The Korean Journal of Applied Statistics
2021, Vol. 34, No. 3, 411–425

DOI: https://doi.org/10.5351/KJAS.2021.34.3.411
Print ISSN1225-066X / Online ISSN 2383-5818

Penalized variable selection in mean-variance accelerated failure time models

Ji Hoon Kwon[a], Il Do Ha[1,b]

[a]Statistics Team, APACE Inc.; [b]Department of Statistics, Pukyong National University

Abstract

Accelerated failure time (AFT) model represents a linear relationship between the log-survival time and covariates. We are interested in the inference of covariate's effect affecting the variation of survival times in the AFT model. Thus, we need to model the variance as well as the mean of survival times. We call the resulting model mean and variance AFT (MV-AFT) model. In this paper, we propose a variable selection procedure of regression parameters of mean and variance in MV-AFT model using penalized likelihood function. For the variable selection, we study four penalty functions, i.e. least absolute shrinkage and selection operator (LASSO), adaptive lasso (ALASSO), smoothly clipped absolute deviation (SCAD) and hierarchical likelihood (HL). With this procedure we can select important covariates and estimate the regression parameters at the same time. The performance of the proposed method is evaluated using simulation studies. The proposed method is illustrated with a clinical example dataset.

Keywords: AFT model, mean-variance model, penalized likelihood, variable selection

1. 서론

Accelerated failure time (AFT) 모형은 비례위험모형(proportional hazards models)의 한 대안 (Lawless, 1982)으로써 사용되며, 로그생존시간과 공변량과의 관계를 선형으로 나타냄으로써 해석이 용이하다. 특히 AFT 모형은 모형의 분포 가정에 위배되더라도 모형 추론 결과가 다소 강건한(robust)한 결과를 준다 (Hutton과 Monaghan, 2002; Ha 등, 2002, 2017). AFT 모형은 통상적으로 생존시간의 평균과 공변량간의 관계를 모형화하여 공변량의 평균 효과를 통계적으로 추론하는데 관심을 보여왔다. 본 논문에서는 생존시간의 평균 뿐만 아니라, 변동성에 영향을 미치는 공변량의 효과를 동시에 추론하는 것에 관심이 있기 때문에 생존시간의 분산에 대한 모형도 고려한다. 즉, 각 개체의 생존시간이 서로 다른 분산을 갖는다고 가정하며, 이러한 모형을 본 논문에서는 평균-분산 AFT 모형이라고 부른다 (Nelder와 Lee, 1998). Wu와 Li (2012), Charalambous 등 (2015) 그리고 Antoniadis 등 (2016)은 완전자료(complete data)에 대한 평균-산포 결합모형에서 벌점화 변수선택(penalized variable selection) 방법들을 제안하여 왔다. Park과 Ha (2018)는 중도절단성(censoring)을 갖는 불완전 자료(incomplete data)에 대한 AFT모형에서 평균 회귀모수에 대한 벌점화 변수선택 절차를 제안하였다. 본 논문에서는 이러한 변수선택 절차를 확장하기 위해, 중도절단된 생존자료에 대한 평균-분산 AFT 모형에서 평균 회귀 모수 뿐만 아니라 분산 회귀 모수에 대한 변수선택을 제안한다.

This work was supported by a Research Grant of Pukyong National University(2019 year).
This paper is a condensed form of the first author's master thesis from the Pukyong National University.
[1] Corresponding author: Department of Statistics, Pukyong National University, 45 Yongso-ro Nam-gu, Busan 48513, Korea. E-mail: idha1353@pknu.ac.kr

Published 30 June 2021 / journal homepage: http://kjas.or.kr

특히, 벌점 함수(penalty function)에 기초하여 벌점 가능도함수(penalized likelihood function)를 사용한 변수선택 방법이 선형모형(linear models), 일반화 선형모형(generalized linear models; Nelder와 Wedderburn, 1972), 콕스의 비례위험모형(cox's proportional hazards models; Cox, 1972) 등에서 폭 넓게 연구되어 왔다. 이러한 변수선택의 장점 중 하나는 중요한 변수를 선택하는 동시에 공변량의 회귀계수를 추정할 수 있다. 특히 중요하지 않은 변수들에 대한 회귀모수들을 0으로 추정한다. 또한 모형에서 중요한 변수만을 포함함으로써 추정과 해석 측면에서 회귀분석의 질을 높일 수 있다. 따라서 많은 공변량을 갖는 회귀 모형에서 중요한 변수를 선택하는 것은 생존 분석(survival analysis)을 포함한 다양한 형태의 자료 분석에서 매우 중요하다 (Ha 등, 2014, 2017). 평균-분산 AFT 모형에서 고정효과에 대한 변수선택을 위해, 본 논문에서는 생존시간의 분포로써 AFT 모형에서 자주 사용되는 로그정규분포(lognormal distribution)를 가정한다. 특히 AFT 모형에서 이러한 로그정규분포의 모수적 가정은 다소 강건한 추론결과를 보여왔다 (Ha 등, 2002, 2017; Park과 Ha, 2018). 변수선택을 위해서는 문헌에서 자주 사용되고 있는 네 가지의 벌점 함수, least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996), adaptive lasso (ALASSO) (Zou, 2006), smoothly clipped absolute deviation (SCAD) (Fan과 Li, 2001), h-likelihood (HL) (Lee와 Oh, 2014)를 고려한다. 이러한 벌점화 함수에 따른 성능을 비교하기 위해 모의실험을 수행한다. 나아가, 북 아일랜드의 한 임상연구에 의한 폐암 자료(lung cancer data) (MacKenzie, 1996)를 통해 제안된 방법을 예증한다. 본 논문의 모든 계산은 R 프로그램을 사용하였다.

본 논문의 구성은 다음과 같다. 2절에서는 평균-분산 AFT 모형에서 벌점화 변수선택 절차 제안한다. 이를 위해 평균-분산 AFT 모형의 기본 개념과 네 가지 벌점화 방법(LASSO, ALASSO, SCAD, HL)을 리뷰한다. 3절에서는 모의실험을 통해, 평균-분산 AFT 모형에서 네가지 벌점화 방법들의 성능을 비교한다. 4절에서는 제안된 방법의 한 예증으로 위에서 제시한 폐암 자료를 사용하여 벌점화 방법들의 추정결과를 비교한다. 5절에서는 연구결과를 토론하고 향후 과제를 제시한다. 마지막으로, 부록에서는 평균-분산 AFT 모형의 변수선택에 필요한 계산절차의 유도를 제시한다.

2. 평균-분산 가속화 실패시간 모형과 변수선택

이 절에서는 먼저 평균-분산 AFT 모형을 기술한 후, 네 가지 벌점함수에 기초한 벌점 가능도 접근법을 사용하여 고려한 평균-분산 AFT 모형에 대한 고정효과의 변수선택 절차를 제안한다. 이를 위해, 벌점 가능도 함수를 구성하는 방법과 변수선택을 위한 회귀 모수 추정 절차를 유도한다.

2.1. 평균-분산 가속화 실패시간 모형

$T_i\,(i = 1,\ldots,n)$를 i번째 개체의 생존시간이라 하자. 평균-분산 AFT 모형은 로그생존시간과 공변량과의 선형적 관계로써 다음과 같이 정의된다.

$$\log T_i = x_i^T\beta + \epsilon_i, \quad i = 1,\ldots,n. \tag{2.1}$$

여기서 회귀계수 $\beta = (\beta_0,\beta_1,\ldots,\beta_{p-1})^T$는 평균모수의 공변량 벡터 $x_i = (1, x_{i1},\ldots,x_{i,p-1})^T$에 대응하는 p 차원의 미지의 회귀계수 벡터이며, ϵ_i는 서로 독립인 랜덤 오차항(random error term)이다. 본 논문에서는 ϵ_i의 분포로 AFT 모형에서 자주 사용되는 정규 분포로써 $\epsilon_i \sim N(0,\sigma_i^2)$을 가정한다. 식 (2.1)에서 σ_i^2과 공변량과의 관계를 다음과 같이 설정하여 개체에 따라 서로 다른 분산을 가정한다 (Charalambous 등, 2015).

$$\sigma_i^2 = \exp\left(z_i^T\alpha\right), \quad \text{즉} \log\left(\sigma_i^2\right) = z_i^T\alpha. \tag{2.2}$$

여기서 회귀계수 $\alpha = (\alpha_0,\alpha_1,\ldots,\alpha_{q-1})^T$는 분산모수의 공변량 벡터 $z_i = (1, z_{i1},\ldots,z_{i,q-1})^T$에 대응하는 q 차원의 미지의 회귀계수 벡터이다. 특히 z_i는 x_i와 같거나 그 일부로 자주 사용된다. 특히 ϵ_i의 정규분포 가정하에서,

모형 (2.1)과 (2.2)는 각각 다음과 같이 표현할 수 있으므로,

$$E(\log T_i) = x_i^T\beta, \quad \text{var}(\log T_i) = \exp\left(z_i^T\alpha\right),$$

모형 (2.2)를 허락하는 모형 (2.1)을 본 논문에서는 평균-분산 AFT 모형이라 부른다. 이 모형은 기존의 AFT 모형에서 제공하는 생존시간의 평균에 영향을 미치는 β를 추론할 수 있을 뿐만 아니라, α를 추정함으로써 생존시간의 분산에 영향을 미치는 α를 또한 동시에 추론 할 수 있는 장점이 있다. 모형 (2.2)에서 모든 개체 i가 동일한 분산(등분산)을 갖는다고 가정하면, 모형 (2.1)은 기존의 고전적인 AFT 모형 (Lawless, 1982; Park과 Ha, 2018)이 된다.

본 논문에서는 T_i에 대응하는 중도절단 시간(censoring time) C_i ($i = 1, \ldots, n$)에 대해 다음과 같은 두 가지를 가정한다 (Zhou, 2005; Park과 Ha, 2018).

- A1 : 공변량 x_i와 z_i가 주어졌을 때, T_i와 C_i는 조건부 독립이며 순서쌍 (T_i, C_i) 또한 조건부 독립이다.
- A2 : 공변량 x_i와 z_i가 주어졌을 때, C_i는 T_i에 대해 조건부 무정보적(conditionally non-informative)이다.

여기서 $i = 1, \ldots, n$이다.

2.2. 벌점화 변수선택

이 절에서는 벌점 가능도 함수를 사용하여 변수선택 절차를 유도하는 방법을 제안한다. 임의의 중도절단(random censoring)하에서 관측 가능한 확률 변수(observable random variables)는 다음과 같이 주어진다.

$$Y_i = \min(\log T_i, \log C_i) \quad \text{and} \quad \delta_i = I(T_i \le C_i).$$

여기서 $I(\cdot)$는 지시함수(indicator function)이며, δ_i는 중도절단 지시함수(censoring indicator)이다.

$f_\theta(\cdot)$와 $S_\theta(\cdot)$를 각각 관심모수 $\theta = (\beta^T, \alpha^T)^T$를 갖는 $\log T_i$에 대한 확률밀도함수(probability density function)와 생존함수(survival function)라 하자. 모형 (2.1)하에서 $\log T_i \sim N(x_i^T\beta, \sigma_i^2)$이므로, 두 가정 (A1과 A2)하에서 모형 (2.1)과 (2.2)의 로그가능도 함수는 다음과 같이 주어진다.

$$\begin{aligned}\ell(\theta; y_i, \delta_i) &= \sum_{i=1}^{n} \log\left[f_\theta(y_i)^{\delta_i} S_\theta(y_i)^{1-\delta_i}\right], \\ &= \sum_{i=1}^{n}\left[-\frac{\delta_i}{2}\left\{\log\left(2\pi\sigma_i^2\right) + m_i^2\right\} + (1-\delta_i)\log\{1-\Phi(m_i)\}\right]. \end{aligned} \tag{2.3}$$

여기서 $m_i = (y_i - x_i^T\beta)/\sigma_i$이며, $\sigma_i^2 = \exp(z_i^T\alpha)$이다.

모형 (2.1)과 (2.2)에서 고정효과인 β와 α에 대응하는 변수선택을 위해서, 본 논문에서는 다음과 같은 벌점 로그가능도 함수를 사용하며 ℓ_p로 표기한다 (Antoniadis 등, 2016).

$$\ell_p = \ell_p(\theta) = \ell(\theta) - n\sum_{j=0}^{p-1} J_{\gamma_1}(|\beta_k|) - n\sum_{k=0}^{q-1} J_{\gamma_2}(|\alpha_k|), \tag{2.4}$$

여기서 $J_\gamma(\cdot)$은 조율모수(tuning parameter) γ_j ($j = 1, 2$)를 갖는 벌점 함수이다. γ_j가 클수록 간단한 모형을 선택하고, γ_j가 작을수록 복잡한 모형을 선택하는 경향이 있다. 본 논문에서는 Wu와 Li (2012)의 제안에서와 같이 $\gamma_1 = \gamma_2 = \gamma$를 사용한다. 하지만 3절의 Table 1의 모의실험 결과에서 보여지는 바와 같이 이러한 조율모수의 제약은 LASSO 방법을 제외하고는 전반적으로 적절한 변수선택 결과를 주었다. 고려된 네 가지의 벌점 함수(LASSO, ALASSO, SCAD, HL)의 형태는 다음과 같다.

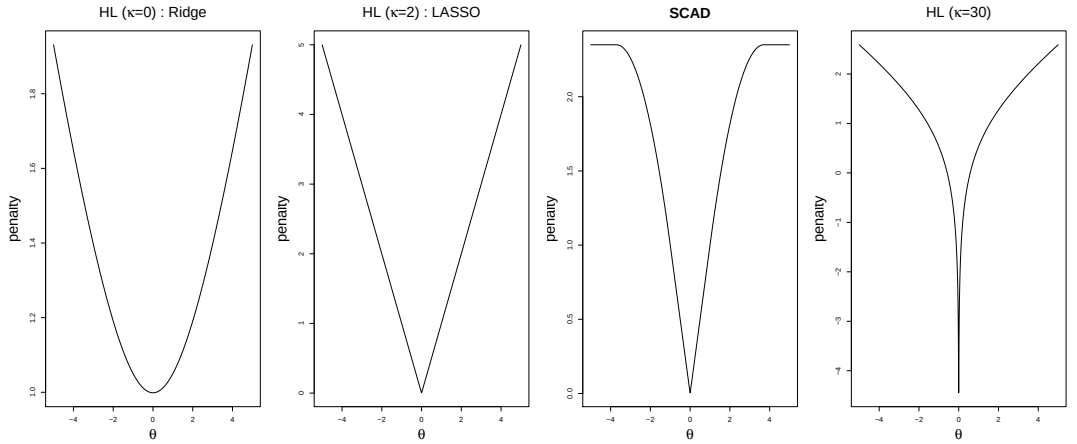

Figure 1: *The penalty functions: LASSO, SCAD, and HL.*

(1) LASSO (Tibshirani, 1996)

$$J_\gamma(|\theta|) = \gamma|\theta|, \tag{2.5}$$

(2) ALASSO (Zou, 2006)

$$J_\gamma(|\theta|) = \gamma|\theta|\omega, \tag{2.6}$$

ω는 기지(known)의 가중치 벡터이다. 본 논문에서는 $\omega = 1/|\hat{\theta}|$로써 $\hat{\theta}$는 평균-분산 AFT 모형에서의no-penalty 추정값을 사용한다 (Wang과 Song, 2011; Park과 Ha, 2018).

(3) SCAD (Fan과 Li, 2001)

$$J'_\gamma(|\theta|) = \gamma I(|\theta| \le \gamma) + \frac{(\omega\gamma - |\theta|)_+}{\omega - 1} I(|\theta| > \gamma), \tag{2.7}$$

여기서 $\omega = 3.7$이며 x_+는 x의 양수 부분이다.

(4) HL (Lee와 Oh, 2014)

$$J_{\gamma,\kappa}(|\theta|) = \frac{\gamma\theta^2}{2a(|\theta|)} + \frac{(\kappa - 2)\log a(|\theta|)}{2\kappa} + \frac{a(|\theta|)}{\kappa}, \tag{2.8}$$

여기서 $a(|\theta|) = [\{8\kappa\gamma\theta^2 + (2-\kappa)^2\}^{1/2} + (2-\kappa)]/4$이다.

특히 SCAD 벌점함수는 다른 세 가지 벌점함수와 달리 단순한 식을 제공하는 미분 함수식을 사용하였다. Figure 1은 $\gamma = 1$에서 세 가지 벌점화 함수(LASSO, SCAD, HL)의 형태를 나타낸 것이다. Figure 1에서 알 수 있듯이, HL은 κ의 값에 따라 그 형태가 변하며, 특히 $\kappa \to 0$인 경우 Ridge, $\kappa = 2$인 경우 LASSO, $\kappa > 2$인 경우 0에서의 값이 음의 무한대 값을 갖는 비유계 형태(unbounded form)를 갖는다. Lee와 Oh (2014)의 제안에 따라 본 논문에서는 식 (2.8)의 HL의 경우 $\kappa = 30$을 사용하였다. 좋은 벌점 함수는 세 가지 오라클(oracle) 성질인 불편성(unbiasedness), 성김성(sparsity), 연속성(continuity)을 만족해야 한다 (Fan과 Li, 2001, 2002). 벌점함수로써 잘 알려진 LASSO 접근법은 이러한 오라클 성질을 만족시키지 못한다. 하지만 나머지 세 벌점함수 (ALASSO, SCAD, HL) 접근법은 세 가지 오라클 성질을 만족하고, 정확한 부분집합 모형을 선택하는 동시에 실제로 0이 아닌 회귀계수를 추정하는 면에서 오라클 성질을 만족한다.

변수선택을 위해서 식 (2.4)의 벌점 로그가능도 함수 ℓ_p를 최대로 하는 θ의 추정량 $\hat{\theta} = (\hat{\beta}^T, \hat{\alpha}^T)^T$을 찾아야 하며, $\hat{\theta}$은 다음과 같이 주어진다.

$$\hat{\theta} = \arg\max_{\theta} \ell_p(\theta).$$

이와 같은 $\hat{\theta}$를 벌점화 최대가능도 추정량(penalized maximum likelihood estimators; PMLEs)이라고 부른다. 본 논문에서는 일련의 추정절차의 논의를 단순화 하기 위해 $\theta = (\beta^T, \alpha^T)^T$를 다음과 같이 표현한다.

$$\theta = (\theta_1, \ldots, \theta_s)^T = (\beta_0, \beta_1, \ldots, \beta_{p-1}, \alpha_0, \alpha_1, \ldots, \alpha_{q-1})^T.$$

여기서 $s = p + q$는 추정해야 할 회귀모수의 총 갯수이다. 그러면 θ의 PMLEs는 다음과 같은 θ_r $(r = 1, \ldots, s)$의 추정 방정식(estimating equations)을 통해 그 해를 얻을 수 있다.

$$\frac{\partial \ell_p}{\partial \theta_r} = \frac{\partial \ell}{\partial \theta_r} - n \sum_{r=0}^{s} [J_\gamma(|\theta_r|)]' = 0. \tag{2.9}$$

하지만 식 (2.9)에서 $|\theta_r|$는 0에서 미분 불가능하므로 우리는 다음과 같은 국소 이차 근사법(local quadratic approximation; LQA) (Fan과 Li, 2001)을 사용한다.

$$[J_\gamma(|\theta_r|)]' = J'_\gamma(|\theta_r|)\mathrm{sgn}(\theta_r) \approx \left\{ \frac{J'_\gamma\left(\left|\theta_r^{(0)}\right|\right)}{\left|\theta_r^{(0)}\right|} \right\} \theta_r, \quad \text{for } \theta_r \approx \theta_r^{(0)}.$$

여기서 $\mathrm{sgn}(\cdot)$은 부호함수이며, $\theta_r^{(0)}$는 θ_r의 실제값(true value)과 가까운 하나의 초기치이다. 따라서 θ의 벌점화 최대가능도 추정방정식은 다음과 같이 간단히 표현할 수 있다.

$$\frac{\partial \ell_p}{\partial \theta} = \frac{\partial \ell}{\partial \theta} - n\{\Sigma_\gamma(\theta)\}\theta = 0. \tag{2.10}$$

여기서

$$\Sigma_\gamma(\theta) = \mathrm{diag}\left\{ \frac{J'_\gamma(|\theta_r|)}{|\theta_r|} \right\} = \left(\Sigma_\gamma(\beta), \Sigma_\gamma(\alpha)\right).$$

는 $(p + q) \times (p + q)$ 차원을 갖는 대각선 행렬이며, $\Sigma_\gamma(\beta)$와 $\Sigma_\gamma(\alpha)$는 각각 $p \times p$와 $q \times q$ 대각선 행렬들이다. 따라서 Ha 등 (2014)의 방법에 의해, ℓ_p에 대한 θ의 $(p + q) \times (p + q)$ 차원의 음의 헤시안 행렬(negative hessian matrix) H_p를 얻을 수 있다.

$$H_p = H_p(\theta) = -\frac{\partial^2 \ell_p}{\partial \theta \partial \theta^T} = \begin{pmatrix} X^T W_1 X + n\Sigma_\gamma(\beta) & X^T W_2 Z \\ Z^T W_2 X & Z^T W_3 Z + n\Sigma_\gamma(\alpha) \end{pmatrix}, \tag{2.11}$$

여기서 X는 β의 $n \times p$ 모형 행렬이며, Z는 α의 $n \times q$ 모형 행렬이다. W_1, W_2, W_3는 모두 $n \times n$ 대각선 행렬로서 각 성분으로 w_{1i}, w_{2i}, w_{3i}를 갖는 가중행렬(weight matrix)이다. 식 (2.11)의 H_p에 대한 자세한 유도는 부록에 주어져 있다.

2.3. 벌점화 변수선택 알고리즘

식 (2.10)의 θ의 추정방정식은 θ에 대해 일반적으로 비선형함수이므로 그 추정량의 해를 구하기 위해 본 논문에서는 뉴튼-랩슨(Newton-Raphson)방법을 사용하며, 이에 대한 추정 절차는 다음과 같다.

$$\hat{\theta}^{(s+1)} = \hat{\theta}^{(s)} + \left[-\ell_p''\left(\hat{\theta}^{(s)}\right)\right]^{-1} \ell_p'\left(\hat{\theta}^{(s)}\right), \quad s = 0, 1, 2, \ldots, \tag{2.12}$$

여기서 $\ell_p'(\theta) = \partial \ell_p(\theta)/\partial\theta$와 $-\ell_p''(\theta) = H_p(\theta)$이며, 이것은 (2.10)과 (2.11)을 각각 나타낸다.

본 논문에서는 $\hat{\theta}$에 대한 표준오차(standard error; SE)를 계산하기 위해 다음의 샌드위치(sandwich) 분산-공분산 행렬을 사용한다 (Fan과 Li, 2001; Ha 등, 2014).

$$\operatorname{cov}\left(\hat{\theta}\right) = \left(H_{\theta\theta} + n\Sigma_\gamma(\theta)\right)^{-1} H_{\theta\theta}\left(H_{\theta\theta} + n\Sigma_\gamma(\theta)\right)^{-1}, \tag{2.13}$$

여기서 $H_{\theta\theta} = -\partial^2\ell/\partial\theta\partial\theta^T$이다. Wang 등 (2007)은 일반화 교차검증(generalized cross validation; GCV) 접근법 (Fan과 Li, 2001, 2002)이 조율모수의 선택 시 얻어지는 모형에서 무시할 수 없는 과대적합(overfitting) 효과를 줄 수 있기 때문에, GCV 접근법이 조율모수 γ를 정확하게 선택 할 수 없다는 것을 밝혔다. 따라서 γ를 결정하기 위해서 본 논문에서는 다음과 같은 Bayesian information criterion (BIC)의 한 형태를 사용한다 (Ha 등, 2014).

$$\mathrm{BIC}(\gamma) = -2\ell\left(\hat{\theta}, \hat{\phi}\right) + \log(n)\mathrm{df}, \tag{2.14}$$

여기서 $\mathrm{df} = \mathrm{tr}[(H_{\theta\theta} + n\sum_\gamma(\theta))^{-1}H_{\theta\theta}]$는 효율적인 자유도(effective degree of freedom)이다.

제안된 변수선택 방법의 적합 알고리즘(fitting algorithm)은 다음과 같다.

Step 1. β와 α의 초기값을 구한다.

Step 2. 내부 루프(inner loop)에서 식 (2.4)의 벌점 로그가능도 ℓ_p를 최대화 시키는 $\hat{\beta}$와 $\hat{\alpha}$를 구하기 위해 뉴튼-랩슨 방법 (2.12)를 사용한다.

Step 3. 외부 루프(outer loop)에서 식 (2.14)의 $\mathrm{BIC}(\gamma)$를 최소화 시키는 γ를 찾는다.

위에서 제시한 알고리즘이 모두 수렴한 후, 식 (2.13)을 이용하여 $\hat{\beta}$와 $\hat{\alpha}$에 대한 추정된 표준오차를 계산한다. LASSO와 ALASSO의 β와 α 초기값으로는 no-penalty를 적합해서 얻어진 추정값을 사용하며, SCAD와 HL의 초기값으로는 LASSO에서 계산된 추정값을 사용한다 (Lee와 Oh, 2014; Ha 등, 2014).

3. 모의실험 연구

이 절에서는 평균-분산 AFT 모형에서 네 가지 변수선택 방법 (LASSO, ALASSO, SCAD, HL)의 성능을 비교 평가하기 위해 모의실험을 수행한다. 여기서 모의실험의 반복횟수는 Fan과 Li (2001, 2002)와 Ha 등 (2014)에서 처럼 100번을 사용한다.

3.1. 모의실험 설계

모의실험 설계 방법으로 Fan과 Li (2001, 2002)의 방법을 사용하며, 모의실험 자료는 평균-분산 AFT 모형 (2.1)과 (2.2)에서 생성하였으며, 실제(true) 회귀계수 β와 α는 4절에서 제시한 폐암 자료에 대해 SCAD를 이용한 변수선택의 결과 추정값에 근거하였다. 여기서 실제 회귀계수는 다음과 같다.

$$\begin{aligned}\beta &= (\beta_0,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5,\beta_6,\beta_7,\beta_8,\beta_9,\beta_{10},\beta_{11},\beta_{12},\beta_{13},\beta_{14},\beta_{15},\beta_{16},\beta_{17},\beta_{18},\beta_{19},\beta_{20},\beta_{21},\beta_{22},\beta_{23},\beta_{24},\beta_{25})^T\\ &= (2.1, \underbrace{0,\ldots,0}_{5}, 1.5, 0.4, 0.5, 1.1, 0, -0.5, -1.2, -1.8, -0.5, 0, 0, -0.3, 0, -0.4, 0, -0.45, \underbrace{0,\ldots,0}_{4})^T,\\ \alpha &= (\alpha_0,\alpha_1,\alpha_2,\alpha_3,\alpha_4,\alpha_5,\alpha_6,\alpha_7,\alpha_8,\alpha_9,\alpha_{10},\alpha_{11},\alpha_{12},\alpha_{13},\alpha_{14},\alpha_{15},\alpha_{16},\alpha_{17},\alpha_{18},\alpha_{19},\alpha_{20},\alpha_{21},\alpha_{22},\alpha_{23},\alpha_{24},\alpha_{25})^T\\ &= (\underbrace{0,\ldots,0}_{5}, 0.5, 0, 0, -0.46, -0.9, \underbrace{0,\ldots,0}_{5}, -0.3, \underbrace{0,\ldots,0}_{9}, 0.7)^T.\end{aligned}$$

Table 1: Simulation results using 100 replications under mean-variance AFT model

Cen.	n	Method	$\hat{\beta}$				$\hat{\alpha}$			
			C(14)	IC(0)	PT	MSE	C(21)	IC(0)	PT	MSE
25%	855	LASSO	10.73	0	0.03	0.010	18.98	0	0.16	0.033
		ALASSO	13.85	0	0.85	0.006	20.75	0	0.79	0.022
		SCAD	11.90	0	0.06	0.006	20.93	0	0.93	0.018
		HL	13.42	0	0.52	0.007	20.76	0	0.80	0.018
	2500	LASSO	11.28	0	0.02	0.003	19.12	0	0.22	0.013
		ALASSO	14.00	0	1.00	0.002	20.99	0	0.99	0.016
		SCAD	13.22	0	0.52	0.002	20.97	0	0.97	0.005
		HL	13.97	0	0.97	0.002	20.87	0	0.87	0.005
50%	855	LASSO	11.01	0	0.06	0.026	18.47	0	0.11	0.042
		ALASSO	13.70	0	0.71	0.016	20.66	0.03	0.66	0.039
		SCAD	12.45	0	0.19	0.012	20.77	0.01	0.78	0.021
		HL	13.61	0	0.65	0.013	20.70	0	0.74	0.024
	2500	LASSO	10.75	0	0.02	0.006	17.96	0	0.03	0.012
		ALASSO	13.98	0	0.98	0.007	20.72	0	0.72	0.031
		SCAD	12.95	0	0.22	0.004	20.82	0	0.83	0.007
		HL	13.90	0	0.90	0.004	20.85	0	0.85	0.008

Cen. = censoring rate; C = average number of coefficients, of the true zeros, correctly set to zero; IC = average number of the true nonzero incorrectly set to zero; PT = probability of choosing the true model; MSE, median of mean square errors; C(14), IC(0), and C(21) indicate corresponding optimal values.

특히 실제 회귀계수 β에 해당하는 공변량 벡터 $x = (1, x^*)$, $x^* = (x_1, \ldots, x_{25})^T$는 다중공선성(multicollinearity) 구조를 만들기 위해 상관계수를 0.5로 하는 AR(1) 모형에서 생성하였으며 (Fan과 Li, 2001, 2002), α에 해당하는 공변량 벡터 $z = (1, z^*)$, $z^* = (z_1, \ldots, z_{25})^T$는 편의상 $x = (1, x^*)$와 동일한 공변량을 사용하였다. x^*에서 중요한 변수는 0이 아닌 β에 대응하는 총 11개($x_6, x_7, x_8, x_9, x_{11}, x_{12}, x_{13}, x_{14}, x_{17}, x_{19}, x_{21}$)의 변수이다. z^*에서 중요한 변수 또한 0이 아닌 α에 대응하는 총 5개($z_5, z_8, z_9, z_{15}, z_{25}$)의 변수이다. 표본의 크기(sample size)는 총 두 가지 경우인 $n = 855$와 $n = 2500$를 고려하였으며, 특히 표본의 크기로 $n = 855$를 사용한 이유는 4절의 폐암 자료의 환자 수와 동일하게 설정하기 위해서이다. 중도절단시간은 지수분포를 사용하여, 25%(낮은 중도절단 비율)와 50%(높은 중도절단 비율)가 되도록 생존자료를 생성하였다.

변수선택의 성능을 평가하기 위해서 PT, MSE, C와 IC를 고려하였다. 여기서 PT는 실제 모형을 제대로 선택 할 확률 probability of choosing the true model (PT)을 나타낸다. Model square error (MSE)는 모형 제곱오차 이다. Fan과 Li (2001)의 방법에 따라 모형 오차로 $\mathrm{MSE}(\hat{\theta}) = (\hat{\theta} - \theta)^T E(XX^T)(\hat{\theta} - \theta)$를 사용하지만, 실제 MSE 값으로는 n개의 MSE값들에 대한 중앙값(median)을 사용하였다. "C"는 β(또는 α)에서 중요하지 않은 변수를 0으로 정확하게 찾아내는 평균 수를 나타내며, 여기서 β의 경우 14, α의 경우 21이 최적의 값이다. 그리고 "IC"는 중요한 변수 (0이 아닌 회귀계수)를 0으로 잘못 찾아내는 평균 수를 나타내며, β와 α 모두 0이 최적의 값이다. 또한 모의실험에서의 추정값 결과로써, 100번의 반복 시행에서 추정된 $\hat{\beta}$와 $\hat{\alpha}$ 각각의 추정치들에 대한 평균(mean), 표준편차(standard deviation; SD), 표준오차들의 평균(mean of the estimated standard errors; SEs)을 계산하였다.

3.2. 모의실험 결과

모의실험 결과는 Table 1과 같다. 먼저 중도절단 비율이 작은 25%하에서 관측된 결과는 다음과 같다.

(i) $\hat{\beta}$에 대해서는 표본크기 n에 상관 없이 "C"에 대해서는 ALASSO가 최적값 14에 가장 가까우며, 다음으로는 HL의 "C"가 최적값에 가깝다. PT에 대해서도 ALASSO가 가장 큰 것을 알 수 있으며, 다음으로는 HL이다. 하지만 두 방법간의 차이는 크지 않다. 특히 LASSO와 SCAD는 "C"와 PT의 결과가 좋지 않다. "IC"는 4가지 방법 모두 0으로써, 0이 아닌 추정값들을 모두 0이 아니라고 잘 추정한다는 것을 알 수 있다. MSE 는 ALASSO가 가장 크고 SCAD 방법이 가장 작다.

(ii) $\hat{\alpha}$에 대해서는 표본이 작을 경우 (n = 855), SCAD가 "C"의 최적값 21에 가장 가까우며, 다음으로는 HL, ALASSO, LASSO 순서이다. PT에 대해서는 SCAD가 가장 큰 것을 알 수 있으며, 다음으로는 HL, ALASSO, LASSO이다. 표본이 큰 경우 (n = 2500) 그 모의실험 결과는 본 논문에서 제시하지는 않았지만 ALASSO가 "C"와 PT 면에서 가장 좋은 결과를 보이고, 다음으로는 SCAD와 HL, LASSO 순서이다. MSE는 여전히 ALASSO가 가장 크고 SCAD 방법이 가장 작게 나타났다.

다음으로 중도절단 비율이 큰 50%하에서의 결과에 따르면, $\hat{\beta}$에서는 25%중도절단 비율에서와 마찬가지로 표본크기에 상관 없이 "C"와 PT 측면에서 ALASSO가 가장 좋은 결과값을 가진다. 다음으로는 HL로써, 두 방법간의 차이는 크지 않으며, SCAD와 LASSO는 결과값이 좋지 않다. $\hat{\alpha}$에서는 표본크기에 상관 없이 "C"와 PT 측면에서 SCAD와 HL이 비슷하게 좋은 결과값을 가지는 것을 알 수 있다.

정리하면, LASSO는 모든 경우에서 "C", PT, MSE가 가장 좋지 않은 결과를 준다. SCAD는 중도절단 비율과 표본크기에 상관 없이 네 가지 방법 중 가장 작은 MSE를 가지고, $\hat{\alpha}$의 "C"와 PT 측면에서 좋은 결과값을 가지지만, $\hat{\beta}$의 측면에서 "C"가 최적값과 많은 차이가 나며, PT역시 작은 값을 가지는 것을 알 수 있었다. ALASSO는 중도절단 비율이 50%로 크고 표본크기가 n = 855로 작은 경우 $\hat{\alpha}$의 "C"와 PT의 결과가 다른 방법에 비해 약간 좋지 않지만, 표본의 크기가 n = 2500으로 커지는 경우 "C"와 PT 모두 크게 개선되는 것으로 나타났다. HL은 전반적으로 좋은 결과를 주지만, n이 작은 경우 $\hat{\beta}$의 "C"와 PT 측면에서 ALASSO보다는 약간 성능이 좋지 않음을 알 수 있다.

나아가, 3.1절의 실제 모수값이 0아닌 평균 및 분산의 회귀계수에 대한 모의실험 추정값의 결과는 Table 2에 제시하였다. 그 결과는 다음과 같이 요약된다.

(i) $\hat{\beta}$에서는 표본크기 n에 상관 없이 작은 중도절단 비율 (25%)을 가질 때는 SCAD의 추정값이 실제 β값에 가장 가깝게 나타난다. 하지만 나머지 세 가지 벌점 함수 역시 SCAD와 큰 차이를 보이지 않는다. 중도절단 비율이 50%로 큰 경우(not shown) SCAD와 HL의 추정값이 거의 유사하지만, SCAD가 실제 β와 매우 가까워서 가장 정확하게 추정한다. 그 다음으로는 ALASSO, LASSO 순서이다. 한편, SE의 측면에서는 네 가지 벌점 함수 모두 n과 중도절단 비율과 상관 없이 SD와 SE는 거의 같은 결과를 보이기 때문에 식 (2.13)에서 제시한 SE가 적절하다는 것을 알 수 있다.

(ii) $\hat{\alpha}$에서는 표본 크기 및 중도절단 비율과 상관 없이 HL의 추정값이 가장 정확하게 나타나며, 다음으로는 SCAD이다. LASSO와 ALASSO의 추정값은 실제 α와 큰 차이를 보이는 경향이 있다. SE의 측면에서는 SCAD와 HL의 SD와 SE가 매우 유사하므로 두 방법은 SE를 잘 추정한다. LASSO와 ALASSO는 중도절단 비율이 50%로 클 때 SD와 SE가 큰 차이를 보이는 경우가 있음을 알 수 있다.

결론적으로 Tables 1–2의 모의실험 결과에 따르면, 평균-분산 AFT 모형의 변수선택에서 LASSO는 전반적으로 좋지 않은 반면, 나머지 세 가지 방법(ALASSO, SCAD, HL)은 LASSO의 단점을 크게 개선하는 것으로 나타난다.

4. 폐암 자료의 적합

2절에서 제안한 방법의 예증을 위해서 북아일랜드의 한 폐암 임상연구로부터 제시된 폐암 자료를 사용한다 (MacKenzie, 1996). 연구 대상자는 총 855명의 폐암 환자이다. 생존시간은 진단일로부터 사망일까지의 시간

Table 2: The simulation results (Mean, SD, and SE) using 100 replications of estimates of non-zero mean coefficients (β) and non-zero variance coefficients (α); n = 855 with 20% censoring

Mean coef.	Method	$\hat{\beta}_0$			$\hat{\beta}_6$			$\hat{\beta}_7$		
		Mean	SD	SE	Mean	SD	SE	Mean	SD	SE
True		2.100			1.500			0.400		
$\hat{\beta}$	LASSO	2.070	0.035	0.036	1.482	0.021	0.023	0.401	0.026	0.024
	ALASSO	2.097	0.033	0.035	1.499	0.020	0.022	0.397	0.026	0.023
	SCAD	2.100	0.032	0.034	1.497	0.020	0.021	0.405	0.025	0.023
	HL	2.098	0.032	0.034	1.496	0.021	0.021	0.404	0.025	0.023
Mean coef.	Method	$\hat{\beta}_8$			$\hat{\beta}_9$			$\hat{\beta}_{11}$		
		Mean	SD	SE	Mean	SD	SE	Mean	SD	SE
True		0.500			1.100			−0.500		
$\hat{\beta}$	LASSO	0.502	0.024	0.025	1.096	0.023	0.024	−0.491	0.025	0.023
	ALASSO	0.498	0.024	0.024	1.097	0.022	0.023	−0.494	0.024	0.022
	SCAD	0.499	0.024	0.023	1.098	0.022	0.022	−0.502	0.024	0.021
	HL	0.500	0.024	0.023	1.098	0.023	0.022	−0.501	0.024	0.021
Mean coef.	Method	$\hat{\beta}_{12}$			$\hat{\beta}_{13}$			$\hat{\beta}_{14}$		
		Mean	SD	SE	Mean	SD	SE	Mean	SD	SE
True		−1.200			−1.800			−0.500		
$\hat{\beta}$	LASSO	−1.193	0.026	0.025	−1.793	0.024	0.026	−0.490	0.022	0.023
	ALASSO	−1.200	0.026	0.024	−1.802	0.023	0.025	−0.492	0.021	0.022
	SCAD	−1.200	0.025	0.023	−1.799	0.023	0.024	−0.499	0.021	0.021
	HL	−1.199	0.026	0.023	−1.799	0.024	0.024	−0.498	0.022	0.021
Mean coef.	Method	$\hat{\beta}_{17}$			$\hat{\beta}_{19}$			$\hat{\beta}_{21}$		
		Mean	SD	SE	Mean	SD	SE	Mean	SD	SE
True		−0.300			−0.4			−0.450		
$\hat{\beta}$	LASSO	−0.292	0.024	0.020	−0.394	0.022	0.021	−0.438	0.020	0.020
	ALASSO	−0.292	0.022	0.019	−0.399	0.021	0.020	−0.443	0.019	0.019
	SCAD	−0.298	0.023	0.018	−0.404	0.021	0.019	−0.448	0.019	0.019
	HL	−0.299	0.023	0.019	−0.401	0.022	0.020	−0.446	0.020	0.019
Var coef.	Method	$\hat{\alpha}_5$			$\hat{\alpha}_8$			$\hat{\alpha}_9$		
		Mean	SD	SE	Mean	SD	SE	Mean	SD	SE
True		0.500			−0.460			−0.900		
$\hat{\alpha}$	LASSO	0.437	0.052	0.043	−0.425	0.056	0.047	−0.871	0.055	0.051
	ALASSO	0.470	0.059	0.046	−0.428	0.060	0.050	−0.913	0.056	0.055
	SCAD	0.520	0.055	0.050	−0.479	0.053	0.058	−0.921	0.054	0.058
	HL	0.509	0.056	0.049	−0.472	0.057	0.056	−0.916	0.056	0.057
Var coef.	Method	$\hat{\alpha}_{15}$			$\hat{\alpha}_{25}$					
		Mean	SD	SE	Mean	SD	SE			
True		−0.300			0.700					
$\hat{\alpha}$	LASSO	−0.246	0.052	0.039	0.641	0.054	0.045			
	ALASSO	−0.239	0.064	0.038	0.679	0.055	0.048			
	SCAD	−0.296	0.063	0.047	0.714	0.052	0.050			
	HL	−0.298	0.055	0.048	0.706	0.053	0.050			

(단위: 월)이며, 연구 종료 시점(study end date)에 사망하지 않은 환자들은 우측 중도절단(right censoring)이다. 855명의 환자 중 182명이 연구 종료 시점까지 사망하지 않았으므로 중도절단 비율은 21.3%이다. 분석에서 사용된 변수들에 대한 자세한 설명은 Table 3과 같다.

Table 3: Explanation of variables for lung cancer data

변수	코딩 및 설명
Survtime	진단일로부터 사망일까지의 시간 (월)
Censoring	폐암에 의한 생존여부 (0 : 생존, 1 : 사망)
Agegrp	나이에 따른 그룹화 (1 : <40; 2 : <50; 3 : <60; 4 : <70; 5 : 80+)
Sex	성별 (0 : 여자, 1 : 남자)
Treat	치료 방법 1 : 일시적인 처방(palliative), 2 : 수술(surgery), 3 : 약물 치료(chemotherapy), 4 : 방사선 치료(radiotherapy), 5 : 약물 치료와 방사선 치료의 병행
Who	활동정도(performance status) 1 : 적당한 일상생활이 가능, 2 : 가벼운 작업이 가능, 3 : 작업 불가능, 4 : 깨어 있는 시간의 50% 이상 걸을 수 있음, 5 : 활동이 침대 또는 의자에 국한
Cell	세포 종류(cell type) 1 : 편평상피 세포(squamous cell), 2 : 소세포(small cell), 3 : 선암(adenocarcinoma), 4 : 그 외
Sod	혈청 나트륨(serum sodium) 농도 1 : ≥136mmol/l, 2 : <136mmol/l, 3 : 결측
Alb	혈청 알부민(serum albumen) 농도 (1 : ≥35g/l, 2 : <35g/l, 3 : 결측)
Met	전이여부(metastases) (1 : 전이 안 됨, 2 : 전이 됨, 3 : 알 수 없음)
Smoking	흡연여부 (1 : 비흡연자, 2 : 흡연자, 3 : 흡연경력 있음, 4 : 결측)

먼저 폐암 생존시간 T가 로그정규 분포를 따르는지 확인하기 위해서 로그정규 위험 그래프(lognormal hazard plot) (Klein과 Moeschberger, 2003)를 사용하였다. 즉 $\Phi^{-1}(1-\hat{S}_0(t))$와 $\log t$의 관계의 그래프가 직선의 형태로 나타난다면 해당 생존자료가 로그정규 분포를 따른다고 할 수 있다. 여기서 $\Phi(\cdot)$는 N(0,1)의 누적분포함수를 나타내며, $\hat{S}_0(t)$는 T의 기저 생존함수(baseline survival function) $S_0(t)$에 대한 카플란-마이어(Kaplan-Meier; KM) 추정값을 나타낸다. 이를 그래프로 표현한 Figure 2는 거의 선형의 형태로 나타나므로, 본 폐암자료의 생존시간이 로그정규 분포를 따른다고 볼 수 있다. 이러한 결과는 2절의 평균-분산 AFT 모형을 사용할 수 있는 근거가 될 수 있다.

Table 4는 평균-분산 AFT 모형에서 변수선택을 위해 추정된 회귀계수와 표준오차의 결과이다. 유의수준 5%를 기준으로 분석한 결과, no-penalty에서 평균 생존시간에 중요하게 영향을 미치는 변수는 총 12개 (treat2, treat3, treat4, treat5, who3, who4, who5, cell2, cell4, sod2, alb2, met2)이며, 생존시간의 분산에 유의한(즉, 중요한) 영향을 미치는 변수는 총 5개(sex1, treat4, treat5, cell3, sod3)로 나타난다. 또한 BIC를 최소로 하는 조율모수의 값은 LASSO가 0.009, ALASSO가 0.005, SCAD가 0.058, HL가 0.011이었다. 전반적으로 LASSO는 다른 세 가지 방법에 비해 no-penalty에서 유의하지 않은 변수를 많이 선택함을 알 수 있다 (Ha와 Lee, 2014; Park과 Ha, 2018).

Table 4: Variable selection under mean-variance AFT model for lung cancer data: $\hat{\beta}$ and $\hat{\alpha}$, estimates of mean and variance coefficients

Covariate	No penalty		LASSO		ALASSO		SCAD		HL	
	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$
Intercept	2.91 (0.34)∗	1.00 (0.50)∗	2.01 (0.09)	0.05 (0.03)	2.20 (0.09)	0.30 (0.07)	2.16 (0.10)	**0 (0)**	2.32 (0.10)	**0 (0)**
agegrp2	−0.14 (0.26)	-0.70 (0.39)	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**
agegrp3	−0.20 (0.25)	-0.63 (0.36)	**0 (0)**	−0.08 (0.05)	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**
agegrp4	−0.17 (0.26)	-0.38 (0.37)	**0 (0)**	0.05 (0.03)	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**
agegrp5	−0.34 (0.29)	-0.31 (0.40)	**0 (0)**	0.16 (0.07)	**0 (0)**	0.09 (0.04)	**0 (0)**	**0 (0)**	**0 (0)**	0.26 (0.10)
sex1	−0.01 (0.08)	0.31 (0.13)∗	**0 (0)**	0.31 (0.07)	**0 (0)**	**0 (0)**	**0 (0)**	0.50 (0.08)	**0 (0)**	0.32 (0.08)
treat2	1.28 (0.26)∗	0.10 (0.41)	1.23 (0.14)	**0 (0)**	1.34 (0.17)	**0 (0)**	1.54 (0.19)	**0 (0)**	1.30 (0.17)	**0 (0)**
treat3	0.42 (0.21)∗	0.15 (0.32)	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**	0.45 (0.21)	**0 (0)**	**0 (0)**	**0 (0)**
treat4	0.36 (0.10)∗	−0.44 (0.15)∗	0.39 (0.07)	−0.33 (0.09)	0.37 (0.07)	-0.32 (0.08)	0.48 (0.09)	−0.46 (0.12)	0.42 (0.08)	−0.41 (0.11)
treat5	0.96 (0.19)∗	−0.88 (0.40)∗	0.58 (0.13)	**0 (0)**	0.76 (0.15)	-0.53 (0.17)	1.15 (0.18)	−0.91 (0.31)	0.82 (0.17)	-0.38 (0.15)
who2	0.00 (0.17)	−0.15 (0.28)	0.32 (0.07)	-0.15 (0.07)	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**
who3	−0.49 (0.17)∗	0.10 (0.28)	−0.18 (0.06)	0.03 (0.03)	−0.49 (0.08)	0.06 (0.04)	−0.54 (0.09)	**0 (0)**	−0.52 (0.09)	**0 (0)**
who4	−1.02 (0.19)∗	0.07 (0.30)	−0.73 (0.09)	**0 (0)**	−1.14 (0.11)	0.13 (0.08)	−1.16 (0.11)	**0 (0)**	−1.11 (0.11)	**0 (0)**
who5	−1.60 (0.32)∗	−0.07 (0.43)	−0.93 (0.16)	**0 (0)**	−1.63 (0.23)	**0 (0)**	-1.79 (0.26)	**0 (0)**	−1.65 (0.24)	**0 (0)**
cell2	−0.65 (0.16)∗	−0.11 (0.25)	−0.26 (0.08)	**0 (0)**	−0.32 (0.09)	**0 (0)**	-0.54 (0.14)	**0 (0)**	−0.30 (0.10)	**0 (0)**
cell3	−0.21 (0.13)	−0.42 (0.21)∗	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**	−0.27 (0.16)	**0 (0)**	**0 (0)**
cell4	−0.23 (0.10)∗	0.04 (0.16)	−0.10 (0.05)	0.15 (0.06)	−0.04 (0.02)	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**	0.16 (0.07)
sod2	−0.23 (0.09)∗	−0.05 (0.13)	−0.24 (0.07)	**0 (0)**	−0.17 (0.05)	**0 (0)**	−0.25 (0.08)	**0 (0)**	−0.24 (0.07)	**0 (0)**
sod3	0.09 (0.18)	−0.71 (0.35)∗	**0 (0)**	−0.06 (0.02)	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**
alb2	−0.45 (0.09)∗	−0.06 (0.14)	−0.41 (0.07)	0.06 (0.04)	−0.37 (0.07)	0.02 (0.02)	−0.40 (0.08)	**0 (0)**	−0.37 (0.08)	**0 (0)**
alb3	−0.22 (0.16)	−0.20 (0.26)	−0.14 (0.07)	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**
met2	−0.61 (0.11)∗	0.01 (0.20)	−0.55 (0.08)	0.01 (0.01)	−0.48 (0.07)	**0 (0)**	−0.45 (0.08)	**0 (0)**	−0.61 (0.09)	**0 (0)**
met3	−0.24 (0.13)	−0.07 (0.22)	−0.16 (0.07)	**0 (0)**	−0.05 (0.02)	**0 (0)**	**0 (0)**	**0 (0)**	−0.23 (0.09)	**0 (0)**
smoking2	−0.26 (0.15)	−0.40 (0.22)	0.00 (0.01)	−0.10 (0.06)	**0 (0)**	-0.08 (0.03)	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**
smoking3	−0.19 (0.15)	−0.15 (0.21)	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**
smoking4	−0.34 (0.33)	0.13 (0.41)	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**	**0 (0)**	0.66 (0.34)	**0 (0)**	**0 (0)**

Note : ∗ = significance at level 5%.

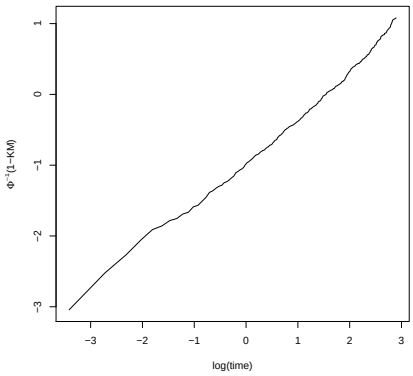

Figure 2: *Plot of* $\Phi^{-1}(1-KM)$ *against log of survival time for lung cancer data; KM is Kaplan-Meier estimate and* $\Phi(\cdot)$ *is CDF of* $N(0,1)$.

Table 5: Accuracy of variable selection for lung cancer data

Method	β		α	
	FN	FP	FN	FP
LASSO	1	5	2	9
ALASSO	1	1	3	5
SCAD	1	0	3	1
HL	2	1	2	1

Note : false negative, FN; false positive, FP.

Table 4의 no-penalty에서 중요(유의)한 변수를 네 가지 변수선택법에서 중요하지 않은 변수(즉, 추정된 회귀계수가 0)로 선택하는 경우를, 본 논문에서는 그 결과가 실제값은 아니지만 단지 정밀도 측면의 분류상 제안된 방법의 성능 비교를 위해서 false negative (FN)이라 하였다. 그러면 no-penalty에서 중요(유의)하지 않은 변수를 네 가지 변수선택법에서 중요한 변수(즉, 추정된 회귀계수가 0이 아님)처럼 잘못 선택하는 경우를 false positive (FP)가 된다. 두 경우 모두 0이 최적값이다. 특히 FN과 FP는 Table 1의 모의실험 결과에서 "IC"와 "C의 반대 경우(complement of C)"를 각각 의미한다. 이러한 결과를 요약한 Table 5에 의하면, 네 가지 방법 중 LASSO 방법이 변수를 가장 정확하지 않게 선택하는 것을 알 수 있다. 특히 FN에서는 네가지 방법이 비슷하지만 FP에서는 LASSO방법이 매우 좋지 않은 결과를 준다. 이러한 결과는 Table 1의 모의실험 결과를 반영하는 것임을 확인할 수 있으며, 특히 Table 5에서 LASSO 방법이 "C" 측면에서 성능이 좋지 않다는 사실을 다시 한번 확증한다고 볼 수 있다.

5. 결론 및 향후과제

본 연구에서는 생존시간의 평균 및 분산에 영향을 미치는 변수에 대해 분석하는 평균-분산 AFT 모형에서 변수선택 방법을 제안하고, 모의실험 및 폐암 임상자료를 사용하여 제안된 방법을 평가 및 예증 하였다. 분석 결과 LASSO 방법은 중요하지 않은 변수를 많이 선택하는 경향을 보였으며, 나머지 세 가지 방법(ALASSO, SCAD, HL)은 LASSO 방법을 개선하여 중요한 변수를 비슷하게 선택하는 것으로 나타났다.

본 논문에서 제시한 두 개의 조율모수를 갖는 벌점 로그가능도 (2.4)에서 변수선택의 편의를 위해 하나의

똑같은 조율모수의 설정을 사용하였다. 하지만 두 개의 조율모수를 사용하는 변수선택 절차를 개발하는 것이 보다 좋은 결과를 줄 것이다. 3절 모의실험에서는 Fan과 Li (2002)의 설계와 같이 설명변수에 대해 표준화를 취하지 않았으며, 4절 폐암자료에서는 모든 설명변수들이 연속형이 아닌 범주형이며 결과 해석상의 문제로 인해 표준화를 또한 취하지 않았다. 하지만, 이러한 비표준화가 제안된 변수선택의 성능에 영향을 미칠 수 있기 때문에 추후연구에서 두 경우를 비교하는 것도 흥미있는 연구가 될 것이다.

한편, 모의실험 설계의 난이도를 파악하기 위해, 영이 아닌 회귀계수들에 해당하는 설명 변수들만 가지고 no-penalty 하에서 평균-분산 AFT 모형을 적합한 결과(oracle 결과라 부름)와 실제 자료 분석에서 사용한 것과 같은 단순한 변수 선택 방법인 hard thresholding(p-value < 0.05 등) 결과와 비교하는 것도 하나의 흥미있는 향후 연구가 또한 될 것이다.

제안된 방법은 로그정규 분포에서의 모수적 벌점 가능도 접근법을 사용했다. 따라서 생존시간의 분포를 완화한 평활법(smoothing method)이나 확률분포를 가정하지 않는 비모수적 평균-분산 AFT 모형으로의 확장하여 제안된 방법의 강건성(robustness)에 대한 추가적인 연구가 필요할 것으로 사료된다. 나아가, 제안된 방법에 변량 효과(random effect)를 허락하여 다변량 생존자료(multivariate survival data) 분석으로 확장한다면 더욱 유용한 모형 접근법이 될 것이다.

Appendix: 평균-분산 AFT 모형에서 음의 헤시안 행렬

평균-분산 AFT 모형의 로그가능도 (2.3)으로 부터, 회귀계수 β_j ($j = 0, 1, \ldots, p-1$)와 α_k ($k = 0, 1, \ldots, q-1$)을 갖는 $\theta = (\beta^T, \alpha^T)^T$에 관한 ℓ의 1차 미분은 다음과 같다.

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \sigma_i^{-1} \left[\delta_i m_i + (1-\delta_i) h(m_i)\right] x_{ij}, \tag{A.1}$$

$$\frac{\partial \ell}{\partial \alpha_k} = \frac{1}{2} \sum_{i=1}^{n} \left[\delta_i \left(m_i^2 - 1\right) + (1-\delta_i) h(m_i) m_i\right] z_{ik}. \tag{A.2}$$

여기서 $\sigma_i^2 = \exp(z_i^T \alpha)$이고, $h(\cdot) = \phi(\cdot)/\{1 - \Phi(\cdot)\}$는 $N(0,1)$의 위험함수이며, $\phi(\cdot)$와 $\Phi(\cdot)$는 각각 $N(0,1)$의 확률밀도함수와 누적분포함수이다. 식 (A.1)과 (A.2)로 부터 θ에 관한 ℓ의 2차 미분은 다음과 같이 표현된다.

$$-\frac{\partial^2 \ell}{\partial \beta_r \partial \beta_s} = \sum_{i=1}^{n} x_{ir} w_{i1} x_{is}, \quad -\frac{\partial^2 \ell}{\partial \beta_r \partial \alpha_s} = \sum_{i=1}^{n} x_{ir} w_{i2} z_{is}, \quad -\frac{\partial^2 \ell}{\partial \alpha_r \partial \alpha_s} = \sum_{i=1}^{n} z_{ir} w_{i3} z_{is}.$$

여기서 $w_{i1} = [\delta_i + (1-\delta_i)\xi(m_i)]/\sigma_i^2$, $w_{i2} = 2w_{i3}/m_i\sigma_i$, $w_{i3} = m_i[2\delta_i m_i + (1-\delta_i)(h(m_i) - m_i\xi(m_i))]/4$이며, $\xi(m_i) = h(m_i)\{m_i - h(m_i)\}$이다. 따라서, 식 (2.10)에 대한 θ의 미분으로 부터 식 (2.11)의 음의 헤시안 행렬 H_p는 다음과 같이 표현됨을 알 수 있다:

$$\begin{aligned} H_p = -\frac{\partial^2 \ell_p}{\partial \theta \partial \theta^T} &= \begin{pmatrix} -\frac{\partial^2 \ell}{\partial \beta \partial \beta^T} & -\frac{\partial^2 \ell}{\partial \beta \partial \alpha^T} \\ -\frac{\partial^2 \ell}{\partial \alpha \partial \beta^T} & -\frac{\partial^2 \ell}{\partial \alpha \partial \alpha^T} \end{pmatrix} + \begin{pmatrix} n\Sigma_\gamma(\beta) & 0 \\ 0 & n\Sigma_\gamma(\alpha) \end{pmatrix} \\ &= \begin{pmatrix} X^T W_1 X + n\Sigma_\gamma(\beta) & X^T W_2 Z \\ Z^T W_2 X & Z^T W_3 Z + n\Sigma_\gamma(\alpha) \end{pmatrix}. \end{aligned}$$

References

Antoniadis A, Gijbels I, Lambert-Lacroix S, and Poggi J (2016). Joint estimation and variable selection for mean and dispersion in proper dispersion models, *Electronic Journal of Statistics*, **10**, 1630–1676.

Charalambous C, Pan J, and Tranmer M (2015). Variable selection in joint modelling of the mean and variance for hierarchical data, *Statistical Modelling*, **15**, 24–50.

Cox DR (1972). Regression models and life-tables. *Journal of the Royal Statistical Society-Series B*, **34**, 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 ID, Lee Y, and Song JK (2002). Hierarchical likelihood approach for mixed linear models with censored data, *Lifetime Data Analysis*, **8**, 163–176.

Ha ID, 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 ID, Jeong JH, and Lee Y (2017). *Statistical Modelling of Survival Data with Random Effects: H-Likelihood Approach*, Springer, Singapore.

Hutton JL and Monaghan PF (2002). Choice of parametric accelerated life and proportional hazard models for survival data: asymptotic results, *Lifetime Data Analysis*, **8**, 375–393.

Klein JP and Moeschberger ML (2003). *Survival Analysis : Techniques for Censored and Truncated Data*(2nd ed), Springer, New York.

Lawless JF (1982). *Statistical Models and Methods for Lifetime Data*, Wiley, New York.

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

MacKenzie G (1996). Regression models for survival data: the generalized time-dependent logistic family, *The Statistician*, **45**, 21–34.

Nelder JA and Lee Y (1998). Joint modeling of mean and dispersion, *Technometrics*, **40**, 168–171.

Nedler JA and Wedderburn RWM (1972). Generalized linear models, *Journal of the Royal Statistical Society A*, **135**, 370–384.

Park E and Ha ID (2018). Penalized variable selection for accelerated failure time models, *Communications for Statistical Applications and Methods*, **25**, 591–604.

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

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

Wang X and Song L (2011). Adaptive lasso variable selection for the accelerated failure models, *Communications in Statistics - Theory and Methods*, **40**, 4372–4386.

Wu L and Li H (2012). Variable selection for joint mean and dispersion models of the inverse Gaussian distribution, *Metrika*, **75**, 795–808.

Zhou M (2005). Empirical likelihood analysis of the rank estimator for the censored accelerated failure time model, *Biometrika*, **92**, 492–498.

Zou H (2006). The adaptive Lasso and its oracle properties. *Journal of American Statistical Association*, **101**, 1418–1429.

Received April 5, 2021; Revised April 13, 2021; Accepted April 14, 2021

평균-분산 가속화 실패시간 모형에서 벌점화 변수선택

권지훈[a], 하일도[1,b]

[a]에이페이스(주) 통계팀, [b]부경대학교 통계학과

요 약

가속화 실패시간모형은 로그 생존시간과 공변량간의 선형적 관계를 묘사해 준다. 가속화 실패시간모형에서 생존시간의 평균뿐만 아니라 변동성에도 영향을 미치는 공변량 효과를 추론하는 것은 흥미가 있다. 이를 위해 생존시간의 평균뿐만 아니라 분산을 모형화 하는 것이 필요하며, 이러한 모형을 평균-분산 가속화 실패시간모형이라 부른다. 본 논문에서는 벌점 가능도함수를 이용하여 평균-분산 가속화 실패시간모형에서 회귀모수에 대한 변수선택 절차를 제안한다. 여기서 벌점함수로서 LASSO, ALASSO, SCAD 그리고 HL (계층가능도)와 같은 네 가지 벌점함수를 연구한다. 제안된 변수선택 절차를 통해 중요한 공변량의 선택 뿐만 아니라 회귀모수의 추정을 동시에 제공할 수 있다. 제안된 방법의 성능은 모의실험을 통해 평가하고, 하나의 임상 예제자료를 통해 제안된 방법을 예증하고자 한다.

주요용어: 가속화 실패시간모형, 평균-분산 모형, 벌점 가능도, 변수선택

이 논문은 부경대학교 자율창의학술연구비(2019년)에 의하여 연구되었음.
본 연구는 제1저자 권지훈의 부경대학교 석사학위논문의 일부를 발췌, 수정한 논문임.
[1]교신저자: (48513) 부산시 남구 용소로 45, 부경대학교 통계학과. E-mail: idha1353@pknu.ac.kr