The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.1.113
29(1), 113–122

Joint model of longitudinal data with informative observation time and competing risk

Yang-Jin Kim[a,1]

[a]Department of Statistics, Sookmyung Women's University

(Received December 14, 2015; Revised January 15, 2016; Accepted January 15, 2016)

Abstract

Longitudinal data often occur in prospective follow-up studies. Joint model for longitudinal data and failure time has been applied on several works. In this paper, we extend it to the case where longitudinal data involve informative observation time process as well as competing risks survival times. We use a likelihood approach and derive an EM algorithm to obtain maximum likelihood estimate of parameters. A suggested joint model allows us to make inferences for three components: longitudinal outcome, observation time process and competing risk failure time. In addition, we can test the association among these components. In this paper, liver cirrhosis patients' data is analyzed. The relationship between prothrombin times measured at irregular visiting times and drop outs is investigated with a joint model.

Keywords: competing risk, Drop out, informative observation process, joint model, longitudinal data, random effect

1. 서론

반응변수를 일정 기간 동안 반복적으로 측정하는 대표적인 자료형태로 경시적 자료와 시계열 자료가 있다. 이 두 자료의 차이점은 관측 대상의 수와 반복 횟수 그리고 연구 목적에 있다. 일반적으로 경시적 자료는 반복 횟수가 그리 커지 않고 고정(fixed)되어 있으며 관측대상은 무한히 커질 수 있다. 반면에 시계열 자료는 관측 대상이 고정되어 있으며 반복 수는 충분히 클 때 적용된다. 즉, 경시적 자료의 관측대상은 전체 관심 있는 모집단에서 추출된 표본인데 반해, 시계열 자료에서는 개개의 관측대상이 관심의 대상이 된다. 따라서 경시적 자료에서 추정된 공변량 효과는 모집단의 효과(population effect)로써 개체들간의 특성을 공유함으로써 추정된다. 이러한 성질덕분에 모형의 가정(예를 들어, 상관관계에 대한 가정)에 대해 시계열자료보다 훨씬 로버스트하다고 알려져 있다 (Diggle 등, 2001). 또한 사회학과 경제학에서 언급하는 패널자료도 경시적 자료와 비슷한 성질을 가지고 있지만 분석하고자 하는 자료의 성질에 따라 분석 방법이 조금씩 차이가 존재하게 된다 (Free, 2004).

본 연구의 주 관심은 경시자료가 서로 다른 시점에서 서로 다른 반복횟수를 가지고 관측될 때, 관측 시점과 관측 중도 절단의 원인에 대한 모형을 경시자료와 함께 고려하고자 한다. 이러한 복잡한 자료 구

This research was supported by the Sookmyung Women's University research grant 2014.
[1]Department of Statistics, Sookmyung Women's University, Chengpa-ro 47-gil 100, Yongsan-Gu, Seoul 04310, Korea. E-mail: yjin@sookmyung.ac.kr

조 패턴하에서 세 가지 요소(경시적 자료, 관측 시점, 관측 종료 원인)를 동시에 고려하기 위해 결합 모형(joint model)이 적용되며 이 세 가지 요소간의 연관성을 위해 랜덤 효과(random effect)가 고려된다. 예를 들어 간경변증를 앓고 있는 환자를 대상으로 신약의 효과를 추정하기 위해 혈액 응고 시간을 반복적으로 측정하는 연구를 고려해보자. 여기서 환자들은 간이식을 받거나 사망함으로써 관측을 중단하게 된다. 만약 이 두 사건이 발생하지 않았다면 그들의 혈액 응고 시간은 연구 종료시점까지 계속 측정될 수 있었을 것이다. 특히 관측 중도 절단은 결측 자료의 원인을 제공하는데 경시적 자료의 결측 패턴은 크게 단조 결측(monotone missing)과 비단조 결측(non-monotone missing)으로 나누게 된다. 이 때, 관측 중도절단은 단조 결측의 예가 된다. 이러한 결측 발생의 확률구조에 대해 다음의 세가지 모형이 사용된다. 결측 확률이 완전 임의로 발생하는 경우는 MCAR(missing completely at random)로 이러한 결측 발생 확률 구조 하에서는 완전하게 관측된 자료만을 사용하여 분석하는 완전 관측 개체 방법(complete case method)을 적용해도 무방하다. 하지만 임상실험에서의 관측 중단이 이러한 가정을 만족하는 경우는 매우 드물다. 따라서 결측 발생 확률에 대한 결측 자료와의 관계를 통해 다음의 두 가지 가정이 고려된다. 구체적으로 결측 발생 확률은 이미 관측된 자료에 의존하지만 결측 자료와는 무관함을 가정하는 MAR(missing at random)과 결측 발생이 관측되지 않은 결측 자료와 연관성이 존재함을 가정하는 MNAR(missing not at random) 또는 NIR(non ignorable) 모형으로 분류할 수 있다. 예를 들어, 간경변증 자료에서 간이식과 사망은 관측 중도 절단의 원인이며 이 사건의 발생은 관측 중단으로 인해 관측 되지 못한 환자의 상태와 연관되었을 가능성이 있다. 결합 모형은 임상실험에서 경시 자료와 생존 자료의 상관관계를 모형화하기 위해 가장 많이 적용되는 방법론이다 (Rizopoulos, 2012). 최초의 결합 모형은 생존 자료의 비례 위험 모형(proportional hazard model)에서 오차를 가지고 측정한 시간 가변 공변량(time-varying covariate)의 효과를 추정하기 위해 제안된 모형이었다 (Wulfsohn과 Tsiatis, 1997). 이에 Henderson 등 (2000)는 경시 자료와 생존 자료에 대한 두 개의 submodel을 결합하기 위해 이 모형을 적용시켰다. Elashoff 등 (2007)은 경시적 자료와 경쟁 위험 모형을 동시에 고려한 결합 모형을 제시하였으며 모든 경쟁 사건의 모형에 동일한 랜덤 효과를 적용하였다. Liu와 Huang (2009)는 경시자료와 재발 사건간의 연관관계를 위한 결합모형을 고려하였다. 따라서 이들 논문은 두 가지 변량의 결합 모형을 고려한데 반해 본 논문에서는 경시적 자료와 측정 시점(재발 사건) 그리고 경쟁 위험 모형을 동시에 결합한 모형을 연구한다. 이러한 결합 모형의 적용을 통해 개별적 분석을 적용할 경우 구할 수 없는 세가지 모형간의 연관관계를 조사할 있다는 점이 이 모형의 장점이라고 할 수 있다. 본 저자가 알고 있는 한 이 세 가지 모형을 동시에 고려한 연구는 국내외에서 아직 발견되지 못했다.

본 논문에서는 다른 결합 모형과 마찬가지로 공통된 특성 또는 연관 관계를 모형화하기 위해 랜덤효과를 적용하였으며 개개인의 효과(subject-specific effect)를 추정할 수도 있다. 2장에서는 경시 자료 분석을 위해 널리 사용되는 방법들을 간략하게 요약하며 3장에서는 본 논문에서 분석할 자료 구조와 결합모형을 소개한다. 4장에서는 결합 모형을 통해 간경변증환자의 혈액 응고 시간, 병원 방문 시점 그리고 관측 중단 원인과의 관계를 분석하기 위해 앞에서 제안한 결합 모형을 적용한다. 5장에서는 향후 관련연구를 소개한다.

2. 경시자료 분석에 대한 간략한 요약

반복적으로 관측되는 경시자료의 분석 목적은 시간에 따라 변화하는 패턴과 공변량의 효과를 추정하는 것이다. 먼저 경시자료 $y_i = (y_{i1}, \ldots, y_{im_i})$, $i = 1, \ldots, n$가 연속형 자료라고 가정하자. 경시적 자료에 대한 가장 초기의 모형으로는 잘 알려진 ANOVA(analysis of variance) 모형의 확장이 있다. 반복측정 자료(repeated measurement data)로 언급되기도 하는데 이 경우엔, 모든 관측 대상의 관측 시점과 관측 횟수가 동일함을 전제로 일변량 반복 측정 ANOVA 모형이 적용되었다. 개인 별 효과에 대

한 개념을 위해 기존의 블럭 효과(block effect) 개념을 적용하였으며 고정 효과(fixed effect) 대신 랜덤효과(random effect)로 간주하여 오차항과 함께 분산을 구성하게 되었다. 또 다른 고전적 방법의 적용 예는 MANOVA(multivariate ANOVA)의 적용으로 다변량 경시자료에 대해 적용되며 multivariate repeated measures ANOVA라고 명명되어진다. 하지만 반복 측정 자료간의 상관 관계에 대한 공분산 구조가 다소 제한적이라는 점과 경시적 자료에서 자주 발생되는 불완전 자료, 즉 결측 자료와 서로 다른 관측 시점을 가질 경우에 위의 고전적 모형은 적용될 수 없다. 또한 이산형 경시적 자료 분석을 위해서 더 다양한 분석 방법의 필요성이 절실하였다. 따라서 최근에 경시자료 분석을 위해 자주 적용되는 혼합 모형(mixed effect model), 주변 모형(marginal model) 그리고 전이 모형(transition model)을 간략하게 설명한다.

앞서 언급된 ANOVA 모형의 확장으로 혼합 모형은 경시적 자료뿐만 아니라 군집 자료(clustered data) 또는 위계 자료(hierarchical data)에서 각각 군집 효과와 층효과를 추정하기 종종 적용된다 (Goldstein, 1995). 특히 개인별 변화패턴을 고려할 수 있다는 점은 성장 곡선의 다양성과 함께 활발하게 연구되었으며 개인별 회귀 계수의 변화는 랜덤 효과로 모형화됨으로써 혼합 모형의 범위를 넓히고 있다. Laird와 Ware (1982)은 Harville (1977)의 혼합 모형에 대한 초기 작업을 확장하여 다음의 혼합 선형 모형을 제시하였다.

$$y_{ij} = x_{ij}\beta + \upsilon_i, \quad \upsilon_i = d_{ij}u_i + w_i(t_{ij}) + \epsilon_{ij},\ i = 1, \ldots, n;\ j = 1, \ldots, m_i, \tag{2.1}$$

여기서 x_{ij}의 효과는 고정 효과 β를 통해 추정되며 d_{ij}는 랜덤 효과 u_i와 관련된 공변량이다. 이 때, y_{ij}은 다음의 세가지 변이(variation)로 구성된다. $u_i \sim N(0, G)$는 i번째 관측대상의 공변량 d_{ij}에 대한 개별효과(subject-specific effect)를 표현하는 랜덤 효과이며 일반적으로 다변량 정규 분포를 가정한다. 두 번째 구성요소인 $w_i(t_{ij})$는 시점간의 연관관계(serial correlation)을 모형화하기 위해 적용되며 예를 들어 $\mathrm{Cor}(t, t+u) = \rho(u) = \exp(-\phi u)$ 또는 $\rho(u) = \exp(-\phi u^2)$와 같은 지수 상관 모형이 고려될 수 있다. 마지막으로 ϵ_{ij}는 측정 오차로, ϵ_{ik}와 ϵ_{il}은 서로 독립이며 평균이 0이고 분산이 σ_e^2인 정규 분포를 가정한다. 따라서 모형 (2.1)은 일변량 반복 측정 ANOVA와 성장 곡선을 모두 포괄하는 모형이 된다. 모수 추정을 위해 EM 알고리즘을 제안하였으며 이후 Fisher scoring 방법과 Newton-Raphson 방법이 적용되었다. 특히 분산 추정량을 위해 제한된 우도 방법(restricted likelihood)을 적용함으로써 불편 추정량을 구할 수 있다.

반면에 이변량, 계수형과 같은 이산형 경시자료에 대해서는 위와 같은 선형 모형으로 변이를 분리하는 것은 쉽지 않다. 이에 다음의 주변 모형(marginal model)이 적용된다. 일반화 선형 모형(GLM)을 확장한 방법으로 반응변수의 전체 확률 분포대신에 다음의 세 가지 성분을 명시할 필요가 있다. (i) 적절한 연결함수를 이용한 공변량의 함수로 평균을 표현: $E(Y_{ij}|x_{ij}) = \mu_{ij}$일 때, $g(\mu_{ij}|x_{ij}) = x'_{ij}\beta$, (ii) 공변량이 주어져 있다는 가정하에서 평균의 함수로써 분산의 명시, $\mathrm{Var}(Y_{ij}|x_{ij}) = \psi v(\mu_{ij})$, 여기서 $v(\mu_{ij})$는 분산 함수이며 ψ는 척도 모수를 의미한다. (iii) 한 개체내 관측치들간의 상관관계(correlation)를 위한 모수 α. 특히, 세 번째 요소에 대해서는 상관계수(correlation coefficient, $\rho(y_{ij}, y_{ik})$)가 일반적으로 사용된다. 하지만 이변량 이산형 자료에 대해서는 평균값의 범위($0 < \mu_{ij} < 1$)의 제약 때문에 상관계수는 적절한 통계량의 역할을 하지 못하며 대신 오즈비(odds ratio)가 적용되기도 한다. Liang과 Zeger (1986)은 다음과 같은 일반화 추정 방정식(generalized estimating equation)을 제안한다.

$$S_\beta(\beta, \alpha) = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta}\right)' V_i(\alpha)^{-1}(Y_i - \mu_i) = \sum_{i=1}^{n} D_i V_i(\alpha)^{-1}(Y_i - \mu_i) = 0,$$

여기서 $\mu_i = (\mu_{i1}, \ldots, \mu_{im_i})$이며 $V_i(\alpha) = \psi A_i^{1/2} R_i(\alpha) A_i^{1/2}$이며 A_i는 대각 행렬로 $\mathrm{diag}(v(\mu_{ij}))$이며 $R_i(\alpha)$는 상관 행렬로 각각의 원소는 $\rho_{ist}(\alpha) = \rho_{its}(\alpha) = \mathrm{Corr}(Y_{is}, Y_{it}; \alpha)$로 구성된다. 실제 V_i에 대한 특별한 정보가 없을 경우 working correlatoin을 적용한다. 예를 들어, $\rho_{ist}(\alpha) = \alpha$인 경우 exchangeable 또는 CS(compound symmetric)은 시간 간격에 무관하게 개체 내 반응 변수들간에 같은 상관 관계를 가짐을 의미하며 $\rho_{ist}(\alpha) = \alpha^{|t-s|}$는 AR(1)로 시간 간격이 멀어질수록 상관 관계 줄어듬을 의미한다. 이러한 GEE는 quasi-likelihood approach (Wedderburn, 1974)를 반복 측정 자료에 확장한 것으로 인식될 수 있다. 특히 공분산 구조에 대한 정확한 명시가 필요로 하지 않는다는 점에서 추정된 $\hat{\beta}$는 로버스트하고 할 수 있다. 따라서 부정확한 분산의 명시를 통해 발생된 미구현화(misspecification)를 보완하기 위해 추정된 β의 분산으로 다음의 샌드위치 분산을 이용하게 된다.

$$\hat{C}_\beta = I_0^{-1} I_1 I_0^{-1},$$

여기서 $I_0 = (1/n) \sum_{i=1}^n D_i'(\hat{\beta}) V_i^{-1}(\hat{\alpha}) D_i(\hat{\beta})$이며 $I_1 = \sum_{i=1}^n D_i'(\hat{\beta}) V_i^{-1}(\hat{\alpha}) \widehat{\mathrm{Cov}}(Y_i) V_i^{-1}(\hat{\alpha}) D_i(\hat{\beta})$으로 추정된 $\hat{\beta}$와 $\hat{\alpha}$ 그리고 표본 분산 $\widehat{\mathrm{Cov}}(Y_i) = (Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)'$를 사용한다.

회귀 계수 β에 대한 해석에 대해서는 혼합 모형과 주변 모형은 서로 다른 성질을 가진다. 주변 모형에서의 β는 같은 $x = x_{ij}$값을 가진 개체들의 집단평균 반응을 보여준다. 즉, β는 모집단 평균 모수(population averaged parameter)이지만 혼합 모형에서의 β의 추정값은 개인 별 수준에서 해석되어진다는 점에서 다른 의미를 가진다. 물론 이산형 자료에 대해서도 앞에서 설명한 혼합 모형이 다음과 같은 구조하에서 적용될 수 있으며 이를 일반화 혼합 모형(generalized linear mixed model; GLMM)이라 한다.

$$g(\mu_{ij}) = x_{ij}'\beta + z_{ij}' b_i,$$

여기서 β는 공변량의 효과를 추정하는 고정 효과이며 b_i는 개체별 효과인 랜덤효과이다. 이 모형에 대한 추론을 위해 penalized quasi likelihood를 적용한다.

마지막으로 전이 모형은 마르코브 모형이라고도 하며 현재 반응치가 과거 관측자료에 의존하는 것을 가정으로 한다. 따라서 이들 관계를 모형에 직접적으로 대입하여 추정할 수 있다. 예를 들어, 연속형 반응자료에 대해 오차항 $\epsilon_{ij} = \alpha \epsilon_{ij-1} + e_{ij}$일 때, 이는 다음과 같은 관계로 표현될 수 있다.

$$Y_{ij} | Y_{ij-1} \sim N\left(x_{ij}'\beta + \alpha\left(Y_{ij-1} - x_{ij-1}'\beta\right), \tau^2\right),$$

여기서 $\alpha = \exp(-\phi)$, $e_{ij} \sim N(0, \tau^2)$, $\tau^2 = \sigma^2(1-\alpha^2)$이 된다. 좀 더 자세한 내용과 다양한 예제를 원하다며 Diggle 등 (2001)과 Fitzmaurice 등 (2004)를 참고하기 바란다.

3. 관측 시점과 관측 중단 원인이 경시 자료와 연관성을 있을 경우에 대한 결합 모형

3.1. 통계적 모형

본 논문에서 연구되는 경시적 자료는 모든 관측 대상자에 대해 $(y_{ij}, t_{ij}, c_i, x_{ij}, z_{ij}, w_i, \delta_i;\ j = 1, \ldots, m_i,\ i = 1, \ldots, n,)$을 관측하게 된다. 여기서 경시 자료 y_{ij}는 관측시점 t_{ij}에서 측정되며 본 연구에서는 연속형 확률변수를 가정한다. 공변량 x_{ij}는 $p \times 1$ 벡터로 경시자료와 관련된 공변량으로 시간 가변 공변량을 포함하며 z_{ij}는 방문 시점(또는 관측 시점)에 유의한 관계를 가질 것으로 예상되는 공변량이며 w_i는 관측 중단 시점에 대한 모형을 위해 적용될 공변량이다. 물론 이들 공변량들간에는 서로 공통된 부분이 존재할 수 있다. 또한, c_i는 관측 중단 시점을 δ_i는 관측 중단 원인을 표시한다. 특히 관측 시점과 관측 중단(drop out)이 반응변수와 연관될 가능성이 있을 때, 이에 대한 검토를 위해 결합 모형

이 제안된다 (Tsiatis와 Davidian, 2004; Henderson 등, 2000). 경시적 자료에 대한 대부분의 결합 모형은 경시적 자료와 생존 자료의 연관관계를 검정하거나 연관 관계가 존재할 경우 발생할 수 있는 가능한 편의를 제거하기 위해 적용되었다. 특히 관측 중단의 원인이 여러 개일 경우 생존 자료분석에서 널리 사용되는 경쟁 위험 모형(competing risk model)을 통해 각 원인과 경시 자료와의 관계를 조사해볼 수 있다. 일반적인 생존 분석 자료는 두 가지 상태를 가지는 모형으로 인식될 수 있으며 여기서 상태 0은 생존 상태(alive)를 상태 1은 사망 상태(dead)가 된다. 경쟁 위험 모형에서는 상태 0에서 전이할 수 있는 상태가 2개 이상인 경우에 이들 중 한개의 상태로의 전이 발생은 다른 상태의 발생을 관측할 수 없게 한다. 이에 대한 다양한 분석 방법이 개발 되고 있으며 특히 공변량의 효과를 추정하기 위해선 원인별 위험 함수(cause specific hazard)를 이용하는 방법과 누적 발생함수(cumulative incidence function)와 1 : 1의 관계로 유도할 수 있는 subdistribution 방법 (Fine과 Gray, 1999)이 적용될 수 있다. 연구 별 목적에 따라 이 두 방법 중 한 가지가 적용되며, 전자의 방법이 발생률에 관심이 있다면 후자의 방법은 일정 기간 사건을 경험한 확률을 구할 때 적용될 수 있다.

본 연구에서 분석한 자료는 경시적 자료, 관측 시점 그리고 결측 시점과 그 원인에 대한 자료를 동시에 고려하는 결합모형을 제안하는 것이다. 첫 번째 구성 요소는 반응 변수에 대한 것으로 x_{ij}와 v_{i1}이 주어져 있을 때, 다음 모형을 가정한다.

$$y_{ij} = \beta^{'} x_{ij} + v_{i1} + \epsilon_{ij}, \tag{3.1}$$

여기서 관측 오차 $\epsilon_{ij} \sim N(0, \sigma_e^2)$이며 β는 공변량의 효과를 추정하는 회귀 계수를 v_{i1}는 랜덤 효과(random effect)로 i번째 개체의 개별 효과(subject-specific effect)를 보여준다. 따라서 식 (3.1)은 경시적 자료에 대한 혼합 모형의 적용이 된다. 두 번째 모형은 관측 시점, $t_i = (t_{i1}, \ldots, t_{im_i})$에 대한 것으로 다음의 강도 함수(intensity function)를 적용한다.

$$\alpha(t_{ij}|z_{ij}, v_{i2}) = \alpha_0(t_{ij}) \exp\left(\gamma' z_{ij} + v_{i2}\right). \tag{3.2}$$

즉, 환자의 방문 시점이 반복적으로 발생하는 경우 재발 사건(recurrent event)로 간주하여 방문 시점에 대한 조건부 확률을 구할 수 있다. 식 (3.2)에서 z_{ij}와 v_{i2}가 주어져 있을 때, 개인의 관측 시점은 조건부 독립(conditionally independent)이라고 가정된다. $\alpha_0(t)$는 기저 강도 함수(baseline intensity function)로 특정 함수형태를 가정하지 않거나 모수적 모형으로 포아송 분포를 사용할 수도 있다 (Cook과 Lawless, 2007). 마지막 모형은 관측 중단 시점(c_i)과 관련 원인(δ_i)을 고려하기 위해 경쟁 위험 모형에서 자주 사용되는 원인별 위험 함수(cause-specific hazard function)를 적용한다 (Kalbfleisch와 Prentice, 2002). $k\,(= 1, \ldots, K)$번째 원인에 대한 위험함수로 공변량 w_i의 효과는 η_k로 추정된다.

$$\lambda_k(c) = \lambda_{0k}(c) \exp\left(\eta_k' w_i + \xi_{1k} v_{i1} + \xi_{2k} v_{i2}\right). \tag{3.3}$$

특히 위 모형에서는 반응변수와 k번째 중단 원인과의 연관관계를 위해 v_{i1}를 공변량으로 적용하였으며 관측 시점과 k번째 중단 원인과의 관계를 위해 v_{i2}을 각각 적용하였다. 따라서, 위 모형하에서 $\xi_{1k} = 0$은 반응 변수와 k번째 관측 중단 원인과의 독립성을 의미하며, $\xi_{2k} = 0$은 관측 시점과 k번째 관측 중단 원인의 독립성을 각각 의미한다. 특히 본 연구에서는 랜덤 효과 (v_{i1}, v_{i2})는 다음과 같은 이변량 정규 분포를 가정한다.

$$\begin{pmatrix} v_{i1} \\ v_{i2} \end{pmatrix} \sim \mathrm{BVN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma\right), \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

이제 $\Psi = (\beta, \alpha_0(\cdot), \gamma, \lambda_{01}(\cdot), \ldots, \lambda_{0K}(\cdot), \eta_k, \xi_{1k}, \xi_{2k}, \Sigma)$는 모형 (3.1)–(3.3)에 포함된 모든 모수 벡터의 집합을 보여준다. 여기서, Σ는 (v_{i1}, v_{i2})의 분산-공분산 행렬로 $\mathrm{Var}(v_{i1}) = \sigma_1^2$, $\mathrm{Var}(v_{i2}) = \sigma_2^2$, 그리고 $\mathrm{Cov}(v_{i1}, v_{i2}) = \sigma_{12}$가 된다.

3.2. 추정

경시 반응 변수는 $y = (y_1, \ldots, y_n)^{'}$, $y_i = (y_{i1}, \ldots, y_{im_i})^{'}$으로 표현되고 관측 시점은 $t = (t_1, t_2, \ldots, t_n)^{'}$, $t_i = (t_{i1}, \ldots, t_{im_i})^{'}$, 그리고 관측 중단 시점과 관련 원인은 $\tilde{c} = \{(c_1, \delta_1), \ldots, (c_n, \delta_n)\}^{'}$이 된다. 즉, $\delta_i = \{1, \ldots, K\}$로 K개의 서로 다른 원인에 의해 관측이 중단될 수 있음을 의미한다. 여기서 우리의 관심은 k번째 원인에 의한 중도 절단이라고 하자. 따라서 다른 원인에 의한 관측 중단은 사건 발생 후 더 이상 위험 그룹에 속하지 않음을 의미한다. 모수 벡터 Ψ를 추정하기 위해 관측 자료에 근거한 우도함수(observed data likelihood)는 다음과 같다.

$$L(\Psi; y, t, \tilde{c}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\prod_{i=1}^{n} f(y_i, t_i, \tilde{c}_i|\psi) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\prod_{i=1}^{n} L_i^Y \times L_i^R \times L_i^{C,k} dv_{i1} dv_{i2}, \tag{3.4}$$

여기서

$$L_i^Y = \prod_{j=1}^{m_i} \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left(-\frac{1}{2\sigma_e^2}\left(y_{ij} - \beta' x_{ij} - v_{i1}\right)^2\right),$$

$$L_i^R = \prod_{k=1}^{m_i} \left[\alpha_0(t_{ik})\exp\left(\gamma' z_{ik} + v_{i2}\right)\right] \exp\left[-\int_0^{c_i} \alpha_0(u)\exp\left(\gamma' z_{ik} + v_{i2}\right) du\right],$$

$$L_i^{C,k}(c) = \left(\lambda_0(c)\exp\left(\eta_k' w_i + \xi_{1k} v_{i1} + \xi_{2k} v_{i2}\right)\right)^{I(\delta_i = k)} \exp\left[-\int_0^{c_i} \lambda_0(s)\exp\left(\eta_k' w_i + \xi_{1k} v_{i1} + \xi_{2k} v_{i2}\right)\right]$$

이다. 하지만 우도 함수 (3.4)는 닫힌 형태(closed form)로 표현될 수 없으므로, 랜덤 효과를 결측 자료(missing data)로 간주한 후 이에 대한 조건부 기대값을 계산하여 모수의 최대 우도 추정값을 구하는 EM 알고리즘을 적용한다. 일반적으로 EM 알고리즘을 적용하기 위해 몬테 칼로 방법(Monte Carlo) 또는 수치 적분(numerical integration)이 사용된다. 본 논문에서는 Gaussian quadrature를 이용한 수치 적분을 이용하였다. 이 방법의 장점은 SAS의 *NLIMIXED* 프로시져를 사용할 수 있다는 점이다. 하지만 초기값에 민감하며 모형이 복잡하면 수렴에 실패하는 경우가 종종 발생한다. 본 연구에서 추정의 편이를 위해 관측시점의 강도 함수 (3.2)와 원인별 위험률 (3.3)에 대해 조각 상수 위험률(piecewise constant hazard)을 적용한다. 예를 들어, 관측 시점에 대한 기저 강도 함수에 전체 시점을 m_1개의 조각 $(Q_1^R, \ldots, Q_{m_1}^R)$을 백분위수를 이용하여 정한 후, 조각 상수 위험률을 가정하게 되며 다음과 같이 조각 상수 기저 강도 함수(piecewise constant baseline intensity)와 누적 기저 강도 함수(cumulative baseline intensity function), $\tilde{\alpha}_0$와 $\tilde{A}_0$를 각각 정의한다.

$$\tilde{\alpha}_0(t) = \alpha_{0k},$$

$$\tilde{A}_0(t) = \sum_{k=1}^{m_1} \alpha_{0k} \max\left\{0, \min\left(Q_k^R - Q_{k-1}^R, t - Q_{k-1}^R\right)\right\},$$

여기서 $Q_{k-1}^R < t \leq Q_k^R$이다.

비슷한 방법으로 관측 종료의 기저 위험함수에 대해 조각 상수 기저함수 $\tilde{\lambda}_0$와 누적 기저함수 $\tilde{\Lambda}_0$를 각각 정의한다 (Kim, 2010). M-step에서 Newton-Raphson 방법을 적용하여 모수의 추정값을 구한 후, 헤시안 행렬(Hessian matrix)을 이용하여 추정된 모수의 분산을 구한다.

4. 자료 분석

312명의 간경변증 환자를 대상으로 병원 진료 시작부터 프로트로빔(혈액 응고시간)을 측정하였다. 한 번 이상 병원을 방문한 286명의 환자들의 병원 방문 시점과 관측 중단원인을 조사하여 이들간의 연관관

Table 4.1. Joint analysis of prothrombin, visiting process and competing risk with death cause

Covariate	Full model: (3.1) + (3.2) + (3.3)			Reduced A: (3.1) + (3.2)			Reduced B: (3.2) + (3.3)		
	Est	Se	p-value	Est	Se	p-value	Est	Se	p-value
Prothrombin									
Intercept	2.363	0.007	<0.001	2.363	0.007	<0.001			
time	0.013	0.007	<0.001	0.013	0.007	<0.001			
trt	−0.006	0.009	0.537	−0.006	0.009	0.476			
gender(Male = 1)	0.030	0.015	0.042	0.031	0.0148	0.0467			
σ_e^2	0.007	0.0002	<0.001	0.007	0.003	<0.001			
Visiting process									
trt	−0.059	0.092	0.519				−0.045	0.091	0.620
gender(Male = 1)	−0.0616	0.143	0.667				−0.059	0.142	0.679
Competing risk (dead)									
trt	0.035	0.286	0.903	−0.071	0.226	0.755	0.070	0.284	0.805
gender(Male = 1)	0.944	0.416	0.024	0.722	0.316	0.023	0.932	0.413	0.025
ξ_1	9.053	2.230	<0.001	14.887	1.700	<0.001			
ξ_2	2.174	0.380	<0.001				2.767	0.379	<0.001
Covariance matrix									
σ_a^2	0.005	0.0005	<0.001	0.005	0.0005	<0.001			
σ_{ab}	0.0253	0.004	<0.001						
σ_b^2	0.384	0.054	<0.001				0.383	0.053	<0.001

계를 조사한다. 286명의 환자 중 123명이 사망하고 29명이 간이식을 받았으며 이 두 사건 중 어떤 사건이 먼저 일어나면 나머지 다른 사건을 경험하지 못하게된다. 따라서 사망과 간이식은 경쟁위험이 된다. 이러한 두 가지 중단 원인을 동시에 고려하기 위해 경쟁 위험 모형의 원인별 위험 함수 모형이 적용되었다. 본 연구에서는 사망을 주 관심으로 하자. 142명의 환자는 D-penicillin을 투여받았으며 144명은 기존의 약을 투여받았다. 반응 변수인 프로트로빔에 유의한 영향을 주는 요인으로 D-penicillin 그룹 여부(D-penicillin 처리 그룹 = 1, 그외 처리 그룹 = 0)와 성별(남성 = 1; 여성 = 0)을 고려하며 시간 관련 추세를 표현하기 위해 선형함수를 가정으로 한다. 본 연구에서는 관측 중단 원인 중에 사망과 프로트롬빈 그리고 방문 시점과의 관계를 조사하고자 한다. Table 4.1에서는 앞 절에서 설명한 결합 모형(full 모형: 경시자료 + 관측 시점 + 관측 중단 시점: (3.1) + (3.2) + (3.3))을 적용한 결과를 보여주며 동시에 축소 모형 A(경시자료 + 관측 중단 시점: (3.1) + (3.3))와 축소 모형 B(관측 시점 + 관측 중단 시점: (3.2) + (3.3))가 함께 비교되었다. 특히 기저 강도 함수와 기저 위험률에 조각 상수 함수를 적용하였으며 여기서 5개의 조각이 사용되었다. 본 자료에서는 조각이 좀 더 작거나($m_1 = 3$) 더 많을 경우($m_1 = 7$)에도 구해진 결과는 큰 차이가 없었다. 축소 모형 A에서는 관측 중단 원인이 사망 사건과 프로트로빔과의 관계만을 고려하였으며 축소 모형 B는 프로트로빔 경시자료는 고려하지 않고 환자의 병원 방문 시점 자료과 사망으로 인한 관측 중단 사건을 고려하였다. 다시 한번 언급하면 세 가지 모형을 동시에 고려한 결합 모형을 통해 세 모형간의(세 변량들간의) 상관 관계를 한꺼번에 추정할 수 있으면, 존재할 수 있는 상관 관계의 누락을 통해 발생할 수 있는 변이를 제거할 수 있다는 장점이 있다.

Figure 4.1은 전체 환자 중 73명의 환자의 프로트롬빈 추세를 보여준다. 그림에서 보다시피 환자마다 병원 방문 시점과 방문 횟수가 다르며 방문 횟수의 범위는 1에서 16이었다. 실제 분석에서 프로트롬빈 시간의 비대칭을 보완하기 위해 로그 변환된 값을 사용하였다. 두 가지 축소 모형과 제안한 결합 모형은 추정량의 유의성에서 비슷한 결론을 보였다. 프로트롬빈 시간에서는 남성이 여성보다 더 길었

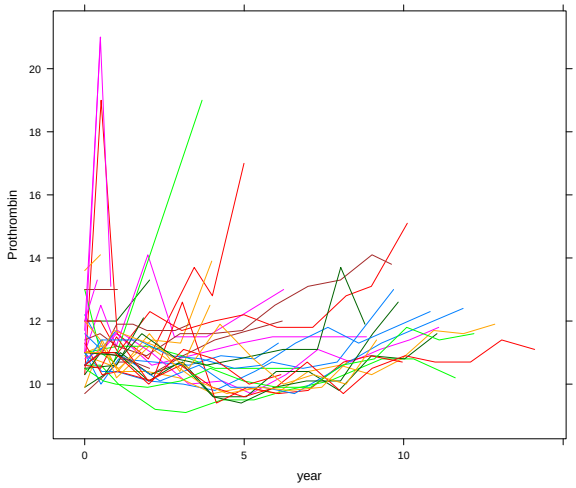

Figure 4.1. Spaghetti plot for 73 subjects.

으며 두 가지 처리 그룹간에는 유의한 차이가 없었다. 병원 방문 시점(visiting)과 관련하여 성별과 처리그룹간에는 유의한 차이가 없었다. 사망 원인으로 인한 관측 중단 사건에 대해서는 처리 그룹간에는 차이가 없었으나 성별간에는 남성이 여성보다 사망 위험률이 더 높았다. 경시적 자료와 사망 사건간의 관계는 매우 강한 상관 관계를 보여($\hat{\eta_1} = 9.053\,(p\text{-value} < 0.001)$), 프로트롬빈이 시간이 길수록 사망 위험률이 높다고 할 수 있었다. 비슷하게, 병원 방문 시점이 빈번할수록 사망할 위험률도 또한 높았다($\hat{\eta_2} = 2.174\,(p\text{-value} < 0.001)$). 또한 모든 공분산 행렬에 포함된 분산($\hat{\sigma}_a = 0.005\,(p\text{-value} < 0.001), \hat{\sigma}_b = 0.384\,(p\text{-value} < 0.001)$)은 유의하며 이는 개인별 효과가 존재함을 의미한다. 또한 공분산도 $\hat{\sigma}_{ab} = 0.0025\,(p\text{-value} < 0.001)$로 경시자료와 방문 시점간에 양의 유의한 상관관계가 존재함을 알 수 있다. 즉 프로트롬빔 시간이 길수록 병원의 방문 횟수가 빈번하다고 해석할 수 있다.

5. 맺음말

본 연구에서는 특수 시계열 자료의 예로 경시적 자료에 관한 논문을 다루었다. 특히 관측 시점과 관측 횟수가 개체마다 다르고 관측 중도 절단이 발생할 경우 이들간의 상호 연관관계를 추정하기 위해 결합모형을 적용하였다. 특히 한 개체내에 반복적으로 측정되는 반응변수와 관측 시점에 대한 모형에서 개체내(subject within) 연관성을 추정하기 위해 이변량 정규 분포를 따르는 이변량 랜덤 효과가 적용되었으며 이러한 랜덤효과는 관측 중단의 원인에 대한 경쟁 위험모형에서 공변량으로 적용되어 변량들간의 관련성을 검정하는 데 사용되었다. 이들 세 가지 요소를 동시에 고려하기 위해 결합모형외의 다른 방법으로 추정 방정식을 확장하는 것이다. Sun 등 (2007)은 적절한 가중치를 사용한 방정식을 이용하였다. 본 논문에서는 관측 중단의 원인이 여러 개인 경우를 모형화하기 위해 경쟁위험 모형이 적용되었으며 특히 원인별 위험 함수(cause-specific hazard function)를 통해 공변량의 유의성과 경시적 자료와 관측 시점과의 연관성을 검토하였다. 경쟁 위험 모형에 대해 빈번하게 사용되는 다른 회귀 모형으로는 Fine와 Gray (1999)가 제안한 subdistribution 모형이 있다. 그들의 모형에서는 이미 경쟁 사건을 경험한 개체를 위험그룹에서 제외시키는 기존의 방법말고 그들이 여전히 위험그룹에 남아있다고 가정하고 그들에게 적절한 가중치를 부여함으로써 경쟁 위험 모형의 특성을 반영하고자 하였다. 본 논문의 향후 연구과제는 이러한 subdistribution 모형과 경시자료를 동시에 고려하는 추정 방정식을 유도하고자 한다.

References

Cook, R. and Lawless, J. (2007). *The Statistical Analysis of Recurrent Events*, Springer.

Diggle, P. J., Heagerty, P., Liang, K.-Y., and Zeger, S. L. (2001). *Analysis of Longitudinal Data*, Oxford Press.

Elashoff, R. M., Li, G., and Li, N. (2007). An approach to joint analysis of longitudinal measurements and competing risks failure time data, *Statistics in Medicine*, **26**, 2813–2835.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk, *Journal of the American Statistical Association*, **94**, 496–509.

Fitzmaurice, G. M., Laird, N. M., and Rotnitzky, A. G. (2004). *Applied Longitudinal Analysis*, Wiley, Hoboken NJ.

Free, E. (2004). *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences*, Cambridge University Press.

Goldstein, H. (1995). *Multilevel Statistical Models*, 2nd Ed, Edward Arnold, London.

Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems, *Journal of the American Statistical Association*, **72**, 320–338.

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data, *Biostatistics*, **1**, 465-480.

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

Kim, Y.-J. (2010). Statistical Analysis of recidivism data using frailty effect, *Korean Journal of Applied Statistics*, **23**, 715–724.

Laird, N. M. and Ware, J. H. (1982). Random effects models for longitudinal data, *Biometrics*, **38**, 963–974.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika*, **73**, 13–22.

Liu, L. and Huang, X. (2009). Joint analysis of correlated repeated measures and recurrent events processes in the presence of death, with application to a study on acquired immune deficiency syndrome, *Applied Statistics*, **58**, 65–81.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data: with Applications in R*, Chapman and Hall/CRC.

Sun, J., Sun, D., and Liu, D. (2007). Regression analysis of longitudinal data in the presence of informative observation and censoring times, *Journal of the American Statistical Association*, **102**, 1397–1406.

Tsiatis, A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview, *Statistica Sinica*, **14**, 809–834.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and Gauss-Newton method, *Biometrika*, **61**, 439–447.

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error, *Biometrics*, **53**, 330–339.

결시적 자료에서 관측 중단을 모형화하기 위해 사용되는 경쟁 위험의 적용과 결합 모형

김양진[a,1]

[a]숙명여자대학교 통계학과

(2015년 12월 14일 접수, 2016년 1월 15일 수정, 2016년 1월 15일 채택)

요 약

경시적 자료는 반복적으로 측정된 다변량 자료의 한 형태로 임상학, 보건학, 경제학에서 자주 발생된다. 시계열자료와 구분되는 가장 큰 특징은 표본수와 공변량 효과의 추정에 있다. 경시적 자료는 일반적으로 시계열 자료보다 더 큰 표본 개체로 이루어져 있으며 연구의 주 관심은 특정 공변량의 효과를 추정하는 데 있다. 또한 시계열 자료보다 반복 측정 횟수가 짧으며 개체마다 다른 관측 횟수와 다른 관측 중단 시점을 가질 수 있다. 본 연구에서는 관측 시점과 관측 종료 원인이 경시자료와 서로 연관된 경우에 결합 모형을 적용함으로써 이들간의 연관성을 분석하고자한다. 따라서 이들 변량간의 연관성을 모형화하기 위해 이변량 랜덤효과가 적용된다. 실제 자료 분석에서는 간경변증 환자의 혈액 응고 수치 시간을 관심 있는 경시적 자료로 환자가 병원 방문시점과 관측 중단원인들간의 상호 연관관계를 규명하고자 하였다. 특히, 중도 절단원인은 사망이나 간이식을 받는 사건일 때 발생하는데 본 연구에서는 사망 사건과의 연관성이 고려되었다. 결과를 통해 혈액 응고 시간이 길고 병원 방문 시점이 빈번할수록 사망할 가능성이 높음을 알수 있었다. 또한 혈액응고 시간이 길수록 병원 방문 횟수가 빈번하였다.

주요용어: 경쟁위험, 경시적 자료, 관측 시점과의 연관성, 관측 중단, 결합모형, 랜덤 효과

본 연구는 2014년도 숙명여자대학교 교내연구비 지원하에서 수행하였음.

[1](04310) 서울특별시 용산구 청파로47길 100, 숙명여자대학교 통계학과. E-mail: yjin@sookmyung.ac.kr