Journal of the Korean Data & Information Science Society
2014, **25**(2), 327–336

http://dx.doi.org/10.7465/jkdi.2014.25.2.327
한국데이터정보과학회지

불완전한 관측틈을 가진 재발 사건 소요시간에 대한 자료 분석[†]

신슬비[1] · 김양진[2]

[1]국민건강보험공단 건강보험정책연구원 · [1]숙명여자대학교 통계학과

접수 2013년 12월 30일, 수정 2014년 2월 5일, 게재확정 2014년 2월 20일

요 약

재발 사건 자료란 연구대상이 같은 종류의 사건을 반복적으로 경험할 때 발생하는 자료이다. 이러한 재발 사건은 사회과학, 자연과학, 공학, 의약학 등 다양한 분야에서 나타날 수 있다. 재발 사건 자료를 분석할 때 연구자의 관심에 따라 사건 발생시간이나 사건 발생간의 소요시간을 이용하여 분석할 수 있다. 이 논문에서는 사건 발생시점간의 소요시간을 이용하여 불완전한 관측을 가진 재발 사건 자료를 분석하고자 한다. 이 자료의 특징은 일부 관측대상들이 일정기간 동안 연구에서 제외되는 관측틈을 갖는다는 것이다. 이 때 관측틈은 불완전한 형태로 나타나게 되는데 그 이유는 관측틈의 시작시점은 알고 있지만 종료시점은 알 수 없기 때문이다. 이러한 미지의 종료시점을 추정하기 위해서 구간 중도 절단 방법이 적용된다. 따라서 종료시점이 추정된 후 프레일티를 포함한 회귀모형을 적용하여 공변량이 사건 재발에 미치는 영향을 알아볼 수 있다. 또한 제안한 방법을 실제자료에 적용하여 관측틈을 고려한 경우와 고려하지 않은 경우를 비교하고자 한다.

주요용어: 구간중도절단, 불완전한 관측, 재발사건자료, 프레일티.

1. 서론

재발 사건 자료 (recurrent event data)란 관측대상이 동일한 사건을 여러 번 경험할 때 발생하는 자료로 사회과학, 자연과학, 공학, 의약학 등 다양한 분야에서 나타난다. 그동안 논의 되었던 재발 사건 자료로는 방광암 환자들의 종양 재발, 골수이식을 받은 백혈병 환자들의 감염 발생, 자동차의 잦은 고장, 낭포성 섬유증 환자의 호흡기 악화, 마약 중독자의 재입원 등이 있다. 재발 사건 분석을 위해 여러 가지 접근 방법이 적용되어 왔다 (Cook과 Lawless, 2007; Kelly와 Lim, 2000; Duchateau 등, 2003). Cook과 Lawless (2007)는 재발 사건 자료 분석에 대해 다양한 분석방법과 예제를 소개하였다. Kelly와 Lim (2000)은 어린이 감염 자료에 관한 재발 자료에 대해 두 가지 시간대 (time scale)를 이용하여 분석하였다. 즉, 사건 발생시간 (total time)과 사건 발생 소요시간 (gap time)에 대해 주변 모형 (marginal model), 조건부 모형 (conditional model)과 계수과정 (counting process)방법을 각각 적용한 결과를 비교하였다.

생존분석기법을 기반으로 재발 사건 자료를 분석할 때 관측대상이 가지고 있는 개별 특성을 모형화하기 위해 프레일티 모형 (frailty model)을 고려할 수 있다 (McGilchrist와 Aisbett, 1991; Sahu 등, 1997; Kim, 2010). McGilchrist와 Aisbett (1991)는 로그정규분포를 따르는 프레일티를 고려했으며, Sahu

[†] 이 연구는 2012년도 정부 (교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행한 기초연구사업임 (과제번호: NRF-2012R1A1A13011350).
[1] (121-710) 서울특별시 마포구 공덕동 254-8, 국민건강보험공단 건강보험정책연구원, 연구원.
[2] 교신저자: (140-742) 서울특별시 용산구 청파로47길 100번지, 숙명여자대학교 통계학과, 교수.
E-mail: yjin@sookmyung.ac.kr

등 (1997)은 감마분포를 따르는 프레일티를 적용시켰다. 재발 사건 자료에 대한 적용 예로 Duchateau 등 (2003)은 어린이 천식 예방에 대한 임상 자료에 대해 두 가지 시간대에서 프레일티 모형을 적합 시킨 결과를 비교하였다. Kim (2010)은 시간 가변 프레일티를 고려한 모형을 제안하였다.

이 논문의 목적은 관측틈 (observation gap)을 가지는 재발 자료에 대한 회귀 모형을 적용하는 것이다. 여기서 관측틈이란 어떤 이유로 관측 대상이 재발 사건을 경험할 수 없는 특정 시간대를 의미한다. 즉, 관측틈이 시작되면 관측 대상은 관측 위험그룹 (risk group)에서 제외되었다가 관측틈이 종료되면 다시 위험그룹으로 복귀하게 된다. 관측틈의 한 예로, 류머티즘 환자의 골절 부상이 있다. 골절 부상을 관심 있는 재발 사건이라고 할 때, 만약 어떤 환자가 입원을 하게 된다면 그 입원기간동안 골절일 가능성은 거의 없을 것이다. 따라서 입원 기간은 골절 재발에 대해서 관측틈으로 간주될 수 있다. 하지만 본 연구에서 다룰 자료는 관측틈에 대한 정확한 정보를 알지 못한 경우를 고려한다. 즉, 관측틈의 시작 시점은 알고 있지만 종료시점을 알 수 없는 경우는 관측 위험그룹으로의 정확한 복귀시점을 알 수 없는 것과 같다. 이러한 불완전한 복귀시점은 위험그룹의 소속여부에 영향을 줄 수 있으며 이는 통계 추론에 영향을 줄 수 있다. 따라서 미지의 종료시점을 추정하는 것은 정확한 추론 결과를 뒷받침한다 (Kim, 2014). 본 논문에서는 관측틈의 종료시점을 추정하기 위해 구간 중도 절단 (interval censoring)을 적용한다. 먼저 공변량과의 연관정도에 따라 비모수 방법과 준모수 방법이 각각 적용된다 (Turnbull, 1976; Pan, 1999). 2절에서는 본 논문에서 다룰 재발 사건 자료와 구간 중도 절단 자료에 대해 알아보고, 3절에서는 불완전한 정보를 추정하는 방법에 대해 알아본다. 4절에서는 실제 자료 분석을 위해 몇 가지 모형들을 적용하고 그 결과를 비교해본다. 5절에서는 제안한 방법의 제한점과 향후 연구 방향에 대해 논의하고자 한다.

2. 불완전한 관측틈을 가진 재발 사건 자료 분석

본 논문에서 사용될 여러 가지 기호를 먼저 정의해보면, i번째 관측대상의 j번째 사건 발생 시간을 t_{ij} $(i = 1, \cdots, n;\ j = 1, \cdots, n_i)$로 정의한다. 먼저 관측틈이 없는 경우를 가정한다면 i번째 관측대상의 $j-1$번째 사건과 j번째 사건간의 발생 소요시간을 $w_{ij} = t_{ij} - t_{ij-1}$로 정의할 수 있다. 여기서 관측 시작 시점은 $t_{i0} = 0$이고 c_i를 중도 절단 시점이라고 할 때, 관측대상이 갖는 마지막 시간간격은 $w_{in_i+1} = c_i - t_{in_i}$이다. 본 연구에서는 시간간격에 대한 위험함수로 준모수 모형과 모수 모형을 함께 고려한다.

$$\lambda_{ij}(w_{ij}) = \lambda_0(w_{ij}) \exp(\boldsymbol{\beta}' z_{ij}) \tag{2.1}$$

$$\lambda_{ij}(w_{ij}) = \lambda_0^{\gamma} \gamma w_{ij}^{\gamma-1} \exp(\boldsymbol{\beta}' z_{ij}) \tag{2.2}$$

식 (2.1)은 비례 위험 모형 (proportional hazard model)을 적용했을 때의 위험함수이며 λ_0는 기저위험함수 (baseline hazard function)이고 $\boldsymbol{z}_{ij}$는 시간가변 공변량과 시간고정 공변량 모두를 포함하는 공변량 벡터이다. 식 (2.2)는 와이블 모형을 적용했을 때의 위험함수로 회귀계수 $\boldsymbol{\beta}$와 기저 분포의 형태를 결정하는 모수 $\lambda = (\lambda_0, \gamma)$가 있다. 본 연구에서는 한 관측대상으로부터 얻어진 소요시간간의 연관관계를 모형화하기 위해 프레일티 효과를 적용한다. 예를 들어, 비례 위험 모형에 대해 다음 두 가지 프레일티 모형을 고려할 수 있다.

$$\lambda_{ij}(w_{ij}) = u_i \lambda_0(w_{ij}) \exp(\boldsymbol{\beta}' \boldsymbol{z}_{ij}) \tag{2.3}$$

$$\lambda_{ij}(w_{ij}) = \lambda_0(w_{ij}) \exp(\boldsymbol{\beta}' \boldsymbol{z}_{ij} + v_i) \tag{2.4}$$

여기서 u_i (또는 v_i)는 프레일티 효과 (frailty effect)로, 식 (2.3)에서 u_i는 양수 값을, 식 (2.4)에서 v_i는 $(-\infty,\ \infty)$ 범위 내의 값을 가지게 된다. 여기서, u_i (또는 v_i)들은 서로 독립이며 동일한 분포를 따른다

고 가정한다. 특히, 식 (2.3)에서 u_i는 평균이 1이고 분산이 θ인 감마분포를, 식 (2.4)에서 v_i는 평균이 0이고 분산이 σ^2인 정규분포를 따른다고 가정한다. 즉,

$$u_i \sim Gamma\left(\frac{1}{\theta},\ \theta\right) \text{ for } \theta = \sigma^2$$

$$v_i \sim Normal(0, \sigma^2)$$

여기서, 분산을 통해 관측개체간의 이질성 (heterogeneity)을 파악할 수 있다. 프레일티를 포함한 자료에 대한 모수 추정을 위해서는 Nielsen 등 (1992), Klein과 Moeschberger (1997), Kim (2013)을 참고하기 바란다.

이제 우리의 관심대상인 관측틈을 포함한 재발 자료에 대해 고려해본다.

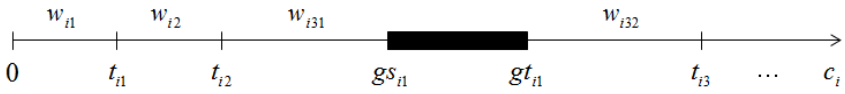

Figure 2.1 Observation gap $[gs, gt]$ at recurrent event data

Figure 2.1에서처럼 두 시점 $t_{ij}-1$와 t_{ij}사이에 관측틈을 포함하고 있다고 하자. 관측틈의 존재 여부를 나타내는 지시변수 q_{ij}와 관측틈의 시간 구간을 나타내는 $\{(gs_{ik}, gt_{ik}), i = 1, \cdots, n, k = 1, \cdots, m_i\}$, 사건 발생 여부를 나타내는 지시변수 δ_{ij}가 필요하다. 여기서 gs_{ik}는 관측틈의 시작시점이며 gt_{ik}는 종료시점이다. 관측틈이 발생한다면, $q_{ij} = 1$이고 j번째 사건의 시간간격을 다음과 같이 두 부분으로 나누어 재 정의할 수 있다.

$$\{(w_{ij1}, \delta_{ij1}), (w_{ij2}, \delta_{ij2})\}, \quad where \begin{cases} w_{ij1} = gs_{ik} - t_{ij} - 1, \delta_{ij1} = 0 \\ w_{ij2} = t_{ij} - gt_{ik}, \delta_{ij2} = 1 \end{cases}$$

여기서 관측틈이 발생한 경우, $(\delta_{ij1}, \delta_{ij2})$에 대해서는 첫 번째 사건은 관측틈에 대한 재발 사건의 중도 절단이므로 $\delta_{ij1} = 0$이 되며 두 번째 사건은 관측틈이 종료된 후 일어나는 재발 사건이므로 $\delta_{ij2} = 1$의 정의를 가진다. 만일 연속적으로 발생하는 사건 사이에 관측틈이 발생하지 않는다면, $q_{ij} = 0$이고 $w_{ij1} = t_{ij} - t_{ij} - 1$,$\delta_{ij1} = 1$이다. 여기서 관측틈의 발생은 재발 사건의 분포와 독립이며 중도 절단 시점과도 독립임을 가정한다.

이때 식 (2.1) 또는 식 (2.2)의 위험함수와 $f(w_{ij}) = \lambda(w_{ij})S(w_{ij})$의 관계를 이용하여 다음의 우도함수 (likelihood function)를 유도할 수 있다.

$$L(\phi) = \prod_{i=1}^{n} \prod_{j=1}^{n_i} \prod_{q=1}^{q_{ij}+1} f(w_{ijq}|\boldsymbol{z}_{ij}, u_i)^{\delta_{ijq}} S(w_{ijq}|\boldsymbol{z}_{ij}, u_i)^{(1-\delta_{ijq})} S(w_{in_{i+1}}|\boldsymbol{z}_{in_i+1}, u_i) \tag{2.5}$$

여기서 $\boldsymbol{\phi} = (\beta, \lambda, \sigma^2)$는 추정해야할 관심 있는 모수 벡터이다. 우도함수 (2.5)에서는 지시변수 q_{ij}을 이용함으로써 관측틈에 대한 정보를 반영하게 된다. 만약 완전한 관측틈 정보를 가지고 있는 경우에는 위 우도함수를 최대화함으로써 모수를 추정할 수 있다. EM 알고리즘의 적용 과정은 가장 많이 적용되는 방법이며 프레일티가 감마 분포와 정규 분포를 따르는지 여부에 따라 여러 가지 계산 알고리즘이 적용될 수 있다 (Klein과 Moeschberger, 1997). 본 논문에서는 R 프로그램의 coxph 명령문을 사용하였다 (coxph(Surv(w,cen) x1+x2+frailty(id,dist=" ")) 또는 survreg (Surv(w,cen) x1+x2+frailty(id,dist=" "))). 이 명령문을 통해 구한 회귀 계수와 분산의 추정치는 Therneau 등 (2000)이 제안한 벌점 가능도 (penalized likelihood) 방법을 적용한 결과이다. 또한 Sahu 등 (1997)은 와이블 모수 모형에 대한 정

규 분포를 따르는 프레일티의 적용을 고려하였다. 프레일티의 추론에 대한 또 다른 방법으로는 계층적 (hierarchical) 우도함수의 적용을 고려할 수 있을 것이다 (Ha와 Cho, 2012; Ha와 Noh, 2013). 특히 최근에 개발된 frailtyHL R-패키지는 매우 유용한 프로그램을 제공한다. 본 연구에서 분석할 자료는 관측틈의 정보가 불완전한 자료로 위에서 언급한 R프로그램의 명령문을 적용하기 전에 관측틈의 종료시간을 추정할 필요가 있다. 다음 절에서는 관측틈의 종료시점 gt_{ik}를 추정하는 방법에 대해 알아보도록 한다.

3. 관측틈의 종료시간에 대한 추정

이 연구에서는 관측틈의 종료시점 gt_{ik}는 재발 사건의 분포와 독립이라고 가정한다. 종료시점을 추정하기 위해 Figure 3.1과 같이 $g_{ik} = gt_{ik} - t_{ij} - 1$를 정의한다. 즉, g_{ik}는 바로 이전 사건 발생 시간 $t_{ij} - 1$부터 관측틈의 종료시점 gt_{ik}까지의 시간간격을 나타낸다. 종료시점 g_{ik}를 이산 확률 변수 (discrete random variable)라고 가정한다면 관련된 확률 질량 함수는 다음과 같이 정의된다.

$$f_{ik}(g_{ik}) = \Pr(g_{ik} = w_l), \quad i = 1, \cdots, n; \; k = 1, \cdots m_i; \; l = 1, \cdots, m.$$

위 확률 분포를 추정하기 위해 관측틈의 발생위치에 따라 g_{ik}는 구간 중도 절단 시간 또는 우 중도 절단 시간으로 간주한다.

먼저 구간 중도 절단으로 고려될 때에는 관측틈이 재발 사건 사이에서 발생하는 경우이며 $g_{ik} \in (tl_{ik}, tr_{ik}) = (gs_{ik} - t_{ij} - 1, t_{ij} - t_{ij} - 1)$로 나타낼 수 있다. 우 중도 절단으로 고려될 경우는 마지막으로 관측된 사건 이후에 관측틈이 발견되었으나 다음 사건이 발생하기 전에 연구가 종료된 경우이다. 이 때 $g_{ik} \in (tl_{ik}, \infty) = (gs_{ik} - t_{ij} - 1, \infty)$이다. 그러나 본 연구에서는 관측틈이 연구 종료 전에 일어났다고 가정하였으며 따라서 모든 g_{ik}는 구간 중도 절단되었다고 가정한다. g_{ik}의 분포를 추정하기 위해 공변량과 관측틈의 종료시간간의 관계에 대한 두 가지 가정이 적용된다.

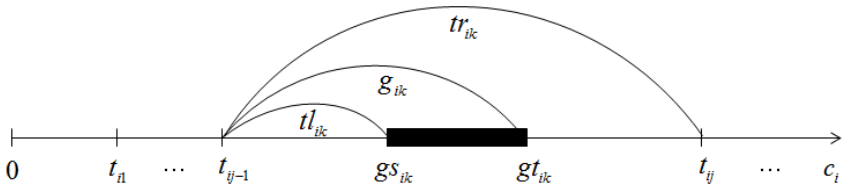

Figure 3.1 Redefinition of termination time of observation gap

3.1. 공변량과 종료 소요 시간이 독립인 경우

먼저 재 정의된 (redefined) 종료 소요 시간 g_{ik}가 공변량 $\boldsymbol{z}_{ij}$와 독립이라고 가정하는 경우이다. g_{ik}들은 서로 독립이고 동일한 확률분포를 가진다고 가정한다. 따라서,

$$f_{ik}(g_{ik}) = \Pr(g_{ik} = w_l) = f_l(w_l) = \Pr(tl_{ik} \le w_l \le tr_{ik})$$

와 같이 표현될 수 있으며, 즉, f_l을 추정하는 것으로 충분하게 된다. f_l을 추정하기 위해 구간 중도 절단 자료의 분포를 추정할 때 자주 사용되는 Turnbull (1976)의 자기 일치 알고리즘 (self consistency algorithm)을 적용할 수 있다. 본 연구에서는 R에서 제공하는 interval 패키지의 icfit함수를 사용하였다. icfit함수를 사용하면 위와 같은 방법으로 $\{w_l,\ l = 1, \cdots, m\}$와 해당되는 발생 확률 분포

$\left\{\widehat{f}_l, l = 1, \cdots, m\right\}$를 추정할 수 있다. 이에 추정된 분포함수 $\widehat{f}_l$로 g_{ik}의 평균을 추정한다.

$$\widehat{E}(g_{ik}) = \frac{\sum_{l=1}^{m} \alpha_{ikl} \widehat{f}_l(w_l) w_l}{\sum_{l=1}^{m} \alpha_{ikl} \widehat{f}_l(w_l)} \quad \text{for} \quad \alpha_{ik} = I(tl_{ik} \le w_l \le tr_{ik})$$

여기서 α_{ik}는 w_l이 구간 (tl_{ik}, tr_{ik}) 내에 포함되는지 여부를 나타내는 지시함수이다. 추정된 $\widehat{E}(g_{ik})$를 이용하여 gt_{ik}를 다음과 같이 추정한다.

$$\widehat{E}(gt_{ik}) = \widehat{E}(g_{ik}) + t_{ij-1} \tag{3.1}$$

따라서 $\widehat{E}(gt_{ik})$는 미지의 관측틈의 종료시점 gt_{ik}의 추정값으로 사용한다.

3.2. 공변량과 종료 소요 시간에 연관관계가 있는 경우

관측틈의 종료 소요 시간 g_{ik}와 공변량 $\boldsymbol{z}_{ik}$의 관계를 모형화하기 위해 비례위험모형을 적용한다. 여기서 개인 별 종료시간은 공변량이 주어져 있을 때 조건부 독립이라고 가정한다.

$$\gamma(g_{ik}; \boldsymbol{\eta}) = \gamma_0(g_{ik}) \exp(\boldsymbol{\eta}' \boldsymbol{z}_{ik}) \tag{3.2}$$

구간 중도 절단 자료 g_{ik}에 대한 모형의 회귀계수 $(\boldsymbol{\eta}, \boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_s))$를 추정하기 위해 R에서 제공하는 intcox 패키지의 intcox함수를 사용한다. Pan (1999)이 제시한 intcox함수는 ICM (iterative convex minorant)을 이용하여 구간 중도 절단 자료에 대해 비례 위험 모형의 회귀 계수를 추정한다. 여기서 구한 결과를 가지고 식 (3.3)과 같이 확률 질량 함수를 추정한다.

$$\widehat{f}_{ikr} = \widehat{f}_r(\boldsymbol{z}_{ik}; \boldsymbol{\eta}, \boldsymbol{\gamma}) = \frac{p_{ikr}}{\sum_{tl_{ik} \le w_l \le tr_{ik}} p_{ikl}}, \; tl_{ik} \le w_r \le tr_{ik} \tag{3.3}$$

여기서 비례 위험 모형인 p_{ikl}은

$$p_{ikl} = \gamma_l \exp[\boldsymbol{\eta}' \boldsymbol{z}_{ik} - \exp(\boldsymbol{\eta}' \boldsymbol{z}_{ik}) \sum_{r=1}^{l} \gamma_r] \tag{3.4}$$

이다. 분포함수가 추정되면 3.1절에서와 유사한 방법으로 종료시점을 다음과 같이 추정할 수 있다.

$$\widehat{E}(g_{ik}|\boldsymbol{z}_{ik}) = \frac{\sum_{r=1}^{s} \alpha_{ir} \widehat{f}_{ikr} w_r}{\sum_{r=1}^{s} \alpha_{ir} \widehat{f}_{ikl}} \quad \text{for } \alpha_{ir} = I(tl_{ik} \le w_r \le tr_{ik})$$

그런 후 추정된 $\widehat{E}(g_{ik}|\boldsymbol{z}_{ik})$를 이용하여 관측틈의 종료시점 gt_{ik}를 추정한다.

$$\widehat{E}(gt_{ik}|\boldsymbol{z}_{ik}) = \widehat{E}(g_{ik}|\boldsymbol{z}_{ik}) + t_{ij-1} \tag{3.5}$$

식 (3.1)과 식 (3.5)를 통해 추정된 관측틈의 종료시간 $\widehat{E}(gt_{ik}|\boldsymbol{z}_{ik})$ (또는 $\widehat{E}(gt_{ik})$)을 계산하여 완전한 관측틈을 구할 수 있다. 이제 4절의 YTOP 자료를 분석하기 위해 2절과 3절에서 언급한 두 가지 위험함수와 두 가지 프레일티 분포가 적용된다.

4. 실제 자료 분석

본 절에서는 앞에서 제시된 방법을 실제 자료에 적용한다. 분석에 사용된 YTOP (Young Traffic Offenders Program)자료는 미국 Missouri 지역에서 운전면허를 취득한 193명의 연령 16∼23세의 젊은 운전자들에 대한 속도위반 기록으로 Missouri State Traffic Violation 자료에서 제공되었다 (Sun 등, 2001). 연구대상은 교통법규를 위반한 젊은 운전자들이며, 관심변수는 운전면허 취득일부터 1995년 7월까지 속도위반을 포함한 법규를 위반한 날짜이다. 전체 193명의 운전자 중 남성운전자는 139명, 여성운전자는 54명이었다. 또한 이들 중에 98명을 랜덤하게 선택하여 일일 교육프로그램인 YTOP를 실시하였다. YTOP는 교통법규위반방지 교육프로그램으로 규정 속도 20mph를 위반했을 때 보호관찰의 일부로서 운전자들에게 참여하도록 시행되었다. 이 프로그램은 운전자들에게 교통사고 생존자들과 인터뷰를 시행하고 여러 가지 비디오를 보여주면서 부적절한 운전습관의 결과를 알리고자 하였다. 운전자들의 속도위반 재발횟수범위는 1∼12회이며 개인당 평균 3.14번의 재발을 경험하였다. 본 분석에서는 YTOP 참여여부 (ytop; 1=참여, 0=참여하지 않음)와 운전자의 성별 (male; 1=남성, 0=여성)이 법규위반 발생 소요시간에 미치는 영향력을 추정한다.

이 자료의 흥미로운 점은 일부 운전자들에게서 불완전한 관측틈이 나타난다는 것이다. 즉, 193명 중 40명은 면허정지를 받게 되며 이 기간 동안 그들은 운전을 할 수 없게 되므로 면허정지 기간이 관측틈이 된다. 특히, 면허정지를 받은 시점에 대한 자료는 제공된 반면에 정지가 해제된 시점은 자료셋에 포함되어있지 않았다. 따라서 면허정지가 해제된 시점을 추정하기 위해 구간 중도 절단 자료 포맷이 적용되었다. 구간 중도 절단된 종료시점의 추정을 위해 앞에서 소개한 두 가지 방법을 적용한다.

4.1. 공변량과 종료시점이 독립인 경우

Table 4.1은 YTOP 프로그램 참여여부와 성별이 면허정지 해제시점 g_{ik}와 독립임을 가정할 때 비례위험 모형과 와이블 위험 모형을 적합 시킨 결과이다. 먼저 비례위험 모형을 적용한 결과를 살펴보자. 프레일티가 감마분포를 따른다고 가정할 때 YTOP의 회귀계수는-0.492 (s.e.=0.115, p-value<0.001)이며, 성별 효과 (male=1)는 0.255 (s.e.=0.123, p-value=0.939)로 추정된다. 즉, YTOP 교육을 받은 운전자의 법규위반 위험률이 교육을 받지 않은 운전자의 위험률의 exp(-0.492)=0.611이며, 남성운전자의 경우 여성운전자일 때보다 법규를 위반할 위험률이 29% (exp(0.255)=1.290) 높은 것으로 나타났다. 프레일티의 분산 추정값 $\hat{\sigma}^2$=0.202 (p-value<0.001)으로부터 개별운전자의 속도위반 성향이 서로 다름을 확인할 수 있다. 만약 프레일티가 정규분포를 가정할 때, YTOP 참여 효과는-0.488 (s.e.=0.115, p-value<0.001)이며 성별 효과는 0.235 (s.e.=0.115, p-value=0.054)로 추정된다. 프레일티의 분산 추정값 $\hat{\sigma}^2$은 0.189 (p-value<0.001)로 개별운전자의 속도위반 성향이 서로 다른 것으로 나타났다. 두 프레일티의 적용 결과는 성별 효과의 유의성을 제외하고 유사하였다. 와이블 모형을 적합 시킨 결과는 비례위험 모형을 적합 시킨 결과와 매우 유사함을 확인할 수 있다. 와이블 모형은 R 프로그램에서 survreg 함수를 적용하고 그 결과 회귀 계수는 AFT (accelerated failure time)모형 하에서 회귀 계수를 해석하는 방법을 사용한다. 즉, 비례 위험 모형에서 회귀 계수는 위험률에 근거한 반면에 와이블 모형에서는 소요시간의 가속 또는 감속에 대한 영향력을 추정하게 된다. 따라서 와이블 모형 하에서 YTOP 참여와 성별의 효과에 대한 추정값은 비례 위험 모형 하에서 추정한 회귀계수와 반대 부호를 가지게 된다. 와이블 모형이 적용되었을 때, YTOP 참여는 법규 위반 재발 소요시간을 연장시키는데 유의한 효과를 보이며 남성운전자에게서 재발 소요시간이 짧아짐을 알 수 있다.

Table 4.1 Termination is independent with covariate

frailty	gamma distribution			normal distribution		
	proportional hazard model					
Covariates	Estimate	S.E.	p-value	Estimate	S.E.	p-value
YTOP	-0.492	0.115	<0.001	-0.488	0.115	<0.001
male	0.255	0.123	0.039	0.235	0.122	0.054
σ^2	0.202		<0.001	0.189		<0.001
	Weibull model					
YTOP	0.494	0.113	<0.001	0.492	0.113	<0.001
male	-0.227	0.120	0.058	-0.247	0.118	0.054
σ^2	0.218			0.177		

4.2. 공변량과 종료시점 간에 연관관계가 있는 경우

공변량 $\boldsymbol{z}_{ik}$가 면허정지가 종료된 시점 g_{ik}와 연관관계가 있을 때 식 (3.2)를 적용하여 추정한다. R에서 제공하는 intcox 패키지의 intcox함수를 이용하여 얻은 $\boldsymbol{\eta}$의 추정값은 $(\hat{\eta}_1, \hat{\eta}_2)$=($0.9472, 0.3586$)이다. 식 (3.3)∼식 (3.5)를 이용하여 종료시점을 추정한다. Table 4.2의 결과는 Table 4.1의 결과와 매우 유사함을 확인할 수 있다.

Table 4.2 Termination is dependent on covariate

frailty	gamma distribution			normal distribution		
	proportional hazard model					
Covariates	Estimate	S.E.	p-value	Estimate	S.E.	p-value
YTOP	-0.500	0.115	<0.001	-0.496	0.115	<0.001
male	0.250	0.123	0.042	0.231	0.122	0.058
σ^2	0.200		<0.001	0.188		<0.001
	Weibull model					
YTOP	0.508	0.115	<0.001	0.506	0.115	<0.001
male	-0.226	0.120	0.059	-0.244	0.119	0.040
σ^2	0.206			0.171		

4.3. 모형 비교

Kim (2014)은 관측틈이 무시된 경우 회귀 계수가 편의된 결과를 모의실험을 통해 보였다. 본 논문에서는 이 결과를 근거로 관측틈이 회귀 계수에 미치는 영향력을 검토하기 위해 비례 위험 모형과 와이블 모형을 적용하였다. 모형Ⅰ은 관측틈을 고려하지 않았을 때의 모형으로 면허정지기간이 위험그룹에 포함된 경우를 의미한다. 모형Ⅱ는 공변량 YTOP 프로그램과 성별이 면허정지 종료시점 gt_{ik}와 독립임을 가정할 때 비모수 방법으로 gt_{ik}을 추정하여 얻은 모형이다. 모형Ⅲ은 공변량과 면허정지 종료시점 gt_{ik}사이에 연관관계가 있을 때 공변량에 의해 추정된 gt_{ik}를 이용한 모형이다.

Table 4.3 Comparison of model I, II and III

proportional hazard model									
model	model I			model Ⅱ			model Ⅲ		
	β	s.e.	p-value	β	s.e.	p-value	β	s.e.	p-value
frailty	gamma distribution								
YTOP	-0.556	0.114	<.001	-0.492	0.115	<.001	-0.500	0.115	<.001
male	0.272	0.118	0.021	0.255	0.123	0.039	0.250	0.123	0.042
frailty	normal distribution								
YTOP	-0.562	0.114	<.001	-0.488	0.115	<.001	-0.496	0.115	<.001
male	0.260	0.117	0.026	0.235	0.122	0.054	0.231	0.122	0.058
Weibull model									
model	model I			model Ⅱ			model Ⅲ		
	β	s.e.	p-value	β	s.e.	p-value	β	s.e.	p-value
frailty	gamma distribution								
YTOP	0.576	0.111	<0.001	0.494	0.113	<0.001	0.508	0.115	<0.001
male	-0.239	0.119	0.044	-0.227	0.120	0.058	-0.226	0.120	0.059
frailty	normal distribution								
YTOP	0.559	0.112	<0.001	0.492	0.113	<0.001	0.506	0.115	<0.001
male	-0.259	0.113	0.022	-0.247	0.118	0.054	-0.244	0.119	0.040

감마분포와 정규분포 하에서 회귀계수의 절대치를 보면 관측틈을 고려하지 않았을 때 (모형Ⅰ)의 추정량은 (-0.556, 0.272)로 관측틈을 고려할 때 (모형Ⅱ,Ⅲ)에 비해 편의된 결과를 보여준다. 즉, 관측틈이 무시될 경우, 관측틈이 소요시간에 포함되어 소요시간의 연장을 가져오며 이는 비례 위험 모형에서 위험률의 감소를 와이블 모형에서는 소요시간의 감속 (deceleration)을 가져오게 된다.

5. 결론 및 토의

관측대상이 동일한 사건을 여러 번 경험하는 경우 이에 대한 적절한 분석방법이 요구된다. 본 연구에서는 사건 발생시점간의 소요시간 (gap time)을 이용하는 방법을 적용하였다. 또한 이 중 일부 관측대상들은 일정기간 관측에서 제외되는 관측틈을 갖는데 이 관측틈은 종료시점을 알 수 없는 불완전한 형태로 나타나기도 한다. 그러므로 미지의 종료시점에 대한 추정이 필요한데 이를 위해 구간 중도 절단 자료 모형이 적용되었으며 종료시점과 공변량과의 관계 여부에 따라 두 가지 방법을 고려하였다. 추정된 종료시점을 이용하여 재발 사건 자료를 분석하였으며 관측대상 내 재발 사건간의 연관성을 알아보기위해 프레일티 효과를 모형에 추가하였다.

제안된 방법은 실제자료인 YTOP 자료에 적용해 보았다. 분석결과, YTOP 프로그램을 통해 운전자의 법규위반 재발률이 줄어듦 (또는 법규위반 재발 소요시간이 연장됨)을 확인할 수 있었고, 남성운전자에 비해 여성운전자의 법규위반 재발률이 낮다 (또는 재발 소요시간이 길다)는 것을 알 수 있었다. 또한 프레일티의 분산 추정을 통해 개별 운전자마다 속도위반 성향이 다르게 나타난다는 것을 확인하였다.

논문에서 제안된 연구와 관련하여 향후 세 가지 연구방향을 다뤄볼 수 있다. 먼저 발생시점간의 소요시간 (gap time)이 아닌 재발 발생시점 (total time)을 이용하는 것이다. YTOP자료에서 반응변수로 발생시점을 고려할 때 성별과 YTOP 프로그램 참여여부가 법규위반에 미치는 영향에 대해서 확인해 볼 수 있다. 다음으로 이변량 재발 사건 (bivariate recurrent event)으로 확장하는 것이다. 본 논문에서는 한 가지 재발 사건만을 고려했으나 위에서 제시된 방법은 두 가지 재발 사건을 가진 경우에도 확장될 수 있다. 또 다른 향후 연구는 본 연구에서 제시된 다양한 가정에 대한 타당성 여부를 검토하는 것이다. 일반적으로 면허정지 기간은 법규위반과 밀접한 관계가 있으므로 본 연구에서 가정한 두 사건간의 독립성은 적절하지 못할 수 있다. 따라서 이를 해결하기 위해 두 변수간의 연관성을 고려할 수 있다. 이룰 위해 두 가지 방법이 고려된다. 첫째, 공통된 공변량을 통해 결합 모형 (joint model)을 고려할 수 있다. 즉, 조건부 독립을 가정하는 것이다. 둘째, 법규위반 횟수를 면허 정지 기간에 대한 회귀 모형의 공변량으로 사용하는 것도 고려할 수 있을 것이다.

References

Cook, J. and Lawless, J. F. (2007). *The statistical analysis of recurrent events*, Springer, New York.

Duchateau, L., Janssen, P., Kezic, I. and Fortpied, C. (2003). Evolution of recurrent asthma event rate over time in frailty models. *Applied Statistics*, **52**, 355-363.

Ha, I. D. and Cho, G. H. (2012). H-likelihood approach for variable selection in gamma frailty models. *Journal of the Korean Data & Information Science Society*, **23**, 190-207.

Ha, I. D. and Noh, M. (2013). A visualizing method for investigating individual frailties using frailtyHL R-package. *Journal of the Korean Data & Information Science Society*, **24**, 931-940.

Kelly, P. J. and Lim, L. (2000). Survival analysis for recurrent event data: An application to childhood infectious diseases. *Statistics in Medicine*, **19**, 13-33.

Kim, Y. (2013). *Survival analysis*, Free academy, Seoul.

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

Kim, Y. (2014). Regression analysis of recurrent events data with incomplete observation gaps. *Journal of Applied Statistics*, in press.

Klein, J. P. and Moeschberger, M. L. (1997). *Survival analysis: Techniques for censored and truncated data*, Springer, New York.

McGilchrist, C. A. and Aisbertt, C. W. (1991). Regression with frailty in survival analysis. *Biometrics*, **47**, 461-466.

Nielsen, G. G., Gill, R. D., Andersen, P. K. and Sorensen, T. I. A. (1992). A counting process approach to maximum likelihood estimator in frailty models. *Scandinavian Journal of Statistics*, **19**, 25-43.

Pan, W. (1999). Extending the Iterative convex minorant algorithm to the Cox model for interval-censored data. *Journal of Computational and Graphical Statistics*, **8**, 109-120.

Ripatti, S. and Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, **56**, 101-1022.

Sahu, S. K., Dey, D. K., Aslanidou, H. and Sinha, D. (1997). A Weibull regression model with gamma frailties for multivariate survival data. *Lifetime Data Analysis*, **3**, 123-137.

Sun, J., Kim, Y. J., Hewett, J., Johnson, J. C., Farmer, J. and Gibler, M. (2001). Evaluation of traffic injury prevention programs using counting process approaches. *Journal of the American Statistical Association*, **96**, 469-475.

Therneau, T., Grambsch, P. and Pankratz, V. (2003). Penalized survival models and frailty. *Journal of Computational and Graphical Statistics*, **12**, 156-175.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped censored and truncated data. *Journal of the Royal Statistical Society B*, **38**, 290-295.

Journal of the Korean Data & Information Science Society
2014, **25**(2), 327–336

http://dx.doi.org/10.7465/jkdi.2014.25.2.327
한국데이터정보과학회지

Statistical analysis of recurrent gap time events with incomplete observation gaps[†]

Seul Bi Shin[1] · Yang Jin Kim[2]

[1]Health Insurance Policy Research Institute, National Health Insurance Corporation
[2]Department of Statistics, Sookmyung Women's University

Received 30 December 2013, revised 5 February 2014, accepted 20 February 2014

Abstract

Recurrent event data occurs when a subject experiences same type of event repeatedly and is found in various areas such as the social sciences, Economics, medicine and public health. To analyze recurrent event data either a total time or a gap time is adopted according to research interest. In this paper, we analyze recurrent event data with incomplete observation gap using a gap time scale. That is, some subjects leave temporarily from a study and return after a while. But it is not available when the observation gaps terminate. We adopt an interval censoring mechanism for estimating the termination time. Furthermore, to model the association among gap times of a subject, a frailty effect is incorporated into a model. Programs included in Survival package of R program are implemented to estimate the covariate effect as well as the variance of frailty effect. YTOP (Young Traffic Offenders Program) data is analyzed with both proportional hazard model and a weibull regression model.

Keywords: Frailty effect, incomplete observation, interval censoring, recurrent event data.

[†] This work was supported by the Korea Research Foundation (MOEHRD, Basic Research Promotion Fund) (No. NRF-2012R1A1A13011350).
[1] Researcher, Institute for Health Insurance Policy Research, National Health Insurance Corporation, Seoul 121-710, Korea.
[2] Corresponding author: Professor, Department of Statistics, Sookmyung Women's University, Seoul 140-742, Korea. E-mail: yjin@sookmyung.ac.kr