정보적 군집 크기를 가진 군집화된 구간 중도 절단 자료 분석을 위한 결합모형의 적용

김양진^{1,a}, 유한나^b

^a숙명여자대학교 통계학과, ^b고려대학교 통계연구소

요약

구간 중도 절단 자료는 감염 자료, 종양 발생 자료등 그 발생 시간을 정확하게 관측할 수 없는 경우에 흔히 발생하는 자료로 정확한 사건 발생 시간대신에 발생 전 마지막 관측시점과 발생 후 첫 번째 관측시점으로 구성된다. 이러한 종류의 자료는 Sun (2006)에 의해 자세하게 논의되었으며 관측 개체간의 독립성 가정 하에서 여러 가지 방법들에 의해 분석되어져 왔다. 본 논문에서는 관측 개체들이 군집으로부터 발생하여 더 이상 독립성 가정이 적절하지 못한 경우를 고려한다. 특히 반응변수인 사건 발생 시간이 군집의 크기와 연관되어 있을 때, 이를 고려하기 위한 결합 모형을 제시한다. 제안된 모형은 림프계 필라리아병의 실제 자료에 적용한다.

주요용어: 구간 중도 절단 자료, 정보적 군집 크기, 프레일티, EM 알고리즘.

1. 서론

구간 중도 절단자료(interval-censored failure time data) 분석을 위해 관측 치들이 독립이라는 가정 하에서 모수적 또는 비모수적 방법들이 적용되어 왔다 (Sun, 2006). 하지만 관측 개체들이 같은 군집에 속한 경우에는 독립성 가정은 적절하지 못한다. 특히 본 논문에서는 이러한 군집화된 자료에서 군집의 크기가 반응변수와 연관되었을 때, 적합한 통계 모형을 제시하고자 한다. 예를 들어 열대성 질환으로 피부가 심하게 붓는 림프계 팔라리아병(Lymphatic Filariasis)은 기생충이 혈관에 기생하는 병으로 특히 어린이들의 지적 또는 신체적 발달을 저하시키고 시각 장애와 신체 불구를 가져올 수 있으며 심하면 목숨을 잃게 하기도 한다. 기생충은 감염자의 혈관에 여러 개의 둥지(nest)를 만들어 기생하게 되는데 기생충학자들은 이러한 둥지를 소멸시키기 위한 효과적인 치료법을 개발하고자 한다. Williamson 등 (2008)은 48명의 림프계 팔라리아병을 앓고 있는 사람들을 대상으로 새로운 치료법의 효과를 분석하기 위해 통계적 방법을 제안하였다. 그들은 둥지의 수가 많을수록 소멸률이 낮음을 발견하였다. 즉 군집의 크기가 클수록 생존률이 더 높았다. 여기서 반응변수가 소멸률이라고 할 때, 반응변수는 둥지의 수와 연관되어 있으며 이러한 현상을 정보적 군집크기(informative cluster size)라고 한다. 정보적 군집 크기를 가진 군집 자료를 분석하기 위해 여러 방법들이 시도되었다. 먼저 중도절단이 없는 일반적 형태의 반응변수에 대해 Catalano와 Ryan (1992)는 군집 크기를 반응변수의 공변량으로 포함시켰으며 Hoffman 등 (2001)은 각 군집 내 한 개체만을 샘플링하는 군집 내 샘플링의 기법(within cluster sampleing; WCR)을 소개하였다. Williamson 등 (2003)은 이 기법을 일반화 추정 방

본 연구는 (MOEHRD, Basic Research Promotion Fund: KRF-2008-331-C00062)와 고려대 통계연구소에 의해 지원받았습니다.

¹ 교신저자: (140-742) 서울시 용산구 효창원길 52, 숙명여자대학교 통계학과, 조교수.
E-mail: yjin@sookmyung.ac.kr

정식(Generalized Estimating Equation; GEE)에 적용하였다. 중도 절단이 포함된 생존 자료에 대해서 Williamson 등 (2008)은 군집 크기를 가중치로 사용하여 WCR을 적용하였으며 Cong 등 (2007)도 같은 가중치를 사용한 가중 점수 함수(weighted score function) 방법을 사용하였다. 비슷한 문제에 대해 베이지안 방법으로 Dunson 등 (2003)은 결합 모형을 적용하였다. 구간 중도 절단 문제에 대해서, 최근에 Zhang과 Sun (2010)은 Williamson 등 (2008)의 방법을 확장한 가중화된 추정방정식을 제안하였다. 본 연구에서는 정보적 군집 크기를 가진 구간 중도 절단 문제를 위해 군집 크기에 대한 모형과 구간 중도 절단 시간의 모형을 결합시키는 결합모형(joint model)을 고려한다. 이를 위해 이 자료의 세가지 특성을 고려해야 한다. 첫째 특성은 군집화된 자료의 연관성에 관한 것이다. 군집화된 자료 분석을 위해 크게 두 가지 방법이 고려될 수 있다. 연관성을 무시한 주변 모형(marginal)에 근거한 로버스트 방법과 프레이리티(frailty) 효과를 포함한 혼합 모형이 적용될 수 있다. 프레이리티는 랜덤효과의 한 형태로 생존 분석 자료 분석에서 개인별 특성을 표현하기 위해 자주 사용된다. 본 논문에서는 군집 내 연관성이 프레이리티로 표현될 때 군집 내 관측치들 간에는 서로 독립임을 가정한다. 즉, 군집 개체간의 공통된 특징이 프레이리티로 표현될 때, 같은 군집 내 관측 개체들은 더 이상 서로 연관되어 있지 않음을 가정한다. 두 번째 자료 특성인 구간 중도 절단 자료에 대해서는 생존 분석에서 널리 적용되는 비례위험 모형과 가법모형대신 모수 분포인 와이블 분포를 가정한다. 이 가정을 통해 SAS와 같은 통계 패키지의 사용이 용이하게 된다. 세 번째 자료 특성은 정보적 군집 크기에 대한 것이다. 이를 위해 군집 크기와 반응 변수와의 연관성을 위해 두 모형에 같은 랜덤 효과를 사용하여 연결하는 결합 모형을 제시하고자 한다. 2장에서는 연관된 통계모형을 제시하고 3장에서는 추론 방법을 소개하며 4장에서는 실제 자료 분석에 적용할 것이다.

2. 모형

본 연구에서 사용할 자료는 구간 중도 절단 자료로 $(L_{ij}, R_{ij}; j = 1, \dots, m_i, i = 1, \dots, n)$ 이며 i 는 군집에 관한 것이며 j 는 군집 내 관측 개체를 표현하기 위해 사용된다. 여기서 L_{ij} 와 R_{ij} 는 실제 관심있는 i 번째 군집에 속한 j 번째 관측 개체의 생존 시간, T_{ij} 을 포함한 두 관측시점이다. 즉, T_{ij} 는 실제로 측정되는 대신 $L_{ij} < T_{ij} < R_{ij}$ 를 만족하는 두 관측 시점인 L_{ij}, R_{ij} 만이 알려져 있다. 이러한 자료를 구간 중도절단 자료(interval-censored failure times)라 한다. 공변량 X_i 는 i 번째 군집이 가지는 군집별 공변량(cluster-specific covariate)을 Z_{ij} 는 군집 내 관측개체의 공변량(subject-specific covariate)을 표현한다.

첫 번째, 군집 크기에 대한 모형으로 순서형 누적 로짓모형(ordinal cumulative logit model)을 적용한다. 1장에서 예시되었던 림프계 팔라리아병에서 기생충의 등지 수는 많아야 6개를 넘지 않는다. 또한 등지 수는 기생충의 생산력(productivity)과 매우 연관되어 있으므로 생산력과 같이 직접적으로 측정할 수 없는 잠재 변수(latent variable)가 적절한 임계치를 통해 등지수로 관측될 수 있을 것이다. 이러한 관계에 의해 순서형 다항 반응변수(ordinal categorical data)가 등지수에 대한 모형으로 적용되었다. 군집별 효과(cluster-specific effect)를 추정하기 위해 랜덤 효과, u_i 를 포함한 다음의 혼합모형(mixed effect model)이 적용된다.

$$\log \frac{\Pr(m_i \leq k | x_i, u_i)}{1 - \Pr(m_i \leq k | x_i, u_i)} = \log \frac{P_{ik}}{1 - P_{ik}} = \alpha_k + x_i' \gamma + u_i, \quad (2.1)$$

여기서 γ 는 공변량의 효과를 추정하는 것으로 두 공변량 값의 오즈비가 모든 단계에서 동일함을 가정한다. 랜덤 효과, u_i 는 평균 0이고 분산 σ_u^2 을 가지는 정규 분포를 가정한다.

두 번째, 군집화된 구간 중도 절단 자료에 대해 본 논문에서는 프레이리티, v_i 가 주어진 경우, 군집 내 반응 변수들이 조건적 독립(conditionally independent)이라는 가정을 사용할 것이다 (Bellamy 등, 2005). 즉 군집 내 관측 개체들간의 생존 시간에 대한 연관성이 프레이리티로 모형화될 때 생존 시간은

독립이라고 가정한다. 본 연구에서는 프레일티, v_i 는 평균이 0이고 분산이 σ_v^2 인 정규 분포를 따른다고 가정한다. i 번째 군집의 프레일티 v_i 가 주어진 경우, i 번째 군집에 속한 j 번째 관측 개체의 생존 시간은 와이블 분포, (λ, p) 를 따른다고 가정한다. 따라서 u_i 와 v_i 가 주어진 경우에, 위험 함수는

$$\lambda_{ij}(t|u_i, v_i) = p\lambda^p t^{p-1} \exp^{-p(\beta z_{ij} + v_i + \eta u_i)} \quad (2.2)$$

으로 정의된다. 여기서 β 는 공변량의 사건 발생에 대한 효과를 추정하며, 회귀 계수 η 는 군집 크기와 연관성 정도를 측정한다. 즉 프레일티 u_i 를 모형 (2.1)과 (2.2)에 동시에 사용함으로써 두 모형을 결합시키고자 한다. 만약 $\eta > 0$ 이라면 군집 크기가 클수록 그 군집 내 개체들의 생존 시간이 더 길며, $\eta < 0$ 는 작은 규모의 군집에 속한 개체들이 더 오랫동안 생존함을 의미한다. 구간 중도절단 자료에 대해 와이블 분포를 적용하여 다음의 우도함수를 구할 수 있다.

$$L_2(\beta, p, \lambda, \sigma_u^2, \sigma_v^2) = \prod_{i=1}^n \prod_{j=1}^{m_i} \{S_{ij}(L_{ij}|v_i, u_i) - S_{ij}(R_{ij}|v_i, u_i)\}, \quad (2.3)$$

여기서 $S_{ij}(t|v_i, u_i) = \exp\{-(\lambda t)^p e^{-p(\beta' x_{ij} + \eta u_i + v_i)}\}$ 이다.

3. 통계적 추론

EM 알고리즘은 우도 함수가 불완전한 자료를 포함하는 경우에 E-Step 을 통해 그 자료를 복원시킨 후 M-Step 에서 우도 함수를 최대화하는 모수를 구하게 된다. 추정되어야 할 모수가 다음과 같이 구성 된다고 하자 $\theta = (\beta, \eta, \lambda_1, \dots, \lambda_q, \alpha, \gamma, \sigma_u^2, \sigma_v^2) = (\theta_1, \theta_2, \theta_3)$, 여기서 $\theta_1 = (\beta, \eta, \lambda_1, \dots, \lambda_q)$, $\theta_2 = (\alpha, \gamma)$, $\theta_3 = (\sigma_u^2, \sigma_v^2)$. i 번째 군집의 우도 함수는 군집 크기와 군집 내 개체들의 구간 중도 절단의 곱으로 다음과 같이 표현될 것이다.

$$L_i(\theta) = \int_{u_i} \int_{v_i} \prod_{j=1}^{m_i} \{S_{ij}(L_{ij}|u_i, v_i) - S_{ij}(R_{ij}|u_i, v_i) \\ \times \prod_{k=1}^K \left[\frac{1}{1 + \exp[-(x_i \gamma + \sum_{l=1}^k \alpha_l)]} - \frac{1}{1 + \exp[-(x_i \gamma + \sum_{l=1}^{k-1} \alpha_l)]} \right]^{r_{ik}} h_1(u_i) h_2(v_i) du_i dv_i. \quad (3.1)$$

또한 $h_1 \sim N(0, \sigma_u^2)$, $h_2 \sim N(0, \sigma_v^2)$ 이다. 여기서 만약 $m_i = k$ 일 때 $r_{ik} = 1$ 이며 그 외 경우엔 $r_{ik} = 0$ 이 된다. 위 우도 함수는 두 번의 적분을 포함하는데 이는 닫힌 형태로 표현될 수 없다. 따라서 위 적분을 수행하기 위해 주로 가우시안 구적법(Gaussian quadrature)이나 몬테 칼로 방법이 사용된다. Metropolis-Hasting 알고리즘과 같은 몬테 칼로 마르코브 체인(MCMC) 방법에서는 E-Step에서 미지의 랜덤 변수들을 추정하기 위해 M 개의 랜덤변수를 생성한 후 그들의 평균을 미지의 값의 추정치로 사용하는 MCEM을 적용할 수 있다. 본 논문에서 가우시안 구적법으로 Gauss-hermite 방법을 이용하여 미지의 랜덤효과를 추정할 것이다. 즉, Q_1 개와 Q_2 개의 미리 정해진 $(u_{q_1}, v_{q_2}, q_1 = 1, \dots, Q_1, q_2 = 1, \dots, Q_2)$ 을 이용하여 우도 함수 (2.3)은 다음과 같이 가중치 합으로 표현될 수 있다.

$$L_i = \sum_{q_1=1}^{Q_1} \sum_{q_2=1}^{Q_2} w_{q_1} w_{q_2} \prod_{i=1}^{m_i} S_{ij}(L_{ij}|u_{q_1}, v_{q_2}) - S_{ij}(R_{ij}|u_{q_1}, v_{q_2}) \\ \times \prod_{k=1}^K \left[\frac{1}{1 + \exp[-(x_i \gamma + u_{q_1} + \sum_{l=1}^k \alpha_l)]} - \frac{1}{1 + \exp[-(x_i \gamma + u_{q_1} + \sum_{l=1}^{k-1} \alpha_l)]} \right]^{r_{ik}} h_1(u_{q_1}) h_2(v_{q_2}).$$

표 1: 림프계 팔라리아 기생충병에 대한 세가지 모형의 적용

Parameter	Model 1	Model 2	Model 3
$\beta(\text{group})$	1.488(0.042)	1.332(0.043)	1.0948(0.118)
η	0.851(0.572)	0.629(0.585)	
σ_u	1.688(0.026)	1.630(0.003)	2.029(<0.001)
σ_v	1.532(0.399)	1.713(0.439)	1.184(0.851)
-2Log-Likelihood	377.5	684.4	383.5

위의 가우시안 구적법은 SAS의 PROC NLMIXED을 이용하여 구현화시킬 수 있다. 일단 미지의 랜덤 효과가 추정되면 Newton-Raphson 방법을 적용하여 최대 우도 추정량을 구할 수 있다.

4. 자료 분석

1장에서 예시된 림프계 팔라리아병은 기생충이 혈관에 여러 개의 등지를 만들어 기생하는 질병으로 특히 아프리카, 남미 등 열대 지방에서 자주 발생된다. 이 질병은 초음파 촬영을 통해 등지 수와 기생충의 생존 여부가 검사되며 기생충 소거를 위해 다양한 치료 방법들이 제시되었다. 본 논문에서 분석할 자료는 47명의 감염자로 구성되어 있으며 두 가지 치료방법이 비교되었다. 22명의 환자는 DEC/ALB의 복합 치료를 받는데 반해 25명의 환자는 DEC만을 처방받았다. 치료가 시작된 후 7, 14, 30, 45, 60, 90, 180, 270, 360일에 기생충의 생존 여부가 검사되었다. 따라서 기생충의 정확한 생존 시간을 관측할 수 없으며 대신 두 시점, 기생충이 여전히 생존한 것으로 검사된 마지막 관측 시점과 기생충이 죽은 것을 발견한 첫 번째 검사 시점으로 구성되는 구간 중도 절단 자료를 사용하게 된다. 본 자료의 특이한 점으로 환자 당 등지 수의 범위가 1에서 5이며 등지 수가 클수록 기생충 소거율이 낮아진다는 것이다 (Williamson 등, 2008). 즉 등지가 많은 환자에 속한 기생충의 생존율이 높았다. 따라서 이러한 생존율과 군집 크기와 관계를 고려하기 위해 본 논문에서 제안한 결합 모형을 적용한다. 즉 등지 수의 모형에 대해

$$\log \frac{\Pr(m_i \leq k | u_i)}{1 - \Pr(m_i \leq k | u_i)} = \log \frac{P_{ik}}{1 - P_{ik}} = \alpha_k + u_i, \quad k = 1, 2, 3, 4$$

가 적용되었다. 여기서 등지 수가 5인 환자수가 작아 등지 수가 4와 5인 그룹을 통합하였다. 따라서 등지 수는 다음 네 개의 그룹(1, 2, 3, ≥ 4)으로 나누게 되었다. 생존 시간에 대해서는 다음의 와이블 위험 모형이 적용되었다.

$$\lambda_{ij}(t | u_i, v_i) = p \lambda^p t^{p-1} \exp^{-p(\beta_0 + \beta \text{group}_i + v_i + \eta u_i)}.$$

표 1은 세 가지 모형에 근거하여 구한 모수의 추정 값과 p -value를 보여준다. Model 1은 제안된 모형을 적용한 결과를 Model 2는 구간 중도절단된 자료에 중간치(midpoint)를 적용한 결과를 Model 3는 군집 크기와 반응변수와의 관계를 무시한 모델로 제안된 모형에서 $\eta = 0$ 을 적용한 것이다. 표에서 보여지는 값은 추정값과 해당되는 p -value이다. $-2\log \text{likelihood}$ 값을 통해 Model 1이 제일 작은 값을 가지며 우도비 검정을 이용하여 Model 1, Model 3을 비교할 때 Model 1이 더 적합함을 알 수 있다 ($-2\text{LR} = 6.0 > \chi^2(1, 0.05) = 3.84$). 또한 치료 방법의 차이에 대한 유의성 ($H_0 : \beta = 0$)에서 Model 1과 Model 3은 서로 다른 결과를 보여주었다. 즉 기생충 등지 수와의 관계를 무시한 Model 3에서는 두 처리 방법이 유의적인 차이가 없다는 결과를 보이는 반면에 제안된 모형의 결과에서는 DEC만을 처방받은 환자들이 더 긴 생존 시간을 가짐을 보여준다. Model 1에서 추정된 $\hat{\eta} = 0.851 > 0$ 은 등지 수가 많을수록 생존시간이 길다는 현상과 일치된 결과를 보여준다. 또한 프레일리 ($H_0 : \sigma_v^2 = 0$)와 랜덤효과 ($H_0 : \sigma_u^2 = 0$)에 대한 p -value를 통해 유의성을 검토할 수 있다. 여기서 군집별 효과를 표현

그림 1: 등지크기에 따라 추정된 생존 함수

하기 위해 적용된 랜덤효과(u)의 분산은 유의적인 반면 군집 내 생존 시간과 연관된 프레일티(v)의 분산은 유의적이지 못하였다. 따라서 생존 시간에 대한 군집별 특성은 서로 비슷한 분포를 가진다고 할 수 있다. 그림 1은 등지 그룹별 생존 함수를 추정하기 위해 Turnbull (1976)이 제시한 자기 일치 알고리즘(self-consistency algorithm)을 적용한 예이다.

5. 향후 연구 과제

본 논문에서는 군집 크기가 생존시간과 연관되어 있는 자료를 분석하기 위해 프레일티를 이용한 결합 모형을 제안하였다. 특히 제안된 모형은 SAS와 같은 상용 프로그램에 의해 쉽게 구현화 될 수 있다는 점에서 그 실용성이 매우 높다고 할 수 있다. 제안된 모형은 특히 구간 중도 절단된 생존 자료에 대해 와이블 분포를 가정하여 분석하였다. 이에 대한 확장으로 비례위험 모형의 적용이 고려될 수 있을 것이다. 우리의 다음 연구는 위의 모형을 좀 더 일반화된 회귀모형에 적용하는 것이며 이에 대한 연구는 현재 진행 중이다 (Kim, 2010). 본 논문에서는 구간 중도 절단 변수가 관심 있는 사건 발생시간과 독립이라는 가정 하에 전개되었다. 현재 생존 분석연구에서는 이러한 가정이 적절하지 못한 경우, 즉 반응 변수와 중도 절단자료가 연관된 경우에 대해 많은 연구가 진행되어 오고 있다 (Liu 등, 2008; Finkelstein 등, 2002). 본 논문에서 제안된 방법은 이러한 문제를 또한 포함시킬 수 있으며 이를 위해서 좀 더 복잡한 모형이 제시되어질 필요가 있을 것이다.

6. Appendix

SAS code to fit the joint model for clustered interval censored data with informative cluster size

```
proc nlmixed data=cluster;

parms beta0=1.0 beta1=1.0 gamma=0.4 i1=1 i2=1.0 b0=-3.1 ;
bounds gamma>0, sd1>0, sd2>0, i1>0 , i2>0;

ebetaxb=beta0+(beta1*group+r1);
alpha=exp(-ebetaxb);
s1=exp(-(t1*alpha)**(gamma));
s2=exp(-(t2*alpha)**(gamma));
ss=exp(-(tt*alpha)**(gamma));
f_t=gamma*alpha*((alpha*tt)**(gamma-1))*ss;

eta=b0+r2;
if (clust=1 and q2=1) then p2=1/(1+exp(-(0-eta)));
else if(clust=1 and q2=2) then p2=1/(1+exp(-(i1-eta)))-1/(1+exp(-(0-eta)));
else if(clust=1 and q2=3) then p2=1/(1+exp(-(i1+i2-eta)))-1/(1+exp(-(i1-eta)));
else if(clust=1 and q2=4) then p2=1-1/(1+exp(-(i1+i2-eta)));
else if(clust>1) then p2=1;

if event=1 then lik=s1-s2;
else lik=s1;

llik=log(lik*p2);

model y~general(llik);
random r1 r2~normal([0,0],[ sd1*sd1 , 0, sd2*sd2]) subject=id;
run;
```

참고 문헌

- Bellamy, S., Li, Y., Ryan, L. M., Lipsitz, S., Canner, M. and Wright, R. (2005). Analysis of clustered and interval censored data from a community-based study in asthma, *Statistics in Medicine*, **34**, 3607–3621.
- Catalano, P. and Ryan, L. M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes, *Journal of the American Statistical Association*, **87**, 651–658.
- Cong, X. J., Yin, G. and Shen, Y. (2007). Marginal analysis of correlated failure time data with informative cluster sizes, *Biometrics*, **63**, 663–672.

- Dunson, D. B., Chen, Z. and Harry, J. (2003). A bayesian approach for joint modeling of clusger size and subunit-specific outcome, *Biometrics*, **59**, 521–530.
- Finkelstein, D. M. Goggins, W. B. and Schoenfeld, D. A. (2002). Analysis of failure time data with dependent interval censoring, *Biometrics*, **58**, 298–304.
- Hoffman, E., Sen, P. and Weinberg, C. (2001). Within-cluster resampling, *Biometrika*, **88**, 1121–1134.
- Kim, Y. J. (2010). *Regression Analysis of Clustered Interval-Censored Data with Informative Cluster Size*, Technical Report
- Liu, L., Huang, X. and O’Quigley, J. (2008). Analysis of longitudinal data in the presence of informative observational times and a dependent terminal event, with application to medical cost data, *Biometrics*, **64**, 950–958.
- Sun, J. (2006). *The Statistical Analysis of Interval-censored Failure Time Data*, Springer-Verlag, New-York.
- Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped censored and truncated data, *Journal of the Royal Statistical Society, Series B*, **38**, 290–295.
- Williamson, J., Datta, S. and Satten, G. (2003). Marginal analysis of clustered data when cluster size is informative, *Biometrics*, **59**, 36–42.
- Williamson, J., Kim, H., Manatunga, A. and Addiss, D. (2008). Modeling survival data with informative cluster size, *Statistics in Medicine*, **27**, 543–555.
- Zhang, X. and Sun, J. (2010). Regression analysis of clustered interval-censored failure time data with informative cluster size, *Computational Statistics and Data Analysis*, **54**, 1817–1823.

2010년 7월 접수; 2010년 8월 채택

Statistical Analysis of Clustered Interval-Censored Data with Informative Cluster Size

Yang-Jin Kim^{1,a}, Hanna Yoo^b

^aDepartment of Statistics, Sookmyung Women's University

^bInstitute of Statistics, Korea University

Abstract

Interval-censored data are commonly found in studies of diseases that progress without symptoms, which require clinical evaluation for detection. Several techniques have been suggested with independent assumption. However, the assumption will not be valid if observations come from clusters. Furthermore, when the cluster size relates to response variables, commonly used methods can bring biased results. For example, in a study on lymphatic filariasis, a parasitic disease where worms make several nests in the infected person's lymphatic vessels and reside until adulthood, the response variable of interest is the nest-extinction times. Since the extinction times of nests are checked by repeated ultrasound examinations, exact extinction times are not observed. Instead, data are composed of two examination points: the last examination time with living worms and the first examination time with dead worms. Furthermore, as Williamson *et al.* (2008) pointed out, larger nests show a tendency for low clearance rates. This association has been denoted as an informative cluster size. To analyze the relationship between the numbers of nests and interval-censored nest-extinction times, this study proposes a joint model for the relationship between cluster size and clustered interval-censored failure data.

Keywords: EM algorithm, frailty effect, informative cluster size, interval censored data.

This work was partly supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD, Basic Research Promotion Fund): KRF-2008-331-C00062 (Y-J. Kim) and this work was also partly supported by the Institute of Statistics, Korea University(Hanna Yoo).

¹ Corresponding author: Assistant Professor, Department of Statistics, Sookmyung Women's University 52 Hyochangwon-gil, Yongsan-gu, Seoul 140-742, Korea. E-mail: yjin@sookmyung.ac.kr