The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.1.021
27(1), 21–30

Cure Rate Model with Clustered Interval Censored Data

Yang-Jin Kim[a,1]

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

(Received October 1, 2013; Revised November 29, 2013; Accepted December 23, 2013)

Abstract

Ordinary survival analysis cannot be applied when a significant fraction of patients may be cured. A cure rate model is the combination of cure fraction and survival model and can be applied to several types of cancer. In this article, the cure rate model is considered in the interval censored data with a cluster effect. A shared frailty model is introduced to characterize the cluster effect and an EM algorithm is used to estimate parameters. A simulation study is done to evaluate the performance of estimates. The proposed approach is applied to the smoking cessation study in which the event of interest is a smoking relapse. Several covariates (including intensive care) are evaluated to be effective for both the occurrence of relapse and the smoke quitting duration.

Keywords: Clustered interval censored data, cure rate model, EM algorithm, frailty effect, smoking cessation data.

1. 서론

구간 중도 절단 자료(interval censored data)는 정확한 사건 발생 시간대신 그 시간을 포함한 두 시점만이 관측 가능한 경우에 발생된다. 종종 주기적인 검진이나 특정 시점을 통해서만 관측 가능한 사건에 대해서 이러한 중도 절단자료 형태가 관측된다. 예를 들어, AIDS 의심 환자의 양성여부는 환자의 병원 방문 시 혈액 체취를 통해서만 확증되게 된다. 이 때, 관심 있는 변수는 AIDS 발현 시점이지만 그 정확한 시점은 관측 될 수 없다. 대신 마지막으로 음성 반응을 진단 받은 시점과 처음으로 양성 반응을 진단 받은 시점을 사용하게 되며 그 두 시점 사이 어디선가 AIDS 가 발현되었음을 짐작하게 된다. 또 다른 예로는, 금연 프로그램의 효과를 평가하기 위해 연구에 참여한 사람을 대상으로 금연 지속 여부를 조사하였다고 하자. 만약 다시 흡연을 시작하였다면 그 흡연 재시작 시점이 관심 있는 변수가 될 것이다. 하지만 이러한 연구에서도 정확한 재흡연 시점을 구하지 못하는 경우가 종종 발생한다. 대신 면접원은 흡연자의 마지막 금연 시점과 다시 흡연을 시작했다고 대답한 면접 시점을 이용하게 된다. 본 연구에서는 후자에 관련된 연구 자료를 분석할 것이다. 특히 이 연구에 참여한 사람들은 여러 다른 도시에 거주하고 이들 도시가 각각의 특성이 존재한다면 이러한 특징은 군집 효과로 표현될 수 있을 것이다. 그러

This research was supported by the research fund of 2012 Sookmyung Women's University.
[1]Professor, Department of Statistics, Sookmyung Women's University, Cheongpa-ro 47 gil Yangsan-gu, Seoul 140-742, Korea. E-mail: yjin@sookmyung.ac.kr

나 전체 연구 대상자들 중에서 흡연을 재시작한 비율이 매우 작아 흡연 재시점에 대해 일반적인 생존 분석 기법의 적용은 부적절할 수 있다. 즉, 보통의 생존 분석에서는 모든 개체가 사건 발생의 위험을 가지고 있기 때문에 충분한 시간이 주어지면 결국 모두 사건을 경험하게 됨을 가정한다. 하지만 가끔 이러한 가정은 적합하지 않을 수 있다. 위의 금연 프로그램의 예에서처럼, 상당한 비율의 연구 대상자들이 계속 금연을 하게 되고 그리고 그들이 남은 일생동안 계속 금연할 경우 위의 가정은 적합하지 못하다. 이와 같은 경우, 일반적으로 생존 분석에서 적용되는 로그 순위 검정과 Cox의 비례 위험 모형을 적용하는 것은 타당한 추론 결과를 제공할 수 없게 된다. Berkson과 Gage (1952)는 혼합 모형에 기반한 완치율 모형(cure rate model)의 적용을 고려했으며 그 이후 많은 학자들에 의해 다양한 관련 연구가 진행되고 있다 (Maller와 Zhou, 1992; Farewell, 1982; Kuk과 Chen, 1992; Sy와 Taylor, 2000; Price와 Manatunga, 2001). 최근에 Kim과 Jhun (2008)은 이러한 완치율 모형을 구간 중도 절단 자료에 적용하였다. 본 연구에서 분석할 자료는 Banerjee와 Carlin (2004)이 베이지안 방법을 적용하여 분석한 자료로 연구 참여자의 프로그램 참여 여부뿐만 아니라 성별, 나이, 흡연량, 흡연 기간과 그들의 거주지에 대한 정보를 포함한다. 특히 그들의 주거 환경은 흡연 여부에 영향을 줄 수 있을 것이라 생각하여 연구 참여자의 주거지를 군집(cluster)으로 간주한다. 따라서 우리가 분석할 자료는 연구 참여자의 흡연 시점에 대한 군집 구간 중도 절단자료(clustered interval censored data)가 된다. 최근에 Xiang 등 (2011)은 혼합 모형을 이용하여 군집 구간 중도 절단 자료를 분석하였다. 본 연구에서는 Kim과 Jhun (2008)의 모형을 확장하여 준모수 모형(semiparametric model)에 기반을 둔 방법을 제시한다. 또한 군집 자료의 특성을 반영하기 위해 공유 프레일티(shared frailty)효과를 적용한다. 2장에서는 적절한 통계 모형을 제시하며 모수추정을 위한 EM 알고리즘의 적용과정은 3장에서 서술되며 간단한 모의실험과 금연 프로그램 자료에 대한 적용 예는 4장과 5장에 각각 주어진다. 6장에서는 연구 결과의 요약과 관련 향후 연구를 제시할 것이다.

2. 통계 모형

$k\,(=1,\ldots,K)$번째 군집에 대해, $\{t_k=(t_{k1},\ldots,t_{kn_k})',\ k=1,\ldots,K\}$은 관심 있는 사건 발생 시간의 벡터이며 여기서 n_k는 군집 크기를 보여준다. 하지만 구간 중도 절단자료 하에서 정확한 발생 시간대신 이를 포함한 구간 $\{(\mathrm{tl}_{ki},\mathrm{tr}_{ki}),\ i=1,\ldots,n_k\}$이 관측된다. 또한, 구간중도 또는 우중도 절단여부를 표시하기 위해 지시함수, $\delta_{ki}=I(\mathrm{tr}_{ki}<\infty)$가 정의된다. 완치율 모형 하에서는 관측 개체는 두 그룹, 사건 발생 취약 그룹(susceptible group)과 비취약 그룹(insusceptible group)으로 나누게 되며 그들의 소속여부는 지시함수로 d_{ki}로 표시한다. 즉, $d_{ki}=1$은 사건 발생 취약 그룹이며 $d_{ki}=0$은 비취약그룹을 의미한다. 따라서 사건 지시함수 $\delta_{ki}=1$은 $d_{ki}=1$인 반면 우중도 절단된 관측개체의 취약 그룹 포함 여부는 알 수 없다. 즉, $\delta_{ki}=0$인 관측 개체의 d_{ki}값 여부는 알 수 없게 된다. 완치율 모형 하에서, 공변량이 고려되지 않을 때, 전체 주변 생존 함수(marginal survival function), $S(t)$는 다음과 같이 표현된다.

$$S(t)=\Pr(d=1)\Pr(T\le t|d=1)+\Pr(d=0)\Pr(T\le t|d=0)=p\tilde{S}_0(t)+(1-p),$$

여기서 $p=\Pr(d=1)$이다. 이제 위 확률을 군집자료에 적용하면 k 군집에 속한 i번째 개체의 사건 발생 확률(incidence probability)을 p_{ki}라고 할 때, 공변량 $x_{ki}=(1,z_{ki})$를 이용하여 로지스틱 회귀 모형을 적용한다. 이 때, 군집 간의 이질성과 군집 내 연관성을 모형화하기 위해 프레일티 효과(frailty effect)를 적용한다. 즉, 프레일티 효과 w_k가 주어질 때, 사건 발생 확률은 다음과 같다.

$$p_{ki}=\Pr(d_{ki}=1|x_{ki},w_k)=\frac{\exp(x_{ki}'b+w_k)}{1+\exp(x_{ki}'b+w_k)}. \tag{2.1}$$

또한 사건 발생 취약그룹에 대한 사건 발생시간에 대한 공변량의 효과를 추정하기 위해 다음의 비례 위험 모형이 적용된다. 여기서도 군집 효과를 반영하기 위해 위에서 사용한 프레일티를 그대로 적용한다. 따라서 다음의 강도함수가 정의된다.

$$\lambda(t_{ki}|z_{ki},w_k)=\lambda_0(t_{ki})\exp\left(\beta' z_{ki}+w_k\right)=\tilde{\lambda}(t_{ki}). \tag{2.2}$$

모형 (2.2)에 근거하여 해당되는 생존 함수는 $\tilde{S}(t_{ki})=\exp(-\tilde{\Lambda}(t_{ki}))=\exp(-\Lambda_0(t_{ki})\exp(\beta' z_{ki}+w_k))$로 표현된다. 여기서 $\tilde{\Lambda}(s)=\int_0^s\tilde{\lambda}(u)du$이며 $\Lambda_0(t)$는 기저 누적 위험함수를 의미한다. 이 연구에서는 프레일티는 평균이 영이고 분산이 σ^2인 정규 분포를 가정한다.

구간 중도절단자료에 대한 공변량의 효과를 추정하기 위해 이 연구에서는 Goetghebeur와 Ryan (2000)이 제시한 approximate likelihood를 적용한다. 즉, 적절한 구간의 시간 간격, $(0=s_0,s_1,\ldots,s_m)$을 선택한다. 여기서 시간 간격은 충분히 좁아 두 사건이 동시에 관측될 수 없을 것이라 가정한다. 우리는 Turnbull (1976)이 제시한 equivalence set을 이용하여 시점을 계산하며 자세한 유도 과정은 Lindsey와 Ryan (1998)과 Turnbull (1976)을 참고하기 바란다. 이렇게 선택된 시간 간격($l=1,\ldots,m$)에 대해 k번째 군집의 i번째 관측개체는 위험 지시 함수(risk indicator function), Y_{kil}와 사건 발생 지시함수(event occurrence indicator function), dN_{kil}를 각각 정의한다. 이 정의 하에서 w_k가 주어져 있다는 조건하에서 식 (2.1)과 (2.2)를 이용하여 조건부 우도 함수를 다음과 같이 정의한다.

$$\begin{aligned}L(b,\beta,\lambda|w)&=\prod_{k=1}^{K}\prod_{i=1}^{n_k}\prod_{l=1}^{m}\left\{\ p_{ki}\tilde{S}(t_{ki})\tilde{\lambda}(t_{ki})\ \right\}^{dN_{kil}d_{ki}}\times(1-p_{ki})^{(1-d_{i\cdot})}\times\left\{p_{ki}\tilde{S}(t_{ki})\right\}^{(1-dN_{kil})\ d_{ki}}\\&=\prod_{k=1}^{K}\prod_{i=1}^{n_k}\prod_{l=1}^{m}\left\{p_{ki}^{d_{ki}}(1-p_{ki})^{1-d_{ki}}\right\}\times\exp\left(-d_{ki}Y_{kil}\lambda_l e^{\beta'\,z_{ki}}\right)\left(\lambda_l e^{\beta'\,z_{ki}}\right)^{d_{ki}dN_{kil}}.\end{aligned} \tag{2.3}$$

우도함수 (2.3)에 대해 로그를 취함으로써 다음과 같이 표현된다.

$$\begin{aligned}l(b,\beta,\lambda|w)=&\sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{l=1}^{m}d_{ki}\left[x'_{ki}b+w_k-\log\left(1+e^{x'_{ki}b+w_k}\right)\right]-\sum_{k=1}^{K}\sum_{i=1}^{n_k}(1-d_{ki})\log\left(1+e^{x'_{ki}b+w_k}\right)\\&-\sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{l=1}^{m}\left[d_{ki}Y_{kil}\lambda_l e^{\beta' z_{ki}+w_k}+d_{ki}dN_{kil}\left(\log\lambda_l+\beta' z_{ki}+w_k\right)\right].\end{aligned}$$

위 로그 우도함수를 최대화시키는 모수값을 구하기 위해선 결측 자료에 대한 처리가 필요하며 이를 위해 다음 절에서는 EM 알고리즘을 적용한다.

3. 모수 추정을 위한 EM 알고리즘의 적용

EM 알고리즘은 결측 자료 또는 불완전하게 측정된 자료를 포함한 우도 함수의 모수 추정을 위해 가장 널리 적용되는 방법이다. 식 (2.3)의 우도 함수에서 $O_{ki}=(\mathrm{tl}_{ki},\mathrm{tr}_{ki},\delta_{ki},z_{ki})$는 관측 가능 자료를 $M_{ki}=(d_{ki},dN_{ki},Y_{ki},w_k)$는 결측 자료를 표현한다. 여기서 추정되어야 할 관심 있는 모수는 $\theta=(b,\beta,\lambda,\sigma^2)$이다. 따라서 완전 자료에 근거한 우도 함수는

$$\begin{aligned}L_c(\theta)=&\prod_{k=1}^{K}\prod_{i=1}^{n_k}\prod_{l=1}^{m}\left(\frac{e^{x'_{ki}b+w_k}}{1+e^{x'_{ki}b+w_k}}\right)^{d_{ki}}\left(\frac{1}{1+e^{x'_{ki}b+w_k}}\right)^{1-d_{i\cdot}}\\&\times\exp\left(-d_{ki}Y_{kil}\lambda_l e^{\beta' z_{ki}+w_k}\right)\left(\lambda_l e^{\beta' z_{ki}+w_k}\right)^{d_{ki}dN_{kil}}\times\frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-\frac{w_k^2}{2\sigma^2}}\end{aligned} \tag{3.1}$$

이며 해당 로그 우도 함수는

$$\begin{aligned} l_c(\theta) &= \sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{l=1}^{m} d_{ki}\left\{x_{ki}'b + w_k - \log\left(1+e^{x_{ki}'b+w_k}\right)\right\} - \sum_{k=1}^{K}\sum_{i=1}^{n_k}(1-d_{ki})\log\left(1+e^{x_{ki}'b+w_k}\right) \\ &\quad + \sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{l=1}^{m} d_{ki}dN_{kil}\left\{\log\lambda_l + \beta' z_{ki} + w_k\right\} - d_{ki}Y_{kil}\lambda_l e^{\beta' z_{ki}+w_k} - \frac{1}{2}\sum_{k=1}^{K}\left(\log\sigma^2 + \frac{w_k^2}{\sigma^2}\right) \\ &= l_1(b) + l_2(\beta,\lambda) + l_3\left(\sigma^2\right) \end{aligned}$$

이다. 위에서 유도한 로그 우도 함수는 우중도 절단 자료의 사건 취약 그룹 포함여부, 구간중도 절단 자료에 의한 사건 발생과 위험 여부와 프레일티 효과에 관련된 결측 자료를 포함한다. E-step에서는 이러한 결측 자료의 조건부 기대값을 계산하기 위해 이단계 알고리즘을 적용한다.

3.1. E-step

완전 자료 $\{O_{ki}, M_{ki};\, k=1,\ldots,K,\, i=1,\ldots,n_k\}$와 현재 추정된 모수값, $\tilde{\theta}$가 주어져 있을 때, 우도함수에 대한 조건부 기대값은 다음과 같다.

$$\begin{aligned} Q(\theta) &= E\left[l(\theta;O,m)|\tilde{\theta},O\right] = E\left[l_1(b)|\tilde{\theta},O\right] + E\left[l_2(\beta,\lambda)|\tilde{\theta},O\right] + E\left[l_3\left(\sigma^2\right)|\tilde{\theta},O\right] \\ &= Q_1(b) + Q_2(\beta,\lambda) + Q_3\left(\sigma^2\right), \end{aligned}$$

여기서

$$\begin{aligned} Q_1(b) &\approx \sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{l=1}^{m} E^*[d_{kil}]\left\{x_{ki}'b + E^*[w_k] - E^*\left[\log\left(1+e^{x_{ki}'b+w_k}\right)\right]\right\} \\ &\quad - \left(1-E^*[d_{ki}]\right)E^*\left[\log\left(1+e^{x_{ki}'b+w_k}\right)\right] = \tilde{Q}_1, \\ Q_2(\beta,\lambda) &\approx \sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{l=1}^{m} E^*[d_{ki}]E^*[dN_{kil}]\left\{\log\lambda_l + \beta' z_{ki} + E^*[w_k]\right\} \\ &\quad - E^*[d_{ki}]E^*[Y_{kil}]E^*\left[e^{w_k}\right]\lambda_l e^{\beta' z_{ki}} = \tilde{Q}_2, \\ Q_3(\sigma^2) &\approx -\frac{1}{2}\sum_{k=1}^{K}\left(\log\sigma^2 + \frac{E^*[w_k^2]}{\sigma^2}\right) = \tilde{Q}_3, \end{aligned}$$

여기서 $E^*(\,\cdot\,) = E(\,\cdot\,|\,\tilde{\theta},O)$를 의미하며 $\tilde{Q}$의 Q에 대한 근사화는 Taylor expansion (Herring 등, 2002)에 근거한다. E-step에서는 $\tilde{Q}_1, \tilde{Q}_2$와 $\tilde{Q}_3$를 계산하기 위해 이단계 알고리즘이 적용된다. 먼저 첫 번째 단계에서 사건 발생률과 구간 중도 절단과 관련된 결측 자료는 그들의 조건부 기대값을 이용하여 계산된다. 그 다음 두 번째 단계에서는 Gauss-Hermite 알고리즘을 이용하여 프레일티와 관련된 항이 계산된다.

단계 1. d_{ki}, dN_{kil}와 Y_{kil}의 조건부 기대값을 계산한다.

A. 각 구간에 대해 d_{kil}을 계산한다.

(a) 구간 중도 절단 자료에 대해, 사건을 경험하였으므로 $E(d_{ki}|\theta^{(h)}, O_{ki}) = 1$인데 반해,

(b) 우 중도 절단 자료에 대해,

$$
\begin{aligned}
&E\left(d_{ki}|\theta^{(h)}, O_i\right) \\
&= \mathrm{pr}\left(d_{ki}=1\,|\,\theta^{(h)}, s_l = \mathrm{tl}_{ki}\right) = \frac{P(d_{ki}=1; s_l=\mathrm{tl}_{ki})}{p(s_l=\mathrm{tl}_{ki})} \\
&= \frac{\mathrm{pr}(d_{ki}=1; b, w_k)\left[\exp\left\{-\sum_{j=1}^{l} Y_{kij}\lambda_j \exp\left(\beta' z_{ki}+w_k\right)\right\}\right]}{1-\mathrm{pr}(d_{ki}=1; b, w_k)+\mathrm{pr}(d_{ki}=1; b, w_k)\left[\exp\left\{-\sum_{j=1}^{l} Y_{kij}\lambda_j \exp(\beta' z_{ki}+w_k)\right\}\right]}.
\end{aligned}
$$

B. 각 구간에 대해 dN_{kil}와 Y_{kil}를 계산한다.

(a)′ 구간 중도 절단 자료에 대해

$$
E\left(dN_{kil}\,|\,\theta^{(h)}, O_{ki}\right) = \frac{\lambda_l \exp\left(\beta' z_{ki}+w_k\right)\left[\exp\left\{-\sum_{j=1}^{l} Y_{kij}\lambda_j\, \exp\left(\beta' z_{ki}+w_k\right)\right\}\right]}{\sum_{tl_{ki}\le s_l\le \mathrm{tr}_{ki}} \lambda_l \exp\left(\beta' z_{ki}+w_k\right)\left[\exp\left\{-\sum_{j=1}^{l} Y_{kij}\lambda_j \exp\left(\beta' z_{ki}+w_k\right)\right\}\right]},
$$

$$
E\left(Y_{kil}\,|\,\theta^{(h)}, O_{ki}\right) = \sum_{r\ge l} E\left(dN_{kir}\,|\,\theta^{(h)}, O_i\right).
$$

(b)′ 우중도 절단 자료에 대해

$$
E\left(dN_{kil}\,|\,\theta^{(h)}, O_{ki}\right) = 0, \qquad E\left(Y_{kil}\,|\,\theta^{(h)}, O_{ki}, s_l \le \mathrm{tl}_{ki}\right) = 1.
$$

단계 2. d_{ki}, dN_{kil}와 Y_{kil}가 주어져 있을 때, 프레일티와 관련된 함수, $f(w_k)$의 조건부 기대값은 다음과 같이 정의된다.

$$
E\left(f(w_k)|\,\theta^{(h)}, O_{ki}\right) = \frac{\int_{-\infty}^{\infty} f(w_k)\, L_c(\theta)\, d\Phi(w_k)}{\int_{-\infty}^{\infty} L_c(\theta)\, d\Phi(w_k)},
$$

여기서 Φ는 평균이 영이고 분산이 σ^2인 정규 분포의 누적 분포함수이며 분모는 완전 우도 함수에 대해 프레일티를 적분한 주변 분포이다. 하지만 이 적분은 완전히 닫힌 형태를 가지지 못하기 때문에 이 논문에서는 7-point Gauss-Hermite algorithm을 적용한다.

3.2. M-step

A. b와 σ^2의 추정,

로지스틱 회귀 계수, b를 추정하기 위해 다음의 스코어 함수가 적용된다.

$$
U = \frac{\partial \tilde{Q}_1}{\partial b} = \sum_{k=1}^{K}\sum_{i=1}^{n_k} x_{ki}\left\{d^*_{ki} - \frac{e^{x'_{ki}b}E^*(e^{w_k})}{1+e^{x'_{ki}b}E^*(e^{w_k})}\right\},
$$

여기서 $d^*_{ki} = E(d_{ki}|\theta^{(h)}, O_i), E^*(e^{w_k}) = E(e^{w_k}|\theta^{(h)}, O_i)$이며 $\tilde{Q}_3$의 미분값을 통해 σ^2의 추정량은

$$
\hat{\sigma}^2 = \frac{1}{K}\sum_{k=1}^{K} E^*\left(w_k^2\right)
$$

으로 유도된다. 여기서 $E^*(w_k^2) = E(w_k^2|\theta^{(h)}, O_{ki})$이다.

B. $\tilde{S}_0$와 $\tilde{\Lambda}$의 업데이트.

안정적인 추정값을 구하기 위해 Sy와 Taylor (2000)의 방법을 따라 $t > s_m$에 대해 $\hat{\tilde{S}}_0(t) = 0$을 가정한다. 따라서

$$\hat{\tilde{S}}_0(t_l) = \exp\left[- \sum_{j:t_j \le t_l} \frac{\sum_{k=1}^{K} \sum_{i=1}^{n_k} e_{kij}}{\sum_{k=1}^{K} \sum_{i=1}^{n_k} r_{kij}} \right]$$

이며 기저 위험 함수는 다음과 같다.

$$\hat{\lambda}_l = \frac{\sum_{k=1}^{K} \sum_{i=1}^{n_k} e_{kil}}{\sum_{k=1}^{K} \sum_{i=1}^{n_k} r_{kil}},$$

여기서 구간 중도 절단 자료에 대해 $E^*(Y_{kil}) = E(Y_{kil} \,|\, \theta^{(h)}, O_{ki})$, $e_{kil} = E^*(dN_{kil}) = E(dN_{kil} \,|\, \theta^{(h)}, O_{ki})$ $r_{kil} = E^*(Y_{kil})\exp(\beta' z_{ki}) E^*(e^{w_k})$이다. 한편 우중도 절단 자료에 대해 $r_{kil} = E^*(Y_{kil})$ $\exp(\beta' z_{ki}) E^*(e^{w_k})$ d^*_{ki}로 정의된다.

C. β에 대한 추정값은 다음 방정식을 최대화함으로써 구할 수 있다.

$$\sum_{k=1}^{K} \sum_{i=1}^{n_k} \sum_{l=1}^{m} \left\{ z_{ki} - \frac{\sum_{j=1}^{K} \sum_{r=1}^{n_k} r_{jrl} z_{jr}}{\sum_{j=1}^{K} \sum_{r=1}^{n_k} r_{jrl}} \right\} e_{kil} = 0.$$

위에서 제시된 EM algorithm은 다음과 같이 요약된다.

(i) 초기값을 구하기 위해 구간 중도 절단 자료에 대해서는 그들의 중간값을 그리고 프레일티는 존재하지 않는다는 가정하에서 모수를 추정한다.

(ii) (i)에서 추정된 초기값을 이용하여 $d_{ik}, dN_{ikl}, Y_{ikl}$에 대한 조건부 기대값과 프레일티 함수, $g(f(w_i))$의 조건부 기대값을 E-step에서 계산한다.

(iii) M-step에서 $\theta = (b, \beta, \lambda, \sigma^2)$를 추정한다

(iv) 수렴할 때까지 (ii)와 (iii)을 반복한다.

추정된 $\hat{\theta}$의 분산을 계산하기위해 관측된 정보행렬(observed information matrix), $E[-l_c''(y, \hat{\theta})]$의 역행렬을 이용하였다.

4. Simulation

앞에서 제시된 방법의 적합 정도를 검토하기 위해 간단한 모의실험을 실시한다. 여기서 반복횟수는 500이며 여러 개의 다른 σ값이 사용되었다. 모의자료를 구하기 위한 과정은 다음과 같다.

단계1: 50개와 100개 군집을 각각 가정한다. 정규 분포를 이용하여 각 군집효과 별 프레일티를 추출한다. 즉, $w_k \sim N(0, \sigma^2)$, $k = 1, \ldots, 50$.

단계2: 각 군집의 크기(m_k)를 결정하기 위해 최대값 6을 가지는 다항 분포를 가정한다, $m_i \sim$ Multi(1, $2, 3, 4, 5, 6$).

Table 4.1. Parameter estimates for three σ values; 500 replicates

	Mean	SE	SEM	CP	Mean	SE	SEM	CP	Mean	SE	SEM	CP
	$\sigma = 0.3$				$\sigma = 0.5$				$\sigma = 0.7$			
	$K = 50$											
$b_0 = -0.3$	−0.301	0.159	0.204	0.882	−0.352	0.160	0.201	0.870	−0.310	0.160	0.243	0.804
$b_1 = 0.5$	0.493	0.168	0.187	0.926	0.504	0.169	0.180	0.924	0.519	0.171	0.178	0.928
$\beta = 0.5$	0.470	0.129	0.145	0.904	0.480	0.128	0.138	0.906	0.473	0.128	0.144	0.914
σ	0.321	0.021	0.028	0.964	0.506	0.128	0.138	0.992	0.671	0.091	0.073	0.890
	$K = 100$											
$b_0 = -0.3$	−0.286	0.111	0.142	0.874	−0.346	0.112	0.152	0.850	−0.350	0.112	0.172	0.802
$b_1 = 0.5$	0.500	0.118	0.126	0.942	0.500	0.118	0.118	0.960	0.484	0.117	0.116	0.938
$\beta = 0.5$	0.477	0.090	0.098	0.906	0.481	0.089	0.089	0.945	0.478	0.088	0.095	0.922
σ	0.318	0.014	0.010	0.954	0.515	0.038	0.035	0.982	0.690	0.068	0.059	0.940

단계3: 각 군집내 개체별로 이변량 공변량 $x_{ki} = (0, 1)$을 같은 비율로 생성한다. $\lambda(t_{ki}) = \exp(0.5x_{ki} + w_k)$와 $\lambda_c(c_{ki}) = 0.1$의 강도함수를 따르는 사건 시간과 중도절단시간을 생성한다. 또한 구간중도 절단 자료를 만들기 위해 Pan (2000)의 방법을 사용하였다.

단계4: 완치여부를 결정하기 위해 로지스틱 회귀모형, $\operatorname{logit}\Pr(y_{ki} = 1) = -0.3 + 0.5x_{ki}$을 이용하여 이변량 반응 변수, y_{ki}를 생성한다.

Table 4.1은 세 가지 다른 분산($\sigma = 0.3, 0.5, 0.7$)하에서 제안된 방법을 이용하여 구한 $(b_0, b_1, \beta, \sigma)$에 대한 추정치들의 평균(Mean), 추정된 표준 오차의 평균(SE), 추정치의 표준 편차(SEM) 그리고 95% CP(coverage probability)를 보여준다. Table 4.1의 결과에 의하면 σ가 커질수록 $\hat{b}_0$와 $\hat{\beta}_1$의 편의가 증가함을 알 수 있다. 하지만 K가 증가할수록 편의가 줄어드는 것 같다. 또 따른 주목할 점은 추정된 표준오차(SE)는 표준 편차(SEM)에 비해 모든 경우에서 작은 값을 가진 다는 점이다. 따라서 분산 추정을 위해 사용된 관측된 정보행렬을 대체할 다른 분산 추정량이 제시될 필요가 있음을 시사한다.

5. 금연 자료에 대한 적용

앞 절에서 제시된 완치율 모형을 금연 프로그램자료에 적용한다. 223명이 참여하였으며 그들은 각각 랜덤하게 두 그룹으로 배치되었다. 첫 번째 프로그램은 smoking intervention(SI) group으로 새로운 교육 프로그램이며 이에 대한 대조군으로 UC(usual care)그룹이 있다. 각 참여자는 대략 일 년에 한 번 면접을 통해 금연 프로그램의 지속 여부를 검사하게 되며 이런 면접은 5년 동안 지속되었다. 특히 다시 흡연을 시작하게 된 경우, 연구 참여자는 정확한 흡연 시작 시점을 기억하지 못한다. 따라서 면접자가 추측할 수 있는 재시작 시점은 마지막으로 금연을 한다고 응답한 면접시점과 다시 흡연을 한다고 대답한 시점의 중간 어느 시점쯤이라는 것이다. 이런 종류의 자료는 구간 중도절단자료의 유형이다. 특히 연구 참여자는 51개의 county에 거주하며 각 county는 도시, 농촌 등 다른 지방색을 가지고 있기 때문에 county효과를 군집효과로 간주하였다. 이 자료의 연구 목적은 새로운 프로그램이 금연지속 그룹에 미치는 영향을 조사하고자 한다. 따라서 SI(= 1)과 UC(= 0)의 효과를 비교할 것이다. 그 이외에도 연구 참여자의 프로그램 참여 전 매일 피우는 담배 개수, 흡연 기간 그리고 성별을 로지스틱 모형과 비례위험 모형에 공변량으로 포함시켰다. 전체 223명 중에 169명이 SI를 54명이 UC 프로그램에 각각 참여하였다. 또한 65명이 재흡연을 시작하였으며 SI 그룹과 UC 그룹에서 각각 45명과 20명으로 그 비율은 0.266과 0.370이었다. Banerjee와 Carlin (2004)는 이 자료에 대해 베이지안 방법을 적용하였으며 특히

Table 5.1. Application of cure rate model to smoking cessation data

	Logistic	Survival hazard
Intercept	−0.081(0.765)	
Treatment(SI/UC)	−0.570(0.349)	−0.348(0.273)
Cigarettes smoked per day	0.023(0.013)	0.011(0.010)
Duration as smoker	−0.035(0.020)	−0.018(0.018)
Sex(male = 0)	0.524(0.302)	0.326(0.254)
Variance estimate $\hat{\sigma}^2$	1.059(0.105)	

공간 상관관계(spatial correlation)를 고려하였다. Table 5.1은 분석 결과를 보여준다. 완치율에 관한 로지스틱 모형에서는 흡연기간만이 유의한 변수였으며 흡연 기간이 짧을수록, 즉 흡연을 시작한 지 얼마 되지 않은 사람일수록 흡연을 다시 시작할 확률이 높게 된다. 흡연 재시작 시점에 대한 모형에서는 유의한 변수가 없었지만 추정된 회귀 계수의 부호에 의하면 SI 그룹이 UC 그룹보다 재흡연율이 낮고 또한 흡연 재시작 시점이 늦음을 알 수 있다. 또한 $\hat{\sigma} = 1.059\,(\text{se} = 0.105)$을 통해 프레일티 효과가 존재하며 이는 재흡연에 county 효과가 존재함을 의미한다. 이 결과는 Xiang 등 (2012)와 다른 결과를 보인다. 그들의 연구에서는 로지스틱 모형과 위험률에 독립적인 두개의 랜덤효과를 가정하였으며 두 효과 모두 유의적이지 못하다는 결론을 가져왔다. 본 연구에서는 그들이 가정한 두 랜덤 효과의 독립성대신 두 모형에 공통된 프레일티를 가정함으로써 재흡연율과 재흡연 시작 시점간의 연관관계를 모형화하고자 하였다. 회귀 계수를 비교할 때, 로지스틱 추정량은 비슷한 값을 보여주는 반면에 사건의 위험률에 대한 계수는 반대 부호를 가지는 공변량의 효과가 있었다.

6. 맺음말

본 연구에서는 낮은 사건 발생률을 가지는 군집 구간 중도절단 자료에 대해 치유율 모형(cure rate model)을 적용하였다. 특히 본 연구에서는 군집의 효과가 치유율과 사건 발생률에 미치는 효과를 모형화하기 위해 정규분포를 따르는 프레일티를 적용하였다. 본 연구에서 제안된 일변량 프레일티의 확장으로 향후 연관된 연구는 치유율과 사건 발생 확률에 각각 다른 프레일티를 적용하기 위해 이변량 정규분포를 적용해볼 수 있을 것이다. 또한 더 나아가 여기서 제안한 근사화 우도함수 방법이 아닌 다른 방법으로 중도 절단 문제를 분석해 볼 수 있을 것이다. 예를 들어, multiple imputation (Pan, 2000)은 그 대안으로 적용될 수 있다. 마지막으로 4장의 모의실험의 결과를 통해 제안된 표준 오차가 과소추정함을 보았다. 이를 해결하기 위해 Louis method가 적용되었지만 그 결과는 제안된 방법과 크게 다르지 않았다. 또 다른 방법으로는 붓스트랩 또는 multiple imputation 을 이용하는 방법 (Kim, 2006)이 고려될 수 있을 것이다.

References

Banerjee, S. and Carlin, B. P. (2004). Parametric spatial cure rate models for interval censored time-to-relapse data, *Biometrics*, **60**, 268–275.

Berkson, J. and Gage, R. P. (1952). Survival curve for cancer patients following treatment, *Journal of the American Statistical Association*, **47**, 505–515.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors, *Biometrics*, **8**, 1041–1046.

Goetghebeur, E, and Ryan, L. (2000). Semiparametric regression analysis of interval-censored data, *Biometrics*, **56**, 1139–1144.

Herring, A. H., Ibrahim, J. G. and Lipsitz, S. R. (2002). Frailty models with missing covariates, *Biometrics*, **58**, 98–109

Kim, Y. (2006). Regression analysis of doubly censored failure time data with frailty, *Biometrics*, **62**, 458–464.

Kim, Y. and Jhun, M. (2008). Cure rate model with interval censored data, *Statistics in Medicine*, **27**, 3–14.

Kuk, A. Y. C. and Chen, C. H. (1992). A mixture model combining logistic regression with proportional hazards regression, *Biometrika*, **79**, 531–541.

Lindsey, J. and Ryan, L. (1998). Methods for interval censored data. Tutorial in biostatistics, *Statistics in Medicine*, **17**, 219–138.

Maller, R. A. and Zhou, S. (1992). Estimating the presence of immune or cured individuals in censored survival data, *Biometrika*, **79**, 731–739.

Pan, W. (2000). Multiple imputation approach to Cox regression with interval censored data, *Biometrics*, **56**, 199–203.

Price, D. L. and Manatunga, A. K. (2001). Modelling survival data with a cured fraction, *Statistics in Medicine*, **20**, 1515–1527.

Sun, J. (1998). Interval Censoring, *Encyclopedia of Biostatistics*, John Wiley and Sons Ltd., 2090–2095.

Sy, J. P. and Taylor, J. M. G. (2000). Estimation in a Cox proportional hazards cure model, *Biometrics*, **56**, 227–236.

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.

Xiang, L., Ma, X. and Yau, K. (2011). Mixture cure model with random effects for clustered interval-censored survival data, *Statistics in Medicine*, **30**, 995–1006.

군집화된 구간 중도절단자료에 대한 치유율 모형의 적용

김양진[a,1]

[a]숙명여자대학교 통계학과

(2013년 10월 1일 접수, 2013년 11월 29일 수정, 2013년 12월 23일 채택)

요 약

치유율 모형(cure rate model)은 위험 그룹의 단조 감소에 대한 가정이 부적절한 경우에 적용될 수 있다. 예를 들어, 생존 분석에서 위험 그룹은 시간이 경과함에 따라 점점 감소하여 무한대의 시간대에는 영으로 수렴하며 이는 곧 생존 함수가 영으로 수렴함을 의미한다. 하지만 이러한 가정이 적합하지 못한 자료가 의약학, 사회학, 경제학 등에서 종종 발생된다. 즉, 어느 시점에 이르러 더 이상의 생존함수는 감소하지 않고 평행선을 보여주는 경우에 로그 순위검정(log rank test)과 Cox's 비례위험모형(proportional hazard model)의 적용은 바람직하지 못한 결론을 가져오게 된다. 이러한 자료에 대해 치유율 모형(cure rate model)에서는 사건 발생 취약 그룹(susceptible group)과 비취약 그룹(insusceptible group)으로 나누어 취약그룹에 대해서만 일반적인 생존 분석 방법을 적용하는 혼합 모형(mixture model)을 적용해왔다 (Berkson과 Gage, 1952). 본 연구에서는 이러한 치유율 모형을 군집화 구간 중도 절단 자료(clustered interval censored data)에 적용해 보고자 한다. 최근에 Kim과 Jhun (2008)은 구간 중도절단자료에 대해 치유율 모형을 적용하였으며 본 연구에서는 그들의 방법을 군집화 자료로 확장할 것이다. 실제 자료 분석의 예로 금연자료를 분석할 것이다.

주요용어: 군집 구간 중도 절단 자료, 금연 자료, 랜덤 효과, EM 알고리즘, 치료율 모형.

이논문은 2012년도 숙명여자대학교 교내 연구지원에 의한 것임.

[1](140-742) 서울시 용산구 청파로47길 100, 숙명여자대학교 통계학과, 조교수. E-mail: yjin@sookmyung.ac.kr