



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

이 학 석 사 학 위 논 문

생존분석모형에서 변수선택 및 응용



통 계 학 과

김 보 현

이 학 석 사 학 위 논 문

생존분석모형에서 변수선택 및 응용

지도교수 하일도, 노맹석

이 논문을 석사 학위논문으로 제출함.

2015년 8월

부 경 대 학 교 대 학 원

통 계 학 과

김 보 현

김보현의 이학석사 학위논문을 인준함.

2015년 8월 21일



위원장	이학박사	장대홍 (인)
위원	이학박사	하일도 (인)
위원	이학박사	노맹석 (인)

목 차

표 차례	ii
그림 차례	ii
I. 서론	1
II. 별점화 방법을 이용한 변수선택	3
2.1 콕스 비례위험모형에서 변수선택	4
2.1.1 모형의 기본개념	4
2.1.2 모형의 형태	4
2.1.3 모형의 추론 및 변수선택	5
2.2 공통 프레이리티모형에서 변수선택	6
2.2.1 모형의 기본개념	6
2.2.2 모형의 형태	6
2.2.3 다단계 가능도 추정법	7
2.2.4 별점화 변수선택법	10
2.3 지분 프레이리티모형에서 변수선택	15
III. 실제 자료 분석	16
3.1 기존 임상 자료	16
3.2 국내 임상 자료	20
3.3 국외 임상 자료	25
IV. 결론 및 제언	30
참고문헌	
부록 : R codes	

표 차례

1. Table 1. PBC자료의 변수설명	17
2. Table 2. PBC자료에서 변수선택을 통한 추정된 회귀계수 및 표준오차	19
3. Table 3. 유방암자료의 변수설명	21
4. Table 4. 유방암자료에서 변수선택을 통한 추정된 회귀계수 및 표준오차	22
5. Table 5. 방광암자료의 변수설명	26
6. Table 6. 방광암자료에서 변수선택을 통한 추정된 회귀계수 및 표준오차	27

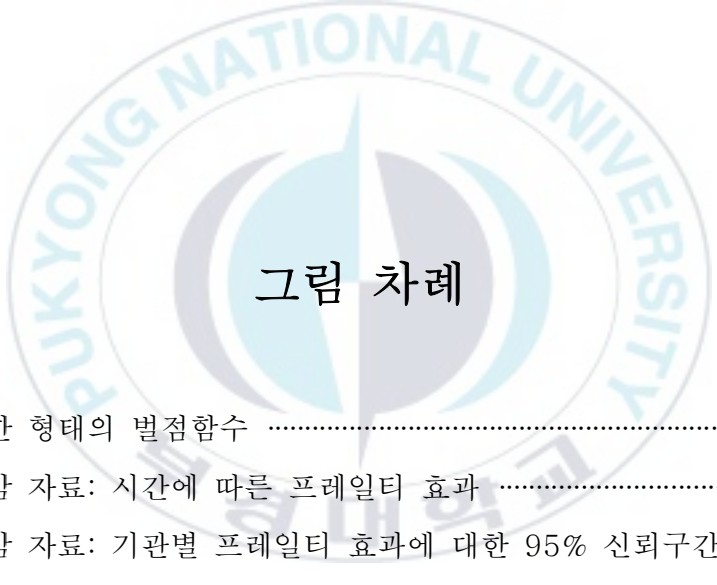


그림 차례

1. Figure 1. 다양한 형태의 별점함수	11
2. Figure 2. 유방암 자료: 시간에 따른 프레일티 효과	24
3. Figure 3. 방광암 자료: 기관별 프레일티 효과에 대한 95% 신뢰구간	29

Variable Selection in Survival Models and Applications

Bo Hyeon Kim

Department of Statistics, The Graduate School,
Pukyong National University

Abstract

Determining relevant variables for a regression model is very important in regression analysis. In ordinary regression models, there are a variety of classical techniques for variable selection such as forward selection, backward elimination and stepwise selection. However, these classical methods can be computationally intensive for a large number of covariates and often suffer from high variability.

Recently, variable selection methods using a penalized likelihood with various penalty functions (e.g. LASSO, SCAD) have been widely studied in simple statistical models such as linear models and generalized linear models. The main advantage of these methods is that they select important variables and estimate the regression coefficients, simultaneously. Thus, they delete insignificant variables by estimating their coefficients as zero.

In this thesis we study how to select proper variables based on the penalized hierarchical likelihood (HL) in survival-hazards models including Cox's proportional hazards models and semi-parametric frailty models. For this purpose we allow three penalty functions, LASSO, SCAD and HL. The computations are based on the R packages including "frailtyHL". Our methods are illustrated with clinical-trial data sets from Korea and Europe. We compare the results from three variable-selection methods, and also discuss their advantages and disadvantages.

key words : Frailty models, H-likelihood, LASSO, SCAD, Variable selection

I. 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, Nelder와 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 폭넓게 연구되고 있다. 벌점화 방법의 주요한 장점은 중요한 공변량을 선택함과 동시에 공변량의 회귀계수를 추정하는 것이다. 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 연구에서는 생존분석모형에서 자주 사용되는 콕스 비례위험모형 및 프레일티 모형 (frailty models)에서의 변수선택 방법들에 관하여 연구한다. 프레일티 모형은 각 개체 (subject)나 군집 (cluster)의 위험률에 대한 콕스의 비례위험모형에 프레일티를 허락한 하나의 확장된 생존분석 회귀모형으로서, 주로 다변량 생존자료 (multivariate survival data)의

회귀분석에 사용되고 있다. 여기서 프레일티는 각 개체의 위험률에 승법적으로 영향을 미치는 관측되지 않는 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. 최근에는 Androulakis 등 (2012)은 이를 역정규 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; Lee와 Nelder, 1996)는 어려운 적분자체를 피할 수 있을 뿐만 아니라, 다양한 변량 효과 모형에서 통계적으로 효율적인 계산절차를 제공한다 (Lee, Nelder와 Pawitan, 2006). 본 연구의 모든 계산절차는 “frailtyHL” R 통계패키지 (Ha, Noh와 Lee, 2012)를 이용하였다.

본 연구의 구성은 다음과 같다. 2장에서는 벌점화 방법 (LASSO, SCAD, HL)을 이용하여 변수선택하는 방법을 살펴보고자 한다. 이를 위해 콕스모형 및 프레일티 모형의 기본 개념을 기술하고 벌점화 변수선택 방법을 리뷰 하고자 한다. 3장에서는 벌점화 방법을 이용하여 국내·외 임상시험자료를 분석하고자 한다. 여기서 자료의 형태 및 성격에 따라 콕스모형, 공통 (shared) 프레일티모형, 및 지분 (nested) 프레일티모형을 대상으로 하여 변수를 선택하고자 한다. 4장에서는 본 연구결과를 토론하고 향후과제를 제시하고자 한다. 마지막으로, 부록에서는 3장의 변수선택 결과에 대한 R코드를 제시한다.

II. 벌점화 방법을 이용한 변수선택

회귀모형의 변수선택에서 설명변수의 수가 매우 많은 경우, 모든 변수들이 실제로 유의한 경우는 드물다. 그래서 유의한 변수들을 선택하기 위해 벌점 가능도 (penalized likelihood) 방법이 많이 연구되어졌다. 벌점 가능도 방법은 기존의 가능도에 적절한 벌점함수 (penalty function)를 붙임으로써 벌점 가능도를 최대화 하는 모수를 추정하다보면 자동적으로 의미 있는 변수가 선택되는 방법이다.

먼저 간단한 선형모형 (즉 고전적 회귀모형)을 고려해 보자.

$$y_i = x_i^T \beta + \varepsilon_i, \quad i = 1, \dots, n, \quad (2.1)$$

여기서 $\varepsilon_i \sim i.i.d N(0, \phi)$.

여기서 x_i 는 p 개의 공변량들의 벡터이며, β 는 대응하는 p 개의 회귀모수 벡터이다. 이때 다음과 같은 $Q_\gamma(\beta)$ 를 최소(즉 $-Q_\gamma(\beta)$ 를 최대)로 하는 회귀계수 β 를 찾는 방법이 벌점가능도 방법이다(Fan과 Lee, 2001).

$$Q_\gamma(\beta) = \frac{1}{2} \sum_{i=1}^n (y_i - x_i^T \beta)^2 + n \sum_{j=1}^p J_\gamma(|\beta_j|),$$

여기서 $J_\gamma(|\beta_j|)$ 는 주어진 벌점함수이고, γ 는 벌점함수의 모양을 결정해주는 조율모수 (tuning parameter)이다.

본 연구에서는 생존 자료를 갖는 콕스모형 및 프레일티 모형에서 벌

점함수를 통한 변수선택의 방법을 리뷰하고 장/단점을 토론한다.

2.1 콕스 비례위험모형에서의 변수선택

2.1.1 모형의 기본개념

콕스모형은 각 개인의 위험률과 공변량간의 관계를 설명하기 위한 대표적인 생존분석 회귀모형으로서, 위험률에 영향을 미치는 공변량(covariates)의 효과 (예: 처리효과 또는 예후요인효과)를 추론하는 것이 주요한 사용목적 중 하나이다. 특히 이 모형은 비례위험성의 가정과 준모수적 (semi-parametric)모형이라는 두 가지 큰 특성을 갖고 있다.

2.1.2 모형의 형태

콕스 회귀모형의 형태는 각 개인의 생존시간 $T_i (i=1, \dots, n)$ 에 대한 위험률과 공변량간의 관계가 아래와 같이 표현된다.

$$\lambda_i(t; x) = \lambda_0(t) \exp(x_i^T \beta), \quad i = 1, \dots, n, \quad (2.2)$$

여기서 $\lambda_0(t)$ 는 미지의 기저 (unknown baseline) 위험함수로서, p 개의 공변량들의 벡터 $x_i = (x_{i1}, \dots, x_{ip})^T$ 의 모든 값들이 0일 때의 위험함수이다. i 번째 개인의 공변량 x_i 에 대응하는 회귀모수들의 벡터 $\beta = (\beta_1, \dots, \beta_p)^T$ 는 주요한 관심추론의 대상이다. 모형(2.2)의 주요한 특징은 다음과 같다.

- (i). 준 모수적 모형 (semi-parametric model): $\lambda_0(t)$ 의 형태는 모르지만 (non-parametric), 공변량의 함수 형태는 $\exp(x^T\beta)$ 로 안다 (parametric).
- (ii). 비례위험성 (proportional hazards; PH): 두 군간 위험률의 비, 즉 상대위험 (relative risk; RR)이 시간에 관계없이 일정한 상수에 비례 (proportional)한다. 특히, 콕스 모형은 PH라는 가정으로 인해 관심회귀모수의 해석이 쉽다.
- (iii). 무 절편항 (No intercept term): $\lambda_0(t)$ 가 절편항 역할을 하기 때문에 $x_i^T\beta$ 는 동일성 (identifiability)문제로 인해 절편항을 포함하지 않는다.

2.1.3 모형의 추론 및 변수선택

콕스 회귀모형의 추론을 위해 1972년 콕스는 기저위험함수에 대한 아무런 정보 없이 회귀모수를 추정하기 위한 평가능도 (partial likelihood)를 소개하였다. Andersen과 Gill (1982)은 counting process 이론을 사용하여, 평가능도를 최대로 하는 회귀모수의 추정량은 일치성과 점근정규성과 같은 대표본 성질을 만족한다는 사실을 증명하였다. Johansen (1983)은 콕스의 평가능도는 단면가능도 (profile likelihood)가 된다는 사실도 보였다.

콕스모형은 2.2절에서 제시한 프레일티 모형의 특별한 경우로서, 프레일티가 없으면 바로 콕스모형이 되기 때문에 2.2절의 프레일티 모형의 변수선택법을 사용할 수 있다. 따라서 (2.5)식에서 언급한 바와 같이 콕스모형에서의 변수선택법은 2.2절에서 논하기로 한다.

2.2 공통 프레일티 모형에서 변수선택

2.2.1 모형의 기본개념

프레일티 모형은 각 개체 (subject)나 군집 (cluster)의 위험률에 대한 콕스의 준모수적 PH모형에 프레일티 (frailty)를 허락한 하나의 확장된 생존분석 회귀모형으로서, 주로 다변량생존자료 (multivariate or correlation survival-time data)의 회귀분석에 사용되고 있다. 여기서 프레일티는 각 개체의 위험률에 승법적으로 영향을 미치는 관측 되지 않는 변량효과를 의미한다. 프레일티의 분포는 통상적으로 로그정규 (log-normal) 또는 감마 (gamma)분포를 지정한다. 특히 콕스모형은 프레일티 모형에서 모든 로그 프레일티 (log-frailty)의 값이 0 (즉 로그프레일티의 분산이 0)인 경우에 해당되므로 프레일티 모형을 이용해서 바로 적합할 수도 있다 (Ha 등, 2012).

2.2.2 모형의 형태

각 개인의 공통된 프레일티 u_i 가 주어질 때, i 번째 개인의 j 번째 관측에 대한 생존시간 T_{ij} 의 조건부 위험률 (conditional hazard rate)은 다음과 같이 표현된다.

$$\lambda_{ij}(t|u_i; x_{ij}) = \lambda_0(t) \exp(x_{ij}^T \beta) u_i, \quad (2.3)$$

여기서 $\lambda_0(t)$ 는 미지의 기저 (baseline) 위험함수이며, β 는 T_{ij} 에 대한 p 개

의 공변량들의 벡터 $x_{ij} = (x_{ij1} \cdots x_{ijp})^T$ 에 대응하는 회귀모수이다. 프레일티 u_i 는 감마분포 또는 로그정규분포를 지정할 수 있다. 만약 모든 개인들의 $u_i = 1$ 이면, 프레일티 모형은 콕스의 비례위험 모형이 된다. 특히 감마 프레일티모형과 로그정규 프레일티 모형은 프레일티 분포에 대해 각각 감마분포와 로그정규분포를 가정하기 때문에, 프레일티 u_i 에 대해서는 평균이 1이고 분산이 α 인 감마분포를, 그리고 로그프레일티 $v_i (= \log u_i)$ 에 대해서는 정규분포, 즉 $v_i \sim N(0, \alpha)$ 를 각각 지정한다.

u_i 의 의미는 $u_i = 1$ 인 사람은 위험률이 표준적인 경향에 있다 (즉 모든 $u_i = 1$ 이면 프레일티 모형은 Cox모형이 된다). 만약 $u_i < 1$ 인 사람은 표준적인 사람보다 위험률이 낮은 경향을 가지며, $u_i < 1$ 인 사람은 표준적인 사람보다 위험률이 높은 경향을 가진다. 이러한 의미는 3.2절의 Figure 2에서 시각적으로 또한 표현된다.

2.2.3 다단계 가능도 추정법

프레일티 모형의 추론을 위해 다양한 방법들이 제안되어 왔다. 대표적인 추론법은 요약하면 다음과 같다.

(i) 몬테카를로 EM (MCEM)을 이용한 주변 가능도 방법 (Vaida와 Xu, 2000), (ii) 벌점 편 가능도 방법 (penalized partial likelihood; PPL, Ripatti와 Palmgren, 2000), (iii) 베이지안 접근법 (Clayton, 1991; Legrand 등, 2005)이 있다. 하지만 이러한 방법들은 어려운 적분계산이 요구되며 추정량의 효율성이 떨어지는 단점이 있다. 따라서 본 연구에서는 Lee와 Nelder (1996)에 의해 제안된 다단계 가능도 (hierarchical likelihood; h-likelihood)를 사용한다. 이 방법은 복잡한 적분계산을 피하

고 통합된 추론틀을 제공하는 장점이 있다 (Lee 등, 2006).

프레일티 모형 (2.3)에 대한 다단계 가능도 (Ha, Lee와 Song, 2001)의 정의는 다음과 같다.

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

여기서

$$\begin{aligned} \ell_0 &= \sum_{ij} \log f(y_{ij}, \delta_{ij} | u_i; \beta, \lambda_0) \\ &= \sum_{ij} \delta_{ij} \{ \log \lambda_0(y_{ij}) + \eta_{ij} \} - \sum_{ij} A_0(y_{ij}) \exp(\eta_{ij}), \end{aligned}$$

여기서 ℓ_0 는 프레일티 u_i 가 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_i 의 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 β 의 추론을 위해 Breslow (1972)와 Ha 등 (2001)의 아이디어에 따라 본 연구에서는 먼저 기저 누적위험함수 (baseline cumulative hazards function)를 다음과 같이 가정하였다.

$$A_0(t) = \sum_{k: y_{(k)} \leq t} \lambda_{0k},$$

여기서 $y_{(k)}$ 는 y_{ij} 들 중 k 번째로 작은 관측되는 생존시간이고, $\lambda_{0k} = \lambda_0(y_{(k)})$ 이다. 이러한 가정 하에서 λ_0 를 제거한 단면 다단계 가능도 (profile h-likelihood), 즉 $h^* \equiv h|_{\lambda_0 = \hat{\lambda}_0}$ 를 사용한다. 따라서 h^* 는 아래와 같이 표현된다 (Ha 등, 2001).

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

여기서

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

는 λ_0 에 의존하지 않으며, $\hat{\lambda}_{0k}(\beta, v) = \frac{d_{(k)}}{\sum_{i,j \in R_{(k)}} \exp(\eta_{ij})}$ 는 $\frac{\partial h}{\partial \lambda_{0k}} = 0$

($k=1, \dots, D$)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.

프레일티 모수 (즉 산포모수) α 의 추론을 위해 조정된 단면 다단계가 능도 (즉 조정된 편 다단계 가능도) $p_{\beta, v}(h^*)$ 를 사용한다 (Ha와 Lee,

2003, 2005).

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

$$\text{여기서 } H(h^*; \beta, v) = -\partial^2 h^* / \partial(\beta, v)^2.$$

2.2.4 별점화 변수선택법

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

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

여기서 h^* 는 λ_0 에 의존하지 않는다. $J_\gamma(|\cdot|)$ 는 조율 (tuning)모수 γ 를 가지는 별점함수이다. 여기서 γ 의 값이 커질수록 공변량을 적게 선택하기 때문에 단순한 모형이 된다. 별점함수의 정의는 다음과 같다.

(i) LASSO (Tibshirani, 1996):

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

(ii) SCAD (Fan과 Li, 2001):

$$J_\gamma = \gamma I(|\beta| \leq \gamma) + \frac{(a\gamma - |\beta|)_+}{a-1} + I(|\beta| > \gamma),$$

여기서 $x_+ = xI(x > 0)$ 으로서 x 의 양수부분을 표시한다.

(iii) HL (Lee와 Oh, 2014);

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

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

위의 세 가지 벌점함수들의 특징은 다음과 같이 요약된다.

- (1) HL벌점함수는 w 의 값에 따라 그 모양이 변화하게 되며, 특히 w 가 0인 경우 Ridge가 되며, w 가 2인 경우 LASSO가 되며, w 가 20보다 큰 경우 0에서의 값이 음의 무한대값을 갖는 형태가 된다 (Figure 1 참조).

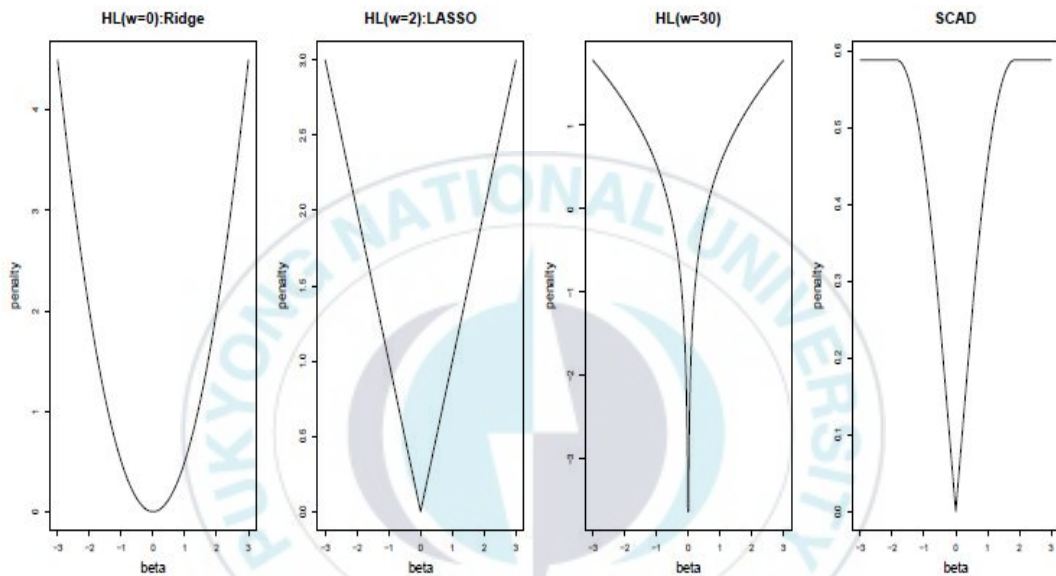


Figure 1. 다양한 형태의 벌점함수

- (2) 좋은 벌점 함수는 세 가지 오라클 (oracle) 성질을 만족한다 (즉, 불편성, 성김성과 연속성; Fan과 Li, 2001, 2002). LASSO는 L_1 벌점으로 가장 공통적으로 사용되는 벌점함수이다. 그러나 이것은 세 가지 성질을 동시에 만족하지는 않는다. Fan과 Li (2001)는 SCAD 벌점이 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.

세 가지 오라클 성질을 요약하면 다음과 같다.

- (i) 성김성 (sparsity) : 추정량이 소수의 변수에서만 0이 아닌 계수를 갖도록 함으로써 모형의 복잡도를 감소시키고 해석력을 높이는 성질을 말한다.
- (ii) 불편성 (unbiasedness): 추정량이 점근적인 의미에서 거의 편이 없는 성질을 말한다
- (iii) 연속성 (continuity): γ 의 변화에 대해 추정량이 연속성을 가진다는 것으로, γ 에 따라 모형이 급작스럽게 변하는 것을 방지하며 최종모형에 의한 예측의 변동성을 줄인다.

(3) Lee와 Oh (2014)는 회귀계수 β 를 변량효과로 보고 혼합효과모형 (mixed effect models)을 적합하여 별점가능도 방법에서와 같은 해를 구하고, 이론적으로 좋은 성질을 가진 새로운 HL별점함수를 제안하였다. Lee 와 Oh (2014)에 의한 HL별점함수 유도의 주요 아이디어는 다음과 같다. 또 하나의 변량효과 u 가 주어졌을 때, β 가 정규분포를 따른다고 가정하였다.

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

그리고 정규 분포의 분산성분에 들어있는 변량효과 u 는 $E(u)=1$ 이고 $var(u)=\omega$ 를 가지는 감마 분포를 따른다고 가정한다. 이러한 변량효과 모형을 가정하고 다단계 가능도를 이용하여 Lee와 Oh (2014)는 식(2.1)에서 HL 별점함수 $J_{\gamma}(|\beta_j|)$ 를 도출하였다.

별점화된 다단계 가능도를 이용하여 (β, v) 를 추정하기 위해서 MPHL (maximum penalized h-likelihood) 추정량을 사용한다. 프레일티 모수 α 가 주어질 때, (β, v) 의 MPHL 추정량은 β 와 v 의 다음의 결합추정방정식 (joint estimating equations)에 의해서 얻어진다.

$$\begin{aligned}\partial h_p / \partial \beta_j &= \partial h^* / \partial \beta_j - n \sum_{j=1}^p [J_\gamma(|\beta_j|)]' = 0, \\ \partial h_p / \partial v &= \partial h^* / \partial v = 0.\end{aligned}$$

따라서 별점화된 다단계 가능도 추정식은 다음과 같이 표현된다 (Ha 등, 2014).

$$\begin{pmatrix} X^T W X + n \sum_{\gamma} & X^T W Z \\ Z^T W X & Z^T W Z + U \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{v} \end{pmatrix} = \begin{pmatrix} X^T w \\ Z^T w + R \end{pmatrix} \quad (2.4)$$

여기서, $\sum_{\gamma} = \text{diag}\{J'_\gamma(|\beta_j|)/|\beta_j|\}$, $W = -\partial^2 h^* / \partial \eta \partial \eta^T$, $U = -\partial^2 l_2 / \partial v^2$, $w = W\eta + (\delta - \mu)$, $\eta = X\beta + Zv$, $\mu = \Lambda_0 \exp(\eta)$, $R = Uv + \partial l_2 / \partial v$ 이며 로그정규 프레일티 모형에서는 $R=0$ 이 된다.

(2.4) 식에서 로그-프레일티 v 가 없으면, (2.4)는 콕스 모형에서 별점화 변수선택 추정식이 됨을 알 수 있다.

$$(X^T W X + n \sum_{\gamma}) \hat{\beta} = X^T w^*, \quad (2.5)$$

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

$$\partial p_\tau(h_p)/\partial \alpha = 0,$$

여기서 $p_\tau(h_p) = [h_p - \frac{1}{2} \log \det \{H(h_p, \tau)/(2\pi)\}]|_{\tau=\hat{\tau}}$, $H(h_p, \tau) = -\partial^2 h_p / \partial \tau^2$

이고, $\hat{\tau}$ 는 $\partial h_p / \partial \tau = 0$ 의 해이다.

한편 조율모수 (tuning parameter) γ 를 선택하기 위하여 다단계 가
능도에 기반한 BIC (Bayesian information criterion) 형태의 한 기준을
이용한다 (Ha 등, 2014).

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

여기서 $e(\gamma) = \text{tr}[\{H_{\beta\beta} + n\sum_{\gamma}\}^{-1}H_{\beta\beta}]$.

$\hat{\beta}$ 에 대한 표준오차 (standard error; SE)는 다음의 sandwich 분산-공분
산행렬로부터 얻어진다.

$$\begin{aligned} \text{Cov}(\hat{\beta}) &= \{-\partial^2 \hat{h}_p / \partial \beta^2\}^{-1} \text{cov}(\partial \hat{h}_p / \partial \beta) \{-\partial^2 \hat{h}_p / \partial \beta^2\}^{-1} \\ &= (H_{\beta\beta} + n\sum_{\gamma})^{-1} H_{\beta\beta} (H_{\beta\beta} + n\sum_{\gamma})^{-1}, \end{aligned}$$

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

2.3 지분 프레일티 모형에서 변수선택

병원급 센터에서 환자를 모집하는 다기관 임상시험 (multi-center clinical trials)에서 각 환자의 사건시간 (event times)들이 관측되는 지분 (nested 또는 multi-level) 디자인 구조인 경우, 두 개의 프레일티 항을 고려할 수 있다. T_{ijk} 를 i 번째 센터에 속해 있는 j 번째 환자에 대한 k 번째 생존시간이라 하자. 두 개의 프레일티 항 v_i 와 v_{ij} 가 주어질 때, 생존 시간 T_{ijk} 의 조건부 위험률로 표현되는 로그정규 지분 프레일티 모형 (Yau, 2001; Ha 등, 2007)의 정의는 다음과 같다.

$$\lambda_{ijk}(t|v_{1i}, v_{2ij}) = \lambda_0(t) \exp(x_{ij}^T \beta + v_{1i} + v_{2ij}),$$

여기서 $v_{1i} \sim N(0, \alpha_1)$ 로서 i 번째 센터의 로그프레일티이고, $v_{2ij} \sim N(0, \alpha_2)$ 로서 i 번째 센터에서 지분된 (nested) j 번째 환자의 로그프레일티이며, v_{1i} 과 v_{2ij} 는 서로 독립이다. 지분 프레일티 모형은 공통 프레일티모형에 변량효과가 하나 더 추가되는 형태이기 때문에, 이에 대한 다단계 가능도의 추론은 쉽게 확장이 가능하다 (Ha 등, 2007, 2015). 따라서 지분모형의 변수선택도 2.2절에서 제시된 공통프레일티 모형에서 변수선택법을 쉽게 확장하여 바로 사용할 수 있다.

Ⅲ. 실제 자료 분석

3.1 기존 임상 자료

예제 1: PBC 자료

원발성 담즙성 간경변증 (PBC; Primary Biliary Cirrhosis)은 만성적인 담즙정체를 특징으로 하는 자가 면역 간질환이다. 흔한 증상은 피로감과 가려움증 등이다. 주로 중년 여성에서 발생하며, 진행 정도에 따라 다양한 정도의 간 섬유화 및 간경변증을 보인다.

PBC의 유병률은 지역마다 다양하여 인구 백만 명당 6.7명부터 402명까지 보고되고 있는데, 영국, 유럽지역 등에 상대적으로 흔하다. PBC는 초기에는 진단 후 수년 내에 간부전으로 진행되는 질환으로 알려져 있었다. 그러나 많은 연구들이 진행되면서 원발성 담즙성 간경변증의 예후가 다양한 것이 밝혀졌고, 초기에 발견하여 적절히 치료하면 양호한 예후를 기대할 수 있다. 질병의 유병률이 낮고 장기간 서서히 진행되는 특성이 있으므로 임상적 관심이 필요한 질환이다 (Kim 등, 2010).

PBC 자료는 424명의 원발성 담즙성 간경변증 환자를 대상으로 10년 동안 디-페니실라민 (D-penicillamine; DPCA)과 위약 (placebo)을 대조한 위약비교 임상시험에서 수집된 자료이다 (Tibshirani, 1997). 112명은 임상 시험에 참여하였지만, 기본 측정 기록이 존재하지 않거나 동의서에 서명하지 않았기 때문에, 실제 분석은 312명의 참가자들을 대상으로 한다.

분석에 사용된 변수는 Table 1에 설명 되어있다.

Table 1. PBC자료의 변수설명

변수	설 명
id	환자번호
futime	등록 후 사망까지의 시간 (일)
status	PBC의 생존여부 (0: 생존, 1:사망)
drug	약의 종류 (1:디-페니실라민, 2: 위약)
age	나이
sex	성별 (0: 남성, 1: 여성)
ascites	복수의 유무 (0: 아니오, 1: 예)
hepato	간종의 유무 (0: 아니오, 1: 예)
spiders	피부혈관기형의 존재 (0: 아니오, 1: 예)
edema	부종의 존재 (0 : 부종이 없거나 부종에 대한 이노치료를 하지 않음, 0.5 : 이노치료 없이 부종이 나타나거나 부종이 치료됨, 1 : 이노치료에도 불구하고 부종이 있음)
bili	혈청 빌리루빈 (serum bilirubin)
chol	혈청 콜레스테롤 농도 (serum cholesterol)
albumin	혈청 알부민 농도 (단위: gm/dl)
copper	소변에서의 copper 농도 (단위: g/day)
alk_phos	alkaline phosphatase 농도 (단위: U/liter)
sgot	SGOT (단위: U/ml)
trig	triglycerides (단위: mg/dl)
platelet	platelets per cubic (단위: ml/1000)
protime	prothrombin time (단위: 초)
stage	조직학적 단계 (1, 2, 3 또는 4)

여기서 생존시간은 futime이고, 중도절단 지시함수는 status이다. id를 제외한 나머지 변수들은 PBC의 생존여부에 영향을 미칠 수 있는 공변량이다.

이 자료는 각 개인에게 한 번만 생존시간이 관측되는 단일생존자료 (univariate survival data)이기 때문에 콕스모형을 대상으로 하여 세 가지 변수선택방법 (LASSO, SCAD, HL)을 사용하였다. 대응하는 추정된 회귀계수와 표준오차는 Table 2에 요약되어 있다.



Table 2. PBC자료: 콕스모형에서 변수선택을 통한 추정된
회귀계수 및 표준오차

Variables	No penalty	LASSO (GCV)	LASSO (BIC)	SCAD (BIC)	HL (BIC)
	Est(SE)	Est(SE)	Coef(SE)	Coef(SE)	Coef(SE)
drug	-0.06(0.11)	0.00(0.00)	0(0)	0(0)	0(0)
age	0.30(0.12)	0.17(0.09)	0.200(0.058)	0.307(0.101)	0.160(0.048)
sex	-0.12(0.10)	-0.01(0.03)	0(0)	0(0)	0(0)
ascites	0.02(0.10)	0.04(0.07)	0.021(0.014)	0(0)	0(0)
hepato	0.01(0.13)	0.00(0.00)	0(0)	0(0)	0(0)
spiders	0.05(0.11)	0.02(0.05)	0(0)	0(0)	0(0)
edema	0.27(0.11)	0.18(0.11)	0.187(0.055)	0.264(0.091)	0.178(0.050)
bili	0.37(0.12)	0.35(0.12)	0.385(0.065)	0.468(0.081)	0.421(0.064)
chol	0.12(0.10)	0.00(0.01)	0(0)	0(0)	0(0)
albumin	-0.30(0.12)	-0.22(0.10)	-0.237(0.063)	-0.306(0.111)	-0.247(0.060)
copper	0.22(0.10)	0.21(0.11)	0.247(0.060)	0.299(0.082)	0.268(0.059)
alk_phos	0.00(0.08)	0.00(0.00)	0(0)	0(0)	0(0)
sgot	0.23(0.11)	0.09(0.08)	0.106(0.042)	0.024(0.011)	0(0)
trig	-0.06(0.09)	0.00(0.00)	0(0)	0(0)	0(0)
platelet	0.08(0.11)	0.00(0.00)	0(0)	0(0)	0(0)
protime	0.23(0.11)	0.09(0.09)	0.150(0.051)	0.008(0.004)	0.110(0.037)
stage	0.39(0.15)	0.21(0.09)	0.255(0.065)	0.379(0.123)	0.242(0.060)

Note: GCV(generalized cross-validation; Tibshirani, 1997),

BIC(Bayesian information criterion; Ha 등, 2014)

분석결과, BIC기준에서 선택된 조율모수 γ 의 값은 LASSO에서는 0.05, SCAD에서는 0.06, HL에서는 $(w, a) = (10, 0.001)$ 이었다. No-penalty 하에서는 공변량 age, edema, bili, albumin, copper, protime, stage이 유의하였다. GCV에 의한 LASSO방법 (Tibshirani, 1997)은 전반적으로 BIC에 의한 LASSO방법과 유사한 결과를 보인다. LASSO(BIC)방법은 16개 공변량중 9개 (age, ascites, ederna, bili, albumi, copper, sgot, protime, stage)가 선택되었고 SCAD(BIC)방법은 8개 (age, ederna, bili, albumi, copper, sgot, protime, stage) 선택되었다. HL(BIC)방법은 7개 (age, ederna, bili, albumi, copper, protime, stage)만을 선택했다. Table 2에 의하면 LASSO(BIC)방법에서는 No-penalty 하에서 유의하지 않은 공변량인 “ascites”도 선택하여 보다 많은 공변량을 변수선택한다. 여기서 SCAD방법과 HL방법이 보다 유의한 변수만을 선택한다는 것을 알 수 있다. 또, SCAD방법이 HL방법보다 변수를 하나 더 선택했다. 그러므로 LASSO방법은 공통 프레일티 모형에서 중요한 변수를 선택하는 데에는 적절하지 못하며, SCAD방법과 HL방법이 적절한 것으로 보여진다.

3.2 국내 임상 자료

예제 2: 유방암 임상자료

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.

본 분석에 사용된 유방암 자료는 전남대학교 의과대학에서 수집된 유방암 환자 54명의 재발생존자료를 이용하였다. 반응변수는 재발시간 (PFS)과 중도절단여부 (PFSdel)이고, 재발시간의 단위는 월 단위이다. 유방암의 재발시간에 영향을 미치는 공변량이 무엇인지를 알아보기 위해 적절한 변수를 선택하여 분석하였다.

Table 3은 분석에 이용된 변수의 설명이다.

Table 3. 유방암자료의 변수설명

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

유방암 자료는 각 개인에게 한 번만 재발시간이 제시된 단일 생존 자료이지만 자료간 이질성이 있을 것으로 기대되기 때문에 공통 프레일티 모형을 적용하여 변수선택 분석법을 시도하였다. 대응하는 추정된 회귀 계수와 표준오차는 Table 4에 요약되어 있다.

Table 4. 유방암자료: 공통 프레일티모형에서 변수선택을 통한 추정된 회귀계수 및 표준오차

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)

BIC기준에서 선택된 조율모수 γ 의 값은 LASSO에서는 0.11, SCAD에서는 0.19, HL에서는 $(w, a) = (3, 0.077)$ 이었다. 프레일티 모수의 추정치는 No-penalty에서 $\hat{\alpha}=1.935$ 로 매우 큰 값을 가진다. 이는 환자들간 생존시간에 이질성이 심하다는 것을 나타낸다 (Figure 2 참조). 공변량 Tvol은

4가지 방법 모두에서 유의했다. LASSO방법은 6개 공변량 중 2가지 공변량이 (T_{max} , T_{vol})가 선택되었고 반면에 SCAD와 HL에서는 1개의 공변량 (T_{vol})만을 선택했다. 여기서 LASSO는 No-penalty에서 유의하지 않은 변수 T_{max} 를 하나 더 선택함을 관측할 수 있다. 이러한 결과는 Table 2와 유사한 결과를 준다.

Figure 2는 No-penalty하에서 시간에 따른 프레일티 효과를 시각화해 주는 그림이다. 우측은 로그-프레일티 추정치에 대한 상자그림이다.



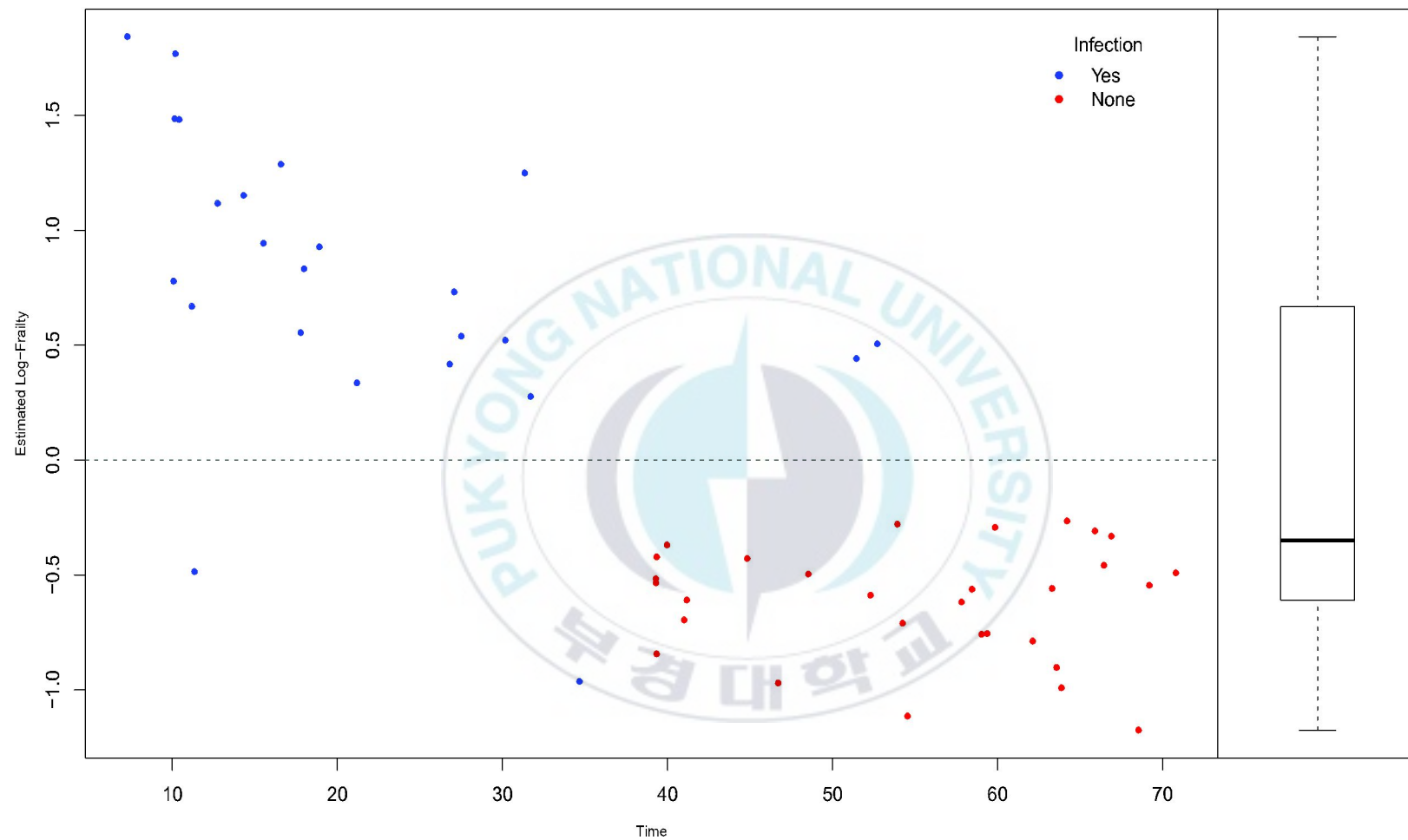


Figure 2. 유방암 자료: 시간에 따른 프레일티 효과

Figure 2에 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 2에서 “none”)는 시간이 많이 경과된 이후에도 재발을 경험하지 않아서 상대적으로 낮은 프레일티 값을 줄 수 있기 때문에, 재발을 경험한 환자보다 경험하지 않은 환자가 훨씬 덜 위험, 즉 덜 연약함 (less frail)함을 알 수 있다.

3.3 국외 임상 자료

예제 3: 방광암 임상자료

이 자료 (Oddens 등, 2013)는 유럽의 EORTC (유럽 암 치료기구)에서 주관하여 13개 국가의 57개의 센터에서 수집된 방광암 환자 1066명의 임상시험자료를 대상으로 하였다.

비근침윤성 방광암 (non-muscleinvasive bladder cancer; NMIBC)은 경요도 절제술 (transurethral resection; TUR)후 장기간의 종양학적 결과는 매우 좋지만, 5년 이내에 30~80%의 환자는 방광 내에 재발을 하게 되며, 1~45%의 환자는 근침윤성 방광암으로 진행하게 된다는 것이 특징이다. 따라서 방광암은 재발이나 진행을 막기 위하여 방광 내 항암요법 혹은 BCG를 이용한 방광 내 면역요법을 시행하게 된다. BCG면역요법의 적절한 치료효과를 위하여 유지요법 기간을 따르는 것이 필요하다 (Kim과 Park, 2014).

본 연구에서는 비근침윤성 방광암환자들이 경요도 절제술 후 재발에

영향을 미치는 위험인자(공변량)에 대하여 알아보하고자 한다. 본 분석에서는 고위험군의 환자 70세이상인 466명을 대상으로 했고, 나라 수는 12개, 기관수는 39개이다.

Table 5는 분석에 이용된 변수의 설명이다.

Table 5. 방광암 자료의 변수설명

변수	설명
timeDFI	수술 후 재발까지의 시간 (일)
statusDFIc	방광암의 재발 여부 (0: 없음, 1: 있음)
Trt dose	BCG면역요법 (1: 1/3 dose BCG, 2: Full dose BCG)
Trt duration	유지요법 (0: 1년, 1: 3년)
Age	나이
Gender	성별 (0: 남자, 1: 여자)
typeBC	방광암의 유형 (0: Primary, 1: Recurrent)
Tumsize	가장 큰 종양의 크기 (mm)
Nbtum	종양의 개수
Tstage	방광암의 임상적 진행단계 (0: pTa , 1: pT1)
Ggrade	방광암의 조직학적 등급 (1:G1, 2:G2 , 3:G3)

먼저 지분 프레일티 모형을 분석한 결과, 국가의 로그-프레일티 분산은 0 ($\hat{\alpha}_1 = 0$)이고 기관의 로그-프레일티 분산은 0.069 ($\hat{\alpha}_2 = 0.069$)로서 국가효과는 존재하지 않았다. 따라서 본 연구에서는 기관프레일티만을 가지고 분석하였다. 분석결과는 Table 6에 제시되어 있다.

Table 6. 방광암자료: 공통 프레일티모형에서 변수선택을 통한 추정된 회귀계수 및 표준오차

Variable	No penalty	LASSO	SCAD	HL
trtdose	-0.175(0.145)	0(0)	0(0)	0(0)
trtduration	-0.338(0.144)	-0.143(0.063)	0(0)	0(0)
age	0.043(0.019)	0.041(0.017)	0.024(0.011)	0.033(0.014)
gender	0.008(0.183)	0(0)	0(0)	0(0)
typeBC	0.628(0.167)	0.343(0.095)	0.602(0.147)	0.201(0.053)
tumsize	-0.002(0.006)	-0.004(0.005)	0(0)	0(0)
nbtum	0.129(0.038)	0.120(0.033)	0.056(0.017)	0.120(0.031)
tstage	-0.144(0.200)	0(0)	0(0)	0(0)
gg1	-0.520(0.252)	-0.054(0.030)	0(0)	0(0)
gg2	-0.214(0.214)	0(0)	0(0)	0(0)

선택된 조율모수 γ 의 값은 LASSO, SCAD, HL에서 각각 0.064, 0.019, $(w, a) = (10, 0.011)$ 이었다. 여기서 로그-프레일티 분산 $\hat{\alpha}$ 은 No-penalty, LASSO, SCAD, HL 하에서 각각 0.069, 0.064, 0.099, 0.111으로 추정되었다. 따라서 프레일티 효과가 강하지는 않은 것으로 평가된다 (Figure 3 참조). No-penalty 하에서 공변량 trtduration, age, typeBC, nbtum, gg1이 유의했다. LASSO방법은 9개 공변량 중 6가지 공변량 (trtduration, age, typeBC, tumsiz, nbtum, gg1)이 선택되었고, 반면에 SCAD와 HL에서는 3개의 공변량 (age, typeBC, nbtum)만을 선택했다. 여기서 LASSO는 No-penalty에서 유의하지 않은 변수 tumsiz를 하나 더 선택함을 관측할 수 있다.

Figure 3은 HL방법에 기초한 39개 기관별 로그-프레일티의 95% 신뢰구간을 보인다. 예상한 바와 같이 39개의 신뢰구간은 모두 0의 프레일티 값을 포함하므로 기관별로 기저위험에 대한 이질성은 전반적으로 거의 없는 것으로 평가된다.

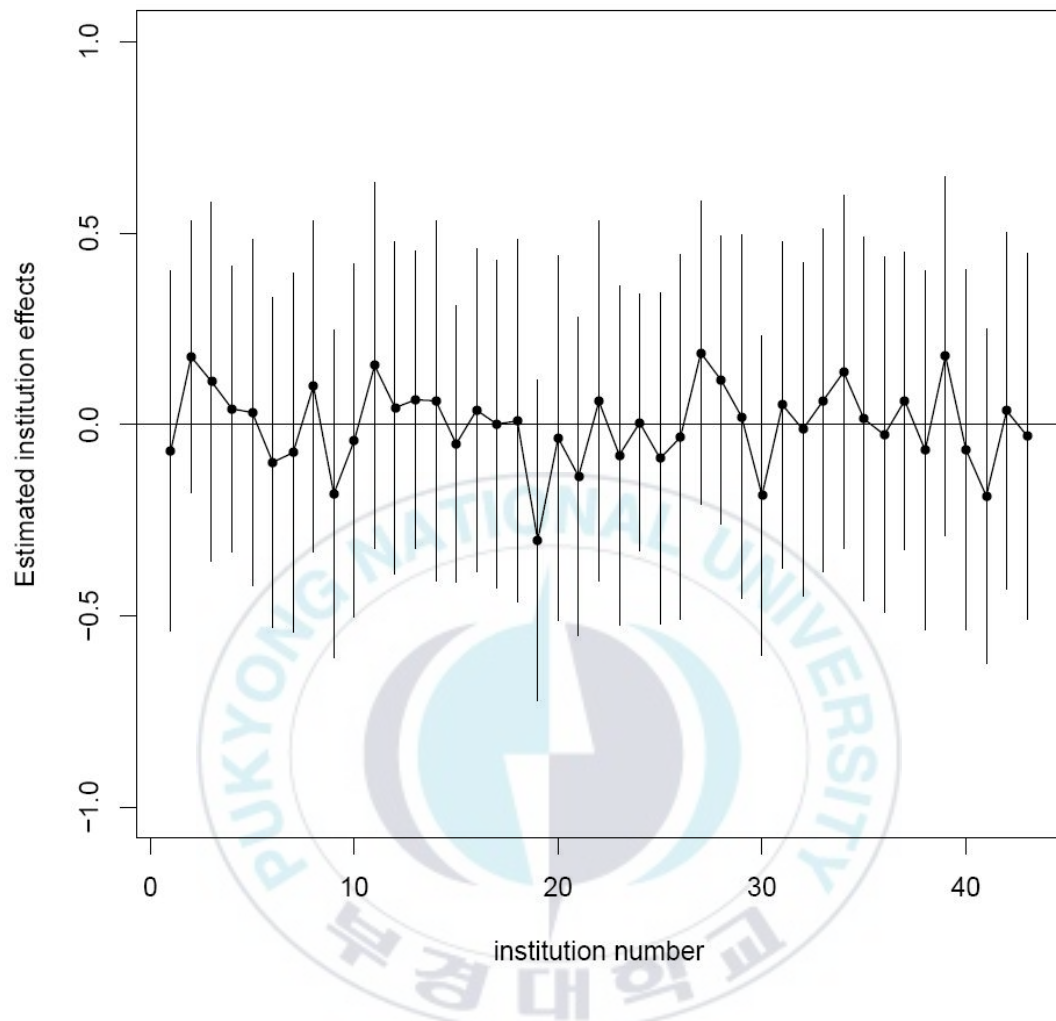


Figure 3. 방광암 자료: 기관별 프레일티 효과에 대한 95% 신뢰구간

IV. 결론 및 제언

본 연구에서는 생존분석모형에서 벌점화 가능도를 이용한 변수선택방법에 대해 논의하고, 국내·외 실제 임상시험자료를 이용하여 이러한 방법들을 적용하여 보았다. 분석결과 LASSO 방법은 중요하지 않은 변수를 포함하여 보다 많은 변수를 선택하는 경향이 있었다. 반면에 SCAD와 HL 방법이 전반적으로 적절한 변수를 비슷하게 선택하는 것으로 나타났다. 특히 HL 방법이 유의한 변수를 가장 잘 선택하는 것으로 분석되었다.

최근 HL 벌점화 방법은 통계학의 전반적인 부분에 사용되고 있으며 그 이론적 성질 또한 여러 논문에서 다루어지고 있다. 특히, 이 방법은 선형모형, 일반화 선형모형, 다변량분석 (주성분분석, 정준상관분석)모형 및 프레이티 모형 등을 이용하여 다변량 자료, 반복측정자료 및 생존자료 등의 분석에 많이 응용되어 변수선택에 있어서 탁월한 성능을 보인 바 있다 (Lee와 Oh, 2014; Lee, 2015). 또한 변수간의 그룹정보가 있는 경우, 그룹정보를 반영하여 변수선택을 하는 방법 또한 매우 최근에 개발되었다 (Lee, 2015).

최근의 생존분석은 또 다른 사건시간 (event times)인 경쟁사건 (competing events)를 가지는 경우로 확장하여 분석되고 있다. 특히 3장에서 제시된 방광암 임상자료는 경쟁사건 (즉 재발 전 사망)도 포함하는 확장된 하나의 생존 자료이다. 따라서 경쟁위험 모형 (competing risks models)에서 적절한 변수를 선택하는 방법에 대해 연구하는 것은 흥미 있는 향후 연구과제가 될 것으로 사료된다.

참고문헌

- [1] Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting processes: a large sample study, *The Annals of Statistics*, **10**, 1100–1120.
- [2] Androulakis, E., Koukouvinos, C. and Vonta, F. (2012). Estimation and variable selection via frailty models with penalized likelihood, *Statistics in Medicine*, **31**, 2223–2239.
- [3] Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, **24**, 2350–2383.
- [4] Breslow, N. E. (1972). Discussion of Professor Cox's paper. *Journal of the Royal Statistical Society B*, **34**, 216–217.
- [5] Clayton D.G. (1991). A Monte Carlo method for Bayesian inference in frailty models, *Biometrics*, **47**, 467–480.
- [6] Cox, D. R. (1972). Regression models and life tables (with Discussion), *Journal of the Royal Statistical Society B*, **74**, 187–220.
- [7] 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.
- [8] Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model, *The Annals of Statistics*, **30**, 74–99.
- [9] 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.
- [10] 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.

- [11] Ha, I. D., Lee, Y. and MacKenzie, G. (2007). Model selection for multi-component frailty models, *Statistics in Medicine*, **26**, 4790–4807.
- [12] Ha, I. D., Lee, Y. and Song, J.-K. (2001). Hierarchical likelihood approach for frailty models, *Biometrika*, **88**, 233–243.
- [13] 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.
- [14] Ha, I. D., Noh, M. and Lee, Y., Lim, J., Lee, J., Oh, H., Shin, D., Lee, S., Seo, J., Park, Y., Cho, S., Park, J., Kim, Y., You, K. (2015). Survival analysis using SRC-Stat statistical package, *The Korean Journal of Applied Statistics*, **28**, 309–324.
- [15] 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.
- [16] Johansen, S. (1983). An extension of Cox's regression model, *International Statistical Review*, **51**, 165–174.
- [17] Kim, K. A., Jeong, S.H. (2010). Clinical features and prognosis of primary biliary cirrhosis in Korea, *The Korean Journal of Hepatology*, **16**, 139–146.
- [18] Kim, K.H. and Park, Y.Y. (2014). Current evidence for the treatment of bladder cancer, *The Ewha Medical Journal*, **37**, 1–9.
- [19] Lee, Y. (2015). Review of mixed-effect models, *The Korean Journal of Applied Statistics*, **28**, 123–136.
- [20] Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion), *Journal of the Royal Statistical Society B*, **58**, 619–678.
- [21] Lee, Y., Nelder, J. A. and Pawitan, Y. (2006). Generalised Linear Models with Random Effects: Unified Analysis via h-Likelihood. London: Chapman

and Hall.

- [22] Lee, Y. and Oh, H. S. (2014). A new sparse variable selection via random-effect model, *Journal of Multivariate Analysis*, **125**, 89–99.
- [23] 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.
- [24] Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models, *Journal of the Royal Statistical Society A*, **135**, 370 - 384.
- [25] Oddens, J., Brausi, M., Sylvester, R., Bono, A., et al. (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.
- [26] Ripatti.S., Palmgren.J.,(2000). Estimation of multivariate frailty models using penalized partial likelihood, *Biometrics*, **56**, 1016–1022.
- [27] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society Series B*, **58**, 267 - 288.
- [28] Tibshirani R (1997). The Lasso method for variable selection in the Cox model, *Statistics in Medicine*, **16**, 385–395.
- [29] Vaida, F. and Xu, R. (2000). Proportional hazards models with random effects, *Statistics in Medicine*, **19**, 3309–3324.
- [30] Yau, K. K. W. (2001). Multilevel models for survival analysis with random effects, *Biometrics*, **57**, 96–102.

부록: R codes

#####자료 불러오기#####

```
rm(list=ls())
setwd("D:\\data")
library(survival)
library(frailtyHL)
source("frailtyVS.R")
pbc<-read.table("pbc.csv",sep=";",header=T)
bre<-read.table("raw_stand.csv",sep=";",header=T)
eortc<-read.table("eortcdata.csv",sep=";",header=T)
```

#1. PBC자료 (3.1절)

##Cox model

```
coxph(Surv(futime, status) ~ drug + age + sex + ascites + hepato
+ spiders + edema + bili + chol + albumin + copper + alk_phos + sgot +
trig+ platelet + protime + stage, data = pbc)
```

Cox model (No penalty)

```
no_pbc<-frailty.vs(Surv(futime, status) ~ drug + age + sex + ascites +
hepato + spiders + edema + bili + chol + albumin + copper + alk_phos +
sgot + trig + platelet + protime + stage +(1|id), model="lognorm",
penalty="lasso", data=pbc,
B=c(-0.06,0.30,-0.12,0.01,0.01,0.05,0.27,0.37,0.12,-0.30,0.22,0,0.23,-0.06,
0.08,0.23,0.39), tun1=0,varfixed=TRUE, varinit=0 )
```

Cox model (LASSO)

```
la_pbc<-frailty.vs(Surv(futime, status) ~ drug + age + sex + ascites +  
hepato + spiders + edema + bili + chol + albumin + copper + alk_phos +  
sgot + trig + platelet + protime + stage +(1|id), model="lognorm",  
penalty="lasso", data=pbcb,  
B=c(-0.06,0.30,-0.12,0.01,0.01,0.05,0.27,0.37,0.12,-0.30,0.22,0,0.23,-0.06,  
0.08,0.23,0.39),tun1=seq(0,0.1,0.01), varfixed=TRUE, varinit=0 )
```

Cox model (SCAD)

```
sc_pbc<-frailty.vs(Surv(futime, status) ~ drug + age + sex + ascites +  
hepato + spiders + edema + bili + chol + albumin + copper + alk_phos +  
sgot + trig + platelet + protime + stage +(1|id), model="lognorm",  
penalty="scad", data=pbcb,  
B=c(0,0.2,0,0.02,0,0,0.18,0.39,0,-0.24,0.25,0,0.11,0,0,0.15,0.25),  
tun1=seq(0,0.1,0.01), varfixed=TRUE, varinit=0 )
```

Cox model (HL)

```
hl_pbc<-frailty.vs(Surv(futime, status) ~ drug + age + sex + ascites +  
hepato + spiders + edema + bili + chol + albumin + copper + alk_phos +  
sgot + trig + platelet + protime + stage +(1|id), model="lognorm",  
penalty="hl", data=pbcb,  
B=c(0,0.2,0,0.02,0,0,0.18,0.39,0,-0.24,0.25,0,0.11,0,0,0.15,0.25),  
tun1=c(2.1,3,10,30,50), tun2=seq(0.001,0.1,0.01), varfixed=TRUE, varinit=0)
```

#2. 유방암자료 (3.2절)

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(0.001,0.3,0.019))
```

#3. 방광암자료 (3.3절)

```
##nested frailty model
```

```
ne_eortc<-frailtyHL(Surv(timeDFI, statusDFIc) ~ trtdose + trtduration + age  
+ gender + typeBC + tumsiz + nbtum + tstage+ gg1 + gg2 +  
(1|institution) + (1|Country) ,data=eortc)
```

```
## Frailty model (No penalty)
```

```
no_eortc<-frailty.vs(Surv(timeDFI, statusDFIc) ~ trtdose + trtduration + age  
+ gender + typeBC + tumsiz + nbtum + tstage + gg1 + gg2 +  
(1|institution), model="lognorm", penalty="lasso", data=eortc,  
B=c(-0.180,-0.094,-0.010,0.199,0.097,-0.003,0.139,0.364,-0.071,0.042),  
tun1=0)
```

Frailty model (LASSO)

```
la_eortc<-frailty.vs(Surv(timeDFI, statusDFIc) ~ trtdose + trtduration + age  
+ gender + typeBC + tumsiz + nbtum + tstage + gg1 + gg2 +  
(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(0,0.2,0.001))
```

Frailty model (SCAD)

```
sc_eortc<-frailty.vs(Surv(timeDFI, statusDFIc) ~ trtdose + trtduration + age  
+ gender + typeBC + tumsiz + nbtum + tstage + gg1 + gg2 +  
(1|institution), model="lognorm", penalty="scad", data=eortc,  
B=c(0,-0.133,0.041,0,0.330,-0.004,0.120,0,-0.040,0), tun1=seq(0,0.2,0.01))
```

Frailty model (HL)

```
hl_eortc<-frailty.vs(Surv(timeDFI, statusDFIc) ~ trtdose + trtduration + age  
+ gender + typeBC + tumsiz + nbtum + tstage + gg1 + gg2 +  
(1|institution), model="lognorm", penalty="hl", data=eortc,  
B=c(0,-0.133,0.041,0,0.330,-0.004,0.120,0,-0.040,0),  
tun1=c(2.1,3,10,30,50), tun2=seq(0.001,0.1,0.01))
```