碩士學位論文

Variable selection in censored regression models

韓國外國語大學校 大學院

統 計 學 科

李 晟 熙

碩士學位論文

Variable selection in censored regression models

중도절단회귀 모형에서의 변수선택법

指導 梁 城 準 敎授

이 論文을 碩士學位 請求論文으로 提出합니다.

2017年 12月

韓國外國語大學校 大學院

統 計 學 科

李 晟 熙

이　論文을 李晟熙의 碩士學位 論文으로 認定함.

2017年　12月 6日

審査委員　_____ (인)

審査委員　_____ (인)

審査委員　_____ (인)

韓國外國語大學校 大學院

요 약

일반적으로 회귀분석은 설명변수와 반응변수간의 함수적 관계를 규명하는 것을 목적으로, 설명변수 X가 주어졌을 때 반응변수 Y의 조건부 평균을 추정하는 것을 목표로 한다. 하지만, 자료에 중도절단이 있는 경우는 모든 자료를 완벽하게 관측할 수 없기 때문에 오차항이 등분산성을 만족하지 못하며 기존에 연구되어 알려져 있는 회귀기법들을 바로 적용할 수 없다. 중도절단이 존재하는 경우의 최소제곱법에 의한 모수추정에 대한 기존 이론들이 대부분 오차항의 등분산 가정 하에서 이루어진 점과 모형의 이분산성이 존재하더라도 LASSO의 변수선택법은 일치성을 만족한다고 알려져 있음을 활용하여 이를 이분산이 존재하는 모형에 적용하고 모의실험을 통해 LASSO와 Adaptive LASSO의 변수선택 성능을 확인하고자 한다. 본 논문에서는 LASSO, Adaptive LASSO의 이분산성 하에서 변수선택 성능을 확인하기 위해 가중최소제곱법의 가중치 설정 방법을 제안하며, 자료변환법과 최소제곱법 활용하여 중도절단의 비율에 따라 추정의 정확성, 변수선택의 성능을 비교한다.

주요용어 : 중도절단, 가중최소제곱법, 자료변환법, LASSO, Adaptive LASSO

목 차

[표] 목 차

[그 림] 목 차

1. 서론

일반적으로 회귀분석은 설명변수와 반응변수간의 함수적 관계를 규명하는 것을 목적으로, 설명변수 X가 주어졌을 때 반응변수 Y의 조건부 평균을 추정하는 것을 목표로 한다. 하지만, 자료에 중도절단이 있는 경우는 기존에 연구되어 알려져 있는 회귀기법들을 바로 적용할 수 없다. 중도절단은 자료의 불완전한 관측을 나타내는 특성 중 하나로, 생존시간을 관측할 때 가장 빈번히 발생한다. 중도절단이 발생하는 원인은 다양하지만 가장 대표적인 원인은 해당 연구의 종료이다. 이러한 문제점을 해결하기 위한 간단한 방법 중 하나는 조건부회귀함수의 값을 보존하는 자료변환법을 활용하는 것인데, Buckley & James(1979), Koul et al. (1981), Leurgans(1987) 등에서 제안된 방법들이 대표적이다.

회귀모형의 모수추정을 위해 가장 많이 사용되는 방법은 최소제곱법(Least Squares Estimation, LSE) 이지만, 자료에 중도절단이 존재한다면 모든 자료를 완벽히 관측할 수 없기 때문에 오차항이 등분산성을 만족하지 못하며 최소제곱법에 의한 모수추정은 제대로 이루어지지 않는다. 하지만, 기존 이론들이 대부분 오차항의 등분산 가정 하에서 이루어졌으며 Jinzhu et al. (2013)에서는 모형의 이분산성이 존재하더라도 LASSO의 변수선택법은 일치성을 만족한다고 알려져 있기에 이를 이분산성이 존재하는 모형에 적용하여 실험을 통해 그 성능을 확인해볼 필요가 있다. 가중최소제곱법(Weighted Least Squares Estimation, WLSE)를 활용하여 모수추정을 실시한다면 추정량의 성능을 향상시킬수 있고 변수선택의 정확성 또한 성능을 향상시킬 수 있을 것이라 예상한다. 이때, 가중최소제곱법의 가중치 설정은 다음 과정의 결과를 고려하였다.

(1) 초기 반응변수 Y에 자료변환법 적용 후 결과를 반응변수로 하는 다중회귀분석 결과의 잔차제곱

(2) (1)의 잔차제곱을 로그변환 후 결과를 반응변수로 하는 2차 다항회귀분석 결과의 반응변수 추정값의 지수변환

가중최소제곱법의 모수추정 성능을 비교하기 위해 최소제곱법에서 변수선택법 성능을 함께 비교하였다. 설명변수의 개수가 많은 경우, 예측력의 향상 및 모형의 단순화를 위한 변수선택의 과정이 필요할 수 있다. 본 논문에서 사용할 변수선택법은 기존에 연구되어진 LASSO, Adaptive LASSO와 같은 벌점화함수를 도입한 축소추정 방식이다. 위 방법들은 최근 활발하게 연구되어져 왔으나 대부분 오차항에 대한 등분산가정 하에서 이루어져왔으나, 자료에 중도절단이 존재하여 자료변환법을 통해 회귀모형을 적합하는 경우, 필연적으로 오차항에 대한 등분산 가정이 만족되지 않게 된다. 이 경우, 가중최소제곱법에 의한 추정량이 좋은 성질을 가지고 있다. 본 논문의 목적은 자료에 중도절단이 포함된 선형회귀모형에서 가중최소제곱법에 사용되는 가중치에 의한 모형과 최소제곱법을 사용한 모형을 여러 변수선택법을 사용하여 모의실험을 통해 오차항의 이분산성 하에서 각 변수선택의 성능을 비교하는 것이며, 가중최소제곱법의 가중치 설정 방법을 제안하였다.

논문 구성은 다음과 같다. 2절에서는 중도절단자료 자료변환법을 소개하고 3절에서는 가중최소제곱법에 사용할 가중치 설정 방법을 제안한다. 4절에서는 변수선택법인 LASSO와 Adaptive LASSO를 소개하며 성능비교를 진행한 모의실험은 5절에서 소개하고자 한다. 6절에서는 결론을 제시하고자 한다.

2. 자료변환법

여러 자료변환법 중, 본 논문에서는 Koul et al.(1981)에서 제안된 자료변환법 방식을 사용하였다. $(Y, X) \in R \times R^p$가 반응변수 및 설명변수이고, $(Y_i, X_i), i = 1, \cdots, n$가 그에 대한 임의표본이라 하자. 이때, 관측하게 되는 변수는 $T_i = \min(Y_i, C_i)$와 $\delta_i = I(Y_i \leq C_i)$가 된다. 여기서 Y는 실제 관심대상인 생존시간이며, C는 중도절단시간을 나타내는 변수이다. 추정의 대상은 조건부회귀함수인 $E(Y|X=x)$이며 Y가 부분적으로만 관측되기 때문에 불편변환법에 의해 다음과 같이 새로운 반응변수를 생성할 수 있다.

$$Y_i^G = \frac{\delta_i T_i}{1 - G(T_i)} \tag{2.1}$$

여기서, G는 C의 분포함수 즉, $G(t) = P(C \leq t)$이다. 그러면, (1) Y와 C는 서로 독립, (2) $P(Y \leq C|X, Y) = P(Y \leq C|Y)$의 조건 하에서

$$E(Y|X=x) = E(Y^G|X=x) \tag{2.2}$$

임을 보일 수 있다. 즉, 조건부회귀함수의 값이 모든 $X=x$에 대해 일치하게 되어 Y^G를 반응변수로 간주할 수 있으며, 중도절단이 존재하지 않는 일반적인 회귀모형을 적용하여 회귀함수를 추정할 수 있게 된다. 실제로는 G도 알려져 있지 않으므로 Kaplan-Meier 추정량 등으로 추정하게 된다.

$$1 - \hat{G}(t) = \prod_{i=1}^{n} \left(1 - \frac{(1-\delta_i) I(T_i \leq t)}{\sum_{j=1}^{n} I(T_j \geq T_i)} \right) \tag{2.3}$$

단, 이 경우는 절단변수의 분포가 설명변수 X에 의존하지 않는 경우에만 적합하다고 할 수 있다.

3. 가중최소제곱법

자료에 중도절단이 존재하여 자료변환법을 통해 회귀모형을 적합하는 경우, 필연적으로 오차항에 대한 등분산 가정이 만족되지 않게 된다. 중도절단에 의한 모형의 이분산성은 다음의 식에서 확인할 수 있다.

$$Var(Y^G|X=x) = Var(Y|X=x) + E(\frac{G(Y)}{1-G(Y)}Y^2|X=x) \qquad (3.1)$$

즉, 원래 모형 $Y=X\beta+\epsilon$에서는 등분산가정이 만족되더라도 변환된 자료에 대한 모형 $Y^G=X\beta+\epsilon^G$은 등분산성이 만족되지 않게 된다. 따라서 최소제곱추정법에 의한 모수추정은 효율이 떨어지게 된다. $\sigma_i^2 = Var(Y_i^G|X_i=x_i)$ 일 때, $w_i=\sigma_i^{-2}$의 가중치를 생각한다면, 다음의 최소화 문제를 통해 모수추정이 가능하다.

$$\hat{\beta}= argmin_{\beta_0,\cdots,\beta_p}\left(\sum_{i=1}^{n} w_i(Y_i^G - x_i\beta)^2\right) \qquad (3.2)$$

이분산성이 존재할 때, 위와 같은 가중최소제곱법에 의한 추정량이 좋은 성질을 가지는 것이 알려져 있다. 본 논문에서 제안할 가중치 설정 방법을 설명하기에 앞서, $r_i = Y_i - \widehat{Y}_i$일 때, 다음 식이 성립하는 것이 M. Davidian and R. J. Carroll(1987)에서 알려져 있다.

$$E(\log(r_i^2)\,|\,X_i=x_i) \approx \log(\sigma_i^2) \qquad (3.3)$$

위 식을 활용하여 분산추정에 의한 가중치 설정 방법을 제안하며, 가중최소제곱법의 가중치 설정은 다음의 1), 2) 과정의 결과인 w_i^*로 하였다.

1) Y^G를 반응변수로 하는 다중회귀분석 결과의 잔차제곱 생성

① $\widehat{Y^G} = X\hat{\beta}$ 다중회귀모형 적합

② $r_i^2 = (Y_i^G - \widehat{Y_i^G})^2$

2) r_i^2의 로그변환을 반응변수로 하는 2차 다항회귀모형 적합 결과의 반응변수 추정값의 지수변환

① $\hat{r_i}$은 1) 결과의 잔차제곱, $r_i^2 = (Y_i^G - \widehat{Y_i^G})^2$

② 반응변수는 $\log(\hat{r_i^2})$, 설명변수는 X인 2차 다항회귀분석 적합

③ 위 회귀분석 결과의 반응변수 추정값을 $\hat{\gamma_i}^*$라 할 때, $r_i^* = \exp(\hat{\gamma_i}^*)$ 인 지수변환 적용

④ $w_i^* = r_i^{*-1}$인 가중치 생성

따라서 본 논문에서는 위 가중치를 모수추정에 사용하여 최소제곱법과의 효율을 비교하고자 한다. 가중최소제곱법을 활용한 모수추정은 다음의 최소화 문제를 통해 가능하다.

$$\hat{\beta}^* = argmin_{\beta_0,\cdots,\beta_p}\left(\sum_{i=1}^{n} w_i^* (Y_i^G - x_i\beta)^2\right) \qquad (3.4)$$

4. 변수선택법

4.1 LASSO

회귀모형을 분석할 때, 일반적으로 사용하는 방법은 잔차제곱합을 최소로 하는 보통최소제곱(Ordinary Least Squares, OLS)을 사용한다. 하지만, 보통최소제곱을 자료분석에 활용할 때 두 가지 문제점이 있다. 첫 번째는 '예측정확성'이다. 보통최소제곱은 작은 편향을 가지지만 분산이 커질 수 있기 때문에 예측정확성을 높이기 위해 회귀계수를 축소(shrinkage)하거나 영향력이 없는 회귀계수 몇 개를 0으로 지정하는 방법을 생각할 수 있다. 이때, 회귀계수 축소에 사용되는 방법이 능형회귀(ridge regression)이며 회귀 계수들에 제약조건 $\sum\beta_j^2 \leq t$을 적용하여 회귀계수를 추정하며 회귀계수를 축소시켜 분산이 작아지는 효과를 볼 수 있지만 완전히 0으로 추정하지는 못한다. 따라서 이 방법은 분산을 줄이기 위해 작은 편향은 포기해야하지만 전체적인 예측정확성을 높일 수 있다. 이때, 회귀계수를 완전히 0으로 추정하지 못하여 생기는 문제점이 두 번째 이유인 '모형의 해석'이다. 일반적인 보통최소제곱에서는 추정된 회귀식에 모든 설명변수를 포함하게 된다. 이는 영향력이 큰 변수들이 무엇인지 판단하기에 어려움을 줄 수 있다. 이때 흔히 사용하는 방법이 모형축소를 위한 변수선택법이다.

본 논문에서 활용할 변수선택법인 LASSO(Least Absolute Shrinkage and Selection Operator)는 Tibshirani(1996)에서 제안된 방법이며 회귀계수들에 제약조건을 주어 계수추정치 크기를 축소시키고 변수선택을 할 수 있는 회귀추정방법이다. 이는 영향력이 없는 회귀계수 값을 0으로 만들어서 계수추정치 크

기 축소와 동시에 추정회귀식의 해석력을 높여줄 수 있다. LASSO 회귀계수추정량은 다음 식을 통해 구할 수 있다.

$$\hat{\beta}^{lasso} = argmin_{\beta_0,\cdots,\beta_p}\left(\sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right) \quad (4.1)$$

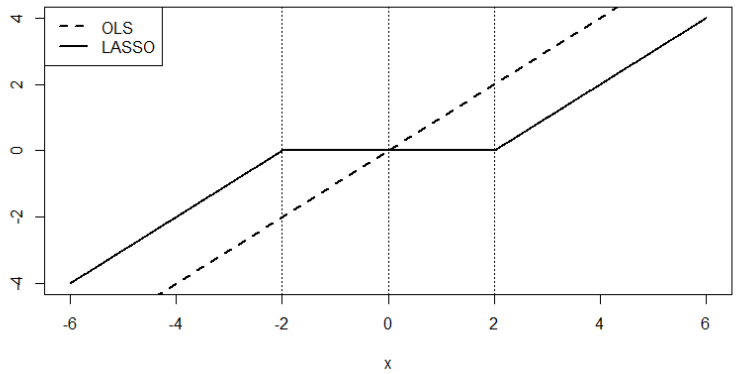
이때, 조절모수 λ는 추정에는 교차타당성(cross-validation, CV)이 사용된다. 교차타당성은 예측오차를 추정하는 데 사용되는 방법이다. 자료를 임의의 k개로 나누어 $k-1$개의 자료는 훈련용 자료(training set)로 사용하고, k개의 자료 중 하나는 검증용 자료(test set)로 사용된다. 즉, $k-1$개의 훈련용 자료로 모형을 만들고 검증용 자료로 모형의 타당성을 검정하며, 이것을 k번 반복하여 평균제곱오차(MSE)를 계산한다. 본 논문에서는 조절모수 λ를 변화하면서 교차타당성을 실시한 후, 가장 작은 평균제곱오차를 가지는 조절모수 λ를 선택하였다.

또한, 설명변수 x_{ij}는 평균이 0, 분산이 1로 표준화되었다고 가정한다면, 모든 j에 대하여 $\sum_{i=1}^{n} x_{ij} = 0$이므로 항상 $\hat{\beta}_0 = \bar{y}$가 되며 β_0은 추정의 대상이 아니게 된다. 따라서 다음의 제약조건이 주어진 최소화 문제를 통해 LASSO 회귀추정량을 구할 수 있다. (Tibshirani,1996)

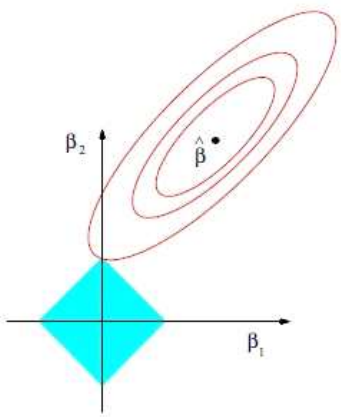
$$\hat{\beta}^{lasso} = argmin_{\beta_1,\cdots,\beta_p}\left(\sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2\right) subject\ to \sum_{j=1}^{p}|\beta_j| \le t \quad (4.2)$$

[그림 4-1]은 LASSO의 회귀계수 축소를 보여준다. 일정한 값까지는 0의 값을 갖다가 일정 값 이상이 되면 보통최소제곱보다 일정한 양이 줄어드는 것을 볼 수 있다. [그림 4-2], [그림 4-3]은 LASSO와 Ridge의 제약조건을 비교한 것으로 왼쪽은 LASSO의 제약조건 $|\beta_1| + |\beta_2| \le t$를 오른쪽은 Ridge의 제약조건

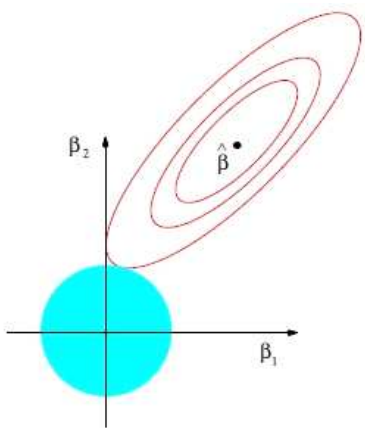
$\beta_1 + \beta_2^2 \leq t$을 의미한다. 각 등고선의 중심에 있는 점은 최소제곱추정량이며 타원은 잔차제곱합을 의미한다. Ridge의 경우, 원과 닿는 부분에서 추정된 회귀계수가 0이 될 수 없지만, LASSO의 경우 사각형과 닿는 부분에서 회귀계수가 0이 될 수 있는 것을 볼 수 있다.



[그림 4-1] LASSO의 회귀계수 축소



[그림 4-2] LASSO의 계수추정



[그림 4-3] Ridge의 계수추정

4.2 Adaptive LASSO

LASSO에서는 λ의 결정에 따라 계수추정의 편향에 영향을 받게 된다. 이는 모든 계수에 동일한 가중치를 적용함으로써 계수추정을 적절히 하지 못하기 때문에 발생하는 문제점이다. LASSO에서 추정의 편향을 줄이기 위한 방법 중 하나는 Hui Zou(2006)에서 제안된 Adaptive LASSO로, 동일한 조절모수가 아닌 $\lambda_j = w_j\lambda$와 같은 다중 조절모수를 사용하는 것이다. 만약, 작은 가중치를 가지는 값이 큰 회귀계수를 선택할 수 있다면, LASSO의 특성을 유지하면서 LASSO의 추정 편향을 줄일 수 있을 것이다. 또한, 더 정확한 회귀계수를 추정함으로써 LASSO의 변수선택 정확성을 향상시킬 수 있을 것이다. Adaptive LASSO의 회귀계수추정량은 다음 식의 최소화를 통해 구할 수 있다.

$$\hat{\beta}^{ad.lasso} = argmin_{\beta_1,\cdots,\beta_p}\left(\sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p} w_j|\beta_j|\right) \qquad (4.3)$$

여기서, $w_j = |\widetilde{\beta}_j|^{-1}$로 구할 수 있으며 이때 $\tilde{\boldsymbol{\beta}}$는 보통최소제곱 또는 LASSO의 초기 추정량을 사용한다. 이 가중치 적용 방법을 활용하면, 초기 추정량 $\widetilde{\beta}_j = 0$이라면 $w_j = \infty$이 되어 $\hat{\beta}_j = 0$이 됨을 알 수 있다. 따라서 값이 큰 회귀계수에 작은 가중치를 적용할 수 있게 된다.

5. 모의실험

모든 모의실험 과정은 오픈 소스 R을 이용하여 진행되었으며, 모의실험에 사용할 기초 모형은

$$y = X\beta + \epsilon$$

이다. $\beta^T = (3, 1.5, 0, 0, 0, 0, 0)$로 설정하였으며 $X \sim N_7(\mathbf{0_7}, \Sigma)$로 지정하였다. 이때, 분산·공분산행렬은 $\Sigma_{ij} = \rho^{|i-j|}$, $\rho = 0, 0.5, 0.9$로 지정하여 X의 상관계수에 따른 차이를 비교하였다. 비교를 위해 중도절단이 없는 경우도 함께 고려하였으며, 절단변수의 분포는 $C_i \sim Unif(-7, 7) + \alpha$로 설정하였다. 절단비율(PC : Percentage of Censoring)이 약 10%, 30%, 50%가 되도록 α의 값을 적절히 선택하였으며, 오차항은 $\epsilon \sim N(0, 1^2)$, $\epsilon \sim \Gamma(2, 4)$로 설정하여 오차항이 대칭인 경우와 비대칭인 경우를 함께 고려하였다. 모의실험에 적용한 반복수는 1,000번이며, 표본의 크기는 100, 200, 400을 동시에 고려하여 표본의 크기가 커짐에 따라 변수선택의 정확성이 증가하는지를 살펴보았다. 또한, 변환된 반응변수 Y^G에서 분포함수 G는 모르는 경우가 일반적이므로 이의 추정을 위해 Kaplan-Meier 추정량 $\hat{G}$을 사용하였다. 변수선택법은 다중회귀분석(결과표의 M_r), LASSO(결과표의 L), Adaptive LASSO(결과표의 A_L)를 사용하였으며, Jianqing Fan and Runze Li(2001)에서 제안된 SCAD(Smoothly Clipped Absolute Deviation)의 성능도 함께 확인하기 위해 모의실험에 포함하여 진행하였으나 Adaptive LASSO와 성능 상 큰 차이가 없었기에 본 논문에서는 SCAD에 대한 결과는 기술하지 않았다. 다중회귀분석의 경우, 다중회귀분석 실

시 후 각 변수의 p-value가 0.05보다 작은 경우를 선택하였다. LASSO와 Adaptive LASSO의 경우, R의 glmnet, parcor 패키지를 활용하여 모의실험을 진행하였고, Adaptive LASSO의 가중치의 초기값 $\widetilde{\beta}_j$은 LASSO의 회귀계수추정량을 사용하였다. 각 변수선택법의 성능을 확인하기 위해 최소제곱법, 가중최소제곱법을 고려하였으며 가중최소제곱법의 가중치 행렬 설정은 본 논문에서 제안한 방법을 사용하였다. 가중치 행렬 생성과정을 반복할수록 성능이 향상될 것이라 예상하여 한 차례 반복 후 동일하게 변수선택법을 적용시켜보았으나 성능에 뚜렷한 향상이 보이지 않아 성능비교 결과에는 포함시키지 않았다.

또한, 변수선택의 성능비교를 위해 기초모형의 $\boldsymbol{\beta}^T = (3, 1.5, 0, 0, 0, 0, 0)$ 중,

① average # of 0 coefficients : 0을 0으로 선택한 평균 개수

② average # of miss selections : 0이 아닌데 0으로 선택한 평균 개수

③ average # of none 0 coefficients : 0이 아닌데 0이 아닌 것으로 선택한 평균 개수

④ 평균제곱오차(MSE) : 추정량의 평균제곱오차(MSE)

를 비교하였다. ①의 경우 영향이 없는 변수를 제대로 선택한 경우로, 5에 가까울수록 좋은 성능을 의미하며, ②의 경우는 변수선택의 오류를 나타내므로 0에 가까울수록 좋은 성능을 의미한다. ③의 경우는 영향이 있는 변수를 제대로 선택한 경우로, 2에 가까울수록 좋은 성능을 의미한다.

각 추정방법에 대한 변수선택 방법의 성능비교 결과를 오차항의 분포가 정규분포일 경우, 중도절단의 비율에 따라 0%, 10%, 30%, 50% 순서대로 [표 5-1]부터 [표 5-4]에 나타내었고, 오차항의 분포가 감마분포일 경우, 중도절단의 비

율에 따라 0%, 10%, 30%, 50% 순서대로 [표 5-5]부터 [표 5-8]에 나타내었다.

[표 5-1] 중도절단이 없는 경우, 변수선택법 성능비교, $\epsilon \sim N(0, 1^2)$

	ρ	Sel.	N=100				N=200				N=400			
			①	②	③	MSE	①	②	③	MSE	①	②	③	MSE
LSE	0	M_r	4.751	0	2	0.011	4.754	0	2	0.005	4.771	0	2	0.003
		L	4.807	0	2	0.021	4.893	0	2	0.013	4.974	0	2	0.009
		A_L	4.831	0	2	0.004	4.972	0	2	0.002	4.998	0	2	0.001
	0.5	M_r	4.719	0	2	0.018	4.732	0	2	0.008	4.760	0	2	0.004
		L	4.837	0	2	0.015	4.936	0	2	0.010	4.972	0	2	0.006
		A_L	4.882	0	2	0.006	4.989	0	2	0.002	5.000	0	2	0.001
	0.9	M_r	4.733	0.003	1.997	0.092	4.736	0	2	0.043	4.753	0	2	0.021
		L	4.574	0	2	0.027	4.700	0	2	0.015	4.774	0	2	0.008
		A_L	4.777	0.022	1.978	0.039	4.944	0.001	1.999	0.011	4.999	0	2	0.005
WLSE	0	M_r	4.149	0	2	0.015	4.385	0	2	0.007	4.546	0	2	0.003
		L	4.048	0	2	0.019	4.687	0	2	0.013	4.921	0	2	0.009
		A_L	4.433	0.002	1.998	0.015	4.868	0	2	0.003	4.993	0	2	0.001
	0.5	M_r	4.159	0	2	0.024	4.431	0	2	0.010	4.570	0	2	0.005
		L	4.282	0.001	1.999	0.016	4.799	0	2	0.010	4.937	0	2	0.006
		A_L	4.577	0.017	1.983	0.027	4.915	0	2	0.005	5.000	0	2	0.001
	0.9	M_r	4.219	0.011	1.989	0.124	4.418	0	2	0.057	4.545	0	2	0.026
		L	4.205	0	2	0.048	4.540	0	2	0.018	4.656	0	2	0.009
		A_L	4.586	0.148	1.852	0.122	4.827	0.024	1.976	0.035	4.982	0	2	0.008

① average # of 0 coefficients, ② average # of miss selections, ③ average # of none 0 coefficients

※ M_r : 다중회귀분석, L : LASSO, A_L : Adaptive LASSO

먼저, 오차항이 정규분포를 따르며 중도절단이 없는 경우인 [표 5-1]의 결과를 살펴보면, 전체적으로 최소제곱법의 성능이 가중최소제곱법 보다 더 좋은 것을 확인할 수 있다. 표본의 크기가 커질수록 Adaptive LASSO의 변수선택 성능이

가장 뛰어나게 확인되었다. 이는 동일 조절모수가 아닌 다중 조절모수를 적용한 효과라고 할 수 있다. 또한, 표본의 크기가 증가하면서 모든 방법들의 변수선택 성능, 추정정확성의 성능이 좋아지는 것을 확인할 수 있다.

[표 5-2] 중도절단이 10%인 경우, 변수선택법 성능비교, $\epsilon \sim N(0, 1^2)$

	ρ	Sel.	N=100				N=200				N=400			
			①	②	③	MSE	①	②	③	MSE	①	②	③	MSE
LSE	0	M_r	4.764	0	2	0.080	4.753	0	2	0.041	4.771	0	2	0.020
		L	4.904	0.025	1.975	0.275	4.984	0.001	1.999	0.190	4.995	0	2	0.131
		A_L	4.525	0.001	1.999	0.065	4.669	0	2	0.032	4.775	0	2	0.014
	0.5	M_r	4.738	0.061	1.939	0.154	4.747	0.001	1.999	0.081	4.751	0	2	0.042
		L	4.937	0.056	1.944	0.316	4.983	0.011	1.989	0.232	5.000	0.002	1.998	0.164
		A_L	4.454	0.035	1.965	0.107	4.602	0	2	0.051	4.741	0	2	0.025
	0.9	M_r	4.772	0.743	1.257	0.742	4.752	0.492	1.508	0.411	4.747	0.201	1.799	0.225
		L	4.760	0.227	1.773	0.419	4.833	0.097	1.903	0.306	4.906	0.027	1.973	0.225
		A_L	4.478	0.452	1.548	0.354	4.481	0.256	1.744	0.220	4.541	0.074	1.926	0.114
WLSE	0	M_r	4.233	0	2	0.029	4.338	0	2	0.011	4.507	0	2	0.005
		L	4.154	0.003	1.997	0.054	4.698	0	2	0.025	4.914	0	2	0.013
		A_L	4.392	0.024	1.976	0.052	4.726	0.006	1.994	0.030	4.971	0.001	1.999	0.025
	0.5	M_r	4.250	0.001	1.999	0.047	4.430	0	2	0.017	4.590	0	2	0.007
		L	4.388	0.001	1.999	0.054	4.819	0	2	0.025	4.921	0	2	0.012
		A_L	4.521	0.112	1.888	0.094	4.855	0.084	1.916	0.081	4.990	0.064	1.936	0.095
	0.9	M_r	4.428	0.121	1.879	0.240	4.489	0.007	1.993	0.094	4.607	0	2	0.039
		L	4.367	0.012	1.988	0.099	4.627	0	2	0.038	4.778	0	2	0.018
		A_L	4.516	0.341	1.659	0.228	4.845	0.458	1.542	0.211	4.984	0.646	1.354	0.272

① average # of 0 coefficients, ② average # of miss selections, ③ average # of none 0 coefficients

※ M_r : 다중회귀분석, L : LASSO, A_L : Adaptive LASSO

[표 5-2]는 절단비율이 10%인 경우, 오차항이 정규분포일 때의 변수선택법 성능비교에 대한 결과이다. 전체적으로 최소제곱법보다 가중최소제곱법의 성능이 더 좋은 것을 확인할 수 있는데, 이는 모형의 이분산성으로 인한 모수추정의 정확성이 떨어지기 때문이라고 할 수 있다. 또한, 이분산성이 존재할 때, 가중최소제곱법의 경우, Adaptive LASSO의 변수선택 성능이 가장 좋으며 표본의 크기가 증가하면서 성능의 증가폭 또한 Adaptive LASSO가 가장 뛰어남을 확인할 수 있다.

[표 5-3]은 오차항이 정규분포, 절단비율이 30%인 경우로, 절단비율이 10%일 때의 결과와 비교해보면, 표본의 크기가 증가하면서 가중최소제곱법의 LASSO의 변수선택 성능이 작은 차이지만 Adaptive LASSO를 역전하는 것을 확인할 수 있다. 이는 일정수준의 중도절단 비율을 넘어서면 Adaptive LASSO의 성능이 저하될 것이라 예상할 수 있다. 이어서 중도절단이 50%인 [표 5-4]의 결과를 살펴보면, 가중최소제곱법의 Adaptive LASSO의 변수선택 성능이 급격히 저하된 것을 확인할 수 있다. 이를 최소제곱법과 비교하면 추정의 정확성은 향상되지만 변수선택의 성능은 현저히 낮은 수치이다. 따라서 일정수준 이상의 중도절단 비율이 존재한다면 조절모수 λ를 정교하게 조절하는 방법을 통해 이를 완화시킬 필요가 있다고 생각된다.

[표 5-3] 중도절단이 30%인 경우, 변수선택법 성능비교, $\epsilon \sim N(0, 1^2)$

	ρ	Sel.	N=100				N=200				N=400			
			①	②	③	MSE	①	②	③	MSE	①	②	③	MSE
LSE	0	M_r	4.758	0.031	1.969	0.165	4.727	0	2	0.097	4.714	0	2	0.055
		L	4.938	0.298	1.702	0.601	4.985	0.162	1.838	0.489	4.999	0.086	1.914	0.396
		A_L	4.364	0.034	1.966	0.158	4.451	0	2	0.087	4.583	0	2	0.047
	0.5	M_r	4.772	0.244	1.756	0.281	4.736	0.051	1.949	0.17716	4.746	0.002	1.998	0.106
		L	4.959	0.289	1.711	0.654	4.991	0.242	1.758	0.572	5.000	0.150	1.850	0.494
		A_L	4.443	0.126	1.874	0.225	4.466	0.026	1.974	0.136	4.602	0.001	1.999	0.078
	0.9	M_r	4.733	1.052	0.948	1.183	4.777	0.755	1.245	0.720	4.729	0.532	1.468	0.468
		L	4.798	0.485	1.515	0.740	4.985	0.399	1.601	0.666	4.943	0.320	1.680	0.598
		A_L	4.399	0.567	1.433	0.536	4.469	0.455	1.545	0.381	4.422	0.296	1.704	0.258
WLSE	0	M_r	4.126	0	2	0.006	4.360	0	2	0.020	4.458	0	2	0.008
		L	4.030	0.007	1.993	0.126	4.673	0	2	0.050	4.924	0	2	0.022
		A_L	4.166	0.039	1.961	0.082	4.573	0.014	1.986	0.040	4.912	0.005	1.995	0.035
	0.5	M_r	4.199	0.009	1.991	0.105	4.404	0	2	0.035	4.574	0	2	0.013
		L	4.229	0.021	1.979	0.156	4.759	0.002	1.998	0.066	4.946	0	2	0.029
		A_L	4.122	0.116	1.884	0.130	4.521	0.084	1.916	0.071	4.893	0.069	1.931	0.064
	0.9	M_r	4.212	0.360	1.640	0.513	4.437	0.078	1.922	0.187	4.642	0.002	1.998	0.071
		L	4.318	0.119	1.881	0.265	4.656	0.011	1.989	0.106	4.825	0	2	0.049
		A_L	4.036	0.364	1.636	0.371	4.419	0.242	1.758	0.181	4.780	0.266	1.734	0.136

① average # of 0 coefficients, ② average # of miss selections, ③ average # of none 0 coefficients

※ M_r : 다중회귀분석, L : LASSO, A_L : Adaptive LASSO

[표 5-4] 중도절단이 50%인 경우, 변수선택법 성능비교, $\epsilon \sim N(0, 1^2)$

	ρ	Sel.	N=100				N=200				N=400			
			①	②	③	MSE	①	②	③	MSE	①	②	③	MSE
LSE	0	M_r	4.750	0.135	1.865	0.302	4.763	0.014	1.986	0.203	4.762	0	2	0.144
		L	4.925	0.817	1.183	1.022	4.994	0.720	1.280	0.980	4.999	0.746	1.254	0.963
		A_L	4.246	0.131	1.869	0.325	4.380	0.014	1.986	0.202	4.524	0	2	0.141
	0.5	M_r	4.773	0.429	1.571	0.421	4.767	0.204	1.796	0.308	4.743	0.056	1.944	0.221
		L	4.942	0.700	1.300	0.985	4.994	0.714	1.286	0.985	5.000	0.771	1.229	1.007
		A_L	4.329	0.233	1.767	0.379	4.411	0.104	1.896	0.276	4.473	0.022	1.978	0.194
	0.9	M_r	4.730	1.242	0.758	1.341	4.739	0.955	1.045	0.921	4.722	0.705	1.295	0.648
		L	4.777	0.716	1.284	0.979	4.930	0.732	1.268	1.022	4.971	0.729	1.271	1.032
		A_L	4.382	0.654	1.346	0.656	4.427	0.545	1.455	0.506	4.423	0.406	1.594	0.390
WLSE	0	M_r	4.000	0.056	1.944	0.245	4.214	0.001	1.999	0.106	4.461	0	2	0.043
		L	4.140	0.329	1.671	0.650	4.661	0.134	1.866	0.435	4.920	0.024	1.976	0.285
		A_L	3.196	0.069	1.931	0.269	3.650	0.002	1.998	0.100	4.128	0	2	0.036
	0.5	M_r	4.001	0.180	1.820	0.335	4.271	0.035	1.965	0.164	4.499	0	2	0.068
		L	4.290	0.388	1.612	0.670	4.804	0.205	1.795	0.510	4.965	0.079	1.921	0.367
		A_L	3.313	0.145	1.855	0.327	3.634	0.018	1.982	0.146	4.108	0	2	0.051
	0.9	M_r	3.973	0.787	1.213	1.187	4.211	0.509	1.491	0.063	4.401	0.274	1.726	0.319
		L	4.474	0.640	1.360	0.801	4.748	0.389	1.611	0.633	4.899	0.259	1.741	0.489
		A_L	3.309	0.490	1.510	0.809	3.514	0.287	1.713	0.438	3.741	0.116	1.884	0.242

① average # of 0 coefficients, ② average # of miss selections, ③ average # of none 0 coefficients

※ M_r : 다중회귀분석, L : LASSO, A_L : Adaptive LASSO

[표 5-5]부터 [표 5-8]은 오차항이 감마분포를 따를 때, 각 절단 비율에 따른 결과이다. 절단 비율과 모수추정방법(최소제곱법, 가중최소제곱법)에 관계없이

오차항이 감마분포를 따를 때, 정규분포일 때보다 좋은 성능을 보여준다. 이는 대부분의 생존자료 분포가 대칭분포가 아니기 때문에 오차항의 비대칭 분포가정이 더 적절하다는 결과로 해석할 수 있다.

[표 5-5] 중도절단이 없는 경우, 변수선택법 성능비교, $\epsilon \sim \Gamma(2,4)$

	ρ	Sel.	N=100				N=200				N=400			
			①	②	③	MSE	①	②	③	MSE	①	②	③	MSE
LSE	0	M_r	4.771	0	2	0.007	4.760	0	2	0.003	4.762	0	2	0.002
		L	4.853	0	2	0.015	4.941	0	2	0.010	4.985	0	2	0.007
		A_L	4.946	0	2	0.003	4.999	0	2	0.001	5.000	0	2	0.001
	0.5	M_r	4.731	0	2	0.011	4.717	0	2	0.006	4.749	0	2	0.003
		L	4.881	0	2	0.011	4.931	0	2	0.007	4.992	0	2	0.005
		A_L	4.945	0	2	0.005	4.999	0	2	0.002	5.000	0	2	0.001
	0.9	M_r	4.749	0.005	1.995	0.091	4.787	0	2	0.042	4.760	0	2	0.021
		L	4.606	0	2	0.028	4.691	0	2	0.015	4.747	0	2	0.008
		A_L	4.795	0.013	1.987	0.034	4.926	0	2	0.010	4.996	0	2	0.005
WLSE	0	M_r	4.120	0	2	0.010	4.406	0	2	0.004	4.541	0	2	0.002
		L	4.105	0	2	0.013	4.711	0	2	0.009	4.946	0	2	0.006
		A_L	4.694	0.007	1.993	0.019	4.971	0	2	0.003	4.999	0	2	0.001
	0.5	M_r	4.159	0	2	0.014	4.429	0	2	0.007	4.560	0	2	0.003
		L	4.356	0	2	0.010	4.796	0	2	0.006	4.959	0	2	0.006
		A_L	4.820	0.021	1.979	0.032	4.963	0.001	1.999	0.007	4.999	0	2	0.001
	0.9	M_r	4.209	0	2	0.076	4.444	0	2	0.035	4.622	0	2	0.015
		L	4.245	0	2	0.025	4.559	0	2	0.012	4.737	0	2	0.006
		A_L	4.770	0.180	1.820	0.129	4.934	0.045	1.955	0.048	4.996	0.005	1.995	0.012

① average # of 0 coefficients, ② average # of miss selections, ③ average # of none 0 coefficients

※ M_r : 다중회귀분석, L : LASSO, A_L : Adaptive LASSO

[표 5-6], [표 5-7]은 각각 절단비율이 10%, 30%이며 오차항이 감마분포를 따를 때의 결과이다. 이를 정규분포의 경우와 비교해보면, Adaptive LASSO를 제외한 나머지 결과에서 변수선택, 모수추정 모두 감마분포의 성능이 더 좋은 것을 확인할 수 있다.

[표 5-6] 중도절단이 10%인 경우, 변수선택법 성능비교, $\epsilon \sim \Gamma(2,4)$

	ρ	Sel.	N=100				N=200				N=400			
			①	②	③	MSE	①	②	③	MSE	①	②	③	MSE
LSE	0	M_r	4.775	0.001	1.999	0.076	4.749	0	2	0.036	4.738	0	2	0.019
		L	4.929	0.023	1.977	0.277	4.979	0.002	1.998	0.177	4.999	0	2	0.126
		A_L	4.565	0.001	1.999	0.064	4.672	0	2	0.028	4.779	0	2	0.014
	0.5	M_r	4.767	0.049	1.951	0.145	4.760	0.002	1.998	0.076	4.759	0	2	0.039
		L	4.935	0.047	1.953	0.307	4.986	0.012	1.988	0.218	4.999	0	2	0.154
		A_L	4.505	0.034	1.966	0.101	4.620	0.001	1.999	0.047	4.750	0	2	0.022
	0.9	M_r	4.735	0.766	1.234	0.767	4.760	0.470	1.530	0.421	4.777	0.182	1.181	0.218
		L	4.742	0.242	1.758	0.428	4.835	0.112	1.888	0.311	4.921	0.040	1.960	0.222
		A_L	4.449	0.444	1.556	0.378	4.521	0.258	1.742	0.223	4.540	0.082	1.918	0.106
WLSE	0	M_r	4.255	0	2	0.018	4.342	0	2	0.007	4.446	0	2	0.003
		L	4.344	0	2	0.034	4.752	0	2	0.016	4.947	0	2	0.009
		A_L	4.723	0.038	1.962	0.082	4.958	0.040	1.960	0.097	4.999	0.020	1.980	0.109
	0.5	M_r	4.356	0	2	0.030	4.459	0	2	0.011	4.536	0	2	0.005
		L	4.564	0	2	0.038	4.822	0	2	0.015	4.956	0	2	0.008
		A_L	4.765	0.172	1.828	0.130	4.967	0.245	1.755	0.169	5.000	0.367	1.633	0.229
	0.9	M_r	4.414	0.051	1.949	0.161	4.527	0	2	0.061	4.571	0	2	0.026
		L	4.415	0.004	1.996	0.066	4.702	0	2	0.026	4.802	0	2	0.011
		A_L	4.719	0.497	1.503	0.237	4.946	0.690	1.310	0.271	5.000	0.924	1.076	0.324

① average # of 0 coefficients, ② average # of miss selections, ③ average # of none 0 coefficients
※ M_r : 다중회귀분석, L : LASSO, A_L : Adaptive LASSO

[표 5-7] 중도절단이 30%인 경우, 변수선택법 성능비교, $\epsilon \sim \Gamma(2,4)$

	ρ	Sel.	N=100				N=200				N=400			
			①	②	③	MSE	①	②	③	MSE	①	②	③	MSE
LSE	0	M_r	4.736	0.022	1.978	0.160	4.749	0.001	1.999	0.089	4.754	0	2	0.053
		L	4.915	0.242	1.758	0.561	4.985	0.141	1.859	0.460	4.998	0.087	1.913	0.385
		A_L	4.345	0.028	1.972	0.153	4.529	0	2	0.079	4.641	0	2	0.045
	0.5	M_r	4.768	0.228	1.772	0.275	4.740	0.040	1.960	0.170	4.739	0.001	1.999	0.103
		L	4.950	0.290	1.710	0.629	4.985	0.205	1.795	0.548	5.000	0.137	1.863	0.474
		A_L	4.430	0.139	1.861	0.222	4.477	0.024	1.976	0.129	4.604	0	2	0.075
	0.9	M_r	4.722	1.062	0.938	1.244	4.749	0.748	1.252	0.708	4.726	0.533	1.467	0.459
		L	4.802	0.509	1.491	0.756	4.895	0.387	1.613	0.665	4.954	0.330	1.670	0.610
		A_L	4.378	0.586	1.414	0.575	4.416	0.428	1.572	0.365	4.499	0.305	1.695	0.267
WLSE	0	M_r	4.160	0	2	0.048	4.341	0	2	0.014	4.453	0	2	0.005
		L	4.083	0.006	1.994	0.094	4.660	0	2	0.033	4.941	0	2	0.015
		A_L	4.374	0.060	1.940	0.094	4.828	0.074	1.926	0.090	4.985	0.056	1.944	0.114
	0.5	M_r	4.218	0.009	1.991	0.086	4.448	0	2	0.024	4.616	0	2	0.009
		L	4.290	0.024	1.976	0.130	4.781	0	2	0.045	4.939	0	2	0.020
		A_L	4.286	0.145	1.855	0.132	4.769	0.178	1.822	0.110	4.953	0.230	1.770	0.136
	0.9	M_r	4.222	0.309	1.691	0.443	4.487	0.039	1.961	0.143	4.647	0	2	0.052
		L	4.397	0.102	1.898	0.222	4.660	0.006	1.994	0.082	4.810	0	2	0.033
		A_L	4.204	0.410	1.590	0.350	4.559	0.344	1.656	0.192	4.886	0.462	1.538	0.196

① average # of 0 coefficients, ② average # of miss selections, ③ average # of none 0 coefficients

※ M_r : 다중회귀분석, L : LASSO, A_L : Adaptive LASSO

[표 5-8] 중도절단이 50%인 경우, 변수선택법 성능비교, $\epsilon \sim \Gamma(2,4)$

	ρ	Sel.	N=100				N=200				N=400			
			①	②	③	MSE	①	②	③	MSE	①	②	③	MSE
LSE	0	M_r	4.800	0.127	1.873	0.283	4.731	0.015	1.985	0.199	4.723	0	2	0.140
		L	4.948	0.733	1.267	0.993	4.994	0.733	1.267	0.976	5.000	0.676	1.324	0.926
		A_L	4.379	0.128	1.872	0.309	4.407	0.011	1.989	0.198	4.491	0	2	0.136
	0.5	M_r	4.763	0.409	1.591	0.404	4.753	0.201	1.799	0.304	4.732	0.039	1.961	0.221
		L	4.963	0.589	1.411	0.947	0.994	0.661	1.339	0.979	4.999	0.692	1.308	0.976
		A_L	4.356	0.219	1.781	0.361	4.426	0.110	1.890	0.275	4.480	0.006	1.994	0.193
	0.9	M_r	4.763	1.274	0.726	1.366	4.741	0.954	1.046	0.907	4.768	0.764	1.236	0.607
		L	4.820	0.757	1.243	1.002	4.915	0.720	1.280	1.019	4.967	0.752	1.248	1.026
		A_L	4.402	0.659	1.341	0.660	4.412	0.537	1.463	0.507	4.413	0.421	1.579	0.378
WLS E	0	M_r	3.976	0.034	1.966	0.219	4.257	0.001	1.999	0.091	4.491	0	2	0.036
		L	4.047	0.246	1.754	0.571	4.646	0.077	1.923	0.372	4.926	0.020	1.980	0.247
		A_L	3.166	0.041	1.959	0.233	3.735	0.001	1.999	0.085	4.266	0	2	0.029
	0.5	M_r	3.982	0.175	1.825	0.318	4.319	0.018	1.982	0.144	4.507	0	2	0.060
		L	4.273	0.313	1.687	0.627	4.782	0.160	1.840	0.453	4.953	0.047	1.953	0.321
		A_L	3.194	0.118	1.882	0.317	3.615	0.011	1.989	0.129	4.147	0	2	0.045
	0.9	M_r	4.046	0.737	1.263	1.088	4.279	0.494	4.506	0.571	4.506	0.199	1.801	0.272
		L	4.454	0.574	1.426	0.770	4.768	0.348	1.652	0.578	4.934	0.190	1.810	0.428
		A_L	3.403	0.441	1.559	0.746	3.651	0.268	1.735	0.396	3.926	0.088	1.912	0.175

① average # of 0 coefficients, ② average # of miss selections, ③ average # of none 0 coefficients

※ M_r : 다중회귀분석, L : LASSO, A_L : Adaptive LASSO

[표 5-8]은 절단비율이 50%이며 오차항이 감마분포를 따를 때의 결과이다. [표 5-4]의 정규분포일 때와 비교해보면 전체적으로 좋은 성능을 나타내지만, 이 경우에도 Adaptive LASSO의 변수선택 성능은 중도절단의 비율이 증가하면

서 급격히 저하되는 현상이 여전히 발생한다. 오차항이 정규분포를 따를 때와 마찬가지로 일정수준 이상의 중도절단 비율이 존재한다면 모수추정 시, 조절모수 λ를 정교하게 조절할 필요가 있다고 할 수 있다.

6. 결론

본 논문에서는 자료의 중도절단으로 인해 발생하는 오차항의 이분산성이 존재할 때, 추정의 정확성과 변수선택의 정확성에 대해 알아보았다. 그로인해 먼저 자료변환법에 대해 설명하였고 가중최소제곱법의 가중치 행렬 설정 방법을 제안하였으며, 변수선택법 중 LASSO와 Adaptive LASSO에 대해 설명하였다. 모의실험을 통해 다양한 상황에서의 다중회귀분석, LASSO, Adaptive LASSO의 추정과 변수선택의 성능을 몇 가지 수치와 평균제곱오차(MSE)를 사용하여 비교하였다. 모의실험 결과, 모형의 예측정확성, 해석을 고려한 LASSO와 Adaptive LASSO의 성능이 일반적인 회귀계수추정 방법인 다중회귀분석보다 뛰어난 것을 확인하였다. 특히, 중도절단에 관계없이 LASSO의 성능은 최소제곱법과 가중최소제곱법 모두에서 변수선택이 높은 성능을 보여주는데 이는 모형의 이분산성이 존재하여도 LASSO의 변수선택은 일치성을 만족한다는 것을 보여준다. 표본의 크기, 절단비율을 고려했을 때, 부분적으로 LASSO의 추정정확성과 변수선택의 성능이 Adaptive LASSO보다 높았지만, 대부분의 조건에서 Adaptive LASSO와 비슷한 수준의 성능을 보여주었다. 하지만, 일정수준 이상의 중도절단 비율을 넘으면 Adaptive LASSO의 변수선택 성능이 저하되는 것을 확인하였다. 이는 Adaptive LASSO의 모수추정 시, LASSO의 모수추정결과를 초기값으로 활용하는 부분에서 발생하는 문제로, LASSO 모수 추정 시, 조절모수 λ를 정교하게 조절한다면 해결될 수 있을 문제로 파악된다. 이를 종합하여, 모형의 오차항에 이분산성이 존재할 경우, 가중최소제곱법을 적용한 축소추정 등 변수선택법에 대한 차후 연구가 필요하다고 생각한다.

참고문헌

[1] Buckley, J & James, I. (1979). Linear regression with censored data. *Biometrika*, 66, 429-436.

[2] Koul, H., Susarla, V & Van Ryzin, J. (1981). Regression analysis with randomly right-censored data. *Annals of Statistics.* 9, 1276-1288

[3] Leurgans, S. (1987). Linear models, random censoring and synthetic data. *Biometrika,* 74, 301-309

[4] Robert Tibshirani. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society.* 58, 267-288

[5] Jinzhu Jia, Karl Rohe and Bin Yu. (2013) The lasso under poisson-like heteroscedasticity. *Statistica Sinica 23,* 99-118

[6] Hui Zou. (2006). The adaptive lasso and its oracle properties. *Journal of American Statistical Association.* 101, 1418-1429

[7] Raymond J. Carroll, C. F. Jeff Wu and David Ruppert. (1988). The effect of estimating weights in weighted least squares. *Journal of the American Statistical Association.* 83, 1045-1054

[8] M. Davidian and R. J. Carroll. (1987). Variance Function Estimation. *Journal of the American Statistical Association.* 82, 1079-1091

[9] Jianqing Fan and Runze Li. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association.* 96, 1348-1360