The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.5.965
28(5), 965–976

Variable Selection in Frailty Models using FrailtyHL R Package: Breast Cancer Survival Data

Bohyeon Kim[a] · Il Do Ha[a,1] · Maengseok Noh[a] · Myung Hwan Na[b] · Ho-Chun Song[c] · Jahae Kim[c]

[a]Department of Statistics, Pukyong National University;
[b]Department of Statistics, Chonnam National University;
[c]Department of Nuclear Medicine, Chonnam National University Hospital

(Received July 27, 2015; Revised July 31, 2015; Accepted August 6, 2015)

Abstract

Determining relevant variables for a regression model is important in regression analysis. Recently, a variable selection methods using a penalized likelihood with various penalty functions (e.g. LASSO and SCAD) have been widely studied in simple statistical models such as linear models and generalized linear models. The advantage of these methods is that they select important variables and estimate regression coefficients, simultaneously; therefore, they delete insignificant variables by estimating their coefficients as zero. We study how to select proper variables based on penalized hierarchical likelihood (HL) in semi-parametric frailty models that allow three penalty functions, LASSO, SCAD and HL. For the variable selection we develop a new function in the "frailtyHL" R package. Our methods are illustrated with breast cancer survival data from the Medical Center at Chonnam National University in Korea. We compare the results from three variable-selection methods and discuss advantages and disadvantages.

Keywords: frailty models, H-likelihood, LASSO, SCAD, Variable selection

1. 서론

회귀분석모형에서 적절한 변수를 선택하는 것은 매우 중요하다. 일반 회귀 모형에서는 전진선택법(forward selection), 후진제거법(backward elimination), 단계적 선택법(stepwise selection)과 같은 변수선택을 위한 다양한 고전적인 방법들이 있다. 그러나 이러한 방법들은 공변량의 개수가 클 때 과도한 계산이 요구되며 종종 높은 변동성을 주는 단점이 있다 (Breiman, 1996; Fan과 Li, 2001).

최근 고전적 회귀모형, 일반화 선형모형(generalized linear models; GLMs; Nelder과 Wedderburn, 1972), 콕스의 비례 위험 모형(Cox's proportional hazards models; Cox, 1972)과 같은 다양한 통계적 모형에서 벌점함수에 기초하여 벌점화 가능도를 이용한 변수 선택 방법이 폭 넓게 연구되고 있다. 벌점화 방법의 주요한 장점은 중요한 공변량을 선택함과 동시에 공변량의 회귀계수를 추정하는 것이

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology, Korea (No. 2010-0021165).
[1]Corresponding author: Department of Statistics, Pukyong National University, Busan 608-737, Korea.
E-mail: idha1353@pknu.ac.kr

다. 그러므로 이 방법들은 0으로 회귀계수를 추정함으로써 중요하지 않은 변수를 삭제한다. 예를 들면 LASSO(least absolute shrinkage and selection operator; Tibshirani, 1996, 1997), SCAD(smoothly clipped absolute deviation; Fan과 Li, 2001, 2002) 그리고 HL(hierarchical likelihood; Lee와 Oh, 2014) 벌점 함수를 이용한 변수선택 방법들이 있다.

본 논문에서는 생존분석모형에서 자주 사용되는 콕스 비례위험모형의 한 확장인 프레일티 모형(frailty models; Clayton, 1991; Hougaard, 2000)에서의 변수선택 방법들에 관하여 연구한다. 여기서 프레일티는 각 개체의 위험률에 승법적으로 영향을 미치는 관측 안되는 변량효과(unobserved random effect)를 의미한다. Fan과 Li (2002)는 감마 프레일티 모형에 대해 SCAD 벌점함수를 이용하여 벌점화된 주변 가능도(penalized marginal likelihood) 방법을 제안했다. 최근에 Androulakis 등 (2012)은 이를 역 정규 분포(inverse gaussian distribution)와 같은 또 다른 프레일티 분포로 확장하였다. 일반적으로 프레일티 모형에 대한 주변 가능도 함수는 프레일티를 제거하는데 있어서 매우 다루기 힘든 적분 계산을 요구한다. 하지만 다단계 가능도(Hierarchical likelihood 또는 h-likelihood; Lee와 Nelder, 1996)는 어려운 적분자체를 피할 수 있을 뿐만 아니라, 다양한 변량 효과 모형에서 통계적으로 효율적인 추론절차를 제공한다 (Lee 등, 2006). 따라서 매우 최근에 Ha 등 (2014)은 다단계 가능도에 근거하여, 다양한 프레일티 구조를 허락하는 일반적인 프레일티 모형에 대해 벌점화 변수선택 방법을 제안하였다. 본 논문에서는 Ha 등 (2014)에 의한 변수선택을 효율적으로 수행하기 위해 "frailtyHL" R 패키지 (Ha 등, 2012)를 기반으로 하여 새로운 함수("frailty.vs( )")를 개발하였다.

본 논문의 구성은 다음과 같다. 2절에서는 프레일티 모형의 개념에 대해 설명하고, 다단계 가능도에 대한 변수선택 방법을 리뷰한다. 3절에서는 frailtyHL 통계패키지를 이용하여 최근에 수집된 유방암 재발 생존자료를 분석한다. 이를 위해 개발된 R 코드를 통한 모형추정 및 변수선택 방법을 설명한다. 4절에서는 변수선택 방법 및 분석결과에 대해 토론한다. 마지막으로 4절의 자료분석에 사용된 변수선택 R 코드를 부록에 제시한다.

2. 프레일티 모형에서 변수선택

2.1. 모형의 기본개념 및 형태

각 개인의 공통된 프레일티 u_i가 주어질 때, $i\,(i = 1, \ldots, q)$번째 개인의 j번째 관측에 대한 생존시간 T_{ij}의 조건부 위험률(conditional hazard rate)은 다음과 같이 표현된다.

$$\lambda_{ij}\left(t|u_i; x_{ij}\right) = \lambda_0(t) \exp\left(x_{ij}^T \beta\right) u_i, \tag{2.1}$$

여기서 $\lambda_0(t)$는 미지의 기저 위험함수(unknown baseline hazard function)이며, β는 T_{ij}에 대한 p개의 공변량들의 벡터 $x_{ij} = (x_{ij1}, \ldots, x_{ijp})^T$에 대응하는 회귀모수이다. 프레일티 u_i는 통상적으로 감마분포 또는 로그정규분포를 지정할 수 있다. 특히 로그정규분포의 경우 로그-프레일티 $v_i = \log u_i$를 정규분포로 지정한다.

즉, 감마 프레일티 모형과 로그정규 프레일티 모형은 프레일티 분포에 대해 각각 감마분포와 로그정규분포를 가정한다. 따라서 감마프레일티 모형에 대해 프레일티 u_i에 대해서는 평균이 1이고 분산이 α인 감마분포를, 그리고 로그 정규프레일티 모형에 대해서는 로그프레일티 $v_i(= \log u_i)$를 정규분포, 즉 $v_i \sim N(0, \alpha)$로 각각 지정한다. 만약 모든 개인들의 $u_i = 1$이면, 프레일티 모형은 콕스의 비례위험 모형이 된다. 프레일티 u_i의 의미는 u_i가 가지는 값에 따라 약간 달라진다. 만약 i번째 개인의 프레일티 u_i의 값이 1 (즉 $u_i = 1$)인 사람은 위험률이 표준적인 경향에 있다고 말한다. 만약 $u_i < 1$인 사람은 표준적인 사람 (즉 $u_i = 1$)보다 위험률이 낮은 경향을 가지며, $u_i > 1$인 사람은 표준적인 사람보다 위험률

이 높은 경향을 가진다 라고 말한다 (Hougaard, 2000). u_i는 관측 불가능 하지만 추정은 가능하기 때문에, 이러한 의미는 실제자료를 다룬 3.2절에서 u_i의 추정치들을 사용하여 Figure 3.1에서 시각적으로도 또한 표현된다.

프레일티 모형은 프레일티라는 관측 안 되는 변량효과를 통해 단변량 생존자료에서는 개체 간 이질성을 설명하고, 다변량 생존자료에서는 개체간 자료의 이질성 뿐만 아니라 개체 내 자료의 의존성(dependency)을 모형화 해 준다. 특히 콕스모형은 프레일티 모형에서 모든 로그 프레일티(log-frailty)의 값이 0 (즉 로그프레일티의 분산이 0)인 경우에 해당되므로, 콕스모형은 프레일티 모형을 이용해서 바로 적합할 수 도 있다 (Ha 등, 2012).

2.2. 다단계 가능도 추정법

프레일티 모형의 추론을 위해 다양한 방법들이 제안되어 왔다. 대표적인 추론법을 요약하면 다음과 같다. (i) 몬테카를로 EM(MCEM)을 이용한 주변 가능도 방법 (Vaida와 Xu, 2000), (ii) 벌점 편 가능도 방법(penalized partial likelihood; Ripatti와 Palmgren, 2000), (iii) 베이지안 접근법 (Clayton, 1991; Legrand 등, 2005)이 있다. 하지만 이러한 방법들은 종종 어려운 적분계산이 요구된다. 따라서 본 논문에서는 Lee와 Nelder (1996)에 의해 제안된 다단계 가능도를 사용한다. 이 방법은 복잡한 적분계산을 피하고 통계적으로 효율적이면서 통합된 추론 틀을 제공하는 장점이 있다 (Lee 등, 2006).

프레일티 모형 (2.1)에 대한 다단계 가능도 (Ha 등, 2001)의 정의는 다음과 같다.

$$h = h\left(\beta, \lambda_0, \alpha\right) = \ell_0 + \ell_1,$$

여기서

$$\begin{aligned}\ell_0 &= \sum_{ij} \log f\left(y_{ij}, \delta_{ij} | u_i; \beta, \lambda_0\right) \\ &= \sum_{ij} \delta_{ij} \left\{\log \lambda_0(y_{ij}) + \eta_{ij}\right\} - \sum_{ij} \Lambda_0(y_{ij}) \exp(\eta_{ij}),\end{aligned}$$

여기서 ℓ_0는 프레일티 u_i가 주어질 때 관측되는 확률변수 (y_{ij}, δ_{ij})들의 조건부 로그가능도(conditional log-likelihood)의 합이고,

$$\ell_1 = \sum_i \log f(v_i; \alpha)$$

는 로그 프레일티 v_i의 로그-가능도의 합이다. 여기서 $\eta_{ij} = x_{ij}^T\beta + v_i$는 위험률에 관한 선형예측식(linear predictor)이다. 다만 y_{ij}는 관측되는 생존시간이며 δ_{ij}는 중도절단 여부를 나타내는 지시함수이다. 하지만 $\lambda_0(t)$의 함수 형태를 전혀 모르기 때문에 관심 모수 β의 추론을 위해 Breslow (1972)와 Ha 등 (2001)의 아이디어에 따라 기저 누적위험함수(baseline cumulative hazards function)를 다음과 같은 계단함수 형태를 가정 한다:

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

여기서 $y_{(k)}$는 y_{ij}들 중 k번째로 작은 관측되는 생존시간이고, $\lambda_{0k} = \lambda_0(y_{(k)})$이다. 이러한 가정 하에서 장애모수(nuisance parameters) λ_0를 제거한 단면 다단계 가능도(profile h-likelihood), 즉 $h^* \equiv h|_{\lambda_0 = \hat{\lambda}_0}$를 사용한다. 따라서 h^*는 아래와 같이 표현 된다 (Ha 등, 2001).

$$h^* = h^*(\beta, \lambda_0, \alpha) = \ell_0^* + \ell_1,$$

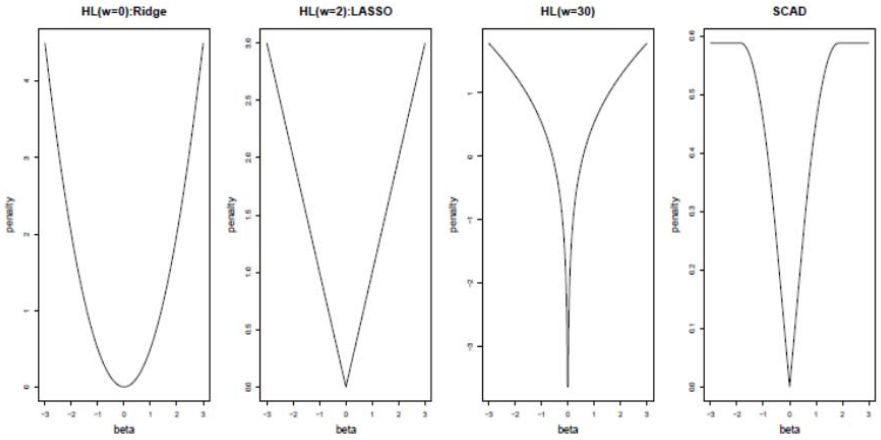

Figure 2.1. Various penalty functions

여기서

$$\begin{aligned}\ell_0^* &= \sum_{ij} \log f\left(y_{ij}, \delta_{ij} | u_i; \beta, \hat{\lambda}_0\right) \\ &= \sum_k d_{(k)} \log \hat{\lambda}_{0k} + \sum_{ij} \delta_{ij}\eta_{ij} - \sum_k d_{(k)}\end{aligned}$$

는 λ_0에 의존하지 않으며, $\hat{\lambda}_{0k}(\beta, v) = d_{(k)}/\sum_{i,j\in R_{(k)}} \exp(\eta_{ij})$는 대응하는 추정방정식 $\partial h/\partial\lambda_{0k} = 0$ $(k = 1, \ldots, D)$으로부터 얻어지는 추정량이다. 여기서 $d_{(k)}$는 $y_{(k)}$에서의 사건들의 수이고 $R_{(k)} = R(y_{(k)}) = \{(i,j) : y_{ij} \geq y_{(k)}\}$는 $y_{(k)}$에서의 위험집합(risk set)이다.

$v = (v_1, \ldots, v_q)^T$를 v_i들의 $q \times 1$ 벡터라 하면, $\tau = (\beta^T, v^T)^T$는 $(p+q) \times 1$ 백터로 표시할 수 있다. 프레일티 모수 (즉 산포모수) α의 추론을 위해 조정된 편 다단계 가능도(adjusted partial h-likelihood) $p_\tau(h^*)$를 사용한다 (Ha와 Lee, 2003, 2005).

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

여기서 $H(h^*;\tau) = -\partial^2 h^*/\partial\tau\partial\tau^T$.

2.3. 벌점화 변수선택법

이 절에서는 Ha 등 (2014)에 의해 제안된 프레일티 모형에서 변수선택을 위한 벌점화된 다단계 가능도 접근법을 요약하여 설명한다. 벌점화 다단계 가능도 h_p (Ha 등, 2014)는 다음과 같이 정의된다.

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

여기서 h_p는 h^*에 의해 λ_0에 의존하지 않는다. $J_\gamma(|\bullet|)$는 조율(tuning)모수 γ를 가지는 벌점함수이다. 여기서 γ의 값이 커질수록 공변량을 적게 선택하기 때문에 단순한 모형이 된다. h_p에 다양한 벌점함수

가 적용가능 하지만 본 논문의 서론에서 언급한 세 가지 벌점함수를 고려한다. 따라서 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 $u(|\beta|) = [\{8w\beta^2/\alpha + (2-w)^2\}^{1/2} + 2 - w]/4$.

위의 세 가지 벌점함수들의 특징은 다음과 같이 요약된다.

(1) 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. 이러한 특징은 SCAD와 구별 된다 (Figure 2.1).

(2) 좋은 벌점 함수는 세 가지 오라클(oracle) 성질을 만족 한다 (즉, 불편성, 성김성(sparsity)과 연속성; Fan과 Li, 2001, 2002). LASSO는 L_1벌점으로 가장 공통적으로 사용되는 벌점함수이다. 그러나 이것은 위의 세 가지 성질을 동시에 만족하지는 않는다. Fan과 Li (2001)는 SCAD 벌점함수가 세 가지 성질을 모두 만족 한다는 것을 보였다. Fan과 Li (2001, 2002)는 또한 위의 SCAD식에서 $a = 3.7$을 취하는 경우 다양한 상황에서 잘 수행된다는 것을 보였다.

(3) Lee와 Oh (2014)는 회귀계수 β를 변량효과로 보고 혼합효과모형(mixed effect models)을 적합하여 벌점가능도 방법에서와 같은 해를 구하고, 이론적으로 좋은 성질을 가진 새로운 HL벌점함수를 제안하였다. Lee 와 Oh (2014)에 의한 HL벌점함수 유도의 주요 아이디어는 다음과 같다. 하나의 변량효과 u가 주어졌을 때, β가 정규분포를 따른다고 가정하였다.

$$\beta|u \sim N(0, u\theta)$$

그리고 정규 분포의 분산성분에 들어있는 변량효과 u는 $E(u) = 1$이고 $\mathrm{var}(u) = \omega$를 가지는 감마분포를 따른다고 가정한다. 이러한 변량효과 모형을 가정하고 다단계 가능도를 이용하여 Lee와 Oh (2014)는 HL 벌점함수 $J_\gamma(|\beta_j|)$를 도출하였다.

벌점화된 다단계 가능도를 이용하여 프레일티 모형에 있는 (β, v)를 추정하기 위해서 MPHL(maximum penalized h-likelihood) 추정량을 사용한다 (Ha 등, 2014). 프레일티 모수 α가 주어질 때, (β, v)의 MPHL 추정량은 β와 v의 다음의 결합추정방정식(joint estimating equations)에 의해서 얻어진다.

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

$$\frac{\partial h_p}{\partial v} = \frac{\partial h^*}{\partial v} = 0.$$

따라서 벌점화된 다단계 가능도 추정식은 다음과 같이 표현 된다 (Ha와 Lee, 2003; Ha 등, 2014).

$$\begin{pmatrix} X^TWX + n\sum_\gamma & X^TWZ \\ Z^TWX & Z^TWZ + U \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{v} \end{pmatrix} = \begin{pmatrix} X^Tw \\ Z^Tw + R \end{pmatrix}, \tag{2.2}$$

여기서 X와 Z는 각각 β와 v의 모형행렬이다. $\sum_\gamma = \text{diag}\{J'_\gamma(|\beta_j|)/|\beta_j|\}$, $W = -\partial^2 h^*/\partial\eta\partial\eta^T$, $U = -\partial^2 l_1/\partial v^2$, $w = W\eta + (\delta - \mu)$, $\eta = X\beta + Zv$, $\mu = \Lambda_0 \exp(\eta)$, $R = Uv + \partial l_1/\partial v$이며 로그정규 프레일티 모형에서는 $R = 0$이 된다. 식 (2.2)에서 로그-프레일티 v가 없으면, 식 (2.2)는 콕스 비례위험모형에서 벌점화 변수선택 추정식이 됨을 알 수 있다:

$$\left(X^TWX + n\sum\nolimits_\gamma\right)\hat{\beta} = X^Tw.$$

프레일티의 모수 α를 추정하기 위해 h_p로부터 $\tau = (\beta^T, v^T)^T$가 제거된 $p_\tau(h_p)$를 이용한다. 대응하는 추정방정식은 다음과 같다.

$$\frac{\partial p_\tau(h_p)}{\partial \alpha} = 0,$$

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.

한편, 조율모수(tuning parameter) γ를 선택하기 위하여 다단계 가능도에 기반한 BIC(Bayesian information criterion)형태의 한 기준을 이용한다 (Ha 등, 2014).

$$\text{BIC}(\gamma) = -2p_v(h_p) + e(\gamma)\log(n),$$

여기서 $e(\gamma) = \text{tr}[\{H_{\beta\beta} + n\sum_\gamma\}^{-1}H_{\beta\beta}]$이다. 조율모수 γ는 BIC를 최소로 하는 하나의 간단한 grid 방법에 의해 선택될 수 있다 (Fan과 Li, 2002). 마지막으로 $\hat{\beta}$에 대한 표준오차(standard error; SE)는 다음의 sandwich 분산-공분산행렬로부터 얻어진다.

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

여기서 $H_{\beta\beta} = -\partial^2\hat{h}/\partial\beta\partial\beta^T$이고 $\hat{h} = h^*|_{v=\hat{v}}$이다.

본 논문에서는 프레일티 분포에 대해 감마분포 등 다른 분포로의 적용이 가능하다. 하지만 편의상 로그정규분포에 기반한 frailtyHL 통계패키지를 사용하여 3절의 실제자료에 변수선택 방법을 적용하는 절차를 소개한다. 특히 로그-프레일티에 대한 정규분포의 지정은 지분(nested)구조나 상관성(correlation)을 갖는 프레일티 모형으로의 확장이 매우 쉬운 장점이 있다 (Ha 등, 2011, 2014).

3. 실제자료 분석

3.1. 자료의 설명

유방암은 유방 조직 안에 악성세포들이 모여 생기는 암이다. 우리나라 여성에서도 자궁암이나 위암에 이어 세 번째로 흔한 것으로 알려져 있으며, 생활양식의 서구화로 인해 갈수록 증가하고 있는 추세다. 유방암에 관해 많은 연구가 진행되어 왔음에도 불구하고, 원인은 여성 호르몬과 유방암이 관련이 있으리라 추측되지만 아직 분명히 밝혀진 것은 없다.

Table 3.1. The explanation of variables in the breast cancer data

변수	설명
ID	환자번호
PFS	수술 후 재발까지의 시간 (단위: 월)
PFSdel	유방암의 재발 여부 (0: 없음, 1: 있음)
Age	나이
Tumor	유방암 종양의 개수 (0: 1개 있음, 1: 2개이상 있음)
Tmax	유방암에서 가장 높은 수치를 보이는 픽셀(pixel)의 값
Tvol	유방암 종양의 부피(metabolic volume)
Nmax	림프절 전이에서 가장 높은 수치를 보이는 픽셀(pixel)의 값
Nvol	림프절 전이의 부피(metabolic volume)

Table 3.2. The breast cancer data: the estimated regression parameters and standard errors via variable selection methods in shared frailty model

Variable	No-penalty	LASSO	SCAD	HL
Tmax	−0.059 (0.392)	0.003 (0.002)	0 (0)	0 (0)
Tvol	0.892 (0.328)	0.447 (0.133)	0.613 (0.171)	0.559 (0.160)
Nmax	0.393 (0.453)	0 (0)	0 (0)	0 (0)
Nvol	−0.275 (0.458)	0 (0)	0 (0)	0 (0)
Tumor	−0.910 (0.884)	0 (0)	0 (0)	0 (0)
Age	0.157 (0.318)	0 (0)	0 (0)	0 (0)
BIC: tuning γ	0	0.11	0.19	$(w, a) = (3, 0.077)$

본 분석에 사용된 유방암 자료는 최근 전남대학교 의과대학병원에서 수집된 유방암 환자 54명의 재발 생존자료(recurrent survival data)이다. 반응변수는 재발시간(PFS)과 중도절단여부(PFSdel)이고, 재발시간의 단위는 월(month)이다. Table 3.1은 분석에 이용된 변수의 설명이다. 본 논문에서 하나의 관심사항은 유방암의 재발시간에 영향을 미치는 공변량이 무엇인지를 알아보는 것이기 때문에 프레일티 모형 적합을 통해 적절한 변수를 선택하려고 한다.

3.2. 모형적합 및 변수선택

3.1절의 유방암 자료는 각 환자에게 한 번만 재발시간이 관측된 단변량 생존자료이다. 하지만 자료 간 이질성이 있을 것으로 기대되기 때문에 공통(shared) 프레일티 모형 (2.1)을 적용하여 2.3절에서 제시된 변수선택 분석법을 적용하였다. 대응하는 추정된 회귀계수와 표준오차는 Table 3.2에 요약되어 있다.

먼저 No-penalty 하에서 프레일티 모수의 추정치는 $\hat{\alpha} = 1.935$로 매우 큰 값을 가진다. 이는 환자들 간 생존시간에 이질성이 심하다는 것을 의미한다. 따라서 이러한 결과는 프레일티의 효과를 반영하므로, Figure 3.1은 No-penalty하에서 시간에 따른 프레일티 효과를 시각화 해 주는 그림이다. 그 우측은 로그-프레일티 추정치에 대한 상자그림(box plot)이다. Figure 3.1에 의하면, 초기에 재발을 경험한 환자는 높은 프레일티 값(매우 연약함)을 주지만, 나중에 재발을 경험한 환자는 낮은 프레일티 값(덜 연약함)을 준다. 따라서 시간이 경과함에 따라 프레일티 값이 감소함을 알 수 있다. 특히 중도절단된 환자(Figure 3.1에서 "None"으로 표시)는 시간이 많이 경과된 이후에도 재발을 경험하지 않아서 상대적으로 낮은 프레일티 값을 줌을 알 수 있기 때문에, 재발을 경험한 환자보다 경험하지 않은 환자가 훨씬 덜 위험, 즉 덜 연약함(less frail)함을 알 수 있다. 또한 이 자료에 의하면 재발여부에 따른 두 집단 간 시간에 따라 이질성이 뚜렷이 구분됨을 알 수 있다.

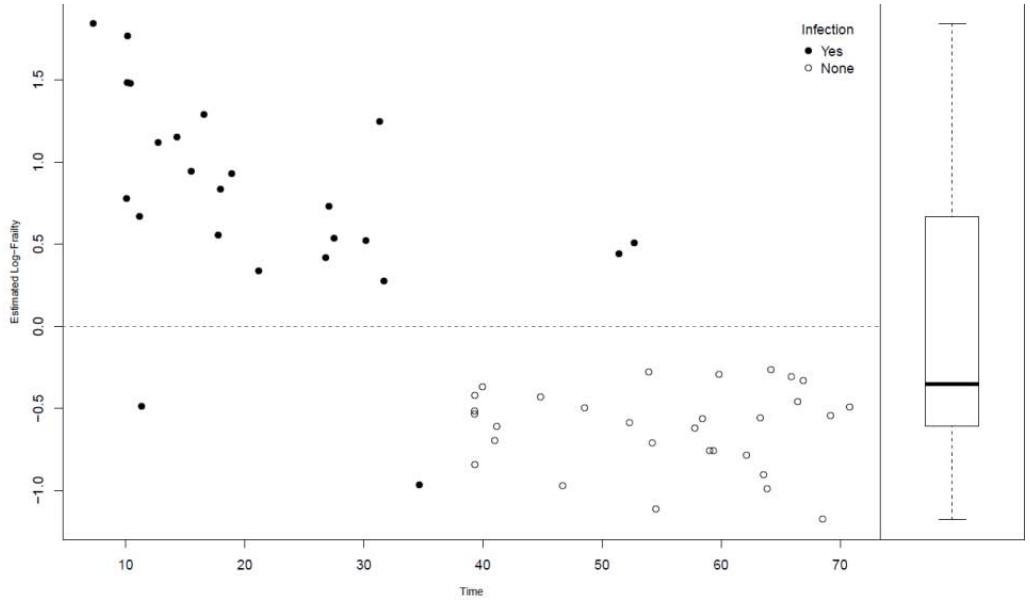

Figure 3.1. The frailty effects over times: the breast cancer data

세 가지 변수선택법을 적용한 결과, BIC기준에서 선택된 조율모수 γ의 값은 LASSO 에서는 0.11, SCAD에서는 0.19, HL에서는 $(w, a) = (3, 0.077)$이었다. Table 3.2에 의하면 공변량 Tvol은 4가지 방법 모두에서 유의했다. LASSO 방법은 6개 공변량 중 2개의 공변량(Tmax, Tvol)을 선택한 반면에, SCAD와 HL에서는 1개의 공변량(Tvol)만을 선택했다. 여기서 LASSO는 No-penalty에서 유의하지 않은 변수 Tmax를 하나 더 선택함을 관측할 수 있다. 이러한 결과는 Ha 등 (2014)의 실제자료의 분석과 유사함을 알 수 있다. Ha 등 (2014)에 의하면, LASSO 방법은 중요하지 않은 변수를 선택하는 경향이 있는 반면에, SCAD와 HL 방법은 전반적으로 적절한 변수를 유사하게 선택하는 것으로 입증하였다. 따라서 프레일티 모형에서는 LASSO보다는 SCAD와 HL 방법을 이용하여 변수선택 하는 것이 적절하다고 할 수 있다.

3.3. 변수선택을 위한 R 코드 및 설명

부가적으로, 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 프레일티 모형의 적합은 "B=(0,0,0,0,0,0)"과 "tun1=0"의 지정을 통해 쉽게 구현된다.

```
############################### frailty model (LASSO) ################################
> library(frailtyHL)
> source("frailtyVS.R")
> setwd("c:\\VS")
> bre<-read.table("raw_stand.csv",sep=",", header=T)
```

```
> la_bre <-frailty.vs(Surv(PFS,PFSdel)~ Tmax + Tvol + Nmax + Nvol + tumor +
+ Age + (1|id),  model="lognorm", penalty="lasso", data=bre,
+ B=c(-0.059, 0.892, 0.393, -0.275, -0.910, 0.157), tun1=seq(0,0.2,0.01))
############################### Output of LASSO #####################################
[1] "Result of variable selection in frailty model"
[1] "==Fitted model=="
[1] "model : lognorm"
[1] "penalty : lasso"
[1] "formula :"
Surv(PFS, PFSdel) ~ Tmax + Tvol + Nmax + Nvol + tumor + Age + (1 | id)
[1] "converge"
[1] "==Fixed coefficients=="
             Estimate Std. Error
Tmax          0.00270    0.00201
Tvol          0.44701    0.13286
Nmax          0.00000    0.00000
Nvol          0.00000    0.00000
tumor         0.00000    0.00000
Age           0.00000    0.00000
[1] "==Dispersion parameter=="
[1] 0
[1] "==Tuning parameter=="
[1] 0.11
[1] "==BIC=="
[1] 171.4812
```

다음으로, SCAD와 HL의 초기치는 LASSO의 추정치를 사용하며 (Ha 등, 2014), 특히 HL은 두 개의 조율모수 “tun1”과 “tun2”를 지정함으로써 구현된다. 보다 자세한 R 코드는 부록에 제시되어 있다.

4. 토론

본 논문에서는 프레일티 모형에서 벌점화 다단계 가능도에 기반하여 변수선택을 위한 R 패키지를 개발하였다. 그 예증을 위해 국내 임상자료에 적용하여 보았다. Table 3.2의 분석결과 LASSO 방법은 중요하지 않은 변수를 선택하는 경향이 있었다. 반면에 SCAD와 HL 방법이 전반적으로 적절한 변수를 유사하게 선택하는 것으로 나타났다 (Ha 등, 2014). 특히 다단계 가능도는 주변 가능도와 달리 프레일티를 바로 추정함으로써 Figure 3.1에서 제시된 바와 같이 자료 분석에 매우 유용한 정보를 제공하는 또 다른 장점이 있다 (Ha 등, 2011).

본 논문에서 사용된 자료는 공변량의 수가 다소 작기 때문에, 차후 공변량의 수(p)가 매우 많은 경우 뿐만 아니라 표본 수보다 큰 경우(즉 $p > n$)에 개발된 패키지를 적용하여 연구하는 것이 필요하다. 현재 개발된 본 패키지의 구현은 “frailtyHL”에 하나의 source 파일이 요구되지만, 본 논문의 교신저자를 통해 이것의 이용이 가능하다. 나아가, 개발된 방법의 효율성을 위해 CRAN에 장착할 예정이다. 마지막으로, 최근에 생존분석은 또 다른 사건시간(event times)인 경쟁사건(competing events)를 가지는 경우

로 확장하여 분석되고 있다. 따라서 경쟁위험 모형(competing risks models)에서 적절한 변수를 선택하는 방법에 대해 연구하는 것은 흥미 있는 향후 연구과제가 될 것으로 사료된다.

부록: R 코드(유방암 자료)

```
### frailty model (No-penalty)
no_bre<-frailty.vs(Surv(PFS,PFSdel)~ Tmax + Tvol + Nmax + Nvol + tumor + Age
+ (1|id), model="lognorm", penalty="lasso", data=bre,
B=c(-0.059, 0.892, 0.393, -0.275, -0.910, 0.157), tun1=(0))

### frailty model (LASSO)
la_bre<-frailty.vs(Surv(PFS,PFSdel)~ Tmax + Tvol + Nmax + Nvol + tumor + Age
+ (1|id), model="lognorm", penalty="lasso", data=bre,
B=c(-0.059, 0.892, 0.393, -0.275, -0.910, 0.157), tun1=seq(0,0.2,0.01))

### frailty model (SCAD)
sc_bre<-frailty.vs(Surv(PFS,PFSdel)~ Tmax + Tvol + Nmax + Nvol + tumor + Age
+ (1|id), model="lognorm", penalty="scad", data=bre,
B=c(0.0027, 0.4470,0,0,0,0), tun1=seq(0,0.3,0.01))

### frailty model (HL)
hl_bre<-frailty.vs(Surv(PFS,PFSdel)~ Tmax + Tvol + Nmax + Nvol + tumor + Age
+ (1|id), model="lognorm", penalty="hl", data=bre,
B=c(0.0027, 0.4470,0,0,0,0), tun1=c(2.1,30,10,30,50), tun2=seq(0.001,0.3,0.01))
```

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.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection, *The Annals of Statistics*, **24**, 2350–2383.

Breslow, N. E. (1972). Discussion of Professor Cox's paper, *Journal of the Royal Statistical Society B*, **34**, 216–217.

Clayton, D. G. (1991). A Monte Carlo method for Bayesian inference in frailty models, *Biometrics*, **47**, 467–480.

Cox, D. R. (1972). Regression models and life tables (with Discussion), *Journal of the Royal Statistical Society 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.

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. and Lee, Y. (2005). Comparison of hierarchical likelihood versus orthodox best linear unbiased predictor approaches for frailty models, *Biometrika*, **92**, 717–723.

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

Ha, I. D., Noh, M. and Lee, Y. (2012). frailtyHL: A package for fitting frailty models with h-likelihood, *The R Journal*, **4**, 307–320.

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**, 28–37.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*, Springer, New York.

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

Lee, Y., Nelder, J. A. and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-Likelihood*, Chapman and Hall, London.

Legrand, C, Ducrocq, V., Janssen, P., Sylvester, R. and Duchateau, L. (2005). A Bayesian approach to jointly estimate centre and treatment by centre heterogeneity in a proportional hazards model, *Statistics in Medicine*, **24**, 3789–3804.

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

Ripatti, S. and Palmgren. J. (2000). Estimation of multivariate frailty models using penalized partial likelihood, *Biometrics*, **56**, 1016–1022.

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

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

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

frailtyHL 통계패키지를 이용한 프레일티 모형의 변수선택: 유방암 생존자료

김보현[a] · 하일도[a,1] · 노맹석[a] · 나명환[b] · 송호천[c] · 김자혜[c]

[a]부경대학교 통계학과, [b]전남대학교 통계학과, [c] 전남대학교병원 핵의학과

(2015년 7월 27일 접수, 2015년 7월 31일 수정, 2015년 8월 6일 채택)

요 약

통계적 모형에서 적절한 변수를 선택하는 것은 회귀분석에서 매우 중요하다. 최근 벌점 함수(예: LASSO 및 SCAD)와 함께 벌점화 가능도를 사용하는 변수 선택 방법들이 선형모형 및 일반화 선형모형과 같은 단순한 통계 모형에서 널리 연구되고 있다. 이러한 방법들의 주요 장점은 중요한 변수를 선택하고 동시에 회귀계수를 추정하는 것이다. 그러므로 이 방법들은 0으로 회귀계수를 추정함으로써 중요하지 않은 변수를 삭제한다. 이 논문에서는 콕스 비례 위험 모형의 한 확장인 준 모수적 프레일티 모형에서 벌점화된 다단계 가능도(h-likelihood; HL)를 기반으로 적절한 변수를 선택하는 방법을 연구한다. 이를 위해 세 가지 벌점 함수 LASSO, SCAD 및 HL을 사용한다. 본 논문에서는 변수선택을 효율적으로 하기 위해 "frailtyHL" R 패키지 (Ha 등, 2012)를 기반으로 하여 새로운 함수를 개발하였다. 개발된 방법의 예증을 위해 전남대 의과대학 병원에서 수집된 유방암 생존자료를 이용하여 세 가지 변수 선택 방법의 결과를 비교하고, 이 변수선택방법들의 상대적 장·단점에 대해 토론한다.

주요용어: 프레일티 모형, 다단계 가능도, LASSO, SCAD, 변수선택

이 논문은 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구 사업임 (No. 2010-0021165).

[1]교신저자: (608-737) 부산광역시 남구 용소로 45, 부경대학교 통계학과. E-mail: idha1353@pknu.ac.kr