The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.7.1459
29(7), 1459–1473

A joint modeling of longitudinal zero-inflated count data and time to event data

Donguk Kim[a,1] · Jihun Chun[a]

[a]Department of Statistics, Sungkyunkwan University

(Received November 28, 2016; Revised December 13, 2016; Accepted December 15, 2016)

Abstract

Both longitudinal data and survival data are collected simultaneously in longitudinal data which are observed throughout the passage of time. In this case, the effect of the independent variable becomes biased (provided that sole use of longitudinal data analysis does not consider the relation between both data used) if the missing that occurred in the longitudinal data is non-ignorable because it is caused by a correlation with the survival data. A joint model of longitudinal data and survival data was studied as a solution for such problem in order to obtain an unbiased result by considering the survival model for the cause of missing. In this paper, a joint model of the longitudinal zero-inflated count data and survival data is studied by replacing the longitudinal part with zero-inflated count data. A hurdle model and proportional hazards model were used for each longitudinal zero inflated count data and survival data; in addition, both sub-models were linked based on the assumption that the random effect of sub-models follow the multivariate normal distribution. We used the EM algorithm for the maximum likelihood estimator of parameters and estimated standard errors of parameters were calculated using the profile likelihood method. In simulation, we observed a better performance of the joint model in bias and coverage probability compared to the separate model.

Keywords: joint model, longitudinal zero-inflated count data, hurdle model, survival data

1. 서론

생물의학 분야에서 생존자료와 경시적 자료(longitudinal data)가 같이 수집되는 경우가 발생한다. 예를 들어 약물에 대한 시험참여자들에 대한 반응을 연구하는 임상시험의 경우, 참여자들로부터 약물 처치에 따른 시간별 반응변수를 일정한 시간동안 반복적으로 측정함과 동시에 사망 또는 병의 재발과 같은 관심있는 사건(event)의 생존시간에 대한 자료도 같이 수집한다. 이렇게 수집된 경시적 자료와 생존자료의 경우, 두 자료의 결과는 서로 연관되어 있다. 연속형 반응변수에 대한 경시적 자료를 분석하기 위해서는 선형혼합모형(linear mixed model; LMM)이 주로 이용되며, 관심있는 사건의 발생시간에 대해서는 비례위험모형(proportional hazards model)과 같은 생존모형(survival model)이 고려된다. 경시적 자료에서 결측 자료가 발생하였을 때, 발생에 어떤 원인이 존재하는 경우 선형혼합모형만으로 경시적 자료를 분석한다면 편향된 결과를 얻을 수 있다 (Prentice, 1982; Little과 Rubin, 2002). 이는 경시적 자료와 생존자료가 서로 연관되어 있고, 경시적 자료의 결측의 발생이 사건의 발생 시간과 관련

This paper was supported by Faculty Research Fund, Sungkyunkwan University, 2013.
[1]Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2, Sungkyunkwan-ro, Jongno-gu, Seoul 03063, Korea. E-mail: dkim@skku.edu

되어있기 때문이다. 이와 같은 무시할 수 없는(non-ignorable) 결측을 가진 경시적 자료의 불편추정량을 얻기 위한 추정방법으로 경시적 자료와 생존자료의 결합모형이 연구되었다 (Henderson 등, 2000; Elashoff 등, 2008). 결합모형의 분석은 주로 경시적 자료, 생존자료, 또는 두 자료 동시에 초점을 맞출 수 있다. 경시적 자료의 결측(missing) 발생에 대한 모형을 고려하는 경우, 경시적 자료 연구에서 흔하게 발생하는 중도탈락(dropout)에 특정 원인이 존재하는 경우라 보고 이를 설명하기 위해 결합모형을 이용한다 (Wulfsohn과 Tsiatis, 1997; Tseng 등, 2005). 또는 생존자료에서 시간에 따라 변화하는 공변량(time dependent covariate)을 경시적 모형으로 설명하기 위해 결합모형을 사용할 수 있다 (Diggle과 Kenward, 1994). 이뿐만 아니라 경시적 자료와 생존자료에 동시에 초점을 두는 경우 두 자료의 연관성을 모형화하기 위해 두 자료를 연결시켜주는 변량효과(random effect)를 이용하여 결합모형을 구성하였다. 이와 같이 경시적 자료와 생존자료 사이의 관련성을 나타내기 위해 결합모형을 이용하여 많은 연구가 이루어졌다 (Wu 등, 2012). 경시적 자료와 생존자료의 연관성을 각 모형 내의 변량효과의 연결을 이용하여 결합한 모형이 연구되어 왔으며, Sousa (2011)는 변량효과를 포함하는 경시적 자료와 생존자료의 결합모형은 random pattern-mixture 모형, random selection 모형 그리고 변량효과모형으로 구분하였다. 본 논문에서는 변량효과모형을 사용한다.

경시적 자료에서는 관심있는 반응변수가 일정한 기간 내에 특정 현상의 발생 개수 또는 발생한 위험의 개수와 같이 가산(count) 또는 정수(integer) 형태인 경우, 일반화선형혼합모형(generalized linear mixed model; GLMM)을 이용할 수 있다. 일반적인 가산 또는 정수 형태의 반응변수에 대한 경시적 자료는 포아송(Poisson) 분포를 가정하여 로그 연결(log link) 함수를 고려한 모형을 사용한다. 그러나 가산자료가 영(0)이 많이 존재 하는 영과잉(zero-inflated) 자료로 관측되는 경우는 일반적인 분포형태를 가지지 않으므로 포아송모형을 영과잉 가산자료에 적용하는 것은 적합하지 않다. 영과잉 가산자료를 분석하기 위한 모형으로 허들모형(hurdle model) (Mullahy, 1986)과 영과잉 가산모형(zero-inflated count model) (Lambert, 1992)이 있다. 영과잉 가산모형의 경우, 영과잉 포아송모형(zero-inflated Poisson model; ZIP)에서 포아송모형 부분에 변량효과를 포함시킴으로써 경시적 자료의 경우로 확장된 형태가 연구되었다 (Hall, 2000). 허들모형의 경우, 영(0)에 대한 부분과 영(0)이 아닌 정수 부분을 나타내는 부분에 대해 서로 독립적인 변량효과들을 고려한 모형이 연구되었으며 (Yau와 Lee, 2001), 반복 측정되는 영과잉 가산자료를 위해 서로 연관된 변량효과들을 고려한 허들모형이 제안되었다 (Min과 Agresti, 2005). 영과잉 가산자료는 여러 분야에서 활용되고 있으며 특히 생물의학 분야에서는 약학연구 (Min과 Agresti, 2005)와 약물남용연구 (Buu 등, 2012) 등에서 연구되어 왔다. 영과잉 가산모형은 가산자료 내에 영(0)이 많이 존재하였을 때에는 적합하지만 영(0)이 적게 존재하는 경우 이항부분의 모수가 무한대로 추정되는 문제가 존재한다. 이와 달리 허들모형은 영(0)이 많거나 작은 경우에도 잘 적합된다. 또한 영과잉 가산모형의 경우 이항부분과 포아송모형 부분의 모수를 동시에 추정해야 하지만 허들모형은 이항부분과 절단된 포아송모형 부분의 모수를 독립적으로 추정할 수 있으므로 계산상 용이하다는 장점을 가지고 있다. 따라서 본 논문에서는 허들모형을 이용하였다.

본 논문에서는 경시적 영과잉 가산자료와 생존자료를 분석하기 위한 결합모형을 연구한다. 제안된 결합모형은 경시적 영과잉 가산자료를 위한 허들모형과 생존자료를 위한 비례위험 모형(proportional hazards model)이 결합된 모형이다. 모수의 추정에 대해서는 Expectation Maximization(EM) 알고리즘을 이용하였고 표준오차(standard error)를 추정하기 위해 프로파일 우도(profile likelihood)로부터 계산된 경험적 피셔 정보(empirical Fisher information)의 역수인 분산-공분산행렬(variance-covariance matrix)을 이용한다. 본 논문의 구성으로 2장에서는 영과잉자료의 허들모형과 생존모형의 결합모형에 대해 설명하고 3장에서는 모수와 표준오차의 추정을 다룬다. 그리고 4장에서는 모의실험을 통하여 허들모형과 생존모형의 결합모형을 개별적인 허들모형과 생존모형 각각에 대해 성능을 비교한다.

2. 모형과 우도함수

2.1. 모형

제안하는 결합모형은 두 개의 부 모형(sub-model)으로 구성되어 있다. 하나는 경시적 자료에 대한 부 모형으로 경시적 영과잉 가산자료를 위한 허들모형이고, 다른 하나는 생존자료에 대한 부 모형으로 변량효과를 포함한 비례위험 모형이다. 여기서 허들모형은 임의절편 영과잉 모형(random intercept zero-inflated model) (Min과 Agresti, 2005)을 사용하였다.

Y_{ij}는 개체 $i(i = 1, \ldots, N)$의 $j(j = 1, \ldots, n_i)$번째 시점의 경시적 가산형 반응변수이며, N은 연구 중 관측된 전체 개체의 수, n_i는 각 개체 i의 측정 횟수를 나타낸다. 허들모형은 혼합모형으로 관측된 변수가 영(0)인지 아닌지를 나타내는 이항모형과 영이 아닌 부분을 설명하는 절삭된(truncated) 모형으로 구성되어 있다. 절삭된 모형이 포아송분포를 따를 때, 포아송 허들모형이 되고 모형의 형태는 다음과 같으며

$$Y_{ij} \sim \begin{cases} 0, & p_{ij}\text{의 확률}, \\ \text{truncated Poisson}(\mu_{ij}), & 1 - p_{ij}\text{의 확률}. \end{cases}$$

확률분포는 다음과 같이 나타낼 수 있다 (Min과 Agresti, 2005).

$$\begin{aligned} P(Y_{ij} = 0) &= p_{ij}, \\ P(Y_{ij} = y_{ij}) &= \frac{1 - p_{ij}}{1 - e^{-\mu_{ij}}} \frac{e^{-\mu_{ij}} \mu_{ij}^{y_{ij}}}{y_{ij}!}, \quad y_{ij} = 1, 2, \ldots \end{aligned} \tag{2.1}$$

그리고 영(0)이 될 확률인 모수 p_{ij}와 절삭된 포아송분포의 평균인 모수 μ_{ij}는 설명변수들에 의해 다음의 모형들로 표현된다.

$$\text{logit}(p_{ij}) = X_{1i}(t)^T \eta + Z_{1i}(t)^T b_{1i}, \qquad \log(\mu_{ij}) = X_{2i}(t)^T \beta + Z_{2i}(t)^T b_{2i}. \tag{2.2}$$

여기서 각 p_{ij}와 μ_{ij}에 대해, $(X_{1i}(t), Z_{1i}(t))$와 $(X_{2i}(t), Z_{2i}(t))$는 각 개체 i의 j번째 시점에 대한 공변량으로, 여기서 $X_{1i}(t)$와 $X_{2i}(t)$는 고정효과에 대한 공변량이고 $Z_{1i}(t)$와 $Z_{2i}(t)$는 변량효과에 대한 공변량이다. η는 $X_{1i}(t)$에 대한 고정효과 모수, β는 $X_{2i}(t)$에 대한 고정효과 모수가 된다. 또한 b_{1i}와 b_{2i}는 각각 p_{ij}와 μ_{ij}의 변량효과이다. 위의 모형을 단순화하기 위해, 식 (2.2)의 고정효과에 대한 두 공변량과 변량효과에 대한 두 공변량이 각각 동일하다고 가정하고($X_{1i}(t) = X_{2i}(t) = X_i^L(t)$, $Z_{1i}(t) = Z_{2i}(t) = Z_i(t)$), 또한 변량효과에 대해 $b_{1i} = \phi b_i$, $b_{2i} = b_i$라고 가정한다. 여기서 모수 ϕ는 p_{ij}와 μ_{ij}의 변량효과의 관계를 나타내고 t는 각 개체 i의 j번째 시점을 나타낸다. 영과잉 가산자료의 분석을 위한 모형으로 식 (2.2)의 모형을 식 (2.3)의 모형으로 단순화하여 제안한다.

$$\text{logit}(p_{ij}) = X_i^L(t)^T \eta + Z_i(t)^T \phi b_i, \qquad \log(\mu_{ij}) = X_i^L(t)^T \beta + Z_i(t)^T b_i. \tag{2.3}$$

생존자료에 대한 부 모형인 변량효과를 고려한 비례위험 모형은 다음과 같다.

$$\lambda\left(t; X_i^S(t), u_i, \gamma\right) = \lambda_0(t) \exp\left(X_i^S(t)\gamma + u_i\right), \tag{2.4}$$

여기서 $\lambda_0(t)$는 기저위험함수(baseline hazard function)이고 $X_i^S(t)$는 고정효과에 대한 공변량, γ는 공변량 $X_i^S(t)$의 고정효과 모수를 나타낸다. u_i는 변량효과로 각 개체의 관측되지 않는 특성을 나타낸다.

위에서 설명한 두 부 모형을 결합시키기 위해, 변량효과모형(random effect model)을 고려하였다. 변량효과 결합모형은 경시적 자료와 생존자료 사이의 관계에 대해 관측되지 않는 변량효과들의 연관성을 가정한다. 제안된 모형에서는 각 부 모형의 변량효과인 b_i와 u_i가 다음과 같이 다변량 정규분포를 따른다고 가정한다 (Elashoff 등, 2008).

$$\theta_i = \begin{pmatrix} b_i \\ u_i \end{pmatrix} \sim \text{MN}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{bb} & \overset{T}{bu} \\ \Sigma_{bu} & \sigma_u^2 \end{pmatrix} \right]. \tag{2.5}$$

2.2. 우도함수

$Y_i(t)$는 개체 i의 j번째 시점 t_{ij}에서 관측된 경시적 영과잉 결과변수이고, $Y_i(t) = Y_{ij}(i = 1, 2, \ldots, N,$ $j = 1, 2, \ldots, n_i)$, 그리고 $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})$로 나타낼 수 있다. 시점 $(t_{i1}, t_{i2}, \ldots, t_{in_i})$의 경우, 각 개체 i의 j번째의 반복 측정된 값에 결측이 발생하거나 중도탈락될 수 있으며, 각 개체 i마다 다른 값을 가질 수 있다. 여기서 결측이나 중도탈락에 대해서는 Missing Not At Random(MNAR)을 가정한다. $S_i = (T_i, \delta_i)$는 각 개체 i에 대한 생존시간 자료이며 T_i는 관측된 생존시간으로 사건발생 시간(failure time) 또는 중도절단 시간(censoring time)을 나타내고, δ_i는 개체 i에 대한 사건발생 여부를 나타내는 지시변수로 만약 $\delta_i = 1$이면 사건이 발생한 것이고, $\delta_i = 0$이면 중도절단(censoring)이 발생한 것이다. 모수 벡터를 $\Delta = \{\eta, \phi, \beta, \gamma, \Sigma, \lambda_0(t)\}$라고 정의할 때, 관측자료 우도함수(observed data likelihood function)는 다음과 같다.

$$\begin{aligned}
L(\Delta; Y, S) &\propto \prod_{i=1}^{N} f\left(Y_i, S_i | \Delta\right) \\
&= \prod_{i=1}^{N} \int_{\theta} f\left(Y_i | S_i, \theta, \Delta\right) f\left(S_i | \theta, \Delta\right) f\left(\theta | \Delta\right) d\theta \\
&= \prod_{i=1}^{N} \int_{\theta} f\left(Y_i | \theta, \Delta\right) f\left(S_i | \theta, \Delta\right) f\left(\theta | \Delta\right) d\theta \\
&= \prod_{i=1}^{N} \int_{\theta} \left[\prod_{j=1}^{n_i} p_{ij}(b)^{I(y_{ij}=0)} \left(\frac{1 - p_{ij}(b)}{1 - e^{-\mu_{ij}(b)}} \frac{e^{-\mu_{ij}(b)} \mu_{ij}(b)^{y_{ij}}}{y_{ij}!} \right)^{1 - I(y_{ij}=0)} \right] \\
&\quad \times \lambda\left(T_i; X_i^S(T_i), u, \gamma\right)^{\delta_i} \exp\left[- \int_0^{T_i} \lambda\left(t; X_i^S(t), u, \gamma\right) dt \right] \\
&\quad \times \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left(-\frac{1}{2} \theta^T \Sigma^{-1} \theta \right) d\theta,
\end{aligned} \tag{2.6}$$

여기서 모든 공변량과 변량효과가 주어졌을 때, Y와 S는 서로 독립임을 가정한다. 위의 관측자료 우도함수는 변량효과 θ가 잠재변수(latent variable)이므로 적분이 용이하지 않다. 자료가 잠재변수를 가진 불완전한 자료인 경우인 경우 최대 우도를 추정하기 위한 방법으로 EM 알고리즘이 이용된다 (Dempster 등, 1977; Wulfsohn과 Tsiatis, 1997).

3. 추정

모수 $\Delta = \{\eta, \phi, \beta, \gamma, \Sigma, \lambda_0(t)\}$의 추정을 위한 EM 알고리즘에서는 관측자료 우도함수 대신에 변량효과

θ_i가 주어진 완전자료 우도함수(complete data likelihood function)를 이용한다.

$$\begin{aligned} L_c(\Delta; Y, S, \theta) \propto & \prod_{i=1}^{N} \left[\prod_{j=1}^{n_i} p_{ij}(b_i)^{I(y_{ij}=0)} \left(\frac{1-p_{ij}(b_i)}{1-e^{-\mu_{ij}(b_i)}} \frac{e^{-\mu_{ij}(b_i)} \mu_{ij}(b_i)^{y_{ij}}}{y_{ij}!} \right)^{1-I(y_{ij}=0)} \right] \\ & \times \lambda \left(T_i; X_i^S(T_i), u, \gamma \right)^{\delta_i} \exp \left[- \int_0^{T_i} \lambda \left(t; X_i^S(t), u, \gamma \right) dt \right] \\ & \times \frac{1}{\sqrt{2\pi|\Sigma|}} \exp \left(-\frac{1}{2} \theta_i^T \Sigma^{-1} \theta_i \right). \end{aligned} \tag{3.1}$$

위의 식 (3.1)에 식 (2.3)에서 정의한 연결함수(link function)를 적용한 완전자료 로그우도함수(complete data log-likelihood function)는 다음과 같다.

$$\begin{aligned} & \log L_c(\Delta; Y, C, \theta) \\ & = l_c(\Delta; Y, S, \theta) \\ & \propto \sum_{i=1}^{N} \left[\sum_{j=1}^{n_i} \left\{ I(Y_{ij}=0) \left(X_i^L(t_{ij})^T \eta + Z_i(t_{ij})^T \phi b_i \right) - \log \left(1 + \exp \left(X_i^L(t_{ij})^T \eta + Z_i(t_{ij})^T \phi b_i \right) \right) \right\} \right. \\ & \quad + \sum_{j=1}^{n_i} \{1 - I(Y_{ij}=0)\} \left\{ \left(X_i^L(t_{ij})^T \beta + Z_i(t_{ij})^T b_i \right) Y_{ij} \right. \\ & \quad \quad - \log \left(\exp \left(\exp \left(X_i^L(t_{ij})^T \beta + Z_i(t_{ij})^T b_i \right) \right) - 1 \right) - \log(Y_{ij}!) \Big\} \\ & \quad + \delta_i \left[\log \lambda_0(t) + X_i^S(t)^T \gamma + u_i \right] - \int_0^{T_i} \lambda_0(t) \exp \left(X_i^S(t) \gamma + u_i \right) dt \\ & \quad \left. + \log \frac{1}{\sqrt{2\pi|\Sigma|}} - \frac{1}{2} \theta_i^T \Sigma^{-1} \theta_i \right]. \end{aligned} \tag{3.2}$$

EM 알고리즘은 관측된 자료와 현 시점의 모수 추정값이 주어졌을 때의 완전자료 로그우도(complete data log-likelihood)의 조건부 기대값을 계산하는 E(expectation)-step과 E-step에서 계산된 조건부 기대값을 최대화시키는 모수값을 찾는 M(maximization)-step의 두 단계로 되어있다. E-step에서는 완전자료 로그우도 식 (3.2)의 조건부 기대값을 계산하는데, 구체적으로는 $l_c(\Delta; Y, S, \theta)$에 존재하는 θ_i의 모든 함수 $k(\theta_i)$에 대해서 계산한다. 조건부 기대값은 다음과 같다.

$$\begin{aligned} E_{\theta_i|Y_i,S_i,\Delta^{(t)}} \left(k\left(\theta_i\right) \right) &= \int k\left(\theta_i\right) f\left(\theta_i | Y_i, S_i, \Delta^{(t)} \right) d\theta_i \\ &= \frac{\int k\left(\theta_i\right) f\left(Y_i | \theta_i, \Delta^{(t)} \right) f\left(S_i | \theta_i, \Delta^{(t)} \right) f\left(\theta_i | \Delta^{(t)} \right) d\theta_i}{\int f\left(Y_i | \theta_i, \Delta^{(t)} \right) f\left(S_i | \theta_i, \Delta^{(t)} \right) f\left(\theta_i | \Delta^{(t)} \right) d\theta_i}. \end{aligned} \tag{3.3}$$

또한 $Q(\Delta; \Delta^{(t)}) = E_{\theta|Y,S,\Delta^{(t)}}(l(\Delta; \theta))$라고 정의했을 때, M-step에서는 모수벡터 $\Delta^{(t+1)} = \text{argmax}_{\Delta} Q(\Delta; \Delta^{(t)})$를 통해 모수 Δ를 갱신(updating)하면서 완전자료 로그우도의 조건부 기대값을 최대화시키는 모수를 찾는다. 모든 적분은 수치적분방법인 가우스 헬머트 구적(Gauss-Hermite quadrature) (Liu와 Pierce, 1994)을 이용하며, 적분 시 총 20개의 구적을 이용한다.

3.1. 모수 추정

모든 모수 $\Delta = \{\eta, \phi, \beta, \gamma, \Sigma, \lambda_0(t)\}$에 대해, 모수 Σ와 누적기저위험함수(cumulative baseline hazard function) $\Lambda_0(t)$는 닫힌 형식(closed form)으로 갱신할 수 있으며, 여기서 $\Lambda_0(t)$은 사건이 발생하였을 때 단계적으로 변하는 계단함수(step function)를 따른다. 모수 η, ϕ, β, 그리고 γ들은 닫힌 형식을 가지지 않기 때문에 뉴튼랩슨 알고리즘(one-step Newton-Raphson algorithm)을 이용해 모수를 갱신하였다. 각 모수에 대한 갱신 과정은 모수가 수렴할 때까지 반복되며 자세한 수식은 부록에 있다.

3.2. 표준오차의 추정

모수 벡터 Δ는 모수부분인 $B = (\eta, \phi, \beta, \gamma, \Sigma)$와 비모수부분인 기저위험함수 $\lambda_0(t)$의 두 개의 부분으로 나눠질 수 있다. 여기에서 우리의 관심은 모수벡터 B에 있으므로, 기저위험함수를 포함한 정보행렬(information matrix)을 계산할 필요가 없다. 따라서 기저위험함수를 다른 관심있는 모수들로 나타낸 프로파일 우도(profile likelihood)를 사용한다. $R(t_j)$가 t_j시점의 위험집합(risk set)이라고 할 때 프로파일된 기저위험함수는 다음과 같다.

$$\lambda_0^{\text{profile out}} = \frac{1}{\sum_{r \in R(t_j)} \exp\left(X_r^S(t_j)^T \gamma + u_r\right)}.$$

모수 벡터 B에 대한 분산-공분산 행렬은 프로파일 우도로부터 구한 피셔정보(empirical Fisher information)의 역수를 취함으로써 계산할 수 있다 (Lin 등, 2004; Murphy와 Vaart, 2000). $l^{(i)}(\hat{B}; Y, S)$는 개체 i의 프로파일 우도에서 구한 관측스코어벡터(observed score vector)로 EM 알고리즘에서 수렴된 모수값 $\hat{B}$가 주어졌을 때의 프로파일 우도로부터 구한다. 모수 벡터 B의 관측정보행렬(observed information matrix)은 다음과 같이 근사된다.

$$\sum_{i=1}^{N} l^{(i)}\left(\hat{B}; Y, S\right) l^{(i)}\left(\hat{B}; Y, S\right)^T. \tag{3.4}$$

4. 모의실험

경시적 영과잉 허들모형과 생존모형으로 이루어진 결합모형을 분리된 개별 모형과 각각 성능을 비교한다. 경시적 영과잉 자료와 생존자료 간에 강한 연관성이 존재한다고 하면 생존자료는 경시적 영과잉 자료의 결측 또는 중도탈락을 설명할 수 있다. 이와 같이 두 자료 간 연관성이 존재할 때 분리된 모형으로 자료를 분석하면 편향된 결과를 가지는 문제가 발생하며, 이를 해결하기 위해 결합모형을 이용하면 분리된 모형보다 작은 편의를 가질 것이다. 제안된 모형에서는 경시적 영과잉 자료와 생존자료의 관계가 각 자료의 변량효과들 간의 연관성에 영향을 받기 때문에, 이를 확인하기 위해 두 자료 간에 연관성이 거의 없을 경우($\rho = 0.2$), 높은 양의 상관관계가 존재할 경우($\rho = 0.85$) 그리고 높은 음의 상관관계를 가질 경우($\rho = -0.85$) 각각에 대해 모의실험을 실시한다. 모의실험을 위한 결합모형의 두 부 모형 중 첫 번째 부 모형은 경시적 영과잉 가산자료에 대한 허들모형으로 다음과 같다.

$$\begin{aligned}
P(Y_{ij} = 0) &= p_{ij}, \\
P(Y_{ij} = y_{ij}) &= \frac{1 - p_{ij}}{1 - e^{-\mu_{ij}}} \frac{e^{-\mu_{ij}} \mu_{ij}^{y_{ij}}}{y_{ij}!}, \quad y_{ij} = 1, 2, \ldots, \\
\text{logit}(p_{ij}) &= \eta_0 + \eta_1 t_{ij} + \eta_2 X_{2i} + \phi b_i t_{ij}, \\
\log(\mu_{ij}) &= \beta_0 + \beta_1 t_{ij} + \beta_2 X_{2i} + b_i t_{ij},
\end{aligned} \tag{4.1}$$

Table 4.1. Comparison of the joint model and the separate model ($\rho = 0.2$, censoring rate = 50%)

	True	Separate model				Joint model			
		Bias	SE	Est.SE	CP	Bias	SE	Est.SE	CP
η_0	0.40	−0.008	0.144	0.146	0.955	−0.010	0.146	0.149	0.960
η_1	0.20	0.011	0.266	0.258	0.950	0.031	0.285	0.287	0.965
η_2	−0.80	−0.002	0.158	0.172	0.965	0.000	0.161	0.175	0.965
β_0	0.80	0.007	0.079	0.088	0.960	0.002	0.082	0.091	0.970
β_1	0.40	−0.036	0.178	0.191	0.980	0.002	0.206	0.232	0.975
β_2	−0.70	−0.009	0.114	0.116	0.955	−0.001	0.117	0.120	0.955
ϕ	0.50	0.084	0.750	0.744	0.970	0.069	0.796	0.728	0.970
σ_b	0.40	−0.002	0.182	0.195	0.930	0.016	0.187	0.207	0.940
γ_1	1.50	−0.001	0.266	0.218	0.915	0.020	0.271	0.224	0.925
γ_2	1.20	−0.029	0.240	0.232	0.935	−0.014	0.243	0.235	0.930
σ_u	0.40	−0.061	0.387	0.334	0.990	−0.014	0.414	0.358	0.980
σ_{bu}	0.08					0.014	0.221	0.285	0.990

여기서 $t_{ij}(t_{ij} = 0, 0.2, 0.4, \ldots, 1)$는 계획된 방문 시간을 나타낸다. X_{2i}는 시험군과 대조군을 나타내는 지시변수로 평균이 0.5인 베르누이분포로부터 각 개체 i에 대해 독립적으로 생성한다. 두 번째 부 모형은 생존시간 자료의 생성을 위해 비례위험 모형을 사용한다.

$$\lambda\left(t; X_{1i}, X_{2i}, u_i, \boldsymbol{\gamma}\right) = \lambda_0(t)\exp\left(\gamma_1 X_{1i} + \gamma_2 X_{2i} + u_i\right), \tag{4.2}$$

여기서 공변량 X_{1i}는 $N(1,\ 0.4)$인 독립적인 정규분포로부터 생성하고 기저위험 $\Lambda_0(t)$는 0.1로 가정한다. 그리고 각 사건이 발생한 시간은 지수분포를 따르는 것으로 가정하며, 모의실험을 위한 중도절단된 시간은 평균이 3인 지수분포로부터 생성하였다. 변량효과인 b_i와 u_i에 대해, $\theta_i = \binom{b_i}{u_i}$로 나타내며, $\theta_i \sim N\left(\binom{0}{0}, \begin{pmatrix}\sigma_b^2 & \sigma_{bu}\\ \sigma_{bu} & \sigma_u^2\end{pmatrix}\right)$를 가정한다.

결합모형에 대한 분석은 (4.1)과 (4.2)의 결합모형을 적용하였으며, 개별적인 분석에 대해서는 경시적 영과잉 모형 (4.1)과 생존모형 (4.2)를 각각 이용한다. 각 모의실험은 표본크기가 200인 자료를 생성해서 총 200번의 몬테칼로 반복실험을 수행하였다. 중도절단비율에 따른 추정의 차이를 확인하기 위해 중도절단비율이 대략 50%, 40%, 30%인 경우로 구분하였으며, 또한 변량효과에 대해서 서로 다른 상관계수의 값을 가진 경우로 구분하여 각각 세 번의 모의실험을 수행하며 상관계수(ρ)에 대한 분산-공분산 행렬은, 약한 상관관계인 $\rho = 0.2$일 때 $\Sigma = \left(\begin{smallmatrix}0.40 & 0.08\\ 0.08 & 0.40\end{smallmatrix}\right)$, 강한 양의 상관관계인 $\rho = 0.85$일 때 $\Sigma = \left(\begin{smallmatrix}0.40 & 0.34\\ 0.34 & 0.40\end{smallmatrix}\right)$ 그리고 강한 음의 상관관계인 $\rho = -0.85$일 때 $\Sigma = \left(\begin{smallmatrix}0.40 & -0.34\\ -0.34 & 0.40\end{smallmatrix}\right)$로 설정하였다. 각 모수에 대한 추정값의 평균과 표준오차 그리고 95% 신뢰구간에 대한 포함확률(coverage probabilities; CP)은 다음의 각 Table에 나타나 있다. Table 4.1–4.9에서 편의(bias)는 모수 추정값의 평균과 모수의 차를 나타내고, SE는 모수추정량의 표준오차를, Est.SE는 프로파일 우도를 이용하여 계산된 표준오차들의 중앙값(median)을 나타낸다. 모의실험의 결과는 다음의 Table 4.1–4.9에 나타낸다.

중도절단이 50% 발생한 경우, Table 4.1의 두 부 모형 간에 상관관계가 약한 경우에는 결합모형이 분리된 모형보다 $\eta_0, \eta_1, \sigma_b, \gamma_1$을 제외한 나머지 7개 추정량은 편의가 작은 결과를 보였다. 두 모형 사이에 큰 상관관계가 존재하는 Table 4.2에서는 대부분의 결합모형의 결과가 분리된 모형보다 편의가 더 작은 것을 확인할 수 있다. 계획된 방문시간의 모수인 η_1과 β_1을 비교할 때, 분리된 모형의 η_1과 β_1에 대한 편의는 각각 −0.054, −0.156인 반면, 결합모형에서는 0.005, −0.046으로 편의가 상당히 줄었음을 확인할 수 있고, σ_b는 −0.042에서 −0.004으로 매우 개선된 결과를 보였다. 단, 허들모형 내의 p_{ij}의 모형과

Table 4.2. Comparison of the joint model and the separate model ($\rho = 0.85$, censoring rate = 50%)

	True	Separate model				Joint model			
		Bias	SE	Est.SE	CP	Bias	SE	Est.SE	CP
η_0	0.40	−0.002	0.147	0.147	0.945	−0.010	0.148	0.149	0.955
η_1	0.20	−0.054	0.267	0.259	0.940	0.005	0.288	0.295	0.965
η_2	−0.80	−0.007	0.156	0.172	0.975	0.004	0.157	0.175	0.970
β_0	0.80	0.022	0.080	0.089	0.955	0.008	0.082	0.091	0.975
β_1	0.40	−0.156	0.175	0.192	0.925	−0.046	0.196	0.229	0.980
β_2	−0.70	−0.033	0.122	0.117	0.910	−0.012	0.122	0.121	0.930
ϕ	0.50	0.050	0.842	0.855	0.985	0.051	0.786	0.778	0.975
σ_b	0.40	−0.042	0.168	0.193	0.895	−0.004	0.183	0.213	0.925
γ_1	1.50	0.000	0.270	0.218	0.900	0.031	0.273	0.224	0.905
γ_2	1.20	−0.027	0.265	0.231	0.930	−0.006	0.274	0.235	0.915
σ_u	0.40	−0.057	0.419	0.334	0.965	0.007	0.449	0.364	0.955
σ_{bu}	0.34					−0.090	0.227	0.287	0.965

Table 4.3. Comparison of the joint model and the separate model ($\rho = -0.85$, censoring rate = 50%)

	True	Separate model				Joint model			
		Bias	SE	Est.SE	CP	Bias	SE	Est.SE	CP
η_0	0.40	−0.024	0.142	0.146	0.945	−0.016	0.142	0.148	0.950
η_1	0.20	0.110	0.270	0.257	0.890	0.048	0.278	0.282	0.940
η_2	−0.80	0.018	0.155	0.170	0.965	0.007	0.157	0.174	0.965
β_0	0.80	−0.024	0.086	0.088	0.960	−0.008	0.087	0.090	0.955
β_1	0.40	0.158	0.178	0.187	0.855	0.050	0.196	0.231	0.965
β_2	−0.70	0.038	0.111	0.113	0.935	0.016	0.113	0.117	0.955
ϕ	0.50	−0.024	0.737	0.687	0.940	0.003	0.695	0.643	0.930
σ_b	0.40	−0.020	0.158	0.179	0.920	0.005	0.168	0.193	0.945
γ_1	1.50	−0.003	0.245	0.218	0.930	0.026	0.250	0.224	0.925
γ_2	1.20	−0.053	0.248	0.231	0.910	−0.028	0.260	0.235	0.910
σ_u	0.40	−0.072	0.369	0.329	0.970	−0.015	0.393	0.359	0.980
σ_{bu}	−0.34					0.095	0.194	0.284	0.975

μ_{ij}의 모형 내의 변량효과의 관계를 나타내는 모수 ϕ의 경우, 분리된 모형과 결합모형이 유사한 값을 보였다. 프로파일 우도에서 구해진 표준오차(Est.SE)는 모수추정량의 표준오차(SE)와 비교할 때 거의 유사함을 확인할 수 있다. 포함확률의 경우 결합모형이 분리된 모형보다 0.95에 더 근접한다. 음의 상관관계를 나타내는 Table 4.3에서는 결합모형의 모수의 추정 결과가 분리된 모형보다 전반적으로 편의가 더 작음을 확인할 수 있다.

Table 4.4–4.6의 중도절단비율이 40%인 경우는 중도절단비율이 50%인 경우와 유사한 경향을 보이며, 결합모형의 대부분의 추정값이 분리된 모형보다 더 작은 편의를 가졌다. 중도절단이 30% 발생한 Table 4.7–4.9에서 $\rho = 0.2$인 경우에는 $\eta_0, \eta_1, \phi, \sigma_b$를 제외하고는 결합된 모형이 더 나은 결과를 보였지만, 상관이 큰 $\rho = 0.85$일 때 모든 모수의 추정 결과는 결합모형이 분리된 모형보다 편의가 더 작았다. 또한, 중도절단비율이 줄어들수록 편의가 감소하였다.

모수별로 살펴보면 시간효과를 나타내는 모수인 η_1과 β_1은 대체적으로 결합모형이 분리된 모형보다 더 작은 편의를 가지는 것을 확인할 수 있고, 다른 모수들에 비해 결합모형이 분리된 모형보다 향상된 정도가 더 두드러졌다. 그룹효과를 나타내는 $\eta_2, \beta_2, \gamma_2$는 결합모형이 분리된 모형에 비해 편의가 더 작았으

Table 4.4. Comparison of the joint model and the separate model ($\rho = 0.2$, censoring rate = 40%)

	True	Separate model				Joint model			
		Bias	SE	Est.SE	CP	Bias	SE	Est.SE	CP
η_0	0.40	−0.011	0.151	0.152	0.955	−0.013	0.152	0.154	0.955
η_1	0.20	0.005	0.286	0.282	0.945	0.025	0.300	0.314	0.965
η_2	−0.80	0.004	0.173	0.185	0.950	0.006	0.174	0.188	0.960
β_0	0.80	0.004	0.087	0.092	0.970	0.001	0.088	0.094	0.970
β_1	0.40	−0.037	0.188	0.207	0.970	−0.003	0.216	0.250	0.965
β_2	−0.70	−0.009	0.130	0.125	0.955	−0.004	0.131	0.128	0.945
ϕ	0.50	0.126	0.953	0.946	0.985	0.137	0.977	0.922	0.975
σ_b	0.40	−0.012	0.195	0.217	0.915	0.004	0.200	0.228	0.920
γ_1	1.90	−0.027	0.278	0.210	0.870	−0.006	0.287	0.215	0.865
γ_2	1.30	−0.056	0.243	0.208	0.895	−0.041	0.250	0.211	0.900
σ_u	0.40	−0.094	0.323	0.253	0.965	−0.060	0.350	0.273	0.955
σ_{bu}	0.08					−0.008	0.188	0.272	1.000

Table 4.5. Comparison of the joint model and the separate model ($\rho = 0.85$, censoring rate = 40%)

	True	Separate model				Joint model			
		Bias	SE	Est.SE	CP	Bias	SE	Est.SE	CP
η_0	0.40	0.001	0.154	0.153	0.945	−0.007	0.154	0.155	0.950
η_1	0.20	−0.065	0.288	0.284	0.940	0.001	0.316	0.330	0.970
η_2	−0.80	−0.011	0.181	0.186	0.950	−0.002	0.180	0.189	0.955
β_0	0.80	0.021	0.084	0.092	0.960	0.008	0.085	0.095	0.970
β_1	0.40	−0.184	0.197	0.212	0.905	−0.060	0.218	0.250	0.980
β_2	−0.70	−0.035	0.131	0.127	0.940	−0.015	0.132	0.131	0.950
ϕ	0.50	0.050	0.984	1.051	0.970	0.033	0.904	0.909	0.980
σ_b	0.40	−0.035	0.195	0.223	0.925	0.006	0.205	0.247	0.940
γ_1	1.90	−0.010	0.287	0.211	0.880	0.030	0.300	0.216	0.855
γ_2	1.30	−0.039	0.262	0.208	0.885	−0.013	0.276	0.212	0.885
σ_u	0.40	−0.055	0.392	0.261	0.930	0.013	0.447	0.291	0.920
σ_{bu}	0.34					−0.082	0.196	0.286	0.975

Table 4.6. Comparison of the joint model and the separate model ($\rho = -0.85$, censoring rate = 40%)

	True	Separate model				Joint model			
		Bias	SE	Est.SE	CP	Bias	SE	Est.SE	CP
η_0	0.40	−0.022	0.149	0.152	0.940	−0.013	0.150	0.154	0.945
η_1	0.20	0.122	0.295	0.282	0.905	0.048	0.304	0.315	0.965
η_2	−0.80	0.016	0.165	0.184	0.955	0.005	0.168	0.188	0.960
β_0	0.80	−0.023	0.091	0.091	0.960	−0.009	0.090	0.093	0.965
β_1	0.40	0.179	0.196	0.201	0.845	0.054	0.215	0.248	0.975
β_2	−0.70	0.036	0.123	0.122	0.935	0.014	0.123	0.126	0.940
ϕ	0.50	0.058	0.885	0.870	0.960	0.081	0.843	0.795	0.960
σ_b	0.40	−0.025	0.176	0.201	0.915	0.002	0.187	0.218	0.950
γ_1	1.90	−0.035	0.263	0.210	0.870	0.006	0.261	0.216	0.900
γ_2	1.30	−0.066	0.245	0.208	0.880	−0.037	0.251	0.211	0.905
σ_u	0.40	−0.108	0.313	0.248	0.920	−0.037	0.341	0.282	0.955
σ_{bu}	−0.34					0.087	0.183	0.274	0.980

Table 4.7. Comparison of the joint model and the separate model ($\rho = 0.2$, censoring rate = 30%)

	True	Separate model				Joint model			
		Bias	SE	Est.SE	CP	Bias	SE	Est.SE	CP
η_0	0.40	−0.015	0.166	0.160	0.960	−0.016	0.165	0.162	0.955
η_1	0.20	0.016	0.312	0.318	0.970	0.031	0.336	0.357	0.970
η_2	−0.80	−0.002	0.191	0.204	0.970	0.000	0.191	0.207	0.975
β_0	0.80	0.002	0.090	0.096	0.970	−0.001	0.091	0.098	0.980
β_1	0.40	−0.051	0.197	0.235	0.980	−0.011	0.223	0.285	0.985
β_2	−0.70	−0.010	0.142	0.138	0.955	−0.004	0.144	0.141	0.960
ϕ	0.50	−0.020	1.014	1.181	0.995	−0.040	1.008	1.089	0.990
σ_b	0.40	−0.003	0.212	0.264	0.930	0.017	0.218	0.281	0.950
γ_1	2.40	−0.068	0.295	0.210	0.815	−0.037	0.312	0.215	0.820
γ_2	1.50	−0.073	0.239	0.195	0.855	−0.053	0.250	0.198	0.860
σ_u	0.40	−0.116	0.284	0.193	0.745	−0.077	0.316	0.214	0.815
σ_{bu}	0.08					0.007	0.199	0.291	1.000

Table 4.8. Comparison of the joint model and the separate model ($\rho = 0.85$, censoring rate = 30%)

	True	Separate model				Joint model			
		Bias	SE	Est.SE	CP	Bias	SE	Est.SE	CP
η_0	0.40	0.003	0.162	0.160	0.950	−0.003	0.163	0.163	0.960
η_1	0.20	−0.078	0.306	0.320	0.940	−0.004	0.347	0.378	0.975
η_2	−0.80	−0.016	0.194	0.205	0.960	−0.007	0.194	0.208	0.960
β_0	0.80	0.020	0.090	0.097	0.960	0.010	0.090	0.099	0.965
β_1	0.40	−0.206	0.208	0.237	0.935	−0.073	0.227	0.278	0.975
β_2	−0.70	−0.037	0.143	0.140	0.950	−0.020	0.143	0.143	0.950
ϕ	0.50	0.052	1.167	1.519	0.990	0.034	1.332	1.480	0.980
σ_b	0.40	−0.036	0.224	0.265	0.935	0.005	0.241	0.295	0.945
γ_1	2.40	−0.046	0.310	0.211	0.835	0.001	0.330	0.216	0.820
γ_2	1.50	−0.050	0.258	0.195	0.845	−0.019	0.272	0.198	0.845
σ_u	0.40	−0.082	0.345	0.201	0.765	−0.021	0.385	0.230	0.860
σ_{bu}	0.34					−0.085	0.217	0.300	0.960

Table 4.9. Comparison of the joint model and the separate model ($\rho = -0.85$, censoring rate = 30%)

	True	Separate model				Joint model			
		Bias	SE	Est.SE	CP	Bias	SE	Est.SE	CP
η_0	0.40	−0.016	0.159	0.160	0.945	−0.010	0.160	0.162	0.955
η_1	0.20	0.135	0.338	0.319	0.915	0.058	0.343	0.361	0.975
η_2	−0.80	0.005	0.183	0.203	0.980	−0.004	0.184	0.207	0.980
β_0	0.80	−0.020	0.090	0.096	0.970	−0.008	0.089	0.098	0.980
β_1	0.40	0.189	0.201	0.225	0.860	0.054	0.219	0.279	0.965
β_2	−0.70	0.031	0.141	0.134	0.915	0.011	0.140	0.138	0.930
ϕ	0.50	0.077	1.269	1.437	0.980	0.097	1.287	1.350	0.995
σ_b	0.40	−0.020	0.207	0.242	0.915	0.011	0.221	0.264	0.935
γ_1	2.40	−0.069	0.287	0.210	0.835	−0.023	0.295	0.216	0.850
γ_2	1.50	−0.079	0.241	0.195	0.875	−0.050	0.251	0.198	0.870
σ_u	0.40	−0.119	0.279	0.196	0.760	−0.057	0.331	0.223	0.810
σ_{bu}	−0.34					0.088	0.193	0.287	0.960

며, 상관관계가 높을수록 더 정확하게 추정되었다. 또한 각 변량효과의 분산인 σ_b, σ_u의 경우에도 결합모형에서 분리된 모형에 비해 편의가 줄어든 결과를 보였다.

모의실험 결과 Table 4.1–4.9로부터 전반적으로 거의 모든 결합모형의 결과가 개별적으로 분석한 모형의 결과보다 편의가 감소된 것을 확인할 수 있으며, 경시적 자료와 생존자료의 두 자료 간의 연관성이 크지 않을 경우에는 결합모형이 분리된 모형보다 대체적으로 편의가 줄어든 결과를 보여주며, 두 자료간의 연관성이 커질 경우 상당히 향상된 결과를 가지는 것을 확인할 수 있다.

5. 결론

우리는 경시적 영과잉 가산자료와 생존자료의 결합모형에 대해 연구하였다. 연구된 모형은 경시적 자료와 생존자료에 대한 결합모형에서 새롭게 구축된 모형으로 기존의 반복적으로 측정되는 연속형 자료 대신 반복적으로 측정되는 영(0)이 많이 관측되는 가산자료에 대해 연구하였다. 영과잉 가산자료를 분석하기 위해 가산자료의 분석에 주로 이용되는 포아송모형 대신 포아송 허들모형을 이용하였다.

어떤 원인으로 발생된 결측자료가 존재하는 경시적 영과잉 가산자료를 분석하기 위해 포아송 허들모형과 비례위험모형을 결합모형의 두 부 모형으로 사용하였으며 두 부 모형 내의 변량효과가 서로 연관있다고 가정하여 결합모형을 구성하였다. 모수의 최대우도추정량을 추정하기 위해 EM 알고리즘을 이용하였고 추정된 모수의 표준오차를 구하기 위해 프로파일 우도를 이용하였다. 모의실험에서는 경시적 영과잉 가산자료를 위한 포아송 허들모형과 생존자료를 위한 비례위험모형의 결합모형과 이 두 모형을 개별적으로 분석하여 각 모형의 성능을 비교하였다. 각 변량효과의 상관관계가 거의 없거나 양 또는 음의 상관관계를 갖는 경우를 고려하여 모의실험을 수행하였다. 약한 상관관계를 가지는 경우 결합모형의 개선도가 크지 않았지만, 각 분리된 모형에 대한 강한 상관관계가 존재할 때는 상관관계를 고려하지 않고 개별적으로 분석한 모형보다 결합모형이 편의 측면에서 상당히 개선됨을 확인하였다. 종합적으로 경시적 영과잉 가산자료의 어떤 원인으로 중도탈락이 존재하는 경우, 경시적 자료의 중도탈락과 관련된 생존자료와의 결합모형을 사용함으로써 보다 유기적인 관계를 통하여 모수 추정을 정확히 할 수 있다고 본다

기존 결합모형 연구에서 고려하지 않았던 영(0)이 많이 존재하는 경시적 가산자료에도 결합모형이 잘 적합함을 확인하였다. 본 논문에서는 영과잉 경시적 가산자료의 분석에 허들모형인 포아송 허들모형을 적용하였다. 일반적으로 가산자료의 경우 과대산포(overdispersion)의 문제를 가지고 있으므로 과대산포가 존재하는 자료에 포아송모형을 적용하면 추정값의 표준오차에 편의가 발생하게 된다. 따라서 영과잉 가산자료의 분석에 포아송 허들모형 대신에 음이항 허들모형을 사용한다면 좀 더 개선된 결과를 얻을 수 있을 것이라 기대한다. 모의실험 결과 각 프로파일 우도로부터 구한 표준오차 추정값이 각 모수추정량의 표준오차에 비해 과소추정 또는 과대추정되는 경우가 존재하였다. 표준오차의 추정에 프로파일 방법 외의 다른 대안적 방법의 적용도 향후 연구에서 고려될 수 있다.

부록: EM 알고리즘

• Expectation-step

E-step의 $(t+1)$번째 반복에서 다음의 조건부 기대값을 계산한다.

$$E_{\theta_i|Y_i,S_i,\Delta^{(t)}}\left(k\left(\theta_i\right)\right) = \int k\left(\theta_i\right) f\left(\theta_i|Y_i,S_i,\Delta^{(t)}\right) d\theta_i$$
$$= \frac{\int k\left(\theta_i\right) f\left(Y_i,S_i,\theta_i|\Delta^{(t)}\right) d\theta_i}{f\left(Y_i,S_i|\Delta^{(t)}\right)}$$

$$= \frac{\int k(\theta_i) f\left(Y_i, S_i, \theta_i | \Delta^{(t)}\right) d\theta_i}{\int f\left(Y_i, S_i, \theta_i | \Delta^{(t)}\right) d\theta_i}$$
$$= \frac{\int k(\theta_i) f\left(Y_i | \theta_i, \Delta^{(t)}\right) f\left(S_i | \theta_i, \Delta^{(t)}\right) f\left(\theta_i | \Delta^{(t)}\right) d\theta_i}{\int f\left(Y_i | \theta_i, \Delta^{(t)}\right) f\left(S_i | \theta_i, \Delta^{(t)}\right) f\left(\theta_i | \Delta^{(t)}\right) d\theta_i},$$

여기서 $k(\theta_i)$는 모든 θ의 함수이고 적분을 계산하기 위해 이변량 가우스 헬머트 구적(Bivariate Gauss-Hermite quadrature)을 이용한다.

$$E_{\theta_i | Y_i, S_i, \Delta^{(t)}}\left(k(\theta_i)\right) = \frac{\sum_l \sum_m w_l w_m \frac{1}{\pi} k\left(b_l^*, u_m^*\right) f\left(Y_i | b_l^*, \Delta\right) f\left(S_i | u_m^*, \Delta\right)}{\sum_l \sum_m w_l w_m \frac{1}{\pi} f\left(Y_i | b_l^*, \Delta\right) f\left(S_i | u_m^*, \Delta\right)}.$$

• **Maximization-step**

모든 모수들에 대한 모수 벡터는 $\Delta = \{\eta, \phi, \beta, \gamma, \Sigma, \lambda_0(t)\}$이다. 여기서 모수 $\eta, \phi, \beta, \gamma$은 닫힌 형식을 가지지 않으므로 각 반복단계에서 뉴튼랩슨 알고리즘을 이용해서 갱신된다. 모수들의 갱신은 다음과 같다. 개체 i와 시점 j를 구별하기 위해 본문 식 (2.2)에서 사용된 t 대신 t_{ij}를 사용한다.

$\eta^{(t+1)} = \eta^{(t)} + I_\eta^{(t)-1} S_\eta^{(t)}$에서 $S_\eta^{(t)}$와 $I_\eta^{(t)}$는

$$S_\eta^{(t)} = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left\{ I\left(Y_{ij} = 0\right) X_i^L\left(t_{ij}\right) - E\left[\frac{\exp\left(X_i^L\left(t_{ij}\right)^T \eta^{(t)} + Z_i\left(t_{ij}\right)^T \phi^{(t)} b_i\right)}{1 + \exp\left(X_i^L\left(t_{ij}\right)^T \eta^{(t)} + Z_i\left(t_{ij}\right)^T \phi^{(t)} b_i\right)} \right] X_i^L\left(t_{ij}\right) \right\},$$

$$I_\eta^{(t)} = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left\{ E\left[\frac{\exp\left(X_i^L\left(t_{ij}\right)^T \eta^{(t)} + Z_i\left(t_{ij}\right)^T \phi^{(t)} b_i\right)}{1 + \exp\left(X_i^L\left(t_{ij}\right)^T \eta^{(t)} + Z_i\left(t_{ij}\right)^T \phi^{(t)} b_i\right)} \right] X_i^L\left(t_{ij}\right) X_i^L\left(t_{ij}\right)^T \right.$$
$$\left. - E\left[\frac{\left\{\exp\left(X_i^L\left(t_{ij}\right)^T \eta^{(m)} + Z_i\left(t_{ij}\right)^T \phi^{(t)} b_i\right)\right\}^2}{\left\{1 + \exp\left(X_i^L\left(t_{ij}\right)^T \eta^{(t)} + Z_i\left(t_{ij}\right)^T \phi^{(t)} b_i\right)\right\}^2} \right] X_i^L\left(t_{ij}\right) X_i^L\left(t_{ij}\right)^T \right\}.$$

$\phi^{(t+1)} = \phi^{(t)} + I_\phi^{(t)-1} S_\phi^{(t)}$에서 $S_\phi^{(t)}$와 $I_\phi^{(t)}$는

$$S_\phi^{(t)} = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left\{ I\left(Y_{ij} = 0\right) E\left[Z_i\left(t_{ij}\right)^T b_i\right] - E\left[\frac{\exp\left(X_i^L\left(t_{ij}\right)^T \eta^{(t)} + Z_i\left(t_{ij}\right)^T \phi^{(t)} b_i\right)}{1 + \exp\left(X_i^L\left(t_{ij}\right)^T \eta^{(t)} + Z_i\left(t_{ij}\right)^T \phi^{(t)} b_i\right)} \right] Z\left(t_{ij}\right) \right\},$$

$$I_\phi^{(t)} = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left\{ E\left[\frac{\exp\left(X_i^L\left(t_{ij}\right)^T \eta^{(t)} + Z_i\left(t_{ij}\right)^T \phi^{(t)} b_i\right) b_i^2}{1 + \exp\left(X_i^L\left(t_{ij}\right)^T \eta^{(t)} + Z_i\left(t_{ij}\right)^T \phi^{(t)} b_i\right)} \right] Z_i\left(t_{ij}\right) Z_i\left(t_{ij}\right)^T \right.$$
$$\left. - E\left[\frac{\left\{\exp\left(X_i^L\left(t_{ij}\right)^T \eta^{(t)} + Z_i\left(t_{ij}\right)^T \phi^{(t)} b_i\right) b_i\right\}^2}{\left\{1 + \exp\left(X_i^L\left(t_{ij}\right)^T \eta^{(t)} + Z_i\left(t_{ij}\right)^T \phi^{(t)} b_i\right)\right\}^2} \right] Z_i(t_{ij}) Z_i(t_{ij})^T \right\}.$$

$\beta^{(t+1)} = \beta^{(t)} + I_\beta^{(t)-1} S_\beta^{(t)}$에서 $S_\beta^{(t)}$와 $I_\beta^{(t)}$는

$$S_\beta^{(t)} = \sum_{i=1}^{N}\sum_{j=1}^{n_i}(1 - I(Y_{ij}=0)) \left\{ Y_{ij} X_i^L(t_{ij}) - E\left[\frac{\exp\left(\exp\left(X_i^L(t_{ij})^T\beta^{(t)} + Z_i(t_{ij})^T b_i\right)\right)\left(\exp\left(X_i^L(t_{ij})^T\beta^{(t)} + Z_i(t_{ij})^T b_i\right)\right)}{\exp\left(\exp\left(X_i^L(t_{ij})^T\beta^{(t)} + Z_i(t_{ij})^T b_i\right)\right) - 1} \right] X_i^L(t_{ij}) \right\},$$

$$\begin{aligned} I_\beta^{(t)} = & \sum_{i=1}^{N}\sum_{j=1}^{n_i}(1 - I(Y_{ij}=0)) \\ & \left\{ E\left[\frac{\exp\left(\exp\left(X_i^L(t_{ij})^T\beta^{(t)} + Z_i(t_{ij})^T b_i\right)\right)\left(\exp\left(X_i^L(t_{ij})^T\beta^{(t)} + Z_i(t_{ij})^T b_i\right)\right)^2}{\exp\left(\exp\left(X_i^L(t_{ij})^T\beta^{(t)} + Z_i(t_{ij})^T b_i\right)\right) - 1} \right] X_i^L(t_{ij}) X_i^L(t_{ij})^T \right. \\ & + E\left[\frac{\exp\left(\exp\left(X_i^L(t_{ij})^T\beta^{(t)} + Z_i(t_{ij})^T b_i\right)\right)\exp\left(X_i^L(t_{ij})^T\beta^{(t)} + Z_i(t_{ij})^T b_i\right)}{\exp\left(\exp\left(X_i^L(t_{ij})^T\beta^{(t)} + Z_i(t_{ij})^T b_i\right)\right) - 1} \right] X_i^L(t_{ij}) X_i^L(t_{ij})^T \\ & \left. + E\left[\frac{\left\{\exp\left(\exp\left(X_i^L(t_{ij})^T\beta^{(t)} + Z_i(t_{ij})^T b_i\right)\right)\exp\left(X_i^L(t_{ij})^T\beta^{(t)} + Z_i(t_{ij})^T b_i\right)\right\}^2}{\left\{\exp\left(\exp\left(X_i^L(t_{ij})^T\beta^{(t)} + Z_i(t_{ij})^T b_i\right)\right) - 1\right\}^2} \right] X_i^L(t_{ij}) X_i^L(t_{ij})^T \right\}. \end{aligned}$$

모수 Σ은 닫힌 형식을 가지므로 갱신은 다음의 과정을 따른다.

$$\frac{\partial E[l(\Delta; Y, S, \theta)]}{\partial \Sigma} = \sum_{i=1}^{N}\left\{ -\frac{1}{2}\Sigma^{-T} + \frac{1}{2}E\left[\Sigma^{-T}\theta_i\theta_i^T{}'\Sigma^{-T}\right] \right\} = 0.$$

주어진 식을 풀면,

$$\Sigma^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} E\left[\theta_i\theta_i^T\right]$$

이다.

위의 식을 다시 정리하면

$$\begin{pmatrix} \Sigma_{bb}^{(t+1)} & \Sigma_{bu}^{(t+1)} \\ \Sigma_{bu}^{(t+1)} & \sigma_u^{2(t+1)} \end{pmatrix} = \frac{1}{N}\sum_{i=1}^{N} E\begin{bmatrix} b_i b_i^T & b_i u_i \\ b_i^T u_i & u_i^2 \end{bmatrix}$$

가 된다.

따라서 모수 Σ은

$$\Sigma_{bb}^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} E\left(b_i b_i^T\right), \quad \Sigma_{bu}^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} E(b_i u_i), \quad \text{and} \quad \sigma_u^{2(t+1)} = \frac{1}{N}\sum_{i=1}^{N} E(u_i^2)$$

로 갱신할 수 있다.

q개의 사건이 발생한 시간(failure time)을 $t_1 \le \cdots \le t_q$로 나타낼 때, $R(t_j)$가 t_j 시점의 위험집합(risk set)이라 하고 t_j에서 발생한 사건의 개수를 d_j라 한다면 누적기저위험함수는 다음과 같다.

$$\Lambda_0^{(t+1)}(t_q) = \sum_{j=1}^{q} \frac{d_j}{\sum_{r\in R(t_j)} \exp\left(X_r^S(t_j)^T\gamma^{(t)}\right) E\left[\exp\left(u_r\right)\right]},$$

여기서 $\boldsymbol{\gamma}$는 닫힌 형식을 가지지 않으므로 뉴튼랩슨 알고리즘에 따라 갱신된다.
$\gamma^{(t+1)} = \gamma^{(t)} + I_\gamma^{(t)-1} S_\gamma^{(t)}$에서 $S_\gamma^{(t)}$와 $I_\gamma^{(t)}$는

$$S_\gamma^{(t)} = \sum_{i=1}^{N} \left\{ G_i X_i^S(T_i)^T - \sum_{t_j \le T_i} \lambda_0^{(t+1)}(t_j) \exp\left(X_i^S(t_j)^T \gamma^{(t)} \right) E\left[\exp(u_i)\right] X_i^S(t_j) \right\},$$

$$I_\gamma^{(t)} = \sum_{i=1}^{N} \sum_{t_j \le T_i} \lambda_0(t_j)^{(t+1)} \exp\left(X_i^S(t_j)^T \gamma^{(t)} \right) E\left[\exp(u_i)\right] X_i^S(t_j) X_i^S(t_j)^T.$$

References

Buu, A., Li, R., Tan, X., and Zuker, R. A. (2012). Statistical models for longitudinal zero-inflated count data with applications to the substance abuse field, *Statistics in Medicine*, **31**, 4074–4086.

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

Diggle, P. J. and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis (with discussion), *Applied Statistics*, **43**, 49–93.

Elashoff, R. M., Li, G., and Li, N. (2008). A Joint model for longitudinal measurements and survival data in the presence of multiple failure types, *Biometrics*, **64**, 762–771.

Hall, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: a case study, *Biometrics*, **56**, 1030–1039.

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

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing, *Technometrics*, **34**, 1–14.

Lin, H., McCulloch, C. E., and Rosenheck, R. A. (2004). Latent pattern mixture models for informative intermittent missing data in longitudinal studies, *Biometrics*, **60**, 295–305.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*(2nd ed.), Wiley, New York.

Liu, Q. and Pierce, D. A. (1994). A note on Gauss-Hermite quadrature, *Biometrika*, **81**, 624–629.

Min, Y. and Agresti, A. (2005). Random effects models for repeated measures of zero-inflated count data, *Statistical Modeling*, **5**, 1–19.

Mullahy, J. (1986). Specification and testing of some modified count data models, *Journal of Econometrics*, **33**, 341–365.

Murphy, S. A. and Vaart, W. (2000). On profile likelihood, *Journal of the American Statistical Association*, **95**, 449–465.

Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model, *Biometrics*, **69**, 331–342.

Sousa, I. (2011). A review on joint modeling of longitudinal measurements and time-to-event, *Revstat*, **9**, 57–81.

Tseng, Y., Hsieh, F., and Wang, J. L. (2005). Joint modeling of accelerated failure time and longitudinal data, *Biometrika*, **92**, 587–603.

Wu, L., Liu, W., Yi, G. Y., and Huang, Y. (2012). Analysis of longitudinal and survival data: joint modeling, inference methods, and issues, *Journal of Probability and Statistics 2012*, Article ID 640153.

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

Yau, K. K. and Lee, A. H. (2001). Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme, *Statistics in Medicine*, **20**, 2907–2920.

경시적 영과잉 가산자료와 생존자료의 결합모형

김동욱[a,1] · 천지훈[a]

[a]성균관대학교 통계학과

(2016년 11월 28일 접수, 2016년 12월 13일 수정, 2016년 12월 15일 채택)

요 약

시간의 흐름에 따라 관측되는 경시적(longitudinal) 자료의 경우, 경시적 자료와 생존(survival) 자료가 종종 동시에 수집된다. 이 때 경시적 자료에서 발생하는 결측이 생존자료와의 연관성으로 인해 발생한 무시할 수 없는 결측(non-ignorable missing)이라면, 경시적 자료분석 방법만으로는 두 자료 간의 연관성을 고려하지 않아 독립변수에 대한 효과는 편향된 결과를 얻게 된다. 이러한 문제를 해결하기 위해서 결측의 원인이 생존시간과 연관되어 있으므로 생존모형을 고려하여 불편추정량을 얻기 위해 경시적 자료와 생존자료의 결합모형에 대한 연구가 이루어져 왔다. 본 논문은 경시적 자료의 형태가 영이 많이 존재하는 영과잉 가산자료(zero-inflated count data)와 생존자료의 결합모형을 연구하였다. 경시적 영과잉 가산자료와 생존자료는 각각 허들모형(hurdle model)과 비례위험모형(proportional hazards model)의 부 모형을 적용하였고, 두 부 모형들의 변량효과가 다변량 정규분포를 따른다는 가정을 통하여 결합하였다. 모수의 최우추정법으로 EM 알고리즘을 활용하였고, 추정된 표준오차를 계산하기 위해 프로파일 우도(profile likelihood)를 이용하였다. 최종적으로 모의실험을 통해 두 부 모형의 변량효과 간 상관관계가 존재하는 경우 결합모형이 개별적 모형보다 편의와 포함확률(coverage probability)의 측면에서 더 우수함을 보였다.

주요용어: 결합모형, 경시적 영과잉 가산자료, 허들모형, 생존자료

이 논문은 성균관대학교의 2013학년도 성균학술연구비에 의하여 연구되었음.

[1]교신저자: (03063) 서울시 종로구 성균관로 25-2, 성균관대학교 통계학과. E-mail: dkim@skku.edu