The Korean Journal of Applied Statistics
2021, Vol. 34, No. 6, 957–968

DOI: https://doi.org/10.5351/KJAS.2021.34.6.957
Print ISSN1225-066X / Online ISSN 2383-5818

A comparison study of inverse censoring probability weighting in censored regression

Jungmin Shin[a, b], Hyungwoo Kim[b], Seung Jun Shin[1, b]

[a]Department of Mathematics, Korea Military Academy; [b]Department of Statistics, Korea University

Abstract

Inverse censoring probability weighting (ICPW) is a popular technique in survival data analysis. In applications of the ICPW technique such as the censored regression, it is crucial to accurately estimate the censoring probability. A simulation study is undertaken in this article to see how censoring probability estimate influences model performance in censored regression using the ICPW scheme. We compare three censoring probability estimators, including Kaplan-Meier (KM) estimator, Cox proportional hazard model estimator, and local KM estimator. For the local KM estimator, we propose to reduce the predictor dimension to avoid the curse of dimensionality and consider two popular dimension reduction tools: principal component analysis and sliced inverse regression. Finally, we found that the Cox proportional hazard model estimator shows the best performance as a censoring probability estimator in both mean and median censored regressions.

Keywords: censored regression, censoring probability estimation, dimension reduction, cox proportional hazard model, local Kaplan-Meier estimator

1. 서론

반응변수가 어떤 사건이 일어날 때까지 걸리는 시간인 경우, 해당 사건이 관측되기 이전에 개체에 대한 추적이 불가능하게 되어 정확한 사건 발생시점을 확인할 수 없을 때, 중도절단(censoring)이 발생했다고 한다. 이러한 중도절단은 의생명 분야에서 흔히 발생한다. 예를 들어, 특정 질병의 발병 혹은 완치까지 걸리는 시간을 반응변수로 두었을 때, 환자가 어떤 이유로 병원 방문이 불가능하게 되면 해당 사건의 발생 시점을 정확히 관측할 수 없는 경우가 종종 발생한다. 이러한 이유로 중도절단자료 분석은 생존분석(survival analysis)이라는 이름으로 더욱 널리 알려져 있다. 중도절단이 발생하게 되면 반응변수에 대한 정보의 손실이 발생하게 되므로, 이를 적절히 해결하지 않으면 올바른 분석을 할 수 없다.

생존분석은 통계학에서 아주 오랫동안 활발하게 연구돼 온 주제이다. 관심 사건이 발생할 때까지 걸리는 시간을 T라고 하고, 이와 관련있는 p개의 공변량을 $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$ 라 하자. 본 연구에서는 반응변수 T와 공변량 $\mathbf{X}$간의 연관성을 탐색하기 위한 방법에 대해 살펴보고자 한다. 연관성 분석에서의 최종 목표는 $\mathbf{X}$가 주어졌을 때 T의 조건부 확률분포, 혹은 조건부 생존함수 $S(t|\mathbf{X})$를 주어진 자료로 부터 추정하는 것이다.

$$S(t|\mathbf{X}) := P(T > t|\mathbf{X}) = 1 - P(T \leq t|\mathbf{X}). \tag{1.1}$$

This work is supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (Grant No. NRF-2018R1D1A1B07043034 and NRF-2019R1A4A1028134).

[1] Corresponding author: Department of Statistics, Korea University, 145 Anam-Ro, Seongbuk-Gu, 02841 Seoul, Korea. E-mail: sjshin@korea.ac.kr

Published 31 December 2021 / journal homepage: http://kjas.or.kr

만약, 중도절단이 없다면 T에 대한 적절한 분포가정을 한 뒤 선형회귀와 같은 일반적인 방법을 통해 조건부 생존함수를 추정할 수 있다. 하지만 중도절단이 발생하면 T를 정확히 관측할 수가 없으므로, 통상적인 회귀모형을 직접 활용할 수 없다. 이 때 가장 널리 쓰이는 방법은 Cox의 비례위험(proportional hazard)모형이다 (Cox, 1972).

Cox의 비례위험모형은 위험률(hazard rate)에 대한 모형을 고려한 방법이다. 공변량 $\mathbf{X}$가 주어져 있을 때, 어떤 특정한 시점 t에서의 위험률 $\lambda(t|\mathbf{X})$는 t시점에서 사건 T가 발생할 가능성으로 해석할 수 있으며 $S(t|\mathbf{X})$와는 다음과 같은 관계가 있음이 잘 알려져 있다.

$$S(t|\mathbf{X}) = \exp\left\{-\int_0^t \lambda\left(u|\mathbf{X}\right)\mathrm{du}\right\}.$$

이제 Cox의 비례위험모형은 위험률에 대하여 다음을 가정한다.

$$\lambda(t|\mathbf{X}) = \lambda_0(t)\exp\left(\boldsymbol{\beta}^T\mathbf{X}\right).$$

여기서 $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$는 추정해야할 모수이며, $\lambda_0(t)$는 기저위험률함수를 나타내는 장애모수이다. 즉, Cox의 비례위험모형은 준모수 모형이다. 일반적인 준모수 모형과는 달리 Cox의 비례위험모형은 기저위험률 함수에 대한 정보가 없어도 $\boldsymbol{\beta}$를 손쉽게 추정할 수 있으며, 이렇게 추정된 $\boldsymbol{\beta}$는 비례위험 가정하에서 이론적으로도 최적의 추정량임을 보일 수 있다. 하지만, 비례위험 가정이 만족하지 않는 경우, Cox의 비례위험모형의 활용은 제한적이다.

Cox 비례위험모형의 대안으로 다음과 같은 회귀 모형을 고려할 수 있다.

$$h(T) = \beta_0 + \boldsymbol{\beta}^T\mathbf{X} + \epsilon. \tag{1.2}$$

여기서 $h(\cdot)$는 반응변수에 대한 변환함수이며, ϵ은 오차항을 의미한다. 모형 (1.2)는 통상적인 선형회귀모형과 유사하며, $\boldsymbol{\beta}$를 통해 반응변수와 공변량 사이의 관계를 해석할 수 있다. 모형 (1.2)는 두가지 방법으로 접근하여 풀 수 있다. 첫번째 방법은, 변환함수 h는 특정하고 ϵ의 분포를 모른다고 가정하는 것인데, 이를 가속실패시간 (accelerated failure time, AFT) 모형이라 한다. 두 번째 방법은 반대로, 변환함수 h는 모르고 ϵ의 분포는 특정하여 푸는 방법이다. 이는 변환(transformation) 모형이라 한다. 두 가지 모형 모두 준모수 모형이며, Cox 비례위험 모형과 더불어 중도절단자료의 연관성 분석에 가장 널리 활용되는 방법이다.

본 연구에서는 $h(T) = \log T$로 둔 다음과 같은 모형을 가정한다. 그 이유는 반응변수 T가 시간을 의미한다면, 양의 값을 가지게 되므로 로그변환은 가장 자연스러운 변환 중에 하나이기 때문이다.

$$\log T = \beta_0 + \boldsymbol{\beta}^T\mathbf{X} + \epsilon. \tag{1.3}$$

우선, 중도절단이 없는 경우를 고려해보자. 이 경우, 통상적인 회귀모형과 마찬가지로, 다음과 같은 목적함수 $Q(\beta_0, \boldsymbol{\beta})$를 최소화 하는 추정량을 고려할 수 있으며, 이를 중도절단회귀(censored regression)이라고 한다.

$$Q(\beta_0, \boldsymbol{\beta}) := E\left\{L\left(\log T - \beta_0 - \boldsymbol{\beta}^T\mathbf{X}\right)\middle|\mathbf{X}\right\}. \tag{1.4}$$

이 때, $L(\cdot)$은 통상적인 손실함수를 나타내며, 최소제곱 손실함수 ($L(u) = u^2$) 혹은 절대손실함수 ($L(u) = |u|$) 등을 고려할 수 있다.

하지만, 중도절단이 있다면 (1.4)의 추정은 간단하지 않다. 확률변수 C를 중도절단 시간을 나타내는 확률변수라 하면, 실제 관측되는 반응변수는 T가 아니라 $Y = \min\{T, C\}$와 중도절단 여부를 나타내는 지시확률변수 $\Delta = I(T \le C)$이기 때문이다.

본 연구에서는 중도절단하에서 중도절단 회귀모형의 회귀계수를 추정하기 위한 대표적인 방법으로 역중도절단확률가중(inverse censoring probability weighting, ICPW) 방법을 고려하였다. ICPW 방법은 다음과 같은 관계를 활용한 방법이다.

$$E\left\{L\left(T-\beta_0-\boldsymbol{\beta}^T\mathbf{X}\right)\middle|\mathbf{X}\right\}=E\left\{\frac{\Delta}{G(Y|\mathbf{X})}L\left(Y-\beta_0-\boldsymbol{\beta}^T\mathbf{X}\right)\middle|\mathbf{X}\right\}. \tag{1.5}$$

여기서, $G(c|\mathbf{X}) = P(C > c|\mathbf{X})$는 중도절단시간 C의 생존함수를 나타낸다. ICPW 방법은 중도절단으로 발생하는 정보의 손실을 가중치를 통해 조정하는 방법이다. 실제로 ICPW 방법은 결측자료 분석에서도 널리 활용되고 있다 (Tsiatis, 2007).

모형 (1.5)에 따라, $G(c|\mathbf{X})$가 알려져 있다면, 중도절단된 반응변수를 활용하여 회귀계수 $(\beta_0, \boldsymbol{\beta})$의 추정량을 손쉽게 구할 수 있다 (Ying 등, 1995; Bang과 Tsiatis, 2002; Zhou, 2006; Shin 등, 2017).

ICPW방법을 활용하면 중도절단자료를 통상적인 회귀모형과 유사한 방식으로 분석할 수 있다는 장점이 있다. 하지만 여기서 한가지 중요한 문제는 $G(c|\mathbf{X})$의 추정이다. ICPW를 활용한 연구는 매우 많지만, $G(c|\mathbf{X})$의 추정이 ICPW를 이용한 회귀계수추정량의 성능에 어떠한 영향을 주는지는 명확히 밝혀진 바가 없다. 본 연구에서는 다양한 종류의 $G(c|\mathbf{X})$를 적용하여 ICPW-기반 회귀계수추정량을 계산하여, 그 성능을 비교하고, ICPW-기반 회귀모형에서 중도절단확률을 선택하는 가이드 라인을 제시하고자 한다.

본 논문의 구성은 다음과 같다. 2절에서는 ICPW-기반 중도절단회귀 추정량을 소개하고, 다양한 중도절단확률 추정법에 대해 알아본다. 3절에서는 다양한 중도절단확률 추정방법들이 ICPW-기반 추정량의 성능에 미치는 성능을 알아보고자, 모의실험을 수행한다. 4절에서는 실제자료 분석 결과를 통해 모의실험에서 확인한 결과를 재검증 하고, 5절에서는 연구 내용에대한 요약과 결론을 제시하고자 한다.

2. 중도절단회귀모형

2.1. 중도절단회귀 추정량

중도절단 반응변수 $y_i = \min\{t_i, c_i\},\ i = 1, \ldots, n$와 중도절단 지시변수 $\delta_i = I(t_i \le c_i)$, 그리고 p개의 공변량을 나타내는 벡터 $\mathbf{x}_i$가 주어져 있다고 가정하자. 모형 (1.5)에 의하여, ICPW를 활용한 중도절단회귀 추정량은 다음의 표본 최적화 문제로 표현된다.

$$\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right)=\arg\min_{\beta_0,\,\boldsymbol{\beta}}\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\hat{G}(y_i|\mathbf{x}_i)}L\left(\log y_i-\beta_0-\boldsymbol{\beta}^T\mathbf{x}_i\right). \tag{2.1}$$

본 논문에서는, 가장 대표적인 최소제곱 손실함수, $L(u) = u^2$와 절대손실함수, $L(u) = |u|$를 고려하였다. 만약, 중도절단확률에 대한 추정량 $\hat{G}(\cdot|\mathbf{x})$이 주어져 있다면, (2.1)는 통상적인 가중회귀 방법으로 풀 수 있다. 중도절단확률 $G(\cdot|\mathbf{x})$를 추정하는 방법에는 여러가지가 있으며 , 이에 대한 내용은 다음 절에 자세히 기술하고자 한다.

2.2. 중도절단확률 $G(\cdot|\mathbf{X})$의 추정

중도절단확률을 추정하는 가장 간단한 방법은 Kaplan-Meier (KM) 추정량을 이용하는 것이다 (Bang과 Tsiatis, 2002). 이 방법은 공변량을 활용하지 않기 때문에 가장 간단한 방법이다. 하지만, 공변량과 중도절단시간 사이에 연관관계가 있다면 KM 추정량은 중도절단확률을 제대로 추정할 수 없다.

두번째 방법은 Cox 비례위험모형을 활용하는 것이다. $\hat{G}(\cdot|\mathbf{x})$는 중도절단시간 변수 C의 생존함수를 추정하는 것과 같으며, $(y_i, 1-\delta_i, \mathbf{x}_i)$를 이용하여 Cox 비례위험모형을 적합시켜 $\hat{G}(\cdot|\mathbf{x})$를 계산할 수 있다. Cox

비례위험모형은 KM 추정량과 달리 공변량과 중도절단시간 사이의 연관관계를 고려한 방법이다. Cox 비례위험모형은 비례위험 가정을 만족해야 한다는 것이 알려져 있다.

세번째 방법으로는 Beran (1981)이 처음 제안한 국소(local) KM 추정량을 이용하는 방법이다. 국소 KM 추정량은 다음과 같이 정의된다.

$$\hat{G}(t|\mathbf{x}) = 1 - \prod_{i=1}^{n} \left\{ 1 - \frac{B_{nk}(\mathbf{x})}{\sum_{k=1}^{n} B_{nk}(\mathbf{x})} \right\}^{\eta_j(t)}.$$

여기서 $\eta_j(t) = I(Y_j \le t, \delta_j = 1)$을 나타내며, $B_{nk}(\mathbf{x})$는 양의 가중치를 나타낸다. 만약 $B_{nk}(\mathbf{x}) = 1/n$이라 두면, 통상적인 KM 추정량이 됨을 알 수 있다. 공변량을 고려하기 위해서는 Nadaraya-Watson 추정량을 이용하여 $B_{nk}(\mathbf{x})$를 다음과 같이 계산할 수 있다.

$$B_{nk}(\mathbf{x}) = \frac{K\left(\frac{\mathbf{x}-\mathbf{x}_k}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)}. \tag{2.2}$$

이 때, $K(\cdot)$는 밀도 커널 함수(표준정규밀도함수 등)이며, $h > 0$는커널함수의 폭을 조절하는 모수이다.

국소 KM 추정량은 비모수적 방법으로 공변량과 중도절단시간 간의 관계를 모형가정 없이 추정할 수 있다는 장점이 있다. 그렇지만, 국소 KM 추정량은 커널 추정량을 기반으로 하기 때문에 공변량의 갯수가 많아지면 활용을 할 수 없다. 이를 해결하기 위한 방법으로, 우선 공변량 $\mathbf{x}$를 저차원 공간으로 축약시킨 뒤 국소 KM 추정량을 계산하는 방법을 고려할 수 있다.

차원축약의 방법으로 다음의 두가지 방법을 고려할 수 있다. 첫번째 방법은 주성분 분석을 이용한 것이다. 주성분 분석은 차원축약의 가장 대표적인 방법이다. 하지만 주성분 분석은 공변량의 정보만을 활용하여 차원축약을 하기 때문에 중도절단시간과의 연관관계에 대한 정보를 잘 보존하지 못할 가능성이 있다. 이에 대한 대안으로 두번째 방법은 충분차원축약을 활용하는 것이다. 충분차원축약은 공변량 자체가 아닌 공변량 속에 있는 반응변수의 정보를 가장 잘 유지하는 저차원 공간을 찾는 방법이며, 대표적인 방법으로는 sliced inverse regression (SIR) (Li, 1991)이 있다.

본 연구에서는 위의 네 가지 방법을 이용하여 $\hat{G}(\cdot|\mathbf{x})$를 추정하고, 중도절단확률 추정량이 중도절단회귀 추정량의 성능에 어떤 영향을 미치는지 살펴보고자 한다.

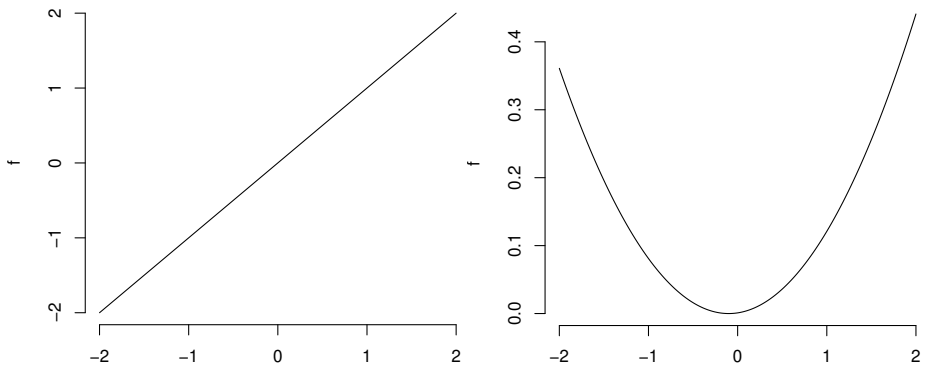
3. 모의실험

모의실험에서는 2절에서 소개한 네 가지 방법을 이용하여 중도절단확률 $\hat{G}(\cdot|\mathbf{x})$를 추정하였다. 네 가지 추정방법은 첫째, Kaplan-Meier(KM) 추정량, 둘째, Cox 비례위험모형을 통한 추정량(CoxPH), 셋째, 주성분 분석을 통한 차원축소 후 Beran (1981)이 제안한 국소(local) KM 추정량(LKM-PCA), 마지막으로는 SIR을 통한 차원축소 후 국소 추정량(LKM-SIR)을 고려하였다. 추가적으로 네 방법의 절대적인 비교를 위하여 중도절단을 무시하고 관찰된 생존시간 Y를 이용한 경우(Naive)와 실제 생존시간(T)를 이용한 경우(Oracle)를 함께 제시하였다. 성능비교는 ICPW-기반 중도절단회귀 회귀계수 추정값에 대한 평균제곱오차(mean squared error, MSE)를 기준으로 하였다. 모의실험은 각 경우마다 100번 반복 수행하였다. 국소 KM 추정량을 통한 추정 시 모형 (2.2)의 초모수 h는 적절한 크기로 조율되었다.

3.1. 데이터 설명 및 실험설계

생존시간 T를 생성하기 위해 다음과 같은 모형을 가정하였다.

$$\log T = 1 + \boldsymbol{\beta}^T \mathbf{X} + \epsilon. \tag{3.1}$$



(a) Linear Model (b) Nonlinear Model

Figure 1: *Regression model for C as function of $\boldsymbol{\gamma}^T\mathbf{X}$.*

한편, C는 다음의 모형에서 생성하였다.

$$\log C = \gamma_0 + f(\mathbf{X};\boldsymbol{\gamma}) + \epsilon'. \tag{3.2}$$

이 때, 두 모형의 ϵ과 ϵ'은 서로 독립이며, 상수항 γ_0는 중도절단비율을 30%로 제어하는 수준으로 결정하였다. 중도절단확률 추정량의 효과를 비교하기 다음과 같은 중도절단시간 회귀함수 f를 고려하였다. (Figure 1 참조)

- (선형) $f(\mathbf{X};\boldsymbol{\gamma}) = \boldsymbol{\gamma}^T\mathbf{X}$.
- (비선형) $f(\mathbf{X};\boldsymbol{\gamma}) = 0.1(0.1 + \boldsymbol{\gamma}^T\mathbf{X})^2$.

각 모형에서 다음과 같은 두가지 조합의 n과 p를 고려하였다: $(p = 4,\ n = 100), (p = 20,\ n = 200)$. p차원 독립변수 $\mathbf{X}$는 다변량 정규분포 $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I})$로 부터 생성하였다. 생존시간 T의 회귀계수는 $\boldsymbol{\beta} = (1, 1, \ldots, 1)$, 중도절단시간의 회귀계수 $\boldsymbol{\gamma}$는 첫 두 변수에 대응되는 값만 1, 나머지는 0 으로 설정하였다. 오차항 ϵ와 ϵ'의 분포는 극단값분포, 로지스틱분포, 표준정규분포 그리고 자유도 5인 t 분포 중 하나로 가정하였으며, 각 오차항에 대해 네 가지 분포를 교차 적용하여 실험을 실시하였다. 따라서 위에서 설명한 총 8가지 실험 가정의 조합에 대해서 각각 모의실험을 진행하였으며, 그 결과를 Table 1– Table 8에 정리하였다.

3.2. 선형모형

선형모형에 대한 결과는 Table 1–Table 4에 정리되어 있다. 네 가지 추정방법을 통해 계산된 회귀계수의 MSE의 중위수 값은 당연하게 중도절단을 무시한 경우와 실제 생존시간을 사용하여 추정한 경우의 사잇값으로 추정되었다. 대부분의 경우에 Cox 비례위험모형을 이용한 중도절단 회귀계수 추정량이 가장 좋은 성능을 보여주었다. 또한 평균적으로 LKM-SIR을 사용한 추정량의 성능이 LKM-PCA을 이용한 추정량보다 더 좋다는 것을 확인할 수 있다. SIR을 통한 차원축소 시 반응변수와 공변량 사이의 관계성을 반영하여 $\mathbf{X}$ 의 차원을 축소하기 때문에 단순히 공변량의 정보만 활용하여 차원을 축소하는 주성분 분석 방법 보다 회귀계수 추정에 있어서 더 좋은 성능을 보여주는 것으로 보인다. 추가적으로 Cox 비례위험모형이 비례위험 가정이 만족되지 않는 상황에서도 다른 추정법보다 우수한 성능을 보여준다. 이는 Cox 비례위험모형 모형이 ICPW 방법을 사용한 추정 문제에서 상당히 범용적으로 사용될 수 있음을 시사한다.

이론적으로는 비모수적 추정방법인 국소 KM 추정량을 사용한 LKM-PCA와 LKM-SIR의 결과가 더 좋을 것으로 예상되지만 실험결과에서는 CoxPH 모형을 통한 추정량이 비례위험 모형의 가정으로부터 강건함을 가졌다는 것을 알 수 있다. 이러한 결과는 실제 데이터를 이용해 성능을 비교한 4절에서도 확인할 수 있다.

Table 1: Median of MSE and its standard error of regression coefficients with squared loss function under the linear model when $n = 100,\ p = 4$

ϵ	ϵ'	Naive	KM	CoxPH	LKM-PCA	LKM-SIR	Oracle
ext	ext	1.15 (0.14)	0.91 (0.17)	0.83 (0.18)	0.73 (0.22)	**0.72** (0.22)	0.65 (0.12)
	logit	1.20 (0.15)	0.84 (0.16)	0.75 (0.15)	**0.64** (0.24)	0.68 (0.20)	0.65 (0.12)
	std	1.07 (0.14)	0.95 (0.15)	0.83 (0.20)	**0.76** (0.26)	0.84 (0.28)	0.65 (0.15)
	$t(5)$	1.14 (0.13)	0.93 (0.18)	0.79 (0.20)	**0.74** (0.29)	0.75 (0.26)	0.65 (0.12)
logit	ext	0.74 (0.15)	0.73 (0.19)	**0.66** (0.21)	0.76 (0.34)	0.68 (0.33)	0.37 (0.13)
	logit	0.78 (0.16)	0.63 (0.19)	**0.56** (0.20)	0.69 (0.45)	0.61 (0.34)	0.37 (0.13)
	std	0.69 (0.12)	0.76 (0.18)	**0.66** (0.22)	0.72 (0.41)	0.71 (0.34)	0.37 (0.13)
	$t(5)$	0.75 (0.14)	0.75 (0.22)	**0.66** (0.22)	0.80 (0.42)	0.78 (0.40)	0.37 (0.13)
std	ext	0.66 (0.12)	0.39 (0.12)	**0.35** (0.11)	0.46 (0.30)	0.41 (0.23)	0.22 (0.08)
	logit	0.68 (0.13)	0.33 (0.11)	**0.31** (0.10)	0.41 (0.28)	0.34 (0.24)	0.22 (0.08)
	std	0.63 (0.10)	0.40 (0.12)	**0.37** (0.14)	0.51 (0.23)	0.44 (0.27)	0.22 (0.08)
	$t(5)$	0.64 (0.12)	0.39 (0.14)	**0.34** (0.13)	0.45 (0.23)	0.39 (0.21)	0.22 (0.08)
$t(5)$	ext	0.67 (0.13)	0.45 (0.14)	**0.39** (0.13)	0.50 (0.28)	0.49 (0.24)	0.26 (0.11)
	logit	0.72 (0.14)	0.40 (0.16)	**0.37** (0.16)	0.45 (0.25)	0.41 (0.24)	0.26 (0.11)
	std	0.62 (0.11)	0.51 (0.15)	**0.45** (0.16)	0.51 (0.33)	0.54 (0.26)	0.26 (0.11)
	$t(5)$	0.72 (0.12)	0.48 (0.15)	**0.42** (0.17)	0.52 (0.27)	0.52 (0.26)	0.26 (0.11)

* ext : 극단분포, logit : 로지스틱 분포, std : 표준정규분포, LKM : Local Kaplan-Meier estimator.

Table 2: Median of MSE and its standard error of regression coefficients with absolute loss function under the linear model when $n = 100,\ p = 4$

ϵ	ϵ'	Naive	KM	CoxPH	LKM-PCA	LKM-SIR	Oracle
ext	ext	0.95 (0.17)	0.74 (0.21)	0.67(0.20)	0.73 (0.22)	**0.65** (0.22)	0.48 (0.13)
	logit	1.03 (0.18)	0.67 (0.19)	**0.59** (0.18)	0.64 (0.24)	0.63 (0.21)	0.48 (0.13)
	std	0.82 (0.44)	0.61 (0.44)	**0.53** (0.38)	0.59 (0.45)	0.56 (0.38)	0.35 (0.26)
	$t(5)$	1.00 (0.16)	0.81 (0.23)	**0.66** (0.24)	0.74 (0.29)	0.70 (0.33)	0.48 (0.13)
logit	ext	0.68 (0.18)	0.72 (0.25)	**0.62** (0.23)	0.76 (0.34)	0.72 (0.35)	0.41 (0.14)
	logit	0.74 (0.19)	0.62 (0.23)	**0.55** (0.26)	0.69 (0.45)	0.68 (0.38)	0.41 (0.14)
	std	0.66 (0.15)	0.75 (0.23)	**0.61** (0.25)	0.72 (0.41)	0.73 (0.39)	0.41 (0.14)
	$t(5)$	0.70 (0.18)	0.70 (0.27)	**0.67** (0.30)	0.80 (0.42)	0.78 (0.43)	0.41 (0.14)
std	ext	0.60 (0.13)	0.42 (0.16)	**0.39** (0.15)	0.46 (0.30)	0.45 (0.28)	0.29 (0.10)
	logit	0.60 (0.14)	0.38 (0.14)	**0.36** (0.13)	0.41 (0.28)	0.39 (0.23)	0.29 (0.10)
	std	0.63 (0.13)	0.46 (0.15)	**0.42** (0.19)	0.51 (0.23)	0.47 (0.28)	0.29 (0.10)
	$t(5)$	0.61 (0.13)	0.44 (0.17)	**0.38** (0.16)	0.45 (0.23)	0.44 (0.22)	0.29 (0.10)
$t(5)$	ext	0.59 (0.13)	0.43 (0.15)	**0.39** (0.14)	0.50 (0.28)	0.49 (0.24)	0.31 (0.10)
	logit	0.64 (0.16)	**0.39** (0.15)	0.40 (0.17)	0.45 (0.25)	0.45 (0.24)	0.31 (0.10)
	std	0.62 (0.13)	0.48 (0.18)	**0.43** (0.22)	0.51 (0.33)	0.52 (0.27)	0.31 (0.10)
	$t(5)$	0.57 (0.28)	0.40 (0.23)	**0.38** (0.34)	0.38 (0.34)	0.40 (0.30)	0.29 (0.17)

3.3. 비선형모형

Table 5 –Table 8은 비선형 모형하에서 진행한 모의실험 결과를 정리한 것이다. 두 번째 모의실험에서도 첫 번째 모의실험의 결과처럼 평균적으로 CoxPH 모형을 가정한 추정방법이 상대적으로 좋은 성능을 보여주었다. 하지만, 선형모형의 모의실험과는 다르게 비선형모형 하에서는 LKM-PCA과 LKM-SIR 방법 사이에는 일관된 우열 관계가 관찰되지 않았다. 이는 비선형회귀함수가 원점에 대해 대칭에 가깝기 때문이다. SIR은

Table 3: Median of MSE and its standard error of regression coefficients with squared loss function under the linear model when $n = 200,\ p = 20$

ϵ	ϵ'	Naive	KM	CoxPH	LKM-PCA	LKM-SIR	Oracle
ext	ext	2.00 (0.14)	1.03 (0.19)	1.03 (0.18)	**0.99** (0.21)	1.00 (0.23)	0.72 (0.11)
	logit	2.03 (0.15)	0.95 (0.18)	0.95 (0.18)	0.94 (0.25)	**0.93** (0.20)	0.72 (0.11)
	std	1.98 (0.14)	**0.96** (0.18)	0.99 (0.16)	1.02 (0.19)	1.00 (0.16)	0.72 (0.11)
	$t(5)$	2.00 (0.14)	1.00 (0.18)	**0.98** (0.17)	1.02 (0.20)	1.00 (0.18)	0.72 (0.10)
logit	ext	1.67 (0.14)	0.98 (0.20)	**0.96** (0.19)	1.02 (0.29)	1.01 (0.35)	0.62 (0.10)
	logit	1.71 (0.15)	0.92 (0.20)	**0.87** (0.19)	0.98 (0.40)	0.99 (0.41)	0.62 (0.10)
	std	1.69 (0.16)	1.17 (0.21)	**1.04** (0.20)	1.17 (0.39)	1.12 (0.39)	0.63 (0.09)
	$t(5)$	1.63 (0.15)	**0.94** (0.22)	0.96 (0.19)	1.07 (0.28)	0.96 (0.28)	0.63 (0.11)
std	ext	1.57 (0.16)	0.49 (0.11)	**0.47** (0.10)	0.50 (0.16)	0.51 (0.22)	0.33 (0.05)
	logit	1.60 (0.16)	0.47 (0.10)	**0.47** (0.10)	0.51 (0.20)	0.51 (0.22)	0.33 (0.05)
	std	1.60 (0.15)	0.53 (0.10)	**0.49** (0.08)	0.55 (0.12)	0.57 (0.12)	0.35 (0.06)
	$t(5)$	1.60 (0.17)	0.50 (0.10)	**0.50** (0.09)	0.58 (0.16)	0.53 (0.13)	0.36 (0.06)
$t(5)$	ext	1.62 (0.15)	0.61 (0.12)	**0.61** (0.14)	0.64 (0.19)	0.68 (0.21)	0.41 (0.08)
	logit	1.68 (0.16)	0.63 (0.13)	**0.61** (0.18)	0.66 (0.22)	0.66 (0.26)	0.42 (0.08)
	std	1.65 (0.16)	0.71 (0.13)	**0.63** (0.12)	0.71 (0.24)	0.69 (0.18)	0.47 (0.09)
	$t(5)$	1.60 (0.61)	0.62 (0.26)	**0.61** (0.25)	0.65 (0.34)	0.64 (0.31)	0.40 (0.17)

Table 4: Median of MSE and its standard error of regression coefficients with absolute loss function under the linear model when $n = 200,\ p = 20$

ϵ	ϵ'	Naive	KM	CoxPH	LKM-PCA	LKM-SIR	Oracle
ext	ext	1.78 (0.19)	0.94 (0.19)	**0.91** (0.18)	0.95 (0.22)	0.95 (0.25)	0.62 (0.10)
	logit	1.79 (0.19)	0.91 (0.19)	**0.87** (0.19)	0.93 (0.26)	0.91 (0.22)	0.62 (0.10)
	std	1.80 (0.18)	0.96 (0.15)	**0.92** (0.14)	1.00 (0.21)	0.98 (0.18)	0.61 (0.10)
	$t(5)$	1.80 (0.17)	0.96 (0.19)	**0.92** (0.18)	0.94 (0.22)	0.98 (0.19)	0.59 (0.10)
logit	ext	1.57 (0.18)	1.05 (0.23)	**0.99** (0.20)	1.07 (0.30)	1.06 (0.37)	0.66 (0.13)
	logit	1.58 (0.19)	1.01 (0.24)	**0.93** (0.22)	1.05 (0.38)	1.05 (0.41)	0.66 (0.13)
	std	1.64 (0.21)	1.09 (0.22)	**1.03** (0.27)	1.28 (0.38)	1.16 (0.33)	0.70 (0.12)
	$t(5)$	1.56 (0.19)	1.02 (0.25)	**0.94** (0.22)	1.08 (0.31)	1.07 (0.31)	0.67 (0.14)
std	ext	1.27 (0.17)	0.62 (0.17)	**0.55** (0.12)	0.62 (0.16)	0.63 (0.22)	0.40 (0.06)
	logit	1.27 (0.18)	0.62 (0.12)	**0.57** (0.12)	0.64 (0.20)	0.63 (0.22)	0.40 (0.06)
	std	1.38 (0.15)	0.62 (0.15)	**0.60** (0.14)	0.69 (0.16)	0.70 (0.15)	0.42 (0.07)
	$t(5)$	1.29 (0.22)	0.62 (0.14)	**0.59** (0.10)	0.71 (0.15)	0.67 (0.13)	0.43 (0.08)
$t(5)$	ext	1.34 (0.22)	0.66 (0.13)	**0.60** (0.1)5	0.70 (0.22)	0.75 (0.21)	0.43 (0.08)
	logit	1.41 (0.19)	0.68 (0.14)	**0.67** (0.14)	0.71 (0.22)	0.72 (0.26)	0.45 (0.08)
	std	1.26 (0.52)	0.71 (0.25)	**0.64** (0.26)	0.72 (0.37)	0.70 (0.27)	0.39 (0.15)
	$t(5)$	1.32 (0.53)	0.67 (0.29)	**0.63** (0.28)	0.68 (0.35)	0.66 (0.33)	0.43 (0.18)

회귀함수가 원점에 대해 대칭인 경우, 성능일 떨어진다는 것이 잘 알려져 있다. 이런 경우는 그 대안으로 sliced average variance estimation (SAVE) (Cook과 Weisberg, 1991)를 활용하면 문제를 해결할 수 있다.

3.4. 모의실험 결과 요약

모의실험 1과 2를 통하여 네 가지 종류의 $G(\cdot|\mathbf{x})$ 추정방법을 적용하여 ICPW기반 중도절단 회귀계수를 추정하고 그 성능을 비교해보았다. 비례위험을 가정한 CoxPH 모형을 통해 구한 $\hat{G}(\cdot|\mathbf{x})$를 사용한 ICPW기반

Table 5: Median of MSE and its standard error of regression coefficients with squared loss function under the nonlinear model when $n = 100,\ p = 4$

ϵ	ϵ'	Naive	KM	CoxPH	LKM-PCA	LKM-SIR	Oracle
ext	ext	1.26 (0.14)	0.85 (0.21)	0.85 (0.20)	0.84 (0.26)	**0.84** (0.25)	0.65 (0.12)
	logit	1.28 (0.15)	0.78 (0.16)	0.77 (0.16)	**0.71** (0.21)	0.73 (0.21)	0.65 (0.12)
	std	1.19 (0.09)	0.92 (0.18)	**0.89** (0.16)	0.94 (0.30)	1.01 (0.38)	0.63 (0.12)
	$t(5)$	1.22 (0.13)	0.82 (0.21)	0.80 (0.19)	0.84 (0.30)	**0.79** (0.28)	0.61 (0.15)
logit	ext	0.86 (0.16)	0.69 (0.24)	**0.68** (0.23)	0.74 (0.48)	0.74 (0.47)	0.37 (0.13)
	logit	0.88 (0.17)	0.58 (0.21)	**0.57** (0.20)	0.70 (0.46)	0.69 (0.46)	0.37 (0.13)
	std	0.75 (0.33)	0.61 (0.31)	0.64 (0.31)	**0.52** (0.31)	0.58 (0.31)	0.25 (0.17)
	$t(5)$	0.83 (0.14)	0.66 (0.22)	0.66 (0.23)	0.64 (0.57)	**0.63** (0.42)	0.40 (0.12)
std	ext	0.80 (0.13)	0.39 (0.11)	0.38 (0.11)	0.48 **0.29**	0.47 (0.32)	0.22 (0.08)
	logit	0.80 (0.14)	0.32 (0.09)	**0.32** (0.10)	0.38 (0.30)	0.38 (0.33)	0.22 (0.08)
	std	0.73 (0.27)	0.34 (0.21)	**0.31** (0.21)	0.63 (0.47)	0.66 (0.37)	0.27 (0.11)
	$t(5)$	0.77 (0.27)	0.35 (0.17)	**0.34** (0.16)	0.42 (0.42)	0.42 (0.34)	0.21 (0.10)
$t(5)$	ext	0.82 (0.14)	0.46 (0.17)	**0.45** (0.16)	0.50 (0.38)	0.51 (0.33)	0.26 (0.11)
	logit	0.86 (0.15)	0.39 (0.17)	**0.39** (0.13)	0.49 (0.40)	0.45 (0.42)	0.26 (0.11)
	std	0.75 (0.14)	0.50 (0.21)	**0.49** (0.22)	0.61 (0.51)	0.64 (0.40)	0.27 (0.10)
	$t(5)$	0.84 (0.17)	0.48 (0.16)	**0.47** (0.17)	0.54 (0.34)	0.50 (0.36)	0.26 (0.12)

Table 6: Median of MSE and its standard error of regression coefficients with absolute loss function under the nonlinear model when $n = 100,\ p = 4$

ϵ	ϵ'	Naive	KM	CoxPH	LKM-PCA	LKM-SIR	Oracle
ext	ext	1.08 (0.16)	0.66 (0.23)	**0.66** (0.22)	0.72 (0.31)	0.75 (0.30)	0.48 (0.13)
	logit	1.11 (0.18)	0.59 (0.20)	**0.58** (0.19)	0.64 (0.26)	0.67 (0.26)	0.48 (0.13)
	std	1.04 (0.46)	**0.76** (0.42)	0.79 (0.40)	0.83 (0.49)	0.82 (0.56)	0.40 (0.22)
	$t(5)$	1.07 (0.15)	0.70 (0.25)	**0.67** (0.25)	0.80 (0.33)	0.77 (0.33)	0.45 (0.19)
logit	ext	0.77 (0.23)	0.63 (0.31)	**0.61** (0.29)	0.69 (0.52)	0.69 (0.49)	0.40 (0.16)
	logit	0.83 (0.19)	0.58 (0.25)	**0.54** (0.22)	0.68 (0.46)	0.66 (0.46)	0.41 (0.14)
	std	0.77 (0.27)	0.67 (0.27)	**0.64** (0.26)	0.68 (0.39)	0.67 (0.37)	0.35 (0.17)
	$t(5)$	0.80 (0.22)	0.74 (0.32)	**0.68** (0.34)	0.74 (0.61)	0.77 (0.41)	0.39 (0.18)
std	ext	0.67 (0.14)	0.41 (0.16)	**0.40** (0.15)	0.49 (0.36)	0.49 (0.30)	0.29 (0.10)
	logit	0.68 (0.17)	0.38 (0.13)	**0.37** (0.13)	0.42 (0.31)	0.43 (0.32)	0.29 (0.10)
	std	0.69 (0.12)	0.40 (0.10)	**0.38** (0.10)	0.43 (0.33)	0.43 (0.15)	0.26 (0.10)
	$t(5)$	0.70 (0.14)	0.43 (0.15)	**0.40** (0.15)	0.43 (0.41)	0.44 (0.40)	0.26 (0.10)
$t(5)$	ext	0.69 (0.21)	0.44 (0.22)	**0.43** (0.17)	0.51 (0.40)	0.52 (0.32)	0.31 (0.12)
	logit	0.71 (0.17)	0.40 (0.14)	**0.40** (0.14)	0.48 (0.40)	0.49 (0.43)	0.31 (0.10)
	std	0.59 (0.15)	0.50 (0.21)	**0.39** (0.20)	0.51 (0.40)	0.58 (0.35)	0.26 (0.11)
	$t(5)$	0.71 (0.16)	0.44 (0.18)	**0.42** (0.18)	0.52 (0.41)	0.51 (0.26)	0.33 (0.10)

중도절단 회귀계수 추정량이 선형모형과 비선형모형 상황 모두에서 평균적으로 가장 좋은 성능을 보여주었다. 이는 Cox비례위험 모형이 비례위험을 가정한 준모수모형임에도 불구하고 생존분석에 많이 사용되는 ICPW 방법을 이용한 회귀분석 상황에서 상당히 강건함을 확인하였다. 두가지 국소 KM 추정량 (LKM-PCA, LKM-SIR)을 사용한 결과를 비교해보면, 공변량 $\mathbf{X}$의 차원을 축소하는 데 있어 반응변수와의 관계를 고려한 충분차원축약 방법인 SIR 방법이 단순히 공변량의 정보만을 활용하여 차원을 축소시키는 방법인 PCA보다 모형을 통한 회귀계수 추정에 있어서 더 우수한 성능을 기대할 수 있다.

Table 7: Median of MSE and its standard error of regression coefficients with squared loss function under the nonlinear model when $n = 200,\ p = 20$

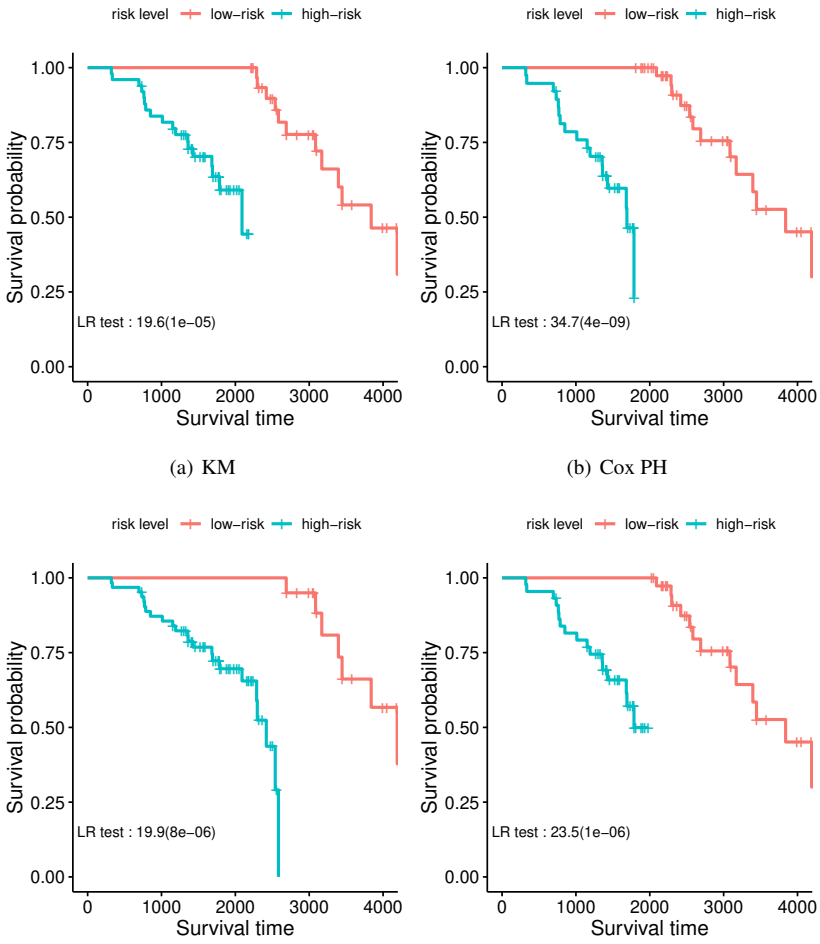
ϵ	ϵ'	Naive	KM	CoxPH	LKM-PCA	LKM-SIR	Oracle
ext	ext	2.67 (0.14)	1.17 (0.22)	1.19 (0.21)	1.15 (0.24)	**1.15** (0.23)	0.72 (0.11)
	logit	2.73 (0.15)	**1.07** (0.20)	1.08 (0.19)	1.11 (0.28)	1.11 (0.28)	0.72 (0.11)
	std	2.70 (0.14)	1.17 (0.24)	1.17 (0.23)	1.13 (0.29)	**1.12** (0.28)	0.72 (0.11)
	$t(5)$	2.71 (0.15)	1.14 (0.23)	1.16 (0.23)	1.15 (0.22)	**1.11** (0.26)	0.72 (0.11)
logit	ext	2.38 (0.14)	1.20 (0.25)	**1.15** (0.25)	1.23 (0.34)	1.22 (0.30)	0.62 (0.10)
	logit	2.40 (0.15)	1.07 (0.22)	**1.06** (0.22)	1.20 (0.45)	1.22 (0.43)	0.62 (0.10)
	std	2.05 (0.11)	**1.14** (0.29)	1.16 (0.27)	1.25 (0.13)	1.16 (0.15)	0.64 (0.10)
	$t(5)$	2.34 (0.15)	**1.13** (0.26)	1.14 (0.25)	1.22 (0.40)	1.27 (0.38)	0.62 (0.10)
std	ext	2.26 (0.15)	0.55 (0.12)	**0.55** (0.12)	0.57 (0.17)	0.58 (0.19)	0.33 (0.05)
	logit	2.29 (0.16)	0.55 (0.12)	**0.55** (0.12)	0.62 (0.20)	0.61 (0.40)	0.33 (0.05)
	std	2.23 (0.15)	0.59 (0.13)	0.58 (0.12)	**0.58** (0.12)	0.61 (0.17)	0.36 (0.06)
	$t(5)$	2.28 (0.16)	0.57 (0.12)	**0.55** (0.12)	0.63 (0.21)	0.63 (0.15)	0.33 (0.05)
$t(5)$	ext	2.30 (0.15)	0.75 (0.18)	**0.73** (0.17)	0.75 (0.28)	0.78 (0.27)	0.42 (0.08)
	logit	2.33 (0.16)	0.70 (0.17)	**0.68** (0.17)	0.80 (0.36)	0.80 (0.38)	0.42 (0.08)
	std	2.31 (0.16)	0.75 (0.20)	**0.75** (0.20)	0.78 (0.24)	0.78 (0.24)	0.42 (0.08)
	$t(5)$	2.28 (0.15)	0.74 (0.15)	**0.72** (0.15)	0.79 (0.21)	0.80 (0.19)	0.42 (0.08)

Table 8: Median of MSE and its standard error of regression coefficients with absolute loss function under the nonlinear model when $n = 200,\ p = 20$

ϵ	ϵ'	Naive	KM	CoxPH	LKM-PCA	LKM-SIR	Oracle
ext	ext	1.90 (0.16)	0.90 (0.21)	0.89 (0.19)	**0.88** (0.20)	0.89 (0.25)	0.59 (0.10)
	logit	1.91 (0.18)	0.91 (0.18)	0.88 (0.15)	0.87 (0.28)	**0.84** (0.24)	0.59 (0.10)
	std	2.58 (0.17)	1.16 (0.20)	**1.09** (0.21)	1.11 (0.28)	1.13 (0.31)	0.59 (0.10)
	$t(5)$	2.55 (0.18)	1.08 (0.23)	1.09 (0.23)	1.11 (0.20)	**1.07** (0.27)	0.60 (0.10)
logit	ext	2.30 (0.18)	1.20 (0.27)	**1.16** (0.29)	1.26 (0.36)	1.27 (0.32)	0.66 (0.13)
	logit	1.65 (0.19)	1.03 (0.19)	**1.03** (0.17)	1.12 (0.33)	1.05 (0.28)	0.66 (0.12)
	std	2.11 (0.19)	1.31 (0.18)	**1.28** (0.19)	1.34 (0.26)	1.37 (0.20)	0.72 (0.13)
	$t(5)$	1.51 (0.89)	0.83 (0.59)	**0.71** (0.59)	0.95 (0.58)	0.79 (0.63)	0.60 (0.39)
std	ext	1.34 (0.66)	0.57 (0.31)	**0.54** (0.31)	0.55 (0.34)	0.55 (0.32)	0.39 (0.21)
	logit	1.34 (0.44)	0.58 (0.22)	**0.56** (0.22)	0.58 (0.24)	0.57 (0.25)	0.39 (0.14)
	std	1.84 (1.09)	**0.52** (0.41)	0.56 (0.42)	0.53 (0.41)	0.53 (0.45)	0.35 (0.23)
	$t(5)$	2.08 (0.21)	**0.70** (0.12)	0.73 (0.17)	0.77 (0.25)	0.76 (0.15)	0.41 (0.07)
$t(5)$	ext	1.39 (0.57)	**0.59** (0.28)	0.60 (0.29)	0.64 (0.28)	0.59 (0.29)	0.40 (0.18)
	logit	1.43 (0.34)	0.68 (0.20)	**0.63** (0.18)	0.68 (0.22)	0.67 (0.26)	0.45 (0.12)
	std	2.17 (0.21)	0.80 (0.18)	**0.75** (0.20)	0.82 (0.27)	0.81 (0.26)	0.45 (0.08)
	$t(5)$	2.11 (0.21)	0.83 (0.16)	**0.79** (0.17)	0.81 (0.24)	0.80 (0.24)	0.45 (0.08)

4. 예시 : PBC data 분석

본 절에서는 일차성 담즙경화증(primary biliary cirrhosis, PBC)환자들의 생존시간을 기록한 데이터를 활용해서 앞서 제시한 네 가지 종류의 $G(\cdot|\mathbf{x})$ 추정방법을 비교하였다. PBC 데이터는 1974년부터 1984년 사이에 미국의 Mayo 클리닉에서 수집되었다. 이 데이터에는 총 312명에 대한 데이터가 기록되어 있으며 그 중 임상연구가 끝나기 전에 죽거나 기록 작성이 단절된 125명의 환자의 데이터를 포함하고 있다. 하지만 본 절의 분석에서는 결측자료를 제외한 276명의 데이터를 사용하여 분석하였다. 생존시간을 설명하는 공변량으로 다



(a) KM (b) Cox PH

(c) LKM-PCA (d) LKM-SIR

Figure 2: *Kaplan-Meier estimates of survival curves using four different* $\hat{G}(\cdot|\mathbf{x})$ *methods for low-risk and high-risk groups in the PBC data.*

음과 같은 네 가지 연속형 변수를 사용하였다. age(일), bili(serum bilirubin in mg/dl), albumin(albumin in g/dl), copper(urin copper in μg/day).

네 가지 종류의 $G(\cdot|\mathbf{x})$ 추정방법의 성능을 비교하기 위해서 먼저 전체 데이터를 학습데이터와 검증데이터로 랜덤하게 나누었으며(학습데이터 : 70%, 검증데이터 : 30%), 학습데이터를 활용하여 중도절단회귀계수 $(\beta_0, \boldsymbol{\beta})$를 추정하였다. 각 추정방법의 예측 성능을 비교하기 위해서 학습데이터에서 추정된 회귀식으로부터 생존시간 추정값의 중위수를 구한뒤 이 예측된 중위수보다 생존시간이 긴 환자는 저위험군으로 반대로 생존시간이 짧은 환자는 고위험군으로 분류하였다. 이 후 검증데이터를 사용하여 Figure 2에서 보는바와 같이 저위험군과 고위험군의 KM 추정량을 비교하였다. 네 가지 추정법 모두 두 그룹의 분리가 잘 이루어 졌으며, Figure 2 (b)에서 보듯이 CoxPH 모형을 사용한 추정법이 상대적으로 두 그룹의 분리가 가장 잘 되었다. 이 결과는 3.4절의 모의실험 결과와 일치한다. 함께 제시된 p값은 log-rank 테스트 결과이다.

5. 결론

본 연구에서는 중도절단하에서 중도절단 회귀모형의 회귀계수를 추정하기 위한 대표적인 방법인 역중도절단 확률가중(ICPW)방법에서 가장 중요한 중도절단 확률 $G(\cdot|\mathbf{x})$를 추정하는데 네 가지 방법을 모의실험과 PBC 데이터에 적용하고 그 결과를 비교분석 하였다.

모의실험 결과 비례위험을 가정한 준모수적 방법인 Cox 비례위험모형을 통해 구한 $\hat{G}(\cdot|\mathbf{x})$를 사용한 ICPW 추정량이 선형모형과 비선형모형 상황 모두에서 평균적으로 가장 좋은 성능을 보여주었다. 이를 통해 중도절단 회귀모형하에서 ICPW 방법의 사용에 있어 Cox 비례위험모형의 강건함을 발견할 수 있었다. 또한, Cox 비례위험모형을 통한 추정방법은 모형의 가정에서 자유로운 비모수적 방법인 국소 KM 추정방법보다 평균적으로 우수한 성능을 보여주었다.

국소 KM 추정량은 커널 추정량을 기반으로 하기 때문에 공변량의 갯수를 줄이기 위해 차원축소 방법인 PCA와 SIR방법을 적용하여 분석하였다. 그 결과 선형모형을 가정한 모의실험에서는 반응변수와 공변량 사이의 관계를 반영한 SIR 방법이 단순히 공변량의 정보만을 활용하여 차원을 축소하는 PCA 방법에 비해서 우수한 성능을 보여주었다.

본 연구를 통하여 중도절단 회귀모형의 회귀계수를 추정함에 있어 Cox 비례위험모형이 비례위험모형의 가정을 필요로 함에도 불구하고 상대적으로 우수한 성능을 보여주었으며, 차원 축소 방법에 있어서는 PCA 방법에 비해 SIR 방법이 더 효과적이라는 사실을 알 수 있다. 향후 연구에서 본 연구에서 사용된 방법론 이외 다양한 중도절단 확률 $G(\cdot|\mathbf{x})$ 추정방법을 비교분석하여 조금더 확장된 중도절단확률 선택의 가이드라인을 제시할 예정이다.

References

Bang H and Tsiatis A (2002). Median regression with censored cost data, *Biometrics*, **58**, 643-–649.

Beran R (1981). Nonparametric regression with randomly censored survival data, *Technical Report*, Universityof California at Berkeley.

Cook RD and Weisberg S (1991). Discussion of "sliced inverse regression for dimension reduction", *Journal of the American Statistical Association*, **86**, 28-–33.

Cox DR (1972). Regression models and life-tables, *Journal of the Royal Statistical Society: Series B*, **34**, 187-–202.

Li KC (1991). Sliced inverse regression for dimension reduction, *Journal of the American Statistical Association*, **86**, 316-–327.

Shin SJ, Zhang HH, and Wu Y (2017). A nonparametric survival function estimator via censored kernel quantile regressions, *Statistica Sinica*, **27**, 457-–478.

Tsiatis A (2007). *Semiparametric Theory and Missing Data*, Springer, New York.

Ying Z, Jung S, and Wei L (1995). Survival anaysis with median regression models, *Journal of the American Statistical Assocication*, **90**, 178-–184.

Zhou L (2006). A simple censored median regression estimator, *Statistica Sinica*, **16**, 1043-–1058.

Received November 6, 2021; Revised November 7, 2021; Accepted November 8, 2021

중도절단 회귀모형에서 역절단확률가중 방법 간의 비교연구

신정민[a,b], 김형우,[b] 신승준[1,b]

[a,]육군사관학교 수학과; [b]고려대학교 통계학과

요 약

역중도절단확률가중(inverse censoring probability weighting, ICPW)은 생존분석에서 흔히 사용되는 방법이다. 중도절단 회귀모형과 같은 ICPW 방법의 응용에 있어서 중도절단 확률의 정확한 추정은 핵심적인 요소라고 할 수 있다. 본 논문에서는 중도절단 확률의 추정이 ICPW 기반 중도절단 회귀모형의 성능에 어떠한 영향을 주는지 모의실험을 통하여 알아보았다. 모의실험에서는 Kaplan-Meier 추정량, Cox 비례위험(proportional hazard) 모형 추정량, 그리고 국소 Kaplan-Meier 추정량 세 가지를 비교하였다. 국소 KM 추정량에 대해서는 차원의 저주를 피하기 위해 공변량의 차원축소 방법을 추가적으로 적용하였다. 차원축소 방법으로는 흔히 사용되는 주성분분석(principal component analysis, PCA)과 절단역회귀(sliced inverse regression) 방법을 고려하였다. 그 결과 Cox 비례위험 추정량이 평균 및 중위수 중도절단 회귀모형 모두에서 중도절단 확률을 추정하는 데 가장 좋은 성능을 보여주었다.

주요용어: 중도절단 회귀, 중도절단확률 추정량, 차원축소, Cox 비례위험모형, 국소 Kaplan-Meier 추정량

이 논문은 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2018R1D1A1B07043034 and NRF-2019R1A4A1028134)

[1]교신저자:(02841) 서울특별시 성북구 안암로 145, 고려대학교 통계학과. sjshin@korea.ac.kr