Journal of the Korean Data & Information Science Society
2016, **27**(2), 499–510

http://dx.doi.org/10.7465/jkdi.2016.27.2.499
한국데이터정보과학회지

다수준 프레일티모형 변수선택법을 이용한 다기관 방광암 생존자료분석[†‡]

김보현[1] · 하일도[2] · 이동환[3]

[1]인제대학교 부산백병원 임상시험센터 · [2]부경대학교 통계학과 · [3]이화여자대학교 통계학과

접수 2016년 2월 4일, 수정 2016년 2월 29일, 게재확정 2016년 3월 21일

요 약

생존분석 회귀모형에서 적절한 변수를 선택하는 것은 매우 중요하다. 본 논문에서는 "frailtyHL" R 패키지 (Ha 등, 2012)를 기반으로 하여 다수준 프레일티 모형 (multi-level frailty models)에서 벌점화 변수선택 방법 (penalized variable-selection method)의 절차를 소개한다. 여기서 모형 추정은 벌점화 다단계 가능도에 기초하며, 세 가지 벌점 함수 (LASSO, SCAD 및 HL)가 고려된다. 개발된 방법의 예증을 위해 벨기에 EORTC (European Organization for Research and Treatment of Cancer; 유럽 암 치료기구)에서 수행된 다국가/다기관 임상시험 자료를 이용하여 세 가지 변수 선택 방법의 결과를 비교하고, 그 결과들의 상대적 장·단점에 대해 토론한다. 특히, 자료 분석 결과에 의하면 SCAD와 HL방법이 LASSO보다 중요한 변수를 잘 선택하는 것으로 나타났다.

주요용어: 다수준 프레일티 모형, 벌점화 다단계 가능도, 벌점화 변수선택.

1. 서론

최근에 일반화 선형모형 (generalized linear models; GLMs, Nelder와 Wedderburn, 1972) 뿐만 아니라 콕스의 비례 위험 모형 (Cox's proportional hazards models; Cox, 1972)과 같은 생존분석 모형에서 벌점함수에 기초하여 벌점화 가능도를 이용한 변수 선택 방법이 폭 넓게 연구되고 있다. 벌점화 방법의 주요한 장점은 중요한 공변량을 선택함과 동시에 공변량의 회귀계수를 추정하는 것이다. 특히 이러한 방법은 중요하지 않은 변수를 0으로 회귀계수를 추정함으로써 대응하는 변수를 삭제한다. 예를 들면 LASSO (least absolute shrinkage and selection operator; Tibshirani, 1996, 1997)와 SCAD (smoothly clipped absolute deviation; Fan과 Li, 2001, 2002) 벌점 함수를 이용한 변수선택 방법이 있으며, 매우 최근에 Lee와 Oh (2014)는 HL (hierarchical likelihood) 벌점 함수를 이용한 변수선택 절차를 제안하였다.

본 논문에서는 "frailtyHL" R 통계패키지 (Ha 등, 2012)를 이용하여 다수준 프레일티 모형 (multi-level frailty models)에서의 변수선택 방법들에 관하여 연구한다. 이에 대한 이론적 절차는 프레일티 모

[†] 이 논문은 2015년도 정부 (교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. NRF-2015R1D1A3A01015663).
[‡] 본 연구는 제1저자 김보현의 부경대학교 석사학위논문의 일부를 발췌, 수정한 논문임.
[1] (333-749) 부산광역시 부산진구 복지로 75, 인제대학교 부산백병원, 임상시험센터, 연구원.
[2] (608-737) 교신저자: 부산광역시 남구 용소로 45, 부경대학교 통계학과, 교수.
E-mail: idha1353@pknu.ac.kr
[3] (120-750) 서울특별시 서대문구 이화여대길 52, 이화여자대학교 통계학과, 조교수.

형에서 Ha 등 (2014)의 벌점화 변수선택 절차에 기초한다. 다변량 생존자료 (multivariate or correlation survival-time data)의 회귀분석에서 가장 많이 사용되는 프레일티 모형은 각 개체 (subject)나 군집 (cluster)의 위험률에 대한 콕스의 준모수적 (semi-parametric) 비례위험모형에 프레일티 (frailty)를 허락한 하나의 확장된 변량효과 (random effect)에 기초한 생존분석 회귀모형이다. 여기서 프레일티는 각 개체의 위험률에 승법적으로 영향을 미치는 관측 되지 않는 변량효과 (unobserved random effect)를 의미한다. 하나의 변량효과의 항만을 갖는 공통된 (shared) 프레일티 모형에 대한 변수선택의 연구는 여러 저자들 (Fan과 Li, 2002; Androulakis 등, 2012)에 의해 개발되어 왔지만, 두 개 이상의 변량효과 항들을 가지는 다수준 프레일티 모형으로의 확장은 별로 연구가 되지 않아왔다. 그 이유 중 하나는 통상적으로 프레일티 모형에 대한 주변 가능도 함수는 프레일티 항을 제거하는데 있어서 매우 다루기 힘든 적분 계산을 요구하기 때문이다. 하지만 다단계 가능도 (hierarchical likelihood or h-likelihood; Lee와 Nelder, 1996)는 어려운 적분자체를 피할 수 있을 뿐만 아니라, 다양한 변량 효과 모형에서 통계적으로 효율적인 계산절차를 제공한다 (Lee 등, 2006). 특히 다수준 프레일티 모형은 생존자료가 지분되는 (nested) 경우에 매우 유용한 모형이며, 통상적으로 2개 이상의 변량효과가 사용된다. 여기서 중요한 변량효과의 항이 있을 경우 이를 무시하는 전통적인 통계분석법은 잘못된 추론결과를 줄 수도 있다 (Goldstein, 1995). 본 논문에서는 Ha 등 (2014)에 의한 변수선택을 효율적으로 수행하기 위해 "frailtyHL" R 패키지 (Ha 등, 2012)를 기반으로 하여 다수준 프레일티 모형에서 새로운 함수 "frailty.vs( )"를 개발하였다.

본 논문의 구성은 다음과 같다. 2절에서는 다수준 프레일티 모형의 개념 및 다단계 가능도를 기술한 후, 다단계 가능도에 기초한 벌점화 변수선택 추정방정식을 유도하고 그 추정절차를 소개한다. 여기서 세 가지 벌점화 방법 (LASSO, SCAD, HL)을 이용하여 변수선택하는 방법을 살펴보고자 한다. 3절에서는 frailtyHL 통계패키지 및 벌점화 방법을 이용하여 벨기에 EORTC에서 수행된 다국가/다기관 임상시험 (multi-country/multi-center clinical trials)을 통한 생존자료를 분석하고자 한다. 이를 위해 개발된 R 코드를 통한 모형추정 및 변수선택 방법을 설명한다. 4절에서는 변수선택 방법 및 분석결과에 대해 토론한다.

2. 다수준 프레일티 모형에서 변수선택

2.1. 모형의 형태

각 국가에 속해 있는 병원급 센터에서 환자를 모집하는 EORTC에 의한 다국가/다기관 임상시험에서 얻어지는 생존자료를 고려하자. 이와 같이 각 환자의 사건시간 (event times)들이 지분 (nested 또는 multi-level) 디자인 구조로 얻어지는 경우, 두 개의 로그프레일티 (log-frailty) 항을 고려할 수 있다. T_{ijk}를 i번째 국가에 속해 있는 j번째 센터에 대한 k번째 생존시간이라 하자. 두 개의 프레일티 항 v_i와 v_{ij}가 주어질 때, 생존시간 T_{ijk}의 조건부 위험률로 표현되는 다수준 프레일티 모형 (Yau, 2001; Ha 등, 2007)은 다음과 같이 정의된다.

$$\lambda_{ijk}(t|v_i, v_{ij}) = \lambda_0(t)\exp(x_{ijk}^T\beta + v_i + v_{ij}), \tag{2.1}$$

여기서 $\lambda_0(t)$는 미지의 기저위험함수 (unknown baseline hazard function)이며, β는 p 개의 공변량들의 벡터 $x_{ijk} = (x_{ijk1}, \ldots, x_{ijkp})^T$에 대응하는 회귀모수이다. 특히 여기서 $v_i \sim N(0, \alpha_1)$로서 i번째 국가의 로그프레일티이고, $v_{ij} \sim N(0, \alpha_2)$로서 i번째 국가에서 지분된 j번째 센터의 로그프레일티이며, v_i와 v_{ij}는 서로 독립이다 (Yau, 2001). 지분 프레일티 모형은 공통 프레일티모형에 변량효과가 하나 더 추가되는 형태이기 때문에, 이에 대한 벌점화 다단계 가능도의 추론은 다음 절에서 보이는 바와 같이 쉽게 확장이 가능하다 (Ha 등, 2007). 로그프레일티의 분포에 대해 정규분포 이외에 로그감마분포 (Ha와

Cho, 2012) 등 다른 분포를 지정할 수 있지만, 정규분포는 지분 프레일티 또는 상관된 프레일티들을 모형화 할 때 매우 유용한 장점이 있다 (Ha 등, 2007, 2011).

2.2. 다단계 가능도의 정의

지분 프레일티 모형 (2.1)에 대한 다단계 가능도 (Ha 등, 2001, 2007)는 다음과 같이 정의된다.

$$h = h(\beta, \lambda_0, \alpha) = \ell_0 + \ell_1 + \ell_2, \tag{2.2}$$

여기서

$$\ell_0 = \sum_{ijk} \log f(y_{ijk}, \delta_{ijk} | v_i, v_{ij}; \beta, \lambda_0) \quad = \quad \sum_{ijk} \delta_{ijk} \{\log \lambda_0(y_{ijk}) + \eta_{ijk}\} - \sum_{ijk} \Lambda_0(y_{ijk}) \exp(\eta_{ijk}),$$

여기서 ℓ_0는 로그 프레일티 v_i와 v_{ij}가 주어질 때, 관측되는 생존시간 y_{ijk}들의 조건부 로그가능도 (conditional log-likelihood)의 합이고,

$$\ell_1 = \sum_i \log f(v_i; \alpha_1) \ \text{ and } \ \ell_2 = \sum_i \log f(v_{ij}; \alpha_2)$$

는 각각 로그 프레일티 v_i와 v_{ij}의 로그 가능도의 합이다. 또한 여기서

$$\eta_{ijk} = x_{ij}^T \beta + v_i + v_{ij}$$

는 모형 (2.1)의 위험률에 관한 선형예측식 (linear predictor)이다. 다만 δ_{ijk}는 중도절단 여부를 나타내는 지시함수 (censoring indicator function)이다. 하지만 $\lambda_0(t)$의 함수 형태를 전혀 모르기 때문에 Ha 등 (2001)에 기초하여, 관심 모수 β의 추론을 위해 먼저 기저 누적위험함수 (baseline cumulative hazards function)를 다음과 같이 가정하였다.

$$\Lambda_0(t) = \sum_{r: y_{(r)} \le t} \lambda_{0r}$$

여기서 y_r는 y_{ijk}들 중 r번째로 작은 관측되는 생존시간이고, $\lambda_{0r} = \lambda(y_{(r)})$ 이다. 이러한 가정 하에서 λ_0를 제거한 단면 다단계 가능도 (profile h-likelihood), 즉 $h^* \equiv h|_{\lambda_0 = \hat{\lambda}_0}$ 를 사용한다. 여기서

$$\widehat{\lambda}_{0r}(\beta, v) = \frac{d_{(r)}}{\sum_{\ (i,j,k) \in R_{(r)}} \exp(\eta_{ijk})}$$

는 $\partial h / \partial \lambda_{0r} = 0$ $(r = 1, \ldots, D)$으로부터 얻어지는 추정량이며, $d_{(r)}$는 $y_{(r)}$에서의 사건들의 수이고 $R_{(r)} = R(y_{(r)}) = \{(i,j,k) : y_{ijk} \ge y_{(r)}\}$ 는 $y_{(r)}$에서의 위험집합 (risk set)이다. 따라서 h^*는 아래와 같이 표현된다 (Ha 등, 2001).

$$h^* = h^*(\beta, \hat{\lambda}_0, \alpha) = \ell_0^* + \ell_1 + \ell_2, \tag{2.3}$$

여기서

$$\ell_0^* = \sum_{ijk} \log f(y_{ijk}, \delta_{ijk} | v_i, v_{ij}; \beta, \hat{\lambda}_0) = \sum_r d_{(r)} \log \hat{\lambda}_{0r} + \sum_{ijk} \delta_{ijk} \eta_{ijk} - \sum_r d_{(r)},$$

는 λ_0에 의존하지 않는다.

나아가, 프레일티 모수 (즉 산포모수) $\alpha = (\alpha_1, \alpha_2)^T$의 추론을 위해 조정된 편 다단계 가능도 (adjusted partial h-likelihood) $p_{\beta,v_1,v_2}(h^*)$를 사용한다 (Ha와 Lee, 2003; Lee 등, 2006).

$$p_{\beta,v_1,v_2}(h^*) = \left[h^* - \frac{1}{2}\log\det\{H(h^*;\beta,v_1,v_2)/(2\pi)\}\right]\Big|_{(\beta,v_1,v_2)=(\hat{\beta},\hat{v_1},\hat{v_2})},$$

여기서 $H(h^*;\beta,v_1,v_2) = -\partial^2 h^*/\partial(\beta,v_1,v_2)(\beta,v_1,v_2)^T$ 이며, v_1과 v_2는 각각 v_i들과 v_{ij}들의 벡터이다.

2.3. 벌점화 변수선택 절차

벌점화된 단면 다단계 가능도 h_p (Ha 등, 2014)는 다음과 같이 정의된다.

$$h_p(\beta,v_1,v_2,\alpha) = h^* - n\sum_{j=1}^{p} J_\gamma(|\beta_j|), \tag{2.4}$$

여기서 (2.3)식의 h^*는 λ_0에 의존하지 않는다. $J_\gamma()$는 조율 (tuning)모수 γ를 가지는 벌점함수이다. 특히 γ의 값이 커질수록 공변량을 적게 선택하기 때문에 단순한 모형이 된다. 본 논문에서 고려하는 세 가지 벌점함수의 정의는 다음과 같다.

(i) LASSO (Tibshirani, 1996):

$$J_\gamma(|\beta|) = \gamma|\beta|.$$

(ii) SCAD (Fan과 Li, 2001):

$$J'_\gamma(|\beta|) = \gamma I(|\beta| \le \gamma) + \frac{(a\gamma - |\beta|)_+}{a-1} + I(|\beta| > \gamma),$$

여기서 $x_+ = xI(x>0)$으로서 x의 양수부분을 표시한다.

(iii) HL (Lee와 Oh, 2014);

$$J_\gamma(|\beta|) \equiv J_{(\alpha,w)} = \log\Gamma\left(\frac{1}{w}\right) + \frac{\log w}{w} + \frac{\beta^2}{2\alpha u} + \frac{(w-2)\log u(|\beta|)}{2w} + \frac{u(|\beta|)}{w},$$

여기서 $u(|\beta|) = [\{8w\beta^2/\alpha + (2-w)^2\}^{1/2} + 2 - w]/4$.

특히 Fan과 Li (2001, 2002)는 위의 SCAD식에서 $a = 3.7$을 취하는 경우 다양한 상황에서 잘 수행된다는 것을 보였다. HL 벌점함수는 w의 값에 따라 그 모양이 변화하게 되며, 특히 w가 0인 경우 ridge가 되며, w가 2인 경우 LASSO가 되며, w가 20보다 큰 경우 0에서의 값이 음의 무한대 값을 갖는 형태가 된다 (Lee와 Oh, 2014).

다수준 프레일티모형 (2.1)에 있는 (β,v_1,v_2)를 추정하기 위한 벌점화된 다단계 가능도 추정 및 변수선택은 Ha 등 (2014)의 방법에 따라 아래의 세 단계의 절차로 수행된다.

단계 1: 모수 및 프레일티의 추정

프레일티 모수 α_1과 α_2가 주어질 때, (β,v_1,v_2)의 MPHL (maximum penalized h-likelihood) 추정량은 식 (2.4)의 h_p에 근거하여 β와 (v_1,v_2)의 다음의 결합 추정방정식 (joint estimating equations)에 의해서 얻어진다.

$$\frac{\partial h_p}{\partial \beta_j} = \frac{\partial h^*}{\partial \beta_j} - n\sum_{j=1}^{p}[J_\gamma(|\beta_j|)]' = 0,$$
$$\frac{\partial h_p}{\partial v_1} = \frac{\partial h^*}{\partial v_1} = 0,$$
$$\frac{\partial h_p}{\partial v_2} = \frac{\partial h^*}{\partial v_2} = 0.$$

따라서 Ha와 Lee (2003)의 결합 추정방정식을 이용하면 벌점화된 다단계 가능도 추정식은 다음과 같이 표현됨을 보일 수 있다.

$$\begin{pmatrix} X^TWX + n\sum_\gamma & X^TWZ_1 & X^TWZ_2 \\ Z_1^TWX & Z_1^TWZ_1 + U_1 & Z_1^TWZ_2 \\ Z_2^TWX & Z_2^TWZ_1 & Z_2^TWZ_2 + U_2 \end{pmatrix} \begin{pmatrix} \hat\beta \\ \hat v_1 \\ \hat v_2 \end{pmatrix} = \begin{pmatrix} X^Tw \\ Z_1^Tw \\ Z_2^Tw \end{pmatrix}, \qquad (2.5)$$

여기서 X, Z_1과 Z_2는 각각 β, v_1와 v_2의 모형행렬이다. $\sum_\gamma = \mathrm{diag}\{J'_\gamma(|\beta_j|)/|\beta_j|\}$는 p차 대각행렬이며, $W = -\partial^2 h^*/\partial\eta\partial\eta^T$, $U_1 = -\partial^2\ell_1/\partial v_1^2$, $U_2 = -\partial^2\ell_1/\partial v_2^2$, $w = W\eta + (\delta - \mu)$, $\eta = X\beta + Z_1v_1 + Z_2v_2$, μ는 $\mu_{ijk} = \Lambda_0(y_{ijk})\exp(\eta_{ijk})$들의 n차 벡터이다. 식 (2.2)의 h에서 로그-프레일티 v_1 또는 v_2 중 하나가 없으면 식 (2.5)는 공통된 (shared)프레일티모형에서 벌점화 변수선택 추정식이 되며, (v_1, v_2) 둘 다 없으면 그것은 콕스 비례위험모형에서 벌점화 변수선택 추정식이 된다.

프레일티의 모수 $\alpha = (\alpha_1, \alpha_2)^T$를 추정하기 위해 h_p로부터 $\tau = (\beta^T, v_1^T, v_2^T)^T$가 제거된 $p_\tau(h_p)$를 이용한다. 대응하는 추정방정식은 각각 다음과 같다.

$$\frac{\partial p_\tau(h_p)}{\partial \alpha_1} = 0 \ \text{ and } \ \frac{\partial p_\tau(h_p)}{\partial \alpha_2} = 0,$$

여기서 $p_\tau(h_p) = [h_p - (1/2)\log\det\{H(h_p,\tau)/(2\pi)\}]|_{\tau=\hat\tau}$, $H(h_p,\tau) = -\partial^2 h_p/\partial\tau\partial\tau^T$이고, $\hat\tau$는 $\partial h_p/\partial\tau = 0$의 해이다.

단계 2: 조율모수의 선택

다음으로, Bayesian information criterion (BIC) 형태의 한 기준을 이용하여 조율모수 (tuning parameter) γ를 선택한다 (Wang 등, 2007; Ha 등, 2014):

$$\hat\gamma = \mathrm{Argmin}_\gamma\{\mathrm{BIC}(\gamma)\},$$

여기서

$$\mathrm{BIC}(\gamma) = -2p_{v_1,v_2}(h_p) + e(\gamma)\log(n),$$

그리고 $e(\gamma) = \mathrm{tr}[\{H_{\beta\beta} + n\sum_\gamma\}^{-1}H_{\beta\beta}]$이다. 여기서 $H_{\beta\beta} = -\partial^2\hat h/\partial\beta\partial\beta^T$이고 $\hat h = h^*|_{(v_1=\hat v_1, v_2=\hat v_2)}$이다. 위의 BIC는 회귀모수 β의 선택에 초점을 두기 때문에, 프레일티 모수 α에 초점을 두는 $p_\tau = p_{\beta,v_1,v_2}$ 대신에 p_{v_1,v_2}를 사용한다 (Ha등, 2014). 조율모수 γ는 BIC를 최소로 하는 하나의 간단한 grid 방법에 의해 선택될 수 있다 (Fan과 Li, 2002).

단계 3: 표준오차의 계산

마지막으로, 단계 1과 2의 수렴 후 $\hat\beta$에 대한 표준오차 (standard error; SE)를 다음의 sandwich 분산-공분산행렬로부터 구한다.

$$\begin{aligned}\mathrm{Cov}(\hat{\beta}) &= \left\{-\frac{\partial^2 \hat{h}_p}{\partial\beta\partial\beta^T}\right\}^{-1} \mathrm{cov}\left(\frac{\partial \hat{h}_p}{\partial\beta}\right)\left\{-\frac{\partial^2 \hat{h}_p}{\partial\beta\partial\beta^T}\right\}^{-1} \\ &= \left(H_{\beta\beta} + n\sum\nolimits_{\gamma}\right)^{-1} H_{\beta\beta}\left(H_{\beta\beta} + n\sum\nolimits_{\gamma}\right)^{-1},\end{aligned}$$

여기서 $\hat{h}_p = h_p|_{(v_1=\hat{v}_1, v_2=\hat{v}_2)} = \hat{h} - n\sum_{j=1}^{p} J_\gamma(|\beta_j|)$이다.

3. 변수선택 실증분석: 다기관 방광암 생존자료

3.1. 자료의 구조 및 설명

본 연구에서 사용된 자료는 벨기에 EORTC에서 주관하여 유럽의 13개 국가의 57개의 센터에서 수집된 방광암 환자 1066명의 임상시험 생존자료를 대상으로 하였다 (Oddens 등, 2013). 초기 경요도 절제술 (transurethral reaction; TUR) 후 비근침윤성 방광암 (non-muscleinvasive bladder concer; NMIBC)은 재발의 위험이 높다는 것이 특징이고, 보다 더 작은 정도로 근침윤성 방광암으로 진행할 수도 있다. 따라서 방광암은 재발이나 진행을 막기 위하여 방광 내 보조항암요법 혹은 BCG (bacille de Calmette-Guerin)를 이용한 방광 내 면역요법을 시행하게 된다. 중·고위험군의 환자들은 항암요법보다는 BCG면역요법의 치료 효과가 더 뚜렷하다. 따라서 적절한 BCG 면역요법의 치료효과를 위하여 유지요법 기간을 따르는 것이 필요하다. 초기에는 6주간의 유지요법이 권장되었다. 후에 BCG 를 추가 투여할수록 재발률을 감소시킨다는 것이 발견되었다. 따라서 유지요법의 기간은 아직까지도 정확히 정해진 것은 없지만, EAU (European Association of Urology) 가이드라인은 적어도 1년간의 유지요법을 할 것을 권장하고있다.

본 연구에서는 BCG의 단점인 독성을 감소시키기 위해 1/3투여와 모두 투여, 유지요법은 1년과 3년으로 설계되었다. 특히 본 연구에서는 비근침윤성 방광암환자들이 경요도 절제술 후 재발에 영향을 미치는 위험인자 (공변량)에 대하여 알아보고자 한다. 본 자료분석에서는 고위험군의 환자 70세이상인 466명을 대상으로 했고, 나라 수는 12개, 기관수는 39개이다. Table 3.1은 분석에 이용된 변수의 설명이다.

Table 3.1 Explanation of variables for the bladder cancer data

Variable	Explanation
timeDFI	Time to first recurrence after surgery (day)
statusDIF	Indicator of recurrence of bladder cancer (0: No, 1: Yes)
Trtdose	Amount of dose of BCG (1: 1/3 dose BCG, 2: Full dose BCG)
Trtduration	Duration of maintenance (0: 1 year, 1: 3 years)
Age	year
Gender	0: Male, 1: Female
TypeBC	Type of bladder cancer (0: Primary, 1: Recurrent)
Tumsize	Largest tumor diameter (mm)
Nbtum	No. of tumors
Tstage	T category of bladder cancer (0: pTa , 1: pT1)
Ggrade	WHO grade of bladder cancer (1: G1, 2: G2 , 3: G3)

Table 3.2는 분석자료에 대한 기초 통계량이다. 여기서 연속형 자료인 경우는 평균 (mean), 중앙값 (median) 및 범위 (range)를, 이산형 자료인 경우 빈도 (frequency) 및 퍼센트를 이용하여 요약하였다. 특히 재발사건에 대한 중도절단 비율 (censoring rate)는 43.35%로 나타났다.

Table 3.2 Basic statistics of the bladder cancer data

Variable	Basic statistic
TimeDFI (day)	
Mean	1314.39
Median	93.5
Range	2.0-4743.0
statusDIF	
No	264(56.65)
Yes	202(43.35)
Trtdose	
1/3 dose BCG	245(52.58)
Full dose BCG	221(47.42)
Trtduration	
1 year	221(47.42)
3 years	245(52.58)
Age	
Mean	75.50
Median	75
Range	70-85
Gender	
Male	382(81.97)
Female	84(18.03)
TypeBC	
Primary	260(55.79)
Recurrent	206(44.21)
Tumsize	
Mean	18.14
Median	15
Range	2-98
Nbtum	
Mean	2.95
Median	2
Range	1-10
Tstage	
pTa	279(59.87)
pT1	187(40.13)
Ggrade	
G1	122(26.18)
G2	202(43.35)
G3	142(30.47)

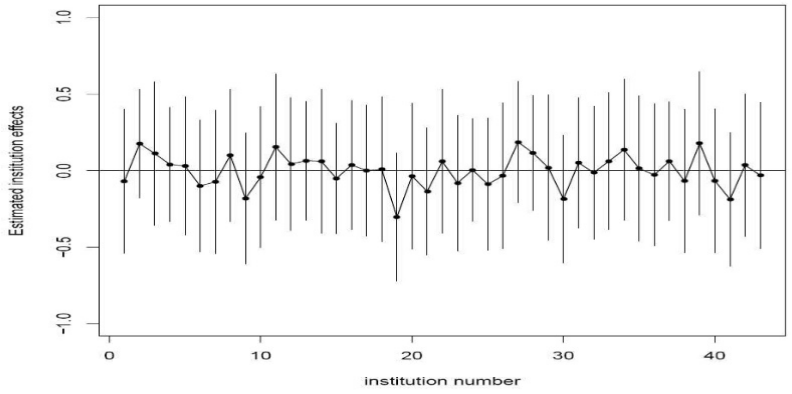

Figure 3.1 Multi-center bladder cancer data: 95% confidence intervals for log-frailties of 39 centers

3.2. 모형적합 및 변수선택

본 논문에서 주요한 관심사항은 다기관 방광암의 재발시간에 영향을 미치는 공변량이 무엇인지를 알아보는 것이기 때문에 다수준 프레일티 모형 적합을 통해 적절한 변수를 선택하려고 한다. 3.1절의 방광암 재발생존자료는 다국가에서 다기관으로 지분된 형태의 자료이기 때문에 다수준 프레일티 모형을 적합하여 세 가지 변수선택법을 적용하였다. 분석 결과에 의하면, BIC기준에서 선택된 조율모수 γ의 값은 LASSO, SCAD, HL에서 각각 0.02, 0.06, (w,α)=(3,0.011)이었다. 여기서 먼저 LASSO에 의하면 국가의 로그-프레일티 분산은 $\hat{\alpha}_1 = 0.00001$ 이고 기관의 로그 프레일티분산은 $\hat{\alpha}_2 = 0.05968$ 으로서 국가효과는 매우 약한것으로 여겨진다. 이러한 국가효과는 No-penalty, LASSO, SCAD에서도 이와 유사한 값을 주었다. 한편 기관의 로그 프레일티 분산 $\hat{\alpha_2}$은 No-penalty, LASSO, SCAD, HL 하에서 각각 0.069, 0.060, 0.068, 0.061 으로 추정되었다.

Table 3.3 Bladder cancer data: Estimated regression coefficients (standard errors) via variable selection in multi-level frailty models

Variable	No-penalty	LASSO	SCAD	HL
trtdose	-0.175 (0.145)	0 (0)	0 (0)	0 (0)
trtduration	-0.338 (0.144)	-0.132 (0.058)	0 (0)	-0.141 (0.062)
age	0.166 (0.073)	0.123 (0.052)	0.016 (0.007)	0.122 (0.052)
gender	0.008 (0.183)	0 (0)	0 (0)	0 (0)
typeBC	0.628 (0.167)	0.344 (0.093)	0.557 (0.151)	0.393 (0.101)
tumsize	-0.002 (0.006)	-0.004 (0.005)	0 (0)	0 (0)
nbtum	0.237 (0.070)	0.198 (0.056)	0.227 (0.068)	0.200 (0.056)
tstage	-0.144 (0.200)	0 (0)	0 (0)	0 (0)
gg1	-0.520 (0.252)	-0.043 (0.024)	0 (0)	0 (0)
gg2	-0.214 (0.214)	0 (0)	0 (0)	0 (0)

본 논문의 주 관심사항인 공변량에 대한 변수 선택결과는 Table 3.3에 제시되어 있다. No-penalty 하에서 공변량 trtduration, age, typeBC, nbtum, gg1 이 유의했다. LASSO방법은 9개 공변량 중 6개의 공변량 (trtduration, age, typeBC, tumsize, nbtum, gg1)이 선택되었지만, SCAD와 HL에서는 3개의 공변량 (age, typeBC, nbtum)만이 선택되었다. 여기서 LASSO는 No-penalty에서 유의하지 않은 변수 tumsize를 하나 더 선택함을 관측할 수 있다. Figure 3.1은 HL방법에 기초한 39개 기관별 로그-프레일티의 95% 신뢰구간 (Ha 등, 2011; Ha와 Noh, 2013)을 보인다. 39개의 신뢰구간은 모두 0의 프레일티 값을 포함하므로 기관별로 기저위험에 대한 이질성은 전반적으로 거의 없는 것으로 평가된다.

3.3. 변수선택을 위한 R 코드 및 설명

3.2절에서 제시된 Ha 등 (2014)의 변수선택을 효율적으로 수행하기 위해 본 논문에서는 “frailtyHL” R 패키지 (Ha 등, 2012)를 기반으로 하여 새로운 함수, 즉 “frailty.vs()”를 개발하였다. 먼저 LASSO 절차는 아래에서 제시된 바와 같이 회귀계수의 초기치 “B”와 조율모수 “tun1”을 지정함으로써 구현된다. LASSO에서 초기치는 No-penalty 하의 회귀 추정치를 사용한다. 여기서 No-penalty하의 다수준 프레일티 모형의 적합은 “B=(0,0,0,0,0,0,0,0,0,0)”과 “tun1=0”의 지정을 통해 쉽게 구현된다.

```
##################### Multi-level frailty model (LASSO) ######################
> library(frailtyHL)
> source("frailtyVS.R")
> eortc<-read.table("eortcdata.csv",sep=",",header=T)
> eortc$age<-(eortc$age-mean(eortc$age))/sd(eortc$age)
> la_eortc<-frailty.vs(Surv(timeDFI, statusDFI) ~ trtdose + trtduration + age +
  +gender + typeBC + tumsize + nbtum + tstage + gg1 + gg2 +
  +(1|Country)+(1|institution), model="lognorm", penalty="lasso", data=eortc,
  B=c(-0.175,-0.338,0.043,0.008,0.628,-0.002,0.129,-0.144,-0.520,-0.214),
  tun1=seq(0,0.1,0.01))
> eortc$nbtum<-(eortc$nbtum-mean(eortc$nbtum))/sd(eortc$nbtum)
######################### Output of LASSO ######################################
[1] "Result of variable selection in frailty model"
[1] "==Fitted model=="
[1] "model : lognorm"
[1] "penalty : lasso"
[1] "formula :"
Surv(timeDFI, statusDFI) ~ trtdose + trtduration + age + gender +
    typeBC + tumsize + nbtum + tstage + gg1 + gg2 + (1|Country) + (1|institution)
[1] "converge"
[1] "==Fixed coefficients=="
            Estimate Std. Error
trtdose      0.00000    0.00000
trtduration -0.13167    0.05831
age          0.12278    0.05159
gender       0.00000    0.00000
typeBC       0.34433    0.09324
tumsize     -0.00395    0.00544
nbtum        0.19820    0.05585
tstage       0.00000    0.00000
gg1         -0.04319    0.02422
gg2          0.00000    0.00000
[1] "==Dispersion parameter=="
[1]  0.00001  0.05968
[1] "==Tuning parameter=="
[1] 0.02
[1] "==BIC=="
[1] 2298.3
```

다음으로, SCAD와 HL의 초기치는 LASSO의 추정치를 사용하며 (Ha 등, 2014), 대응하는 R코드는 다음과 같다. 특히 HL은 두 개의 조율모수 “tun1”과 “tun2”를 지정함으로써 구현된다.

```
#################### Multi-level frailty model (SCAD) ############################
> sc_eortc<-frailty.vs(Surv(timeDFI, statusDFI) ~ trtdose + trtduration + age
  +gender + typeBC + tumsize + nbtum + tstage + gg1 + gg2
  +(1|Country)+(1|institution), model="lognorm", penalty="scad", data=eortc,
  B=c(0,-0.132,0.123,0,0.344,-0.004,0.120,0,-0.043,0), tun1=seq(0,0.1,0.01))
#################### Multi-level frailty model (HL) ##############################
> hl_eortc<-frailty.vs(Surv(timeDFI, statusDFI) ~ trtdose + trtduration + age
 +gender + typeBC + tumsize + nbtum + tstage + gg1 + gg2
 +(1|Country)+(1|institution), model="lognorm", penalty="hl", data=eortc,
 B=c(0,-0.132,0.123,0,0.344,-0.004,0.120,0,-0.043,0),
 tun1=c(2.1,3,10,30,50), tun2=seq(0.001,0.1,0.01))
```

4. 결론 및 토론

본 논문에서는 다수준 프레일티 모형에서 벌점화 다단계 가능도에 기반하여 변수선택을 위한 R 패키지를 개발하였다. 예증을 위해 다국가/다기관 임상시험 생존자료에 적용하였다. Table 3.3의 분석결과 LASSO 방법은 중요하지 않은 변수를 선택하는 경향이 있는 반면, SCAD와 HL 방법이 전반적으로 적절한 변수를 유사하게 선택하는 것으로 나타났다. 이러한 경향은 다른 생존자료의 분석에도 유사한 결과를 보였다 (Ha 등, 2014).

3절 실증분석에서 프레일티의 분산은 0을 포함하기 때문에 통상적인 카이제곱 분포에 근거하여 우도비 검정을 할 수 없으므로, 0에서의 경계값 (boundary value)을 고려하는 mixture 카이제곱 분포에 근거하여 국가나 기관간 강도의 유의성을 검토 할 필요가 있다 (Ha 등, 2011). Table 3.3에서는 주어진 모형에서 변수선택방법을 예증한 것이지만, 차후 모형선택 방법을 개발하는 것도 하나의 흥미있는 연구대상이 될 것으로 사료된다. 본 논문에서 사용된 생존자료는 공변량의 수가 다소 작기 때문에, 차후 공변량의 수(p) 자체가 매우 많은 경우 뿐만 아니라, 공변량 수가 표본 수보다 큰 경우 (즉 $p > n$)에 개발된 패키지를 적용하여 연구하는 것이 필요하다고 사료된다 (Lee, 2015). 이러한 다차원 자료 (high-dimensional data)의 경우 HL 방법이 보다 효율적인 변수선택을 제공하는 장점이 있다 (Lee 등, 2011; Lee와 Oh, 2014). 현재 개발된 본 패키지의 구현은 "frailtyHL"에 하나의 source 파일이 요구되지만, 본 논문의 교신저자를 통해 그 이용이 가능하다.

References

Androulakis, E., Koukouvinos, C. and Vonta, F. (2012). Estimation and variable selection via frailty models with penalized likelihood. *Statistics in Medicine*, **31**, 2223-2239.

Cox, D. R. (1972). Regression models and life tables (with Discussion). *Journal of the Royal Statistical Society, Series B*, **74**, 187-220.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348-1360.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics*, **30**, 74-99.

Goldstein, H. (1995). *Multilevel statistical models*, Arnold, London.

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**, 199-207.

Ha, I. D. and Lee, Y. (2003). Estimating frailty models via Poisson hierarchical generalized linear models. *Journal of Computational and Graphical Statistics*, **12**, 663-681.

Ha, I. D., Lee, Y. and Song, J. K. (2001). Hierarchical likelihood approach for frailty models. *Biometrika*, **88**, 233-243.

Ha, I. D., Lee, Y. and MacKenzie, G. (2007). Model selection for multi-component frailty models. *Statistics in Medicine*, **26**, 4790-4807.

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.

Ha, I. D., Noh, M. and Lee, Y. (2012). FrailtyHL: A package for fitting frailty models with h-likelihood. *The R Journal*, **4**, 28-36.

Ha, I. D. Pan, J., Oh, S. and Lee, Y. (2014). Variable selection in general frailty Models using penalized h-Likelihood. *Journal of Computational and Graphical Statistics*, **23**, 1044-1060.

Ha, I. D., Sylvester, R., Legrand, C. and MacKenzie, G. (2011). Frailty modelling for survival data from multi-centre clinical trials. *Statistics in Medicine*, **30**, 2144-2159.

Lee, S. (2015). A note on standardization in penalized regressions. *Journal of the Korean Data & Information Science Society*, **26**, 505-516.

Lee, W., Lee, D., Lee, Y. and Pawitan, Y. (2011). Sparse canonical covariance analysis for high-throughput data. *Statistical Applications in Genetics and Molecular Biology*, **10**, 1-24.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society B*, **58**, 619-678.

Lee, Y., Nelder, J. A. and Pawitan, Y. (2006). *Generalised linear models with random effects: Unified analysis via h-likelihood*, Chapman and Hall, London.

Lee, Y. and Oh, H. S. (2014). A new sparse variable selection via random-effect model. *Journal of Multivariate Analysis*, **125**, 89-99.

Oddens, J., Brausi, M., Sylvester, R., Bono, A., Bono, A., Beek, C.V.D., Andel, G.V., Gontero P., Hoeltl, W., Turkeri, L., Marreaud, S., Collette, S. and Oosterlinck, W. (2013). Final results of an EORTC-GU cancers group randomized study of maintenance Bacillus Calmette-Gue´rin in intermediate- and highrisk Ta, T1 papillary carcinoma of the urinary bladder: One-third dose versus full dose and 1 year versus 3 years of maintenance. *European Urology*, **63**, 462-472.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society B*, **58**, 267-288.

Tibshirani, R. (1997). The LASSO method for variable selection in the Cox model. *Statistics in Medicine*, **16**, 385-395.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society A*, **135**, 370-384.

Wang, H., Li, R. and Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, **94**, 553-568.

Yau, K. K. W. (2001). Multilevel models for survival analysis with random effects. *Biometrics*, **57**, 96-102.

Journal of the Korean Data & Information Science Society
2016, **27**(2), 499–510

http://dx.doi.org/10.7465/jkdi.2016.27.2.499
한국데이터정보과학회지

Analysis of multi-center bladder cancer survival data using variable-selection method of multi-level frailty models[†‡]

Bohyeon Kim[1] · Il Do Ha[2] · Donghwan Lee[3]

[1]Clinical Trial Center, Busan Paik Hospital of Inje University
[2]Department of Statistics, Pukyong National University
[3]Department of Statistics, Ewha Womans University

Received 4 February 2016, revised 29 February 2016, accepted 21 March 2016

Abstract

It is very important to select relevant variables in regression models for survival analysis. In this paper, we introduce a penalized variable-selection procedure in multi-level frailty models based on the "frailtyHL" R package (Ha *et al.*, 2012). Here, the estimation procedure of models is based on the penalized hierarchical likelihood, and three penalty functions (LASSO, SCAD and HL) are considered. The proposed methods are illustrated with multi-country/multi-center bladder cancer survival data from the EORTC in Belgium. We compare the results of three variable-selection methods and discuss their advantages and disadvantages. In particular, the results of data analysis showed that the SCAD and HL methods select well important variables than in the LASSO method.

Keywords: Multi-level frailty models, penalized hierarchical-likelihood, penalized variable selection.

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. NRF-2015R1D1A3A01015663).
[‡] This paper is a condensed form of the first author's master thesis from the Pukyong National University, Busan, Korea.
[1] Researcher, Clinical Trial Center, Busan Paik Hospital, Inje University, Busan 614-735, Korea.
[2] Corresponding author: Professor, Department of Statistics, Pukyong National University, Busan 608-737, Korea. E-mail: idha1353@pknu.ac.kr
[3] Assistant professor, Department of Statistics, Ewha Womans University, Seoul 120-750, Korea.