碩士學位論文

# Variance reduction via guided Non-Parametric regression in censored data

韓國外國語大學校 大學院

統 計 學 科

金 峻 永

碩士學位論文

# Variance reduction via guided Non-Parametric regression in censored data

중도절단자료에서 유도된 비모수 회귀모형을 통한 분산 축소

指導 梁 城 準 敎授

이 論文을 碩士學位 請求論文으로 提出합니다.

2018年 6月


韓國外國語大學校 大學院

統 計 學 科

金 峻 永

이　論文을 金峻永의 碩士學位 論文으로 認定함.

2018年　6月 5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 $X$가 주어졌을 때 반응변수 $Y$의 조건부 평균을 추정하는 것을 목표로 한다, 하지만, 자료에 중도절단이 존재하는 경우에는 자료를 부분적으로만 관측하게 되기 때문에 기존에 알려져 있는 회귀방법론들의 직접적인 적용이 불가능하다. Koul et al.(1981)에서는 반응변수의 조건부 평균을 보존함으로써 기존 회귀방법론의 적용이 가능케 하는 간단한 자료변환법이 제안되었으나, 이는 회귀함수에 대한 추정량의 분산의 증가로 인하여 추정의 질이 저하되는 문제를 초래하게 된다. 본 논문에서는 이러한 문제의 해결을 위하여 Martins-Filho.et.al(2008)에서 제안된 Guided Non-Parametric Regression 방법론을 고려한다. 방법론의 적용을 위해 필요한 초기 추정량에 대한 몇 가지 방법들을 제안하고, 모의실험을 통하여  그 성능을 기존 자료변환법과 비교하였다.



주요용어 : Guided Non-Parametric Regression, 자료변환법

# 목 차

# [표 목 차]

# [그 림 목 차]

# 1. 서론

일반적으로 회귀분석은 설명변수와 반응변수간의 함수적 관계를 규명하는 것을 목적으로 설명변수 $X$가 주어졌을 때 반응변수 $Y$의 조건부 평균을 추정하는 것을 목표로 한다. 하지만, 자료에 중도절단이 존재하는 경우는 기존에 알려져 있는 회귀방법론들의 직접적인 적용이 불가능하다. 중도절단은 자료의 불완전한 관측을 나타내는 특성 중 하나이며, 생존 시간을 관측할 때 주로 발생한다. 중도절단의 원인은 다양하지만, 가장 주요 원인은 해당 연구의 중단이다. 이러한 불완전 관측을 해결하기 위해 조건부회귀함수의 값을 보존하는 자료변환법을 활용할 것이다. 이 변환법은 Buckley & James(1979), Koul et al.(1981), Leurgans(1987) 등에서 제안된 방법들이 대표적이다.

Koul et al.(1981)에서 소개한 자료변환법을 이용하게 되면 생존시간이 관측된 자료들은 팽창이 일어나고, 중도절단 자료들은 0의 값을 가지게 된다. 이로 인해 분산이 증가하고 추정의 질을 저하시키는 문제를 발생시킨다. 본 논문에서는 Guided Non-Parametric Regression방법을 활용하여 이 문제를 해결하고자 한다. 이 방법은 Beran,R.(1981), Glad, I. K. (1998), Martins-Filho.et.al (2008) 등에서 제안되었고, 이는 회귀함수에 대한 초기추정량을 제시하고, 결과로써 주어지는 잔차에 다시 회귀모형을 적합하여 수정하는 방식으로 최종 추정량을 제안하는 방법이다.

본 논문의 목적은 Guided Non-Parametric Regression 방법을 활용하여 중도절단이 포함되어 있는 자료에서 기존에 소개 되었던 자료변환법과 초기 추정량을 포함하여 새롭게 제안된 방법을 고려하여 모의실험을 통해 각 성능을

비교하는 것이다.

논문은 다음과 같이 구성하였다. 2절에서는 중도절단자료 변환법과 그 문제점을 소개하고 3절에서는 Guided Non-Parametric Regression에 대해 소개한다. 4절은 초기 추정량에 대한 여러 방법을 제안한다. Guided Non-Parametric Regression 방법을 적용시켜 방법론의 성능비교를 한 모의실험은 5절에서 이루어졌다. 6절에서는 결론을 다룬다.

## 2. 자료변환법

조건부회귀함수의 값을 보전하는 자료변환법을 생각했을 때, 비교적 간단하고 적용이 쉬운 Koul et al.(1981)에서 제안된 방식을 고려하였다. $(Y, X) \in R \times R^p$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 $Y$는 실제 관심대상인 생존시간이며, $C$ 는 중도절단시간을 나타내는 변수이다. 추정의 대상은 조건부회귀함수인 $E(Y|X=x)$이며 $Y$가 부분적으로만 관측되기 때문에 기존에 알려져 있는 회귀방법론에 바로 적용할 수 없다. 그리하여 다음과 같은 새로운 반응변수를 고려한다.

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

여기서, $G$는 $C$의 분포함수($G(t) = P(C \leq t)$)이다. 그러면, $Y$와 $C$는 서로 독립, $P(Y \leq C|X, Y) = P(Y \leq C|Y)$의 조건하에서

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

임을 보일 수 있다. 즉, 조건부회귀함수의 값이 모든 $X=x$에 대해 일치하게 되어 $Y^G$을 새로운 반응변수로 간주하여 회귀함수에 대한 추정이 가능해지며, 중도절단이 존재하지 않는다고 가정한 일반적인 회귀모형방법론을 적용하여 회귀함수를 추정할 수 있게 된다. 실제로는 $G$도 알려져 있지 않으므로

Kaplan-Meier 추정량 등으로 추정할 수 있다. 하지만 이는 절단변수의 분포가 설명변수 $X$에 의존하지 않는 경우에만 적합하다고 할 수 있다. 따라서 설명변수 $X$에 의존하는 경우는 Beran,R.(1981)에서 제안된 추정량 등을 사용할 수 있다. 본 논문에서는 우선 절단변수의 분포가 $X$에 의존하지 않는다고 가정한다. 새로운 반응 변수($Y^G$)에 대한 조건부분산은 다음과 같이 계산될 수 있다.

$$Var(Y^G|X=x) = Var(Y|X=x) + E\left(\frac{G(T)}{1-G(T)}Y^2|X=x\right) \qquad (2.3)$$

즉, 변환된 $Y^G$의 조건부분산은 이분산성이 존재할 뿐 아니라, $Y$의 조건부분산보다 커지는 경향이 있다. 여기서 증가분을 살펴보면 $E\left(\frac{G(T)}{1-G(T)}Y^2|X=x\right)$ 의 $\frac{G(T)}{1-G(T)}Y^2$ 은 $Y(Y>0)$에 대한 증가함수임을 알 수 있다. 이것은 생존시간의 값이 클수록 분산의 증가폭도 커짐을 의미한다. 보통은 회귀함수추정량의 분산이 반응변수의 크기에 영향을 받지 않는 경우가 많다. 예를 들어, Local Constant Estimation을 생각해보자. 이는 다음과 같이 표현할 수 있다.

$$\hat{m}(x) = \frac{\sum K_h(X_i-x)Y_i}{\sum K_h(X_i-x)}$$

여기서 $(X_i, Y_i)$는 서로 독립이고, $K$는 Kernel함수, $h$는 Bandwidth이다. $Y_i' = Y_i + C$라 할 때, $\hat{m'}(x) = \hat{m}(x) + C$이며, 두 추정량의 분산은 같다. 하지만, 자료에 중도절단이 존재하는 경우를 고려해 보자.

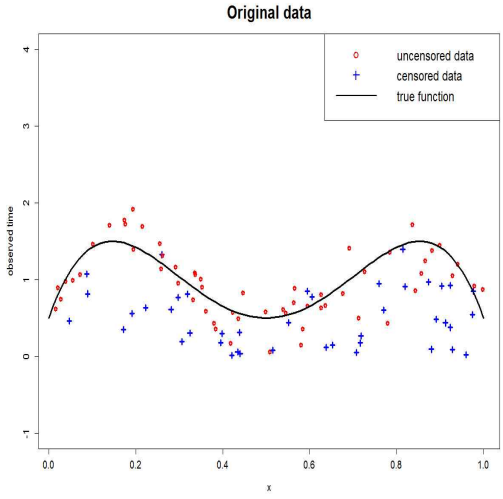
1) $Y_i^{'} = Y_i + C$

2) $C_i^{'} = C_i + C$

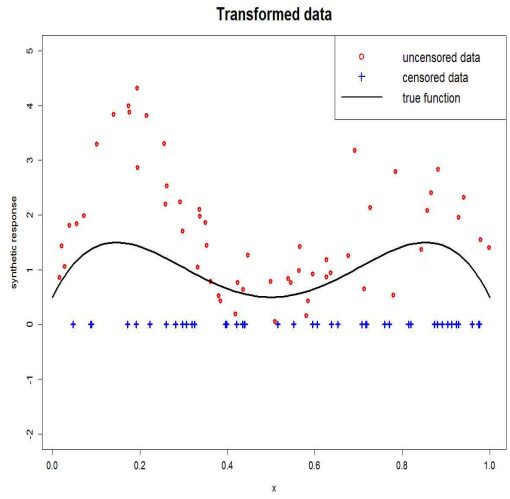
3) $T_i^{'} = T_i + C$

4) $\delta_i^{'} = \delta_i$

위와 같은 조건의 값은 (2.1)에서 소개한 변환법을 활용하여 자료 변환하여야 한다. 변환된 자료는 (2.3)의 성질 때문에 $\widehat{m_G}(x)$와 $\widehat{m}(x)$의 분산이 달라지는 것을 확인할 수 있다. 즉, 생존시간이 상대적으로 길게 나타나는 경우 추정량의 분산의 증가폭이 더 커지므로 추정의 질을 저하시키는 문제를 야기할 수 있다.




[그림 2.1] 중도절단자료 변환 전




[그림 2.2] 중도절단자료 변환 후

**[그림 2.1]**과 **[그림 2.2]**는 중도절단이 존재하는 자료를 변환하기 전과 변환 후의 분포를 나타낸 그림이다. **[그림 2.1]**을 살펴보면 중도절단이 일어나

지 않는 자료는 실제 함수 주변에 분포하며, 중도절단이 일어난 자료는 그보다 다소 적은 값들을 가짐을 알 수 있다. **[그림 2.2]**를 살펴보면 변환 후 관측된 자료는 팽창이 일어나고, 중도절단이 일어난 자료는 0의 값을 가지게 되는 걸 확인할 수 있다. 이로 인해 추정의 질이 저하되고, 분산이 증가하는 문제가 발생한다. 이런 문제를 해결하고자 Guided Non-Parametric 방법을 활용하고자 하였고, 이를 위한 초기 추정량에 대한 여러 방법들을 제안한다.

## 3. Guided Non-Parametric Regression

Guided Non-Parametric Regression은 Martins-Filho.et.al (2008)에서 제안했으며 Additive 접근법을 고려한 방법이다. 관측된 자료 $(Y_i, X_i), i = 1, ..., n$가 있다고 가정하자. 조건부기댓값으로 표현하면 다음과 같다.

$$r(x) = E(Y|X=x) = E(Y - r(x,\theta)|X=x) + r(x,\theta)$$

여기서, $r(x) = E(Y|X=x)$는 실제 회귀 함수이고 $r(x,\theta)$은 유한모수에 의존하는 함수이다. 이 때, $\hat{\theta}$을 $\theta$의 추정량이라 하면 $r$에 대한 추정량은 다음과 같이 표현할 수 있다.

$$\hat{r}(x) = r(x,\hat{\theta}) + s_{\hat{\theta}}(x)$$

여기서, $s_{\hat{\theta}}(x) = \hat{r}(x) - r(x,\hat{\theta}) = E(Y - r(x,\hat{\theta})|X=x)$으로 표현될 수 있다. 초기 추정의 대상이 되는 함수는 꼭 유한 모수에 의존하는 함수일 필요는 없다. 임의의 함수 $r(x)$에 대하여

$$r(x) = E(Y - r(x)|X=x) + r(x)$$

가 성립한다. $r(x)$에 대해서는 위와 같이 모수적 방법을 사용할 수도 있으나 본 논문에서는 비모수적 방법을 활용할 것이다. 초기 추정량 $r(x)$에 대해서는 여러 가지 방법을 제안할 것이다.

본 논문에서는 추정량의 분산을 효과적으로 줄일 수 있도록 초기 추정량

$r(x)$에 대해 여러 가지 접근법을 제안하고 모의실험을 통하여 비교하여 가장 효과적으로 분산을 축소할 수 있는 방법을 찾을 것이다.

## 4. 초기 추정량에 대한 방법

앞서 제기한 문제의 해결을 위해 다음과 같은 여러 방법론을 제안하였다.

### 4.1. 모수적 초기 추정량

첫 번째 방법은 Martins-Filho.et.al (2008)에서 제안한 방법을 그대로 적용하는 것을 고려하였다. 먼저 중도절단자료에서 제안되었던 자료변환법을 이용하여 새로운 반응변수 $Y^G$를 계산 후 $Y^G$와 설명변수 $X$를 이용하여 모수모형에 적합 시킨다. 적합한 모수모형의 잔차를 초기 추정량 $r_{x_P}$로 지정한다. 초기 추정량 $r_{x_P}$를 변수 $T_i(=\min(Y_i, C_i))$에서 뺀 후 새로운 반응변수 $Y_i^{'G}$를 계산 후 $Y_i^{'G}$와 $X_i$를 사용하여 비모수모형에 적합 시킨다. (단, 모수모형의 적합을 위해서는 3차 다항함수를 이용한다.)

① 중도절단 자료를 이용 $Y^G$를 생성

② $Y^G$와 $X_i$를 이용 모수모형에 적합

③ 적합 모형의 잔차를 초기 추정량 $r_{x_P}$로 설정

④ 초기 추정량 $r_{x^3}$를 $T_i(=\min(Y_i, C_i))$에서 뺀 후 새로운 $T^{'}$을 생성

⑤ 새로운 $T^{'}$을 이용하여 자료변환법에 의해 $Y_i^{'G}$을 생성

⑥ $Y_i^{'G}$와 $X_i$를 이용 비모수모형에 적합

## 4.2. 비모수적 초기 추정량

초기 추정량이 변환법에 의한 분산의 증가에서 자유롭게 할 수 있는 접근법을 고려하였다. 이 접근법은 중도절단자료에서 중도절단이 일어나지 않은 자료를 이용한 방법이다. 이는 절단비율이 클 때는 효율적이지 않으나, 절단비율이 작은 경우 잃게 되는 정보가 제한적이므로 비교적 효율적인 초기 추정이 가능할 수 있다. 초기추정량을 모수적 방법을 생각할 수 있으나, 여기서는 비모수적 방법을 고려하였다. 이는 $Y_i - r(x_i)$의 크기를 상대적으로 작게 하여 변환에 의한 분산 증가를 완화시키려는 의도이다.

① 중도절단이 일어나지 않은 자료를 이용 비모수모형에 적합
② 적합한 모형의 잔차를 초기 추정량 $r_{x_{NP}}$로 설정
③ 초기 추정량 $r_{x_{NP}}$를 $T_i(=\min(Y_i, C_i))$에서 뺀 후 새로운 $T^{'}$을 생성
④ 새로운 $T^{'}$을 이용하여 자료변환법에 의해 $Y_i^{'G}$을 생성
⑤ $Y_i^{'G}$와 $X_i$를 이용 비모수모형에 적합

## 4.3. 중위수 초기 추정량

세 번째는 전체 생존시간의 중심위치를 측정하여 이를 초기 추정량으로 활용하는 것을 고려한 방법이다. 이는 독립변수의 값을 고려하지 않는다는 단점이 있으나, 그 계산은 비교적 간단하다는 장점이 있다. 조건부 평균에 대한 추정이므로 $E(Y)$를 추정하는 것이 자연스러우나, 중도절단으로 인하여 그 추정은 어려우므로 중위수 추정량으로 대체한다.

① 생존시간의 중위수를 계산

② 계산된 중위수를 초기 추정량 $r_{x_m}$로 설정

③ 초기 추정량 $r_{x_m}$를 $T_i(=\min(Y_i, C_i))$에서 뺀 후 새로운 $T^m$을 생성

④ 새로운 $T^{'}$을 이용하여 자료변환법에 의해 $Y_i^{'G}$을 생성

⑤ $Y_i^{'G}$와 $X_i$를 이용 비모수모형에 적합

## 4.4. True Mean Function

마지막은 True Mean Function을 고려하였다.. True Mean Function를 초기 추정량 $r_{x_s}$로 설정하는 방법으로 $T_i(=\min(Y_i, C_i))$에서 $r_{x_s}$뺀 후 새로운 반응변수 $Y_i^{'G}$를 계산 후 $Y_i^{'G}$와 $X_i$를 사용하여 비모수모형에 적합 시킨다. 여기서 True Mean Function을 설정한 모형의 값을 사용하였다. 이는 본 논문에서 제안하는 방법들이 True Mean Function을 알고 있다고 가정하는 상황과 비교하여 얼마나 효율적인지를 보고자 시도하였다.

① True Mean Function를 초기 추정량 $r_{x_s}$로 설정

② 초기 추정량 $r_{x_s}$를 $T_i(=\min(Y_i, C_i))$에서 뺀 후 새로운 $T^m$을 생성

③ 새로운 $T^{'}$을 이용하여 자료변환법에 의해 $Y_i^{'G}$을 생성

④ $Y_i^{'G}$와 $X_i$를 이용 비모수모형에 적합

여기서 계산된 $Y_i^{'G}$은 기존 자료변환법인 $\frac{\delta_i T_i}{1-G(T_i)}$을 활용하였고, 초기 추정량의 설정만 변경하는 여러 방법을 제안하였다. 본 논문에서는 제안된 방법을 Guided Non-Parametric Regression 방법론에 적용시켜 성능을 비교할 것이다.

## 5. 모의실험

모의실험은 오픈 소스인 R을 이용하여 진행되었으며, 모의실험에 사용할 2가지 모형은

$$① \ Y_i = 5 + \sin(2\pi X_i) + \epsilon_i, \ ② \ Y_i = 5 + \sin(4\pi X_i) + \epsilon_i$$

이다. 함수의 변화의 정도에 따른 성능의 차이를 보고자 두 경우를 고려하였다. 먼저 $X_i \sim U(0,1)$로 설정하였다. 비교를 위해 중도절단이 없는 경우도 함께 고려하였으며, 절단 변수의 분포는 $C_i \sim N(5,1) + \alpha$로 설정하였다. 절단비율이 약 10%, 30%, 50%가 되도록 $\alpha$의 값을 적절히 조정하였다. 오차항은 $\epsilon_i \sim N(0,1^2)$로 설정하였으며, 표본의 크기는 100, 200, 400을 동시에 고려하였다. 총 반복수는 200으로 설정한 후 모의실험을 진행하였다. 또한, 변환된 반응변수 $Y^G$에서 분포함수 $G$는 모르는 경우가 일반적이므로 이의 추정을 위해 Kaplan-Meier 추정량 $\hat{G}$을 사용하였다. 초기 추정량은 모수함수를 활용한 모수적 추정량, 중도절단이 일어나지 않은 자료를 활용한 비모수적 추정량, 생존시간의 중위수를 활용한 추정량을 활용하는 방법을 제안하였다.

비모수 회귀분석을 위해 R의 'np' 패키지를 활용하여 모의실험을 진행하였고, 'np' 패키지를 활용하기 위해서는 Bandwidth의 지정이 필요하다. 이를 위해 'KernSmooth' 패키지의 'dpill'함수를 활용하여 Bandwidth를 설정하였으며, 'dpill'함수는 Plug-in방식을 활용한 Bandwidth를 지정하게 된다. $GNR_M$은 중위수의 계산이 필요하므로 'Survival' 패키지를 이용하여 중위수를 계산하였다. 제안한 방법의 성능을 비교하기 위해 설정한 2가지 모형의 $Bias^2$(편향$^2$),

Variance(분산), MSE(평균제곱오차)를 계산하여 제시하였다.

$Bias^2$(편향$^2$)은 실제 값에 대한 추정 값의 오차를 나타내므로 0에 가까울수록 좋은 성능을 의미하며, Variance(분산)는 중심위치로부터 얼마나 퍼져 있는지를 나타내는 척도이며, MSE(평균제곱오차)는 실제 값과 추정 값과의 차이를 나타낸다. 두 값은 작을수록 좋은 성능을 의미할 것이다. 초기 추정량에 대한 방법의 성능비교 결과를 모형 $Y_i = 5 + \sin(2\pi X_i) + \epsilon_i$일 경우, 중도절단의 비율에 따라 0%, 10%, 30%, 50% 순서대로 **[표 5.1]**부터 **[표 5.4]**에 나타내었고, 모형 $Y_i = 5 + \sin(4\pi X_i) + \epsilon_i$일 경우, 중도절단의 비율에 따라 0%, 10%, 30%, 50% 순서대로 **[표 5.5]**부터 **[표 5.8]**에 나타내었다. 또한, **[그림 5.1]**과 **[그림 5.2]**는 각 모형에서 중도절단 10%에서의 제안한 방법에 대한 Mean Plot을 나타내었다.

## 5.1. 모형 $Y_i = 5 + \sin(2\pi X_i) + \epsilon_i$

**<표 5.1> 중도절단이 0%인 경우, $Y_i = 5 + \sin(2\pi X_i) + \epsilon_i$**

| | $N = 100$ | | | $N = 200$ | | | $N = 400$ | | |
|---|---|---|---|---|---|---|---|---|---|
| **방법** | $Bias^2$ | $Var$ | $MSE$ | $Bias^2$ | $Var$ | $MSE$ | $Bias^2$ | $Var$ | $MSE$ |
| $KSV$ | 0.0046 | 0.0680 | 0.0726 | 0.0031 | 0.0363 | 0.0395 | 0.0019 | 0.0179 | 0.0197 |
| $GNR_P$ | 0.0009 | 0.0676 | 0.0685 | 0.0011 | 0.0349 | 0.0359 | 0.0009 | 0.0166 | 0.0175 |
| $GNR_{NP}$ | 0.0001 | 0.0853 | 0.0854 | 0.0000 | 0.0393 | 0.0393 | 0.0000 | 0.0185 | 0.0185 |
| $GNR_M$ | 0.0000 | 0.0758 | 0.0758 | 0.0031 | 0.0367 | 0.0398 | 0.0009 | 0.0189 | 0.0198 |
| $GNR_T$ | 0.0000 | 0.0655 | 0.0655 | 0.0000 | 0.0317 | 0.0317 | 0.0001 | 0.0139 | 0.0140 |

$KSV$ : 기존 자료변환법 $GNR_P$ : 모수적 초기 추정량 $GNR_{NP}$ : 비모수적 초기 추정량
$GNR_M$: 중위수 초기 추정량 $GNR_T$: True Mean Function

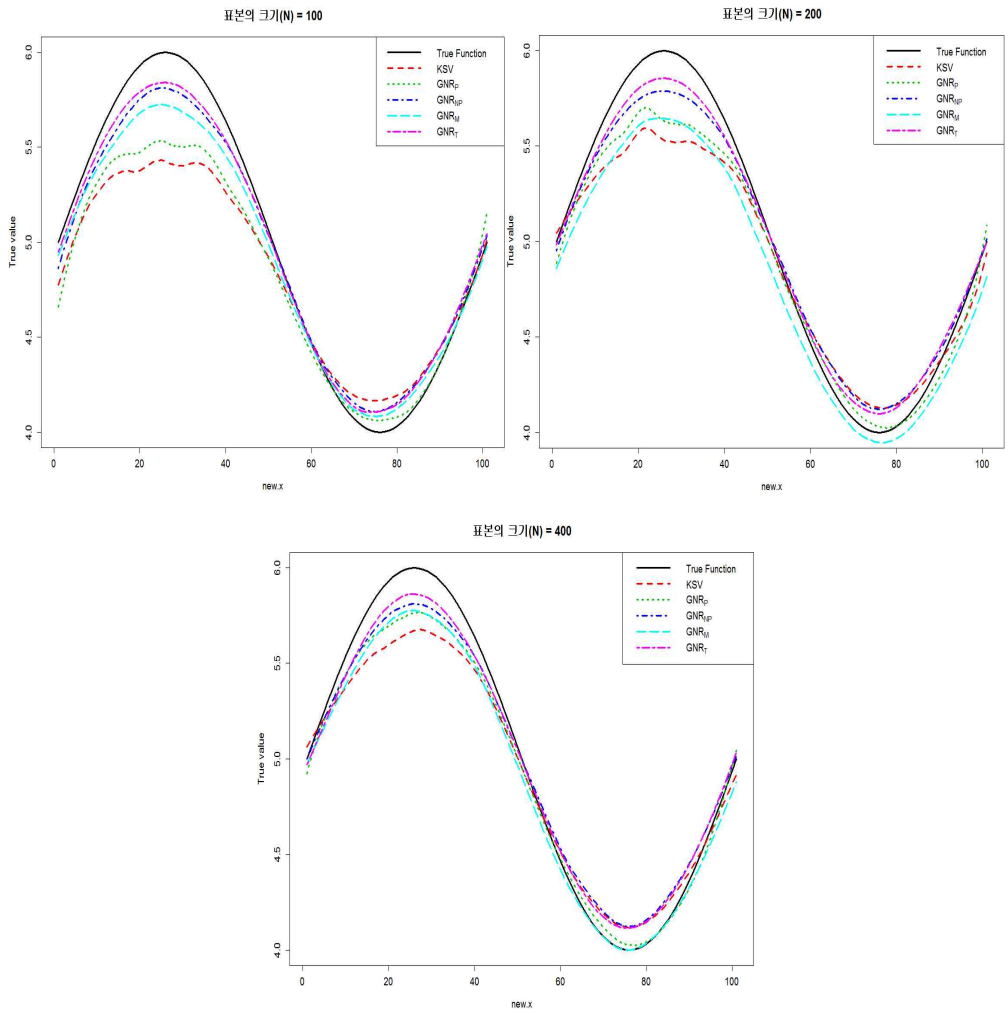
먼저, 모형 $Y_i = 5 + \sin(2\pi X_i) + \epsilon_i$에서 중도절단이 없는 경우인 **<표 5.1>**의 결과를 살펴보면, 전체적으로 $GNR_T$의 성능이 $KSV$와 $GNR_P$, $GNR_{NP}$, $GNR_M$의 성능보다 뛰어난 걸 확인할 수 있으며, 표본의 크기가 증가하면서 모든 방법의 성능이 좋아지는 걸 알 수 있다. 이는 중도절단이 없는 경우에는 표본의 크기와 상관없이 가장 뛰어난 성능을 가지는 방법은 $GNR_T$이다.

**<표 5.2> 중도절단이 10%인 경우, $Y_i = 5 + \sin(2\pi X_i) + \epsilon_i$**

| | $N = 100$ | | | $N = 200$ | | | $N = 400$ | | |
|---|---|---|---|---|---|---|---|---|---|
| **방법** | $Bias^2$ | $Var$ | $MSE$ | $Bias^2$ | $Var$ | $MSE$ | $Bias^2$ | $Var$ | $MSE$ |
| $KSV$ | 0.0856 | 0.3914 | 0.4770 | 0.0466 | 0.1897 | 0.2364 | 0.0287 | 0.1135 | 0.1422 |
| $GNR_P$ | 0.0582 | 0.4110 | 0.4692 | 0.0261 | 0.1959 | 0.2220 | 0.0133 | 0.1133 | 0.1266 |
| $GNR_{NP}$ | 0.0130 | 0.0881 | 0.1011 | 0.0133 | 0.0404 | 0.0537 | 0.0130 | 0.0200 | 0.0330 |
| $GNR_M$ | 0.0132 | 0.0873 | 0.1005 | 0.0114 | 0.0441 | 0.0555 | 0.0057 | 0.0262 | 0.0319 |
| $GNR_T$ | 0.0085 | 0.0738 | 0.0823 | 0.0074 | 0.0375 | 0.0450 | 0.0091 | 0.0177 | 0.0268 |

$KSV$ : 기존 자료변환법 $GNR_P$ : 모수적 초기 추정량 $GNR_{NP}$ : 비모수적 초기 추정량
$GNR_M$: 중위수 초기 추정량 $GNR_T$: True Mean Function

**<표 5.2>**는 중도절단이 10%인 경우, 모형 $Y_i = 5 + \sin(2\pi X_i) + \epsilon_i$ 때의 성능 비교에 대한 결과이다. 전체적으로 살펴보면 표본의 크기와 상관없이 $GNR_T$의 성능이 가장 뛰어난걸 알 수 있으며, 중도절단이 없는 경우와 비교해보면 $GNR_{NP}$와 $GNR_M$의 성능이 향상되었다는 걸 확인할 수 있으며, $GNR_T$의 성능과 큰 차이가 없으므로 효율적인 방법이라 할 수 있다. 반면, $GNR_P$의 성능은 저하된다는 것을 확인할 수 있으며, 이는 $GNR_P$은 중도절단이 없는 경우에는 효율적이지만 중도절단이 존재하는 경우에는 효율적이지 못한 방법임을 의미한다.




**[그림 5.1] 모형** $Y_i = 5 + \sin(2\pi X_i) + \epsilon_i$, **<중도절단10%>에서 각 방법에 대한 Mean Plot**

**[그림 5.1]**은 모형 $Y_i = 5 + \sin(2\pi X_i) + \epsilon_i$, 중도절단이 10%인 경우로, 제안한 초기 추정량 방법에 대한 Mean Plot을 그린 것이다. **[그림 5.1]**을 **<표 5.2>**의 결과와 비교해보면 전체적 성능은 $GNR_T$가 가장 뛰어나다.

또한 $GNR_{NP}$와 $GNR_M$의 성능이 향상 됐으며, $GNR_T$의 성능과 큰 차이가 없음을 확인 할 수 있다.

**<표 5.3> 중도절단이 30%인 경우, $Y_i = 5 + \sin(2\pi X_i) + \epsilon_i$**

| | $N = 100$ | | | $N = 200$ | | | $N = 400$ | | |
|---|---|---|---|---|---|---|---|---|---|
| **방법** | $Bias^2$ | $Var$ | $MSE$ | $Bias^2$ | $Var$ | $MSE$ | $Bias^2$ | $Var$ | $MSE$ |
| $KSV$ | 0.4329 | 1.5949 | 2.0278 | 0.2242 | 0.9166 | 1.1408 | 0.1523 | 0.4794 | 0.6317 |
| $GNR_P$ | 0.4101 | 1.7152 | 2.1253 | 0.2031 | 0.9592 | 1.1622 | 0.1220 | 0.5030 | 0.6250 |
| $GNR_{NP}$ | 0.0721 | 0.1177 | 0.1898 | 0.0683 | 0.0637 | 0.1320 | 0.0700 | 0.0311 | 0.1012 |
| $GNR_M$ | 0.0790 | 0.1184 | 0.1974 | 0.0479 | 0.0736 | 0.1215 | 0.0303 | 0.0436 | 0.0739 |
| $GNR_T$ | 0.0261 | 0.0794 | 0.1055 | 0.0202 | 0.0313 | 0.0515 | 0.0233 | 0.0200 | 0.0433 |

$KSV$ : 기존 자료변환법 $GNR_P$ : 모수적 초기 추정량 $GNR_{NP}$ : 비모수적 초기 추정량
$GNR_M$: 중위수 초기 추정량 $GNR_T$: True Mean Function

**<표 5.3>**은 모형 $Y_i = 5 + \sin(2\pi X_i) + \epsilon_i$에서 중도절단이 30%인 경우로, 제안한 각 방법에 대한 결과를 비교해보면 $KSV$, $GNR_P$의 성능은 절단비율이 10%일 때보다 확연히 저하된 것을 알 수 있다. 반면, $GNR_M$의 성능은 중도절단이 10%일 때와는 다르게 표본의 크기가 커질수록 향상되는 것을 확인할 수 있다. 이는 일정수준의 중도절단 비율을 넘어서면 $GNR_M$을 제외한 $KSV$, $GNR_P$, $GNR_{NP}$의 성능이 저하될 것이라고 예상되며, $GNR_M$의 성능은 $GNR_T$의 성능과 큰 차이가 없으므로 $GNR_T$과 $GNR_M$은 가장 효과적으로 분산을 축소할 수 있을 것으로 예상된다.

**<표 5.4> 중도절단이 50%인 경우, $Y_i = 5 + \sin(2\pi X_i) + \epsilon_i$**

| | $N = 100$ | | | $N = 200$ | | | $N = 400$ | | |
|---|---|---|---|---|---|---|---|---|---|
| **방법** | $Bias^2$ | $Var$ | $MSE$ | $Bias^2$ | $Var$ | $MSE$ | $Bias^2$ | $Var$ | $MSE$ |
| $KSV$ | 1.2987 | 3.0397 | 4.3384 | 0.8785 | 2.3342 | 3.2127 | 0.5143 | 1.3412 | 1.8555 |
| $GNR_P$ | 1.3059 | 3.2217 | 4.5276 | 0.8749 | 2.4341 | 3.3089 | 0.4810 | 1.4174 | 1.8984 |
| $GNR_{NP}$ | 0.2131 | 0.2078 | 0.4209 | 0.1935 | 0.1444 | 0.3379 | 0.1734 | 0.0720 | 0.2454 |
| $GNR_M$ | 0.1136 | 0.1707 | 0.2844 | 0.0887 | 0.1234 | 0.2121 | 0.0604 | 0.0767 | 0.1371 |
| $GNR_T$ | 0.0314 | 0.0645 | 0.0959 | 0.0253 | 0.0634 | 0.0887 | 0.0319 | 0.0440 | 0.0759 |

$KSV$ : 기존 자료변환법 $GNR_P$ : 모수적 초기 추정량 $GNR_{NP}$ : 비모수적 초기 추정량
$GNR_M$: 중위수 초기 추정량 $GNR_T$: True Mean Function

모형 $Y_i = 5 + \sin(2\pi X_i) + \epsilon_i$, 중도절단이 50%인 **<표 5.4>**의 결과를 확인해보면 $KSV$, $GNR_P$와 $GNR_{NP}$의 성능이 급격히 저하된 것을 알 수 있다. 반면, $GNR_M$의 성능은 중도절단이 증가할수록 확연히 좋아지는 걸 확인할 수 있다. 또한, 앞서 예상했듯이 가장 효율적으로 분산을 축소할 수 있는 방법은 $GNR_T$ 이며, $GNR_M$ 역시 효율적으로 분산을 축소할 수 있을 것이다.

## 5.2. 모형 $Y_i = 5 + \sin(4\pi X_i) + \epsilon_i$

**<표 5.5> 중도절단이 0%인 경우, $Y_i = 5 + \sin(4\pi X_i) + \epsilon_i$**

| | $N = 100$ | | | $N = 200$ | | | $N = 400$ | | |
|---|---|---|---|---|---|---|---|---|---|
| **방법** | $Bias^2$ | $Var$ | $MSE$ | $Bias^2$ | $Var$ | $MSE$ | $Bias^2$ | $Var$ | $MSE$ |
| $KSV$ | 0.0154 | 0.0964 | 0.1117 | 0.0074 | 0.0493 | 0.0567 | 0.0052 | 0.0245 | 0.0297 |
| $GNR_P$ | 0.0143 | 0.1026 | 0.1169 | 0.0072 | 0.0518 | 0.0590 | 0.0051 | 0.0255 | 0.0307 |
| $GNR_{NP}$ | 0.0065 | 0.0866 | 0.0931 | 0.0042 | 0.0409 | 0.0451 | 0.0042 | 0.0197 | 0.0239 |
| $GNR_M$ | 0.0086 | 0.1046 | 0.1133 | 0.0063 | 0.0507 | 0.0570 | 0.0037 | 0.0260 | 0.0297 |
| $GNR_T$ | 0.0000 | 0.0655 | 0.0655 | 0.0000 | 0.0317 | 0.0317 | 0.0001 | 0.0139 | 0.0140 |

*KSV* : 기존 자료변환법 $GNR_P$ : 모수적 초기 추정량 $GNR_{NP}$ : 비모수적 초기 추정량
$GNR_M$: 중위수 초기 추정량 $GNR_T$: True Mean Function

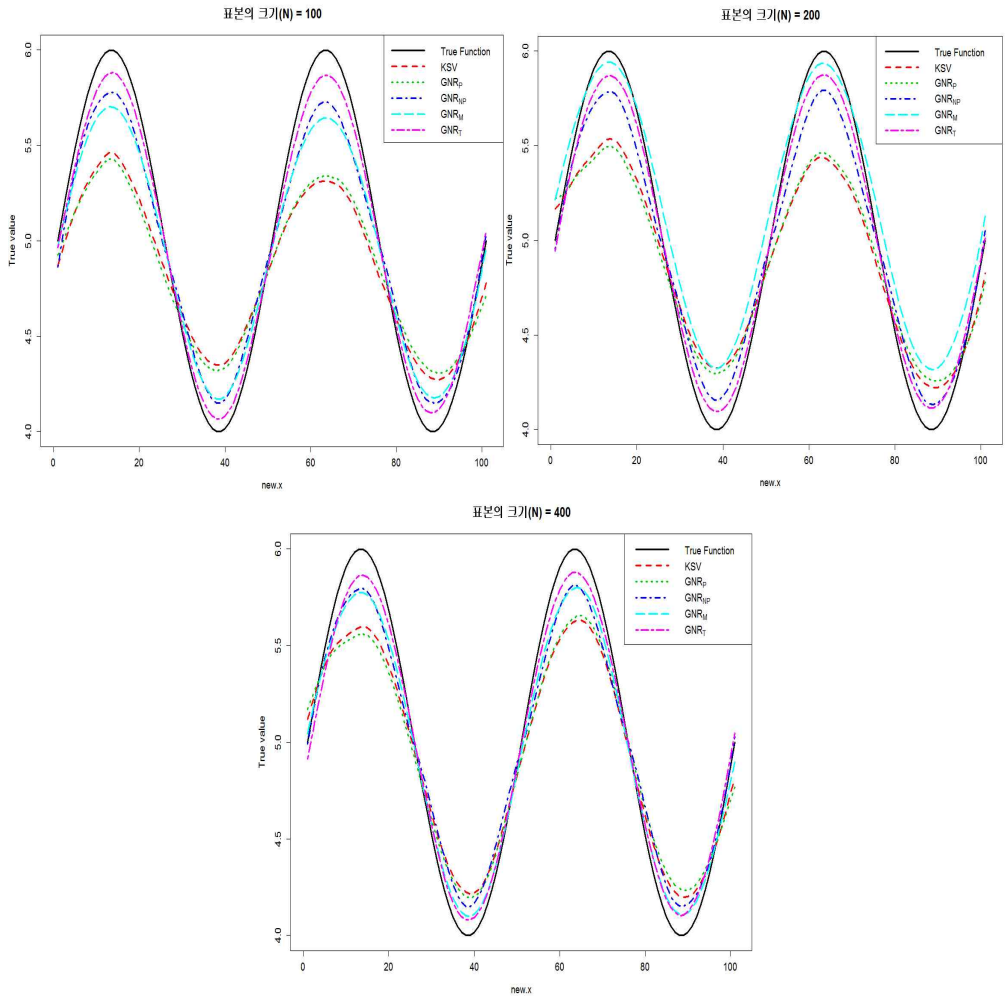
모형 $Y_i = 5 + \sin(4\pi X_i) + \epsilon_i$에서 중도절단이 없는 경우인 **<표 5.5>**의 결과이다. 결과를 살펴보면 $GNR_T$이 제안한 다른 방법보다 뛰어난 걸 확인할 수 있다. **<표 5.1>**의 결과와 비교해보면 $GNR_{NP}$의 성능이 향상된 걸 확인할 수 있으며, $GNR_P$의 성능은 저하된 것을 확인할 수 있다.

**<표 5.6> 중도절단이 10%인 경우, $Y_i = 5 + \sin(4\pi X_i) + \epsilon_i$**

| | $N = 100$ | | | $N = 200$ | | | $N = 400$ | | |
|---|---|---|---|---|---|---|---|---|---|
| **방법** | $Bias^2$ | $Var$ | $MSE$ | $Bias^2$ | $Var$ | $MSE$ | $Bias^2$ | $Var$ | $MSE$ |
| $KSV$ | 0.1301 | 0.4364 | 0.5665 | 0.0906 | 0.2776 | 0.3682 | 0.0471 | 0.1563 | 0.2035 |
| $GNR_P$ | 0.1308 | 0.4661 | 0.5969 | 0.0916 | 0.2925 | 0.3842 | 0.0503 | 0.1644 | 0.2147 |
| $GNR_{NP}$ | 0.0269 | 0.0866 | 0.1136 | 0.0223 | 0.0451 | 0.0674 | 0.0215 | 0.0204 | 0.0419 |
| $GNR_M$ | 0.0272 | 0.1189 | 0.1461 | 0.0180 | 0.0668 | 0.0848 | 0.0109 | 0.0354 | 0.0463 |
| $GNR_T$ | 0.0062 | 0.0781 | 0.0843 | 0.0064 | 0.0428 | 0.0493 | 0.0071 | 0.0224 | 0.0295 |

$KSV$ : 기존 자료변환법 $GNR_P$ : 모수적 초기 추정량 $GNR_{NP}$ : 비모수적 초기 추정량
$GNR_M$: 중위수 초기 추정량 $GNR_T$: True Mean Function

**<표 5.6>**는 중도절단이 10%인 경우, 모형 $Y_i = 5 + \sin(4\pi X_i) + \epsilon_i$ 때의 성능 비교에 대한 결과이다. 전체적으로 살펴보면 표본의 크기와 상관없이 $GNR_T$가 가장 좋은 성능을 보여준다. 또한, 모형 $Y_i = 5 + \sin(2\pi X_i) + \epsilon_i$와는 다르게 $GNR_{NP}$이 성능이 크게 향상된 걸 알 수 있다. 이는 모형 $Y_i = 5 + \sin(4\pi X_i) + \epsilon_i$가 Trend의 경향이 크므로 요약한 정보를 사용한 $GNR_M$보다는 모든 정보를 활용한 $GNR_{NP}$이 더 좋은 성능을 가진다고 할 수 있다.




**[그림 5.2] 모형** $Y_i = 5 + \sin(4\pi X_i) + \epsilon_i$, $<$중도절단$10\%>$**에서 각 방법에 대한 Mean Plot**

**[그림 5.2]**는 모형 $Y_i = 5 + \sin(4\pi X_i) + \epsilon_i$, 중도절단이 10%인 경우로, 제안한 초기 추정량 방법에 대한 Mean Plot을 그린 것이다. **[그림 5.2]**를 <표 5.6>의 결과와 비교해보면 전체적으로는 $GNR_T$의 성능이 가장 좋으며, 모형

$Y_i = 5 + \sin(2\pi X_i) + \epsilon_i$와는 다르게 표본의 크기가 커질수록 $GNR_M$의 성능보다는 $GNR_{NP}$의 성능이 좋다는 걸 확인할 수 있다.

**<표 5.7> 중도절단이 30%인 경우, $Y_i = 5 + \sin(4\pi X_i) + \epsilon_i$**

| | $N = 100$ | | | $N = 200$ | | | $N = 400$ | | |
|---|---|---|---|---|---|---|---|---|---|
| **방법** | $Bias^2$ | $Var$ | $MSE$ | $Bias^2$ | $Var$ | $MSE$ | $Bias^2$ | $Var$ | $MSE$ |
| $KSV$ | 0.3665 | 1.6866 | 2.0531 | 0.2726 | 0.8609 | 1.1334 | 0.2152 | 0.4547 | 0.6699 |
| $GNR_P$ | 0.3696 | 1.8174 | 2.1871 | 0.2749 | 0.9172 | 1.1921 | 0.2217 | 0.4884 | 0.7101 |
| $GNR_{NP}$ | 0.0787 | 0.1183 | 0.1970 | 0.0693 | 0.0667 | 0.1360 | 0.0680 | 0.0313 | 0.0993 |
| $GNR_M$ | 0.0960 | 0.1538 | 0.2497 | 0.0560 | 0.0943 | 0.1503 | 0.0395 | 0.0614 | 0.1009 |
| $GNR_T$ | 0.0134 | 0.0787 | 0.0921 | 0.0124 | 0.0481 | 0.0605 | 0.0169 | 0.0316 | 0.0485 |

$KSV$ : 기존 자료변환법 $GNR_P$ : 모수적 초기 추정량 $GNR_{NP}$ : 비모수적 초기 추정량
$GNR_M$: 중위수 초기 추정량 $GNR_T$: True Mean Function

**<표 5.7>**은 모형 $Y_i = 5 + \sin(4\pi X_i) + \epsilon_i$에서 중도절단이 30%인 경우의 결과이다. 제안한 각 방법에 대한 성능을 비교해보면 모형 $Y_i = 5 + \sin(2\pi X_i) + \epsilon_i$의 결과와 마찬가지로 $KSV$, $GNR_P$의 성능은 절단비율이 10%일 때보다 확연히 저하된 것을 확인할 수 있다. 하지만, $GNR_M$의 성능은 **<표 5.3>**의 결과와 다르게 저하된 걸 알 수 있다. 이는 일정수준의 중도절단 비율을 넘어서면 $GNR_{NP}$을 제외한 $KSV$, $GNR_P$, $GNR_M$의 성능이 저하될 것이라고 예상되며, $GNR_{NP}$의 성능은 $GNR_T$의 성능과 큰 차이가 없으므로 $GNR_T$과 $GNR_{NP}$은 가장 효과적으로 분산을 축소할 수 있을 것으로 예상된다.

**<표 5.8> 중도절단이 50%인 경우,** $Y_i = 5 + \sin(4\pi X_i) + \epsilon_i$

| | $N = 100$ | | | $N = 200$ | | | $N = 400$ | | |
|---|---|---|---|---|---|---|---|---|---|
| **방법** | $Bias^2$ | $Var$ | $MSE$ | $Bias^2$ | $Var$ | $MSE$ | $Bias^2$ | $Var$ | $MSE$ |
| $KSV$ | 1.3696 | 3.1851 | 4.5547 | 0.8551 | 1.9052 | 2.7603 | 0.5680 | 1.2172 | 1.7852 |
| $GNR_P$ | 1.3609 | 3.3788 | 4.7397 | 0.8550 | 2.0400 | 2.8950 | 0.5721 | 1.2974 | 1.8695 |
| $GNR_{NP}$ | 0.2270 | 0.2027 | 0.4298 | 0.1847 | 0.1381 | 0.3228 | 0.1597 | 0.0721 | 0.2318 |
| $GNR_M$ | 0.1572 | 0.1725 | 0.3297 | 0.1050 | 0.1361 | 0.2411 | 0.0760 | 0.0991 | 0.1751 |
| $GNR_T$ | 0.0196 | 0.0722 | 0.0918 | 0.0151 | 0.0580 | 0.0731 | 0.0204 | 0.0472 | 0.0677 |

$KSV$ : 기존 자료변환법 $GNR_P$ : 모수적 초기 추정량 $GNR_{NP}$ : 비모수적 초기 추정량
$GNR_M$: 중위수 초기 추정량 $GNR_T$: True Mean Function

모형 $Y_i = 5 + \sin(4\pi X_i) + \epsilon_i$, 중도절단이 50%인 **<표 5.8>**의 결과를 살펴보면 $GNR_{NP}$의 성능이 급격히 저하된 것을 확인할 수 있다. 반면, $GNR_M$의 성능은 중도절단 30%인 경우보다 향상된 걸 확인할 수 있다. 중도절단 50%인 경우에는 함수의 정보가 적어 요약한 정보를 활용한 $GNR_M$이 더 좋은 방법이라고 할 수 있다. 위의 예상과는 다르게 $GNR_T$과 $GNR_M$이 가장 효과적으로 분산을 축소할 수 있을 것이다.

설정한 2가지 모형의 결과를 종합해보면, 자료를 변환 하게 되면 분산의 증가로 인해 추정의 질을 저하시키는 문제가 발생한다. 이는 변환 전 위에서 제시한 방법에 의해 초기 추정량을 적절히 설정해주면 위와 같은 문제를 해결할 수 있다. 모형과 상관없이 제안한 방법 중 $GNR_T$의 성능은 표본의 크기와 절단비율

에 상관없이 가장 뛰어나며, 효율적으로 분산을 축소할 수 있을 것이라고 예상된다. $GNR_M$은 모형 $Y_i = 5 + \sin(2\pi X_i) + \epsilon_i$에서 일정수준 이상의 절단비율을 넘어서면 $GNR_T$을 제외한 방법 중 가장 좋은 성능을 보인다. 하지만 모형 $Y_i = 5 + \sin(4\pi X_i) + \epsilon_i$에서는 $GNR_M$의 성능보다는 $GNR_{NP}$이 좋은 성능을 보인다. 이는 모형 $Y_i = 5 + \sin(4\pi X_i) + \epsilon_i$기 Trend의 변화가 심하므로 $GNR_M$은 지나치게 요약한 정보를 사용하게 되며, 이로 인해 성능이 저하되고, 함수의 정보를 더 활용한 방법인 $GNR_{NP}$의 성능이 더 좋아짐을 확인할 수 있다. 반대로 모형 $Y_i = 5 + \sin(2\pi X_i) + \epsilon_i$의 경우에는 함수의 정보를 이용한 추정량에 비해 단순한 $GNR_M$이 오히려 더 좋은 성능을 가진다.

## 6. 결론

본 논문에서는 중도절단자료를 자료변환법에 위해 변환했을 때, 분산이 증가와 추정의 질이 저하하는 문제점이 발생하였으며, 이러한 문제점을 해결하고자 Guided Non-Parametric Regression방법을 활용하였다. 또한 제안한 여러 가지 방법에 대한 성능을 알아보았다. 먼저 자료변환법에 대해 설명하였으며, 자료변환법을 사용하여 변환된 자료에서 발생하는 문제점을 설명하였다. 분산 축소를 위해 여러 가지 방법을 제안하였다. 모의실험을 통해 다양한 상황에서의 제안한 $KSV$, $GNR_P$, $GNR_{NP}$, $GNR_M$, $GNR_T$을 Guided Non-Parametric Regression 방법론을 활용하여 성능 비교하였고, 비교를 위해 평균제곱오차(MSE), 분산(Variance), 편향$^2$($Bias^2$)을 계산하여 제시하였다. 모의실험 결과, 제안한 여러 방법 중 $GNR_T$의 성능은 모형, 표본의 크기, 절단비율과 상관없이 다른 방법의 성능에 비해 뛰어난 것을 확인하였다. $GNR_T$을 제외한 방법 중 모형을 고려했을 경우, 모형 $Y_i = 5 + \sin(2\pi X_i) + \epsilon_i$에서는 절단비율이 0%일 경우에 $GNR_P$의 성능은 높았지만, 일정수준의 이상의 중도절단 비율을 넘으면 $GNR_M$의 성능이 좋아지는 것을 확인하였다. 또한, 그림을 살펴봐도 일정수준의 이상의 중도절단 비율을 넘으면 $GNR_M$이 좋은 성능을 가지고 있음을 확인할 수 있다. 모형 $Y_i = 5 + \sin(4\pi X_i) + \epsilon_i$에서는 $GNR_{NP}$의 성능이 좋은 걸 확인하였다. 이는 모형 $Y_i = 5 + \sin(4\pi X_i) + \epsilon_i$가 Trend의 변화가 심하므로 요약한 정보를 활용하는 $GNR_M$보다는 함수의 정보를 더 활용하는 $GNR_{NP}$가 더 좋은 성능이 있음을 보였다.

초기 추정량($r_x$) 설정에 따라 성능이 저하될 수도 있고, 향상될 수 도 있다. 만약 자료변환에 의해 분산 증가 문제가 발생한다면 초기 추정량 설정을 적절히 해주면 해결할 수 있다. 차후에는 제안한 초기 추정량에 대한 방법 이외에도 추가적으로 다른 방법에 대한 연구가 필요하다고 생각한다.

# 참고문헌


[1] Bender, R., Augustin, T. & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. Stat. Med. 24, 1713-1723.

[2] Beran,R.(1981). Non-parametric regression with randomly censored survival data. Technical Report, Univ. California, Berkeley.

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

[4] Dabrowska, D. M. (1987). Non-0parametric regression with censored survival time data. Scand. J. Stat. 14, 181-197.

[5] Dabrowska, D. M. (1992). Variable bandwidth conditional Kaplan-Meier estimate. Scand. J. Stat. 19, 351-361.

[6] Fan, J. & Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. Ann. Statist. 20, 2008-2036.

[7] Glad, I. K. (1998). Parametrically guided non-parametric regression. Scand. J. Stat. 25, 649-668.

[8] Gozalo, P. & Linton, O. (2000). Local nonlinear least squares: using parametric information in non-parametric regression. J. Econometrics 99, 63-106.

[9] Koul, H., Susarla, V. & Van Ryzin, J. (1981). Regression analysis with randomly right censored data. Ann. Statist. 9, 1276-1288.

[10] Lai, T. L. & Ying, Z. (1991). Large-sample theory of a modified Buckley-James estimator for regression analysis with censored data. Ann.

Statist. 19, 1370-1402.

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

[12] Liang, H. & Zou, G. (2008). Improved AIC selection strategy for survival analysis. Comput. Statist. Data Anal. 52, 2538-2548.

[13] Martins-Filho, C., Mishra, S. & Ullah, A. (2008). A class of improved parametrically guided non-parametric regression estimators. Econom. Rev. 27, 542-573.

[14] Martins-Filho, C. & Yao, F. (2006). A note on the use of V and U statistics in non-parametric models of regression. Ann. Inst. Statist. Math. 58, 389-406.

[15] Mays, J. E., Birch, J. B. & Starnes, B. A. (2001). Model robust regression: combining parametric, non-parametric and semi-parametric methods. J. Non-parametric. Stat. 13, 245-277.