J. (2011). A note on MAR, identifying restrictions, model comparison, and sensitivity analysis in Pattern-Mixture models with and without covariates for incomplete data, *Biometrics*, **67**, 810–818.

Pattern-Mixture Model of the Cox Proportional Hazards Model with Missing Binary Covariates

Tae Mi Youk[1] · Juwon Song[2]

[1]Department of Statistics, Korea University; [2]Department of Statistics, Korea University

(Received September 30, 2011; Revised March 12, 2012; Accepted March 27, 2012)

Abstract

When fitting a Cox proportional hazards model with missing covariates, it is inefficient to exclude observations with missing values in the analysis. Furthermore, if the missing-data mechanism is not Missing Completely At Random(MCAR), it may lead to biased parameter estimation. Many approaches have been suggested to handle the Cox proportional hazards model when covariates are sometimes missing, but they are based on the selection model. This paper suggest an approach to handle Cox proportional hazards model with missing covariates by using the pattern-mixture model (Little, 1993). The pattern-mixture model is expressed by the joint distribution of survival time and the missing-data mechanism. In the pattern-mixture model, many models can be considered by setting up various restrictions, and different results under various restrictions indicate the sensitivity of the model due to missing covariates. A simulation study was conducted to show the sensitivity of parameter estimation under different restrictions in a pattern-mixture model. The proposed approach was also applied to mouse leukemia data.

Keywords: Cox proportional hazards model, missing covariates, pattern-mixture model, sensitivity analysis.

[2]Corresponding author: Associate Professor, Department of Statistics, Korea University, Anam-Dong, Seongbuk-Gu, Seoul 136-701, Korea. E-mail: jsong@korea.ac.kr