응용통계연구 (2010)
**23**(4), 715–724

# 프레일티를 이용한 재범 자료의 연구

김양진[1]

[1]숙명여자대학교 통계학과

(2010년 4월 접수, 2010년 7월 채택)

**요 약**

재발 사건 자료(recurrent event data)는 연구 대상이 같은 종류의 사건을 여러 번 경험할 때 발생되는 자료 형태이다. 재발 사건간에 연관관계를 위해 프레일티가 사용된다. 프레일티 효과는 랜덤효과의 한 형태로 개인별 특성을 표현하기 위해 생존 분석에서 널리 적용되어 왔다. 본 논문에서는 이러한 개인별 효과가 시간에 따라 변할 수 있음을 가정하여 시간 가변 프레일티를 적용한다. 본 논문에서는 적용 사례로 범죄 재범 자료를 분석한다. 특히 일부 관측 대상들은 일정 기간 동안 연구에서 제외되는 불연속성을 경험하게 되며 이는 위험그룹(risk group)의 새로운 정의가 필요하다. 모수 추정을 위해 조각 상수 위험 함수가 사용되며 EM 알고리즘이 적용된다.

주요용어: 재범률, 이변량 프레일티 효과, 불연속 관측, EM 알고리즘, Gauss-Hermite 알고리즘.

## 1. 서론

현대의 편리한 문명은 눈부신 과학발전에 더불어 나날이 발전되었으며 그러한 물질적인 풍요로움을 더 많은 인류가 누리기 위해 많은 사회적 제도가 모색되고 있다. 그럼에도 불구하고 현 사회에는 피할 수 없는 사회적 불안이 과거에 비해 더욱 더 커져만 가고 있다. 예를 들어 외래로부터의 공격과 전염병의 확산이라는 과거 시대의 사회적 불안과는 달리 현재의 불안함은 빈부 격차와 같은 사회적 불평등 문제, 복잡한 개인 문제는 과거에는 존재하지 않았던 여러 가지 유형의 범죄를 유도하게 되었다. 이러한 범죄는 경제, 교육 등에 많은 영향을 주고 있다. 예를 들어 2007∼2008년에 급속하게 증가한 성범죄는 전자팔찌 도입, 호신 도구 판매 급증, 경비업체 호황을 가져왔다. 특히 국민들의 공공안전에 대한 위기감이 고조되었으며, 검찰, 경찰 등 정부에 대한 불신감도 증가하게 되었다. 여기서 우리가 우려하는 가장 큰 문제점은 이러한 범죄의 대다수가 상습적이라는 사실이다. 최근 연구에 의하면 국내의 재범률은 꾸준히 늘고 있다. 더욱 심각한 것은 강력 범죄의 경우 증가율이 이보다 더욱 높다는 것이다. 이러한 높은 재범률은 교정 정책에 대한 의문을 초래하게 되었으며 적절한 형사정책 및 사회제도의 시행 및 수정의 필요성을 제시한다 (주희종, 2001). 이를 위해 전반전인 재범에 관한 면밀한 연구가 필요하다. 본 연구의 목적은 두 번 이상의 법규위반자를 대상으로 그들의 반복적인 법규 위반의 분포 형태와 그 재범에 연관된 요인을 알아내기 위한 방법론을 개발하고자 한다. 특히 범죄률은 개인의 유전적, 환경적 배경에 의해 많은 영향을 받으므로 이러한 개인적 특성을 추정하기 위해 프레일티(frailty)효과를 적용한다. 특히 개인적 특성의 시간적 가변성을 조사하기 위해 시간 가변(time-varying) 프레일티를 고려할 것이다.

본 논문에서는 특정 집단의 재범(recidivism)자료에 대해 위 목적들을 달성하기 위해 사건 역사 분석(event history analysis)에 근거한 재발 사건 자료 분석(recurrent event data analysis)을 적용할

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

표 **2.1.** 그룹별 교통 법규 위반 횟수

| 그룹 | | 법규 위반 횟수 | | |
|---|---|---|---|---|
| | | 범위 | 중위수 | 평균 |
| 교화 프로그램 참가 여부 | YTOP | (1, 12) | 3.0 | 3.165 |
| | NON-YTOP | (1, 10) | 3.0 | 3.147 |
| 성별 | 남자 | (1, 12) | 3.0 | 3.405 |
| | 여자 | (1, 6) | 2.0 | 2.519 |

것이다 (Cook과 Lawless, 2007). 재발 사건(recurrent event)이란 연구 대상이 같은 종류의 사건들을 반복적으로 경험한 자료를 의미한다. 예를 들어, 빈번한 자동차의 고장(warranty claims), 교통법규의 반복적인 위반, 심장병 환자들의 심장 발작(heart attack), 암 환자의 재발(tumor recurrence), 취업자의 반복적인 이직 등 재발 사건 자료는 사회학, 공학, 의약학 등 전 분야에서 일어날 수 있는 자료이다.

본 연구에서는 재범을 경험한 위법자들을 대상으로 후향 연구(retrospective study)를 통해 얻어진 자료로 그들의 재범률의 변화와 재범률에 미치는 요인과 개인별 특성 존재 여부를 검토하고자 한다. 이를 위해 본 논문에서는 다음과 같은 사항을 중점적으로 다룰 것이다. 첫째, 재범률이 시간에 따라 어떤 변화를 보였는가, 즉, 재범률의 추세를 2절에서 조사하며, 3절에서는 재범을 방지하기 위해 적용된 교화 프로그램이 재범률에 어떤 영향을 주는지를 조사할 것이다. 이 때, 개인별 특성을 고려하기 위해 프레일티 효과가 적용된다. 4절에서 개인적 특성이 시간에 따라 변할 수 있다는 가정하에 시간 가변 프레일티를 모형에 포함시킨다. 본 논문에서 분석될 자료의 특이성은 일부 관측 대상들이 일정 기간 동안 위험 그룹에서 제외되는 불연속적인 관측 시점을 가진다는 것이다. 5절에서는 본 연구와 연관된 향후 연구 문제가 논의된다.

## 2. 자료 소개와 추세분석

Kim과 Jhun (2008)과 Sun 등 (2001)은 미국 미주리주에 거주하는 15~24세 나이의 운전면허 소지자들에 대한 교통법규 위반(traffic conviction)의 재발과 연관된 연구를 시행하였다. 분석할 자료는 192명의 가속 위반과 신호 위반 등 교통법규를 위반한 청장년이 연구대상이며 그들의 교통 법규 위반에 관련된 모든 기록을 포함한다. 본 자료는 Missouri State Traffic Violation database에 의해 제공되었으며 1995년 7월까지 192명에 대한 후행 연구(retrospective study)에 근거한다. 그 중 97명이 1987년 후반기에 시행된 교화 프로그램인 YTOP(Young Traffic Offenders Program)에 참가하였으며, 192명 중 138명이 남자였다. 이 연구의 관심 변수는 운전 면허증을 취득한 날로부터 교통 법규 위반할 때까지 걸리는 시간이며, 따라서 여러 번의 위반 시 그의 자료는 재발 사건으로 간주된다. 따라서 $i$번째 관측 대상에 대해 재발 사건 시간은 $(t_{i1}, t_{i2}, \ldots, t_{in_i},\ i = 1, \ldots, 192)$이고 서로 다른 연구 종료 시점, $\tau_i$을 가질 수 있다. 최장기 관측 시간(follow-up time)은 11년이며 최다 위반 횟수는 12번, 평균 위반 횟수는 3.11번이다. 표 2.1은 각 그룹별 교통법규 위반 횟수에 대한 자료이다. 여기서 교화 프로그램은 하루 동안 대학병원에 머무르면서, 교통사고로 인해 중장기간 동안 입원한 환자들과 면회를 통해 환자들이 느끼는 고통과 불편함을 인식하게 하고, 담당 의사와의 면담을 통해 환자의 미래 건강에 대한 적응 능력에 대한 정보를 알게 함으로써 교통법규 위반의 심각성을 인식할 수 있게 하기 위해 개발된 것이다.

위 표에 의하면 YTOP 참가여부에 따른 법규 위반 횟수는 크게 다르지 않음을 알 수 있다. 두 집단 비교를 위한 윌콕슨 순위 검정법(Wilcoxon rank test)을 적용한 결과 두 그룹 간에는 유의한 차이가 없음을 보여 준다 ($p$-value = 0.3098). 하지만 여기서 YTOP 참가에 대한 적절한 이해가 필요하다. 예를 들어, 두 집단 비교 검정을 적용하기 위해선 일반적으로 연구 시작 시점에 두 그룹으로 나누어 실험

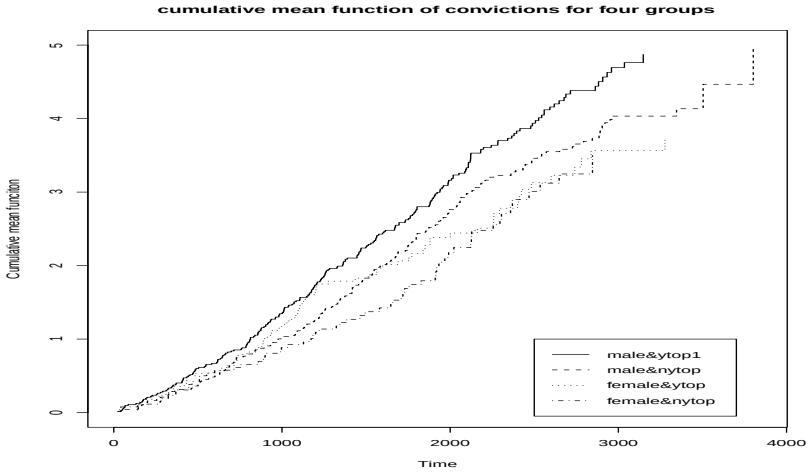


그림 **2.1.** Cumulative mean functions

한 후 그 결과를 비교한다. 하지만 본 연구에서는 모든 연구자들이 NON-YTOP로 시작하다가 그 중 일부분이 YTOP에 참가하게 되며 YTOP 참여 시점이후로는 YTOP 그룹이 된다. 따라서 YTOP 참여 여부는 시간 가변 공변량(time-varying covariate)인데 반해 성별과 같은 변수는 시간에 따라 변하지 않는 시간 고정 공변량이다. 이에 관련된 연구는 생존분석과 반복 측정 자료 등 경시적 자료 연구에서 많은 실용적 결과를 가져왔다 (Klein, 1992; Nielsen 등, 1992; Diggle 등, 2002). 본 자료의 가장 큰 특징은 관측 대상 일부가 운전 면허 정지를 당하게 되어 일정 기간동안 운전을 할 수 없다는 것이다. 그들은 이 기간동안 위험 그룹에 속하지 않기 때문에 자료 분석 시 반영되어야 한다. 이와 비슷한 예로써 중범죄를 범한 재범자가 재수감 될 경우 그 기간 동안 그는 위험 그룹에서 제외되어야 한다. YTOP 자료에서는 192명 중 28명이 운전정지 처분을 받았으며 그 횟수의 범위는 $0 \le K_i \le 9$였다. $S_{ik} = [\mathrm{SL}_{ik}, \mathrm{SU}_{ik}]$는 $i$번째 관측 대상의 $k$번째 운전 정지 기간을 표현한다. 일반적인 생존 분석에서는, $i$번째 관측대상의 연구종료시점이 $\tau_i$일 때, 위험 지시함수(risk indicator)는 $Y_i(t) = I(\tau_i \ge t)$와 같이 정의된다. 그러나 YTOP 자료에서는 일부 관측대상자들이 운전 면허 정지로 불연속적인 관측시점을 가지기 때문에, $Y_i(t) = I(\tau_i \ge t, t \notin \bigcup_k S_{ik})$와 같이 다시 정의된다. $N_i(t) = \sum_j^{n_i} I(t_{ij} \le t)$로 $i$번째 관측대상이 $t$ 시점까지 겪은 재발 사건수를 의미한다. 여기서 재범률을 알아보기 위해 네 그룹의 누적 평균이 다음의 식을 이용하여 추정되었다.

$$M(t) = \frac{\sum_{i=1}^{n} Y_i(t) N_i(t)}{\sum_{i=1}^{n} Y_i(t)} = \frac{N(t)}{Y(t)},$$

여기서 $N(t)$는 $t$ 시점까지 일어난 모든 재발 사건수를 $Y(t)$는 $t$ 시점에서의 위험그룹에 속한 개체수를 위미한다. 그림 2.1은 이러한 불연속적인 관측을 고려하여 YTOP 참여 여부와 성별에 의한 네 그룹의 교통법규 재범의 누적 평균 함수(cumulative mean function)를 추정한 것이다. YTOP에 참여한 남자그룹이 가장 높은 누적 평균을 보여주며, 남자가 여자보다 더 많은 교통 법규 위반을 하였다. 주목할 점은 YTOP 참여한 여자그룹들은 초기에는 높은 재범률을 보이다가 시간이 경과함에 따라 점점 감소하는 경향을 보인다. 다음 절에서는 새롭게 정의된 위험 지시함수와 함께 개인별 특성을 고려한 회귀모형이 고려된다.

## 3. 랜덤 효과를 포함한 회귀 모형

본 절에서는 교화 프로그램의 효과와 성별의 효과를 알아 보기 위해 회귀 모형을 고려한다. 또한 재발사건간의 연관성과 개인별 특징을 모형에 포함한다. $t_{ij}$는 $i$번째 관측대상의 $j$번째 재범시점이고 관심 있는 공변량은 $p \times 1$ 벡터인 $x_i'(t) = (x_{i1}(t), \ldots, x_{ip}(t))$로 표현되며 시간에 따라 변화하는 시간 가변 공변량을 포함한다. 재발 사건 분석을 위해 여러 가지 분석방법들이 적용될 수 있다. 분석할 자료 형태에 따라 연구 시점 후 재발건수에 근거한 방법(Methods based on counts)과 재발 사건간의 시간(Methods based on gap times)에 근거한 방법으로 나눌 수 있다. 연구 목적에 따라 위의 두 자료 형태 중 하나를 결정한 후 다음의 분석 방법들을 각각 적용한다. 전자의 자료 형태하에서는 포아송 과정(Poisson Process)을 가정으로 한 모수적 방법과 연관성을 무시한 로버스트 방법(Robust method)이 적용될 수 있다. 후자의 자료 형태에서는 랜덤 효과를 포함한 강도함수를 이용할 수 있을 것이다. 본 연구에서는 재범의 패턴을 연구하기 위해 전자의 방법을 이용할 것이다. 즉 포아송 모형하에 다음의 위험 함수를 적용한다.

$$\alpha_i(t) = \alpha_0(t)\exp\left(x_i'(t)\beta + b_i\right),$$

여기서 $\alpha_0$는 기저 위험함수이며 $\beta$는 공변량의 효과를 추정한다. 또한 개인별 특성을 모형화하기 위해 프레일티 효과 $b_i$가 포함된다. $b_i$는 관측 대상의 재발 사건간의 상관 관계를 표현할 뿐만 아니라 개인의 성질 및 유전 요소, 환경적 요인을 반영하는 개인별 효과(subject-specific effect)를 표현한다. 특히 개인의 환경적, 유전적 요인은 현재 재범률 연구에서 그 중요성이 더욱 더 강조되고 있다.

개인별 연구 종료시점($\tau_i, i = 1, \ldots, n$)이 재발 사건 발생과 독립이라는 가정하에서 모수 추정을 위한 우도 함수는 다음과 같이 유도된다.

$$\begin{aligned} &L(\beta, \alpha_0) \\ &= \int \prod_{i=1}^{n} \prod_{j=1}^{n_i} \{\alpha_0(t_{ij})\exp(x_{ij}'(t_{ij})\beta + b_i)\} \exp\left(-\int_0^{\tau} Y_i(s)\alpha_0(s)\exp\left(x_i'(s)\beta + b_i\right)ds\right) f(b_i)db_i, \end{aligned} \quad (3.1)$$

여기서 $\tau = \max(\tau_1, \ldots, \tau_n)$이며 프레일티 효과는 다음의 정규 분포를 따른다고 가정한다 $b_i \sim N(0, \sigma^2)$. 본 논문에서는 기저 위험 함수 추정을 위해 조각 상수 위험 함수(piecewise constant hazard function)를 적용한다 (Kim, 2007; Lawless와 Zhan, 1998). 이를 위해 전체 관측 대상의 재범시점에 근거하여 적절하게 $L$개의 구간으로 나누게 된다. 즉, 선택된 $0 = a_0 < a_1 < \cdots < a_L = \tau$를 이용하여 구간 $I_l = [a_{l-1}, a_l)$, $l = 1, \ldots, L$이 결정되면 그 구간 내에서는 같은 강도를 가짐을 가정한다. 기저 위험함수는 다음과 같이 다시 정의된다.

$$\tilde{\alpha}_0(t) = \alpha_l, \quad t \in I_l.$$

누적 기저함수는 불연속 관측 유무에 따라 다음과 같이 정의된다. 모든 관측 대상에 대해 일반화하기 위해 관측 대상을 나타내는 첨자 $i$를 당분간 제외할 것이다. 불연속 기간이 $\{S_k = [\mathrm{SL}_k, \mathrm{SU}_k],\ k = 1, \ldots, K\}$일 때, 불연속 관측 지속기간을 $\mathrm{SDIF}_k = \mathrm{SU}_k - \mathrm{SL}_k$이다. 이 때, $\mathrm{DUR}_l(t) = \max(0, \min(a_l - a_{l-1}, t - a_{l-1}))$이라면

$$\tilde{A}_0(t) = \begin{cases} \displaystyle\sum_{l=1}^{L} \alpha_l \mathrm{DUR}_l(t), & K = 0, \\ \displaystyle\sum_{k=1}^{K}\sum_{l=1}^{L} \alpha_l \widetilde{\mathrm{DUR}}_{lk}(t), & K > 0, \end{cases}$$

여기서 $\widetilde{\mathrm{DUR}}_{lk}(t)$는

$$\begin{cases} \mathrm{DUR}_l(t) - \mathrm{SDIF}_k, & (\mathrm{SL}_k, \mathrm{SU}_k) \in I_l, \\ \mathrm{SL}_k(t) - a_l, & a_l < \mathrm{SL}_k < a_{l+1} < \mathrm{SU}_k, \\ \mathrm{SU}_k(t) - a_l, & \mathrm{SL}_k < a_l < \mathrm{SU}_k < a_{l+1} \end{cases}$$

로

$$\begin{aligned} &L(\beta, \tilde{\alpha}_0) \\ &= \int \prod_{i=1}^{n} \prod_{j=1}^{n_i} \left[\exp\left(x_{ij}'(t_{ij})\beta + b_i\right) \tilde{\alpha}_0(t_{ij})\right] \exp\left\{ -\int_{s \notin \cup S_{ik} \cap [0,\tau]} \exp(x_{ij}(s)\beta + b_i)\tilde{\alpha}_0(s)ds \right\} f_0(b_i)db_i. \end{aligned} \quad (3.2)$$

위 우도함수는 프레일티에 대한 적분을 포함하며 이는 닫힌 형태(closed form)로 유도될 수 없다. 따라서 EM 알고리즘의 E 단계에서 Guass-Hermite 방법에 근거한 수치적분 방법 (Liu 등, 2004)과 몬테칼로 표본을 이용한 방법 (Vaida과 Xu, 2000)이 적용될 수 있다. 본 논문에서는 Guass-Hermite 방법에 근거하여 식 (3.2)의 근사화된 우도함수를 유도한다. $Q$개의 구적점(quadrature point) $u_1, \ldots, u_Q$을 이용하여, $L = \prod_{i=1}^{n} L_i$은 다음의 항들의 곱으로 표현된다.

$$\begin{aligned} L_i &= \sum_{q=1}^{Q} \prod_{j=1}^{n_i} \exp(l_{ijq}) f_\theta(u_q) w_q \\ l_{ijq} &= x_{ij}(t_{ij})'\beta + u_q + \log\tilde{\alpha}_0(t_{ij}) - \exp\left(z_{ij}(t_{ij})'\beta + u_q\right) \tilde{A}_0(\tau_i), \end{aligned}$$

여기서 $u_q = \sqrt{2}z_q$와 $w_q = \sqrt{2}\eta_q \exp(z_q^2)$이며 $z_q$와 $\eta_q$는 여러 수치 적분 관련 책과 통계 패키치에서 얻을 수 있다 (Abramowitz과 Stegun, 1972). 관심 있는 모수 추정을 위해 로그 우도함수를 모수에 대해 미분함으로써 스코어 함수(score function)와 정보 행렬(information matrix)을 유도한 후 EM의 M 단계에서 Newton-Raphson 방법을 이용하여 모수를 추정한다. 본 논문에서는 SAS의 `Proc NLIMIXED`를 이용하여 모수를 추정하였다. 위에서 제안된 추론 과정을 YTOP 자료에 적용한다. 관심 있는 공변량은 교화 프로그램 참가여부와 성별이며 각각 다음과 같이 정의된다.

$$X_1(t) = \begin{cases} 1, & t > \text{교화 프로그램 참가 시점}, \\ 0, & \text{otherwise}. \end{cases} \qquad X_2 = \begin{cases} 1, & \text{남자}, \\ 0, & \text{여자}. \end{cases}$$

조각 개수의 효과를 검토하기 위해 $4, 5, 6$개의 조각을 각각 적용하였다. 표 3.1은 세 경우에 추정된 $\hat{\alpha}, \hat{\beta}$ 그리고 $\hat{\sigma}^2$와 해당되는 표준 오차들을 보여준다. 표 3.1의 값들을 통해 세 모형에 근거하여 추정된 공변량과 프레일티 분산 추정치들은 서로 크게 다르지 않음을 알 수 있다. 우도비 검정을 이용하여 세 모형을 비교할 때, $\chi^2_{0.05}(2) = 5.99$, $\chi^2_{0.10}(2) = 7.81$이므로 위 우도 함수값에 근거하여 5개와 6개 조각을 가진 모형에는 큰 차이가 없음을 알 수 있다. 따라서 5개의 조각을 가진 모형이 최적의 모형으로 고려된다. $\hat{\beta}_1 = -1.194$ 값을 통해 프로그램 참여 여부 그룹은 재범을 감소시키는 데 유의적 ($p$-value $<$ 0.001)인데 반해 성별에는 차이가 없는 것 같다 ($p$-value $= 0.461$). 프레일티의 분산추정값 $\hat{\sigma}^2 = 1.297$ ($p$-value $< 0.001$)을 통해 개인별 성향이 서로 다름을 확인할 수 있으며 그림 3.1은 관측 개체별 추정된 프레일티 효과 $\hat{b}_i$를 보여준다.

## 4. 시간 가변 프레일티를 적용한 재발 사건 분석

본 장에서는 3장에서 적용한 일변량 프레얼티를 확장한 이변량 프레일티 효과(bivariate frailty effect)를 고려한다. 많은 경시적 자료에서 개인의 특성은 시간에 따라서 변화를 보여줄 수 있다. 교통 법

표 **3.1.** 조각의 갯수별 모수 추정의 비교

| $L$ | 4 | 5 | 6 |
|---|---|---|---|
| $\alpha_1$ | 0.036(0.007) | 0.029(0.007) | 0.022(0.005) |
| $\alpha_2$ | 0.133(0.027) | 0.094(0.022) | 0.050(0.011) |
| $\alpha_3$ | 0.285(0.062) | 0.237(0.055) | 0.141(0.032) |
| $\alpha_4$ | 1.263(0.297) | 0.538(0.129) | 0.328(0.077) |
| $\alpha_5$ | | 1.866(0.486) | 1.235(0.321) |
| $\alpha_6$ | | | 3.521(1.052) |
| $\beta_1$ | −1.194(0.155) | −1.352(0.159) | −1.361(0.161) |
| $\beta_2$ | −0.163(0.223) | −0.194(0.245) | −0.210(0.247) |
| $\sigma^2$ | 1.297(0.224) | 1.652(0.273) | 1.695(0.283) |
| −2 Log-Likelihood | 3869.4 | 3826.4 | 3825.1 |

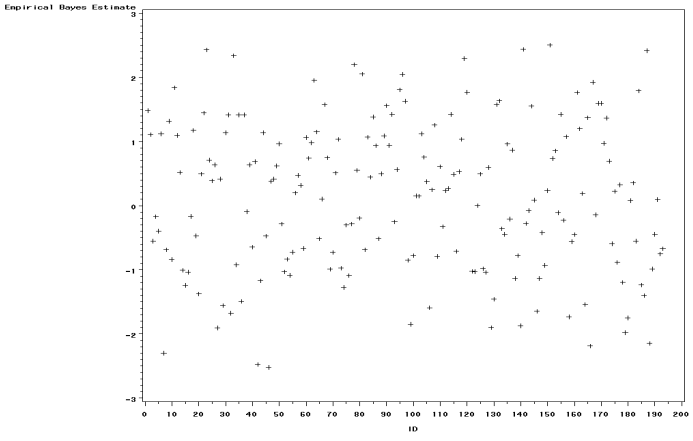


그림 **3.1.** 추정된 일변량 프레일티 효과

규에 관련된 재발 사건에서 개인적 재발성향이 시간에 따라 어떻게 변화하는가를 조사하기 위해 시간 가변 프레일티 모형을 적용한다. 3장에서 적용한 프레일티는 $b_0 + b_1 t$로 확대되며 위험 함수는

$$\alpha_i(t) = \alpha_0(t)\, \exp\left(x_i'(t)\beta + b_{i0} + b_{i1}t\right),$$

이다. 여기서 $(b_0, b_1)$는 이변량 정규 분포가 가정되며 평균이 영이고 분산-공분산 행렬 $\Sigma$은 다음과 같이 정의된다.

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}.$$

조각 상수 위험 함수하에서 우도함수 (3.2)는 다음과 같이 다시 정의될 수 있다.

$$L(\beta, \tilde{\alpha}_0) = \int \prod_{i=1}^{n} \prod_{j=1}^{n_i} \left[\exp\left(x_{ij}'(t_{ij})\beta + b_{i0} + b_{i1}t_{ij}\right) \tilde{\alpha}_0(t_{ij})\right] \exp\left\{ - \int_{s \notin \cup S_{ik} \cap [0,\tau]} \exp\left(x_{ij}(s)\beta + b_{i0} + b_{i1}s\right) \tilde{\alpha}_0(s) ds \right\} f(b_{i0}, b_{i1}) db_i.$$

표 **4.1.** 시간 가변 프레일티의 적용

| | $\sigma_{12} \neq 0$ | $\sigma_{12} = 0$ |
|---|---|---|
| $\beta_1$ | $-1.285(0.176)$ | $-1.352(0.159)$ |
| $\beta_2$ | $-0.078(0.235)$ | $-0.194(0.245)$ |
| $\sigma_1^2$ | 1.098(0.360) | 1.136(0.269) |
| $\sigma_2^2$ | 0.048(0.022) | 0.050(0.015) |
| $\sigma_{12}$ | 0.011(0.074) | |
| $-2$ Log-Likelihood | 3781.7 | 3781.8 |

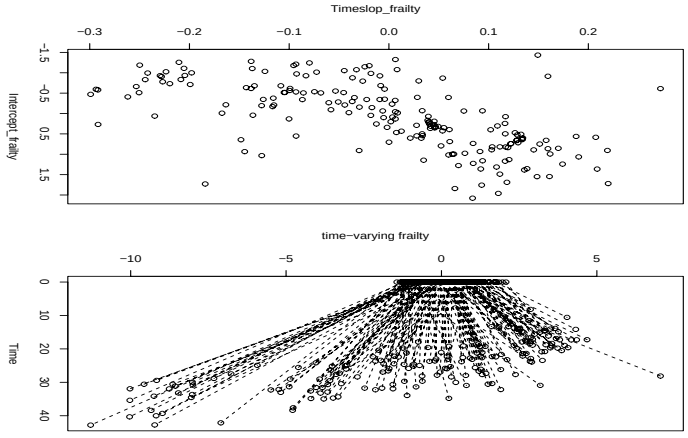


그림 **4.1.** 추정된 이변량 프레일티 효과

표 4.1은 5개의 조각하에서 구한 이변량 프레일티를 적용한 결과이다. 표 3.1과 4.1의 로그 우도값을 이용한 로그우도검정을 통해 이변량 프레일티 모형이 더 적합하다고 할 수 있다 ($3826.4 - 3781.7 = 44.7 > \chi^2_{0.05}(2) = 5.99$). 또한 두 프레일티간의 상관관계는 $H_0 : \sigma_{12} = 0$의 검정결과가 사용되며 표 4.1을 통해 유의하지 않은 결과를 보여준다 ($p$-value $= 0.102$). 즉, 초기에 큰 값의 프레일티를 가진 관측 개체가 계속 증가하거나 감소하지는 않는다. 하지만 $\sigma_1^2$과 $\sigma_2^2$을 통해 개인별 추세가 서로 다름을 알 수 있으며 $\hat{\sigma}_1^2$이 $\hat{\sigma}_2^2$ 보다 더 큰 값을 가짐으로, 연구 초기에는 서로 다른 개인적 특성을 가진데 반해 그 변화 정도는 개인간에 아주 다르다고 할 수 없음을 의미한다. 그림 4.1의 왼쪽 그림은 $(b_{i0}, b_{i1})$으로 다시 한 번 이들간의 무상관 관계를 확인할 수 있다. 또한 $\hat{b}_{i0}$와 $\hat{b}_{i1}$의 범위를 통해 $\hat{b}_{i0}$이 더 넓게 퍼져 있음을 알 수 있다. 그림 4.1의 오른쪽 그림은 $(t, \hat{b}_{i0} + \hat{b}_{i1}t)$, $0 < t < \tau_i$을 보여준다. 또한 이 그림을 통해 $\tau_i$, 즉 연구 참여 기간이 길수록 개인적 재발률이 점점 감소함을 알 수 있다. 물론 한 관측 대상(오른쪽 상반, id $= 4$, $\hat{b}_{4,1} = 7.047$, $\tau_4 = 28.12$, $K_4 = 0$)은 계속 증가하는 재범성향을 보여준다. 공변량의 효과에 대해서는 일변량 프레일티와 비슷한 결론을 얻었다.

## 5. 향후 연구과제

범죄 재범연구를 위해 재발 사건 자료 분석 방법이 적용되었다. 본 논문에서 분석된 자료에서는 반복적인 교통 법규 위반과 간헐적인 연구그룹으로부터의 탈퇴에 의한 불연속적인 관측이 존재하였다. 따라서 연구 목적은 각 시점에 대해 위험그룹에 대한 새로운 정의와 함께 개인별 재범 특성을 구하고자 하였다. 이를 위해 조각 상수 위험함수가 적용되었으며 이에 근거한 시간 가변 프레일티를 모형화함으로써 개인

적 변화 추세를 추정하고자 하였다. 이 논문에서 제안된 연구와 연관하여 다음의 두 가지 연구 문제가 다루어질 필요가 있다.

첫째 연구 방향은 이변량 재발 사건 자료에의 확장에 관한 것이다. 본문에서 다루어진 자료는 교통법규 위반에 관한 것이며 이는 여러 가지 요인에 의한 것이다. 예를 들어 과속 위반과 신호 위반은 가장 많이 일어나는 교통 법규 위반의 예가 된다. 한 관측 대상이 이 두 가지 사건을 각각 여러 번 경험한 경우 이러한 자료 형태를 이변량 재발 사건(bivariate recurrent event data)이라고 한다. 이를 위해 이변량 프레얼티가 다시 적용된다. 본 논문의 목적에 근거하여 이변량 재발 사건 자료를 고려한 경우, 더 복잡한 모형이 정의되어질 필요가 있다.

본 논문에서 분석된 자료에서 운전면허 정지 기간 $(\mathrm{SL}_{ik}, \mathrm{SU}_{ik})$이 알려져 있다고 가정하여 분석되었다. 실제 자료에서는 운전면허 정지 기간에 대해 불완전한 정보만이 주어져 있다. 구체적으로 면허 정지 기간의 시작 시점 $\mathrm{SL}_{ik}$만이 알려져 있으며 종료 시점 $\mathrm{SU}_{ik}$은 알려져 있지 않다. 본 논문에서는 적절한 종료시점 추정을 위해 운전 면허 정지 시작 시점과 그 이후 발생하는 재발 사건과의 중간지점을 운전 면허 정지 기간의 종료 시점으로 간주하였다. 실제 운전면허 정지 기간의 종료시점이 시작시점과 이후 발생하는 재발사건에 의해 둘러싸여 있으며 종료 시점은 구간 중도 절단(interval censored data)으로 간주할 수 있다. Kim과 Jhun (2008)은 한 차례의 운전면허 정지기간이 있을 경우 일변량 구간 중도 절단하에서 로버스트 방법을 이용하여 회귀계수를 추정하였다. 따라서 본 논문에서는 여러 번의 운전 면허정지를 포함하기 때문에 다변량 구간 중도 절단(multivariate interval censored data)의 개념이 적용되어야 한다.

세 번째 연관된 연구는 정보적 중도절단(informative censoring)에 관련된 문제이다. 앞 절에서 언급한 바대로 연구 참여 기간이 길수록 개인적 재범률이 점점 감소함을 알 수 있다. 이는 재발 사건자료에서 빈번하게 발생되는 문제로 연구 중단 사건이 재범과 연관된 경우 이들 간의 관계는 반드시 추론과정에서 고려되어야 한다. 이에 관련된 문제는 여러 통계학자에 대해 연구되어왔다 (Liu 등, 2004; Huang와 Wang, 2004).

## 참고문헌


주희종 (2001). 가석방 출소자의 재범에 관한 한,미간 비교 연구, <한국 공안 행정 학회>, **12**, 317–342.

Abramowitz, M. and Stegun, I. (1972). *Handbook of Mathematical Functions*, Dover, New York.

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

Diggle, P. J., Heagerty, P., Liang, K. Y. and Zeger, S. (2002). *Analysis of Longitudinal Data*, Oxford Express.

Huang, C. Y. and Wang, M. C. (2004). Joint modeling and estimation for recurrent event processes and failure time data, *Journal of the American Statistical Association,* **99**, 1153–1165.

Kim, Y. (2007). Analysis of panel count data with dependent observation time, *Communications in Statistics,* **35**, 983–990.

Kim, Y. and Jhun, M. (2008). The analysis of recurrent event data with incomplete observation gaps, *Statistics in Medicine*, **27**, 1075–1085.

Klein, J. P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm, *Biometrics*, **48**, 795–806.

Lawless, J. F. and Zhan, M. (1998). Analysis of interval-grouped recurrent event data using Piecewise constant rate functions, *Canadian Journal of Statistics,* **26**, 549–565

Liu, L., Wolfe, R. A. and Huang, X. (2004). Shared frailty models for recurrent events and terminal event, *Biometrics,* **60**, 747–756.

Nielsen, G. G., Gill, R. D., Andersen, P. K. and Sørensen, T. I. A. (1992). A counting process approach to maximum likelihood estimation in frailty models, *Scandinavian Journal of Statistics,* **19**, 133–137.

Sun, J., Kim, Y., 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.

Vaida, F. and Xu, R. (2000). Proportional hazard model with random effects, *Statistics in Medicine,* **19**, 3309–3324.

# Statistical Analysis of Recidivism Data Using Frailty Effect

Yang-Jin Kim[1]

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

(Received April 2010; accepted July 2010)

**Abstract**

Recurrent event data occurs when a subject experience the event of interest several times and has been found in biomedical studies, sociology and engineering. Several diverse approaches have been applied to analyze the recurrent events (Cook and Lawless, 2007). In this study, we analyzed the YTOP(Young Traffic Offenders Program) dataset which consists of 192 drivers with conviction dates by speeding violation and traffic rule violation. We consider a subject-specific effect, frailty, to reflect the individual's driving behavior and extend to time-varying frailty effect. Another feature of this study is about the redefinition of risk set. During the study, subject may be under suspension and this period is regarded as non-risk period. Thus the risk variables are reformatted according to suspension and termination time.

Keywords: Observation gap, recidivism, time-varying frailty, recurrent event data, YTOP.

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