Summary of Censored Regression: Local Linear Approximations and Their Applications

비모수 A조: 안재형, 문혜성, 김도은, 고경민, 도아 티트 짱

December 2020

본 보고서는 Fan and Gijbels [1994]의 내용을 정리하고 해당 논문에 제시된 방법론을 적용한 모의실험 (simulation) 결과를 요약한다. 보고서는 다음과 같이 구성된다. 2장에선 방법론을 요약하여 정리하고 3장에서 해당 방법론에 대한 부연설명을 기술한다. 4장에선 모형과 관련한 점근적 결과 (asymptotic result)를 살펴보고 5장에서 모의설계 데이터 (simulated data)와 실제 데이터 (real data)에 대한 모의실험 결과를 제시한다. 모의실험에 사용된 R 코드는 부록에 첨부된다.

1 Introduction

Fan and Gijbels [1994]는 중도 절단 (censored) 데이터가 존재하는 상황에서 종속변수와 독립변수의 관계를 비모수 회귀 모형을 통해 추정하는 방법론을 제시한다. 중도 절단 데이터는 관측값에 대한 정보를 일부만 알고 있는 경우를 뜻한다. 예를 들어, 생존 분석에서 환자의 생존 시간에 대한 정보는 해당 연구가 진행된 기간에 한해서만 수집될 수 있다. 만약 일부 환자가 연구 기간내에 사망하지 않고 생존해 있다면 해당 환자에 대한 정확한 생존 시간은 알 수 없다. 이러한 환자의 생존 시간은 마지막으로 측정된 생존 시간으로 중도 절단되어 기록되기 때문에 관측값에 대한 일부 정보만 알고 있다고 표현한다. 특히 생존 분석과 같이 특정 값보다 큰 값에 대해 중도 절단된 데이터를 우측 중도 절단 데이터 (right censored data)라 한다. 본 보고서는 종속변수가 우측 중도 절단 되어 있는 상황을 상정한다. Fan and Gijbels [1994]는 이러한 중도 절단 데이터를 다루기 위한 방법으로 적절한 데이터 변환 (transformation)을 제시한다. 변환된 데이터에 대해 비모수 회귀 모형을 적합함으로써 최종 회귀 모형이 도출된다. 회귀 모형은 다른 커널 (kernel) 기반 추정법과 달리 경계 수정 (boundary modification)이 필요없는 등 좋은 성질을 갖는다는 것이 알려진 (Fan [1992]) local linear regression smoothers 모형이 사용된다. 이 때, 변환과 비모수 회귀모형 적합에 사용되는 대역폭 (bandwidth)은 데이터의 희소성 (sparsity)에 따라 자동적으로 값이 조정되는 적응적 가변 대역폭 (adaptive variable bandwidth)이 사용된다.

2 Summary of the Methodology

앞으로의 논의는 다음의 회귀모형에 기반한다.

$$Y = m(X) + \sigma(X)\epsilon$$

Where,

Y : Survival time,
X : Associated covariate,
$m(\cdot)$: Unknown regression curve,
$\sigma(\cdot)$: Conditional variance representing the possible heteroscedacity

이 때, X와 ϵ은 상호 독립이고 $E(\epsilon) = 0$, $var(\epsilon) = 1$이라 가정한다. C를 생존시간 (survival time) Y에 대해 중도 절단된 시간 (censoring time)이라고 할 때, Y와 C는 주어진 X에 대해 조건부 독립이라고 하자. 데이터셋은 $\{(X_i, Z_i, \delta_i) : i = 1, \cdots, n\} \sim (X, Z, \delta)$ 와 같이 표현된다. 이 때, $Z = min(Y, C)$이고 $\delta = I(Y \leq C)$를 나타낸다. 데이터셋은 X_i 값에 따라 오름차순 정렬되어 있다고 하자.
알려지지 않은 회귀 곡선 (unknown regression curve) $m(\cdot)$에 대한 추정은 다음의 네가지 과정에 따라 이루어진다.

1. Transformation of the data

   어떤 정수 k와 음이 아닌 가중치 함수 (non-negative weight function) K가 주어졌을 때, 관측값 $\{(X_i, Z_i, \delta_i)\}$를 다음의 식에 따라 $\{(X_i, Y_i^\star)\}$로 변환한다.

$$Y_i^\star = \delta_i Z_i + (1 - \delta_i) \frac{\Sigma_{j:Z_j > Z_i} Z_j K\left(\frac{X_i - X_j}{(X_{i+k} - X_{i-k})/2}\right)\delta_j}{\Sigma_{j:Z_j > Z_i} K\left(\frac{X_i - X_j}{(X_{i+k} - X_{i-k})/2}\right)\delta_j} \tag{1}$$

2. Application of the local linear regression technique

   변환된 데이터 $\{(X_i, Y_i^*)\}$에 대해 local linear regression smoothers 모형을 적합한다. 이 때, 평활 모수 (smoothing parameter) k는 (1)에 사용된 값과 동일하다.

$$\hat{m}(x) = \sum_{i=1}^{n} w_i(x) Y_i^\star / \sum_{i=1}^{n} w_i(x) \tag{2}$$

   with

$$w_i(x) \stackrel{\text{def}}{=} K(\frac{x - X_i}{\hat{h}_k(x)})[s_{n,2} - (x - X_i)s_{n,1}]$$

where

$$s_{n,l}=\sum_{i=1}^{n} K(\frac{x-X_i}{\hat{h}_k(x)})(x-X_i)^l,\ l=0,1,2.$$

적응적 가변 대역폭 (adaptive variable bandwidth) $\hat{h}_k$는 다음과 같이 정의된다.

$$\hat{h}_k(x) = (X_{l+k} - X_{l-k})/2 \tag{3}$$

이 때, l은 x에 가장 근접한 관측값 X_l의 인덱스를 의미한다.

3. Selection of the smoothing parameter k by cross-validation

   평활 모수 k는 교차 검증 (cross-validation)에 기반해 결정된다. 어떤 정수 k가 주어지면 (1))을 통해 $\{(X_i, Y_i^\star)\}$를 얻은 후 i번째 관측값 (X_i, Y_i^*)을 제외한 데이터셋에 대해 (2)를 적합해 $\hat{m}_{-i}(x)$를 얻는다. 추정된 모형을 이용해 다음의 CV 함수를 계산한다.

$$CV(k) = \sum_{i=1}^{n} (Y_i^\star - \hat{m}_{-i}(X_i))^2 \tag{4}$$

   $\hat{k}$은 $\{CV(k) : k = 1, ..., [(n-1)/2]\}$를 최소화하는 k값으로 채택된다.

$$\hat{k} = \underset{k}{\operatorname{argmin}}\ CV(k), \quad k = 1, ..., [(n-1)/2]$$

   이 때, $[a]$ 기호는 a의 가장 큰 정수 값을 산출한다.

4. Calculation of the local linear smoother

   추정된 $\hat{k}$을 이용해 (1)의 데이터 변환과 (2)의 회귀 모형 적합을 시행함으로써 최종 모형이 도출된다.

3 Further Explanation of the Methodology

이 장에선, 2장에서 소개한 방법론에 대한 부연 설명을 기술한다. 특히 데이터 변환 (transformation)과 local linear regression smoothers 방법론에 대한 구체적인 설명이 제시된다.

3.1 Transformation of the Data

$\phi_1(\cdot,\cdot)$과 $\phi_2(\cdot,\cdot)$를 중도 절단되지 않은 (uncensored) 관측값과 중도 절단 (censored) 관측값에 대한 변환 함수라 하자. 이에 따른 변환은 다음과 같이 나타낼 수 있다.

$$\begin{aligned} Y^\star &= \phi_1(X, Y) \quad \text{if uncensored} \\ &= \phi_2(X, C) \quad \text{if censored} \\ &= \delta\phi_1(X, Z) \ + \ (1-\delta)\phi_2(X, Z) \end{aligned} \tag{5}$$

이러한 변환을 "ideal transformation"이라 칭한다. 왜냐하면 이는 ϕ_1과 ϕ_2를 알고 있는 것을 가정하기 때문이다. 하지만 현실에선 실제 변환 함수를 알 수 없기 때문에 변환 함수에 대한 추정이 요구된다. Buckley and James [1979]는 다음과 같은 변환을 제안한다.

$$\phi_1(x, y) = y \text{ and } \phi_2(x, y) = E(Y|Y > y, X = x) \tag{6}$$

이는 $Y_0^\star = E(Y|\delta, Z, X)$로 표현될 수 있다. 이 변환은 $E(Y - Y_0^\star)^2 \leq E(Y - Y^\star)^2$의 성질을 가짐으로 실제 종속변수 (original response)와 가장 근접하다는 점에서 "best restoration"이라 칭할 수 있다. 하지만 이 방법은 조건부 기대값 계산을 위해 모형에 대한 강한 가정이 요구된다. 또한 알려지지 않은 회귀 모형에 의존하기 때문에 평활 모수 k를 구하기 위해 필요되는 계산 비용이 매우 크다. 이를 극복하기 위해 Fan and Gijbles [1994]는 다음의 추정된 변환 (estimated transformation)을 제안한다.

$$Y_i^\star = \frac{\Sigma_{j:Z_j > Z_i} Z_j K\left(\frac{X_i - X_j}{(X_{i+k} - X_{i-k})/2}\right)\delta_j}{\Sigma_{j:Z_j > Z_i} K\left(\frac{X_i - X_j}{(X_{i+k} - X_{i-k})/2}\right)\delta_j}$$

이 변환은 조건부 기댓값을 비모수적 방법인 Nadaraya-Watson 추정량 (Nadaraya [1964]; Watson [1964])의 형태로 추정한다. 즉, 중도 절단 관측값 (X_i, Z_i)에서 X_i의 이웃한 점들 중 중도 절단되지 않은 관측값의 Z값이 Z_i보다 큰 경우들을 음이 아닌 함수 K를 가중치로 가중평균하여 Z_i값을 대체한다. 이에 따라 중도 절단되지 않은 관측값은 변환되지 않고 (i.e., $Y_i^* = Z_i$, if $\delta_i = 1$) 중도 절단된 관측값에 대해서만 변환이 이루어져 변환된 데이터셋 $\{(X_i, Y_i^*) : i = 1, \cdots, n\}$을 얻는다.

3.2 Local Linear Regression Smoothers

이 장에선, 변환된 데이터셋 $\{(X_i, Y_i^*) : i = 1, \cdots, n\}$에 대한 회귀모형 $m(\cdot)$을 추정한다. 고정된 포인트 x의 이웃한 점 z에 대해 알려지지 않은 함수 (unknown function)를 근사한 식이 다음의 선형식을 따른다고 가정하자.

$$m(z) \approx m(x) + m^{'}(x)(z - x) \stackrel{\text{def}}{=} a + b(z - x) \tag{7}$$

이에 따라 $m(x)$를 찾는 문제는 상수항 a의 값을 찾는 문제와 동일해진다 ($m(x) = E(m(z)|z = x) = a$). 이 때, x의 이웃한 점과 그 가중치는 평활 모수 k에 기반한 대역폭 h_k과 커널 (kernel)함수 K에 의해 결정된다고 하자. $m(x)$의 추정은 a와 b에 대해 다음 식을 최소화함으로써 이루어진다.

$$\sum_{i=1}^{n} (Y_i^* - a - b(X_i - x))^2 K\left(\frac{x - X_i}{\hat{h}_k(x)}\right) \tag{8}$$

앞선 가정에 따라 $m(x)$의 추정량은 $\hat{a}$을 구하는 것과 같음을 알 수 있다. 이에 대한 추정은 다음과 같이 이루어진다.

$$\hat{m}(x) = \hat{a} = \sum_{i=1}^{n} w_i(x) Y_i^\star / \sum_{i=1}^{n} w_i(x) \tag{9}$$

with

$$w_i(x) \stackrel{\text{def}}{=} K(\frac{x - X_i}{\hat{h}_k(x)})[s_{n,2} - (x - X_i)s_{n,1}]$$

where

$$s_{n,l} = \sum_{i=1}^{n} K(\frac{x - X_i}{\hat{h}_k(x)})(x - X_i)^l, \ l = 0, 1, 2.$$

이 때, $\hat{h}_k(x)$는 (3)의 정의를 따르며 (4)를 최소화하는 값으로 결정된다.
위의 추정법은 Stone [1977]과 Cleveland [1979]에 의해 처음 소개되었다. Fan [1992]는 이 방법이 대부분의 비모수 회귀 모형과 달리 경계 수정 (boundary modification)이 필요 없을 뿐 아니라 random 이나 fixed design그리고 clustered나 uniform design과 같은 다양한 design에 적용가능한 좋은 속성이 있음을 보인다.

4 Asymptotic Result

이 장에선, 적응적 가변 대역폭 (adaptive variable bandwidth) (3)과 추정된 변환 (estimated transformation) (1)에 기반한 local linear regression smoothers 모형 (2)의 점근적 성질을 살펴본다.

Adaptive variable bandwidth $\hat{h}_k(x)$의 점근적 성질

다음의 정리는 가변 대역폭 (variable bandwidth) $\hat{h}_k$가 $k/(nf_X(x))$와 근사하게 행동한다는 것을 보인다. 이 때, $f_X(\cdot)$은 X의 주변확률밀도함수 (marginal probability density function)를 의미한다.

Theorem 4.1 *Suppose that $f_X(\cdot)$ is positive and continuous on a compact interval $[a, b]$ and that $k_n \to \infty$ such that $k_n/n \to 0$. Then $\hat{h}_{kn}(x)[k_n/(nf_X(x)](1+o_p(1))$ uniformly in $x \in [a, b]$.*

이를 통해 $\hat{h}_k(x)$가 $f_X(\cdot)$와 반비례 관계라는 것을 알 수 있다. 즉, X의 밀도가 낮은 경우 $\hat{h}_k(x)$의 값은 크고 X의 밀도가 높은 경우 $\hat{h}_k(x)$의 값은 작음으로 데이터의 희소성 (sparsity)에 따라 그 값이 조정됨을 알 수 있다.

Ideal transformation에 기반한 local linear regression smoothers 모형의 점근적 성질

Let,

- $\hat{m}(x; \phi_1, \phi_2)$ be the regression estimator (2) based on the ideal transformation (5),

- $K(\cdot)$ be a compactly supported probability density function with mean 0,
- $c_K = \int_{-\infty}^{\infty} v^2 K(v) dv$
- $d_K = \int_{-\infty}^{\infty} K^2(v) dv$

이 때, K는 uniformly Lipschitz continuous하다고 가정한다. 이에 따라 다음의 정리를 얻는다.

Theorem 4.2 *Suppose that $f_X(\cdot), m^{''}(\cdot)$ and $\sigma^*(\cdot)$ are bounded functions, continuous at the point x, and that $f_X(x) > 0$. If $k_n \to \infty$ such that $k_n/n \to 0$, then, conditionally on the covariates $\{X_1, \cdots, X_n\}$,*

$$\sqrt{k_n}(\hat{m}(x;\phi_1,\phi_2) - m(x) - m''(x)c_K h_k^2(x)/2) \to N(0, d_K \sigma^{*2}(x)), \tag{10}$$

이 때, $h_k(x) = k_n/(n f_X(x))$ 이다.

이를 통해 ideal transformation된 데이터셋에 대한 local linear regression smoothers 모형은 점근적 정규성 (asymptotic normality)를 갖는다는 것을 알 수 있다 ($m''(x)c_K h_k^2(x)/2$ 만큼의 편의 존재).

Estimated transformation에 기반한 local linear regression 모형의 점근적 성질

Let,

- $\hat{m}(x;\hat{\phi}_1,\hat{\phi}_2)$ be the regression estimator (2) based on the estimated transformation (1),
- Where $\hat{\phi}_1, \hat{\phi}_2$ estimate ϕ_1 and ϕ_2.

$\hat{m}(x;\phi_1,\phi_2)$의 일치성을 보이기 위한 필요조건은 $\hat{\phi}_1(t,z)$와 $\hat{\phi}_2(t,z)$가 x의 인접 이웃 t와 특정 구간안에 존재하는 z에 대해 uniformly consistent 하다는 것이다. 이를 식으로 표현하면 다음과 같다.

$$\beta_n(x) = \max_{j=1,2} \left\{ \sup_{z\in(0,\tau_n),\, t\in(x\pm\tau)} |\hat{\phi}_j(t,z) - \phi_j(t,z)| \right\} = o_p(1) \tag{11}$$

with $\tau_n > 0$ and $\tau > 0$.
$\hat{\phi}_j (j = 1, 2)$의 정의를 확장해서 나타내면 다음과 같다.

$$\begin{aligned} \hat{\phi}_j(t,z) &= \hat{\phi}_j(t,z), \quad \text{if } z \le \tau_n \\ &= z \qquad\qquad \text{elsewhere} \end{aligned} \tag{12}$$

이 정의를 바탕으로 꼬리 (tail)에서 일치 추정량이 가져야할 조건을 다음과 같이 나타낼 수 있다.

$$\kappa_n(x) = \max_{j=1,2} \left\{ \sup_{t\in(x\pm\tau)} E\left(1_{[Z>\tau_n]} |Z - \phi_j(t,Z)| \,\middle|\, X = t \right) \right\} = o(1) \tag{13}$$

이에 기반해 다음의 정리를 도출할 수 있다.

Theorem 4.3 *Assume that the conditions of Theorem 5.1 hold. Then*

$$\hat{m}(x;\hat{\phi}_1,\hat{\phi}_2) - \hat{m}(x;\phi_1,\phi_2) = O_p(\beta_n(x)+\kappa_n(x)) \tag{14}$$

provided that K is uniformly Lipschitz continuous and has a compact support.

Theorem 4.2와 4.3의 결과, 어떤 일치 추정량 $\hat{\phi}_1$과 $\hat{\phi}_2$에 대해 $\hat{m}(x;\hat{\phi}_1,\hat{\phi}_2)$가 $m(x)$의 일치 추정량임을 알 수 있다.

5 Simulation Result

이 장에선 Fan and Gijbels [1994]에 제시된 모의설계 데이터 (simulated daat)와 실제 데이터 (real data)에 대한 모의 실험을 재현하고 그 결과를 해설한다.

5.1 Simulation study

Data
다음의 환경을 통해 200개의 관측값이 생성되었다고 하자.

$$Y_i = 4.5 - 64X_i^2(1-X_i)^2 - 16(X_i-0.5)^2 + 0.25\epsilon_i$$

$$X_i \overset{\text{iid}}{\sim} \text{Uniform}[0,1],\ \epsilon_i \overset{\text{iid}}{\sim} N(0,1), X_i \perp\!\!\!\perp \epsilon_i$$

Where,

$$(C_i|X_i=x) \overset{\text{independent}}{\sim} exponential(c(x))$$

이 때, $c(x)$는 censoring time의 조건부 평균으로, 다음과 같이 설정된다.

$$\begin{aligned} c(x) &= 3(1.25-|4x-1|), \quad \text{if } \ 0 \le x \le .5, \\ &= 3(1.25-|4x-3|), \quad \text{if } \ .5 < x \le 1 \end{aligned}$$

이에 따라 200개의 관측값 중 대략 40%의 관측값이 중도 절단 (censored) 되어있는 상황이다.

Result
모의실험 결과는 Figure 1을 통해 확인할 수 있다. Figure1-(c)의 Cross-validation Curve는 $k=5$일 때 $CV(k)$값이 최소가 된다. 이에 따라 $\hat{k}=5$가 최적의 평활 모수로 채택됐다.
Figure1-(d)는 Figure1-(b)의 변환되지 않은 데이터 (observed simulated data)에 (2)를 적합한 (노란색) 회귀선과 (1)의 변환을 거쳐 (2)를 적합한 (빨강색) 회귀선 그리고 실제 (검정색) 회귀선을 비교하여 나타낸다. 그 결과 변환하지 않은 데이터를 사용한 경우에 변환한 데이터를 사용한 경우보다 실제 회귀선을 과소 추정 (under estimate) 하는 것을 확인할 수 있다. 이는 식 (1)을 통한 데이터 변환의 필요성을 강조한다.

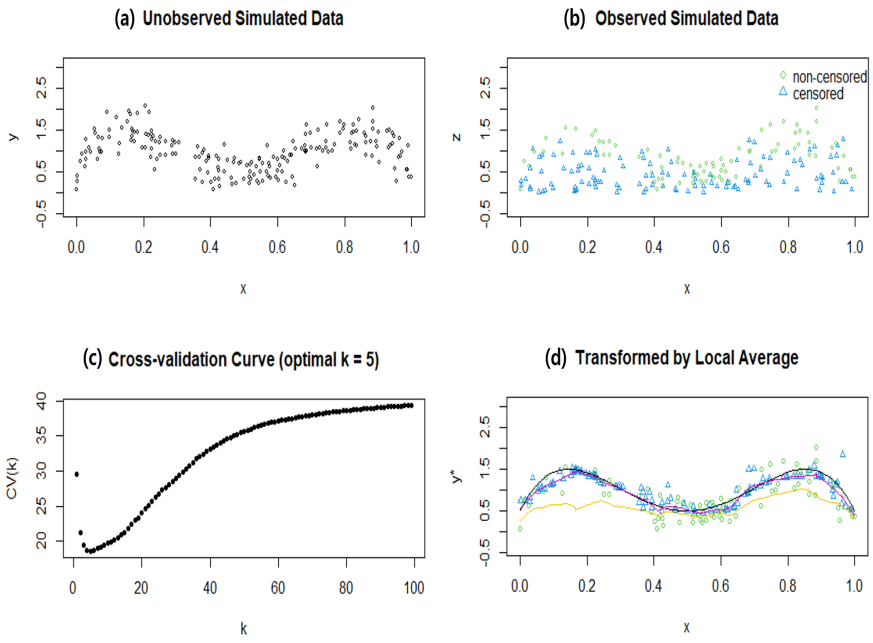

Figure 1: Simulated Data Set. The triangle indicates the censored observations, and the uncensored observations are presented by circle. The individual panels show (a)unobserved simulated data; (b)observed simulated data; (c) cross-validation curve; (d) data transformed by local average using an optimal $\hat{k} = 5$ and local linear smoothers: black curve-the true regression function, red curve-the smoother based on the transformed data, yellow curve-the smoother based on the observed data in (b)

5.2 Real Data : Stanford Heart Transplant data

Data

R의 survival 패키지에 있는 스탠포드 심장 이식 프로그램 데이터를 이용하여 생존시간을 반응변수 (response), 나이를 설명변수 (covariate)로 설정하여 논문에 제시된 방법을 적용해 보았다. 1967년부터 1980년 2월 사이에 이 프로그램에 참여하여 심장 이식을 받은 환자들 중 1980년 2월 이후 생존한 환자들을 censored 데이터로 간주했다. 157개의 데이터 중 55개가 censored 데이터이다.

Result

Figure2를 통해 모의실험 결과를 확인할 수 있다. Figure2-(b)의 Cross-validation curve는 $k = 22$일 때 $CV(k)$값이 최소가 된다. 따라서 $\hat{k} = 22$가 최적의 평활 모수로 채택됐다.

Figure2-(c)는 Fan and Gijbles (1994)에 제시된 생존시간과 나이의 제안된 관계 (suggested relationship)를 이용해 계산한 (검정색) 회귀선($y = 3.009 - 0.093(x -$

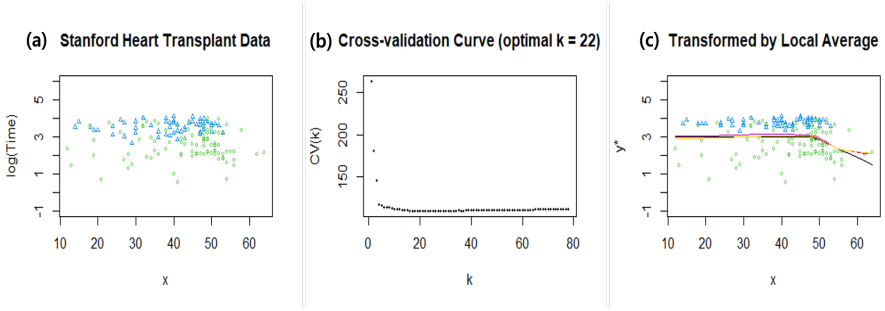

Figure 2: Stanford Heart Transplant Data Set. The triangle indicates the censored observations, and the uncensored observations are presented by circle. The individual panels show (a)log-survival time plotted against age; (b)cross validation curve; (c) data transformed by local average using an optimal $\hat{k} = 22$ and local linear smoothers: black curve-the suggested relationship, red curve-the smoother based on the transformed data, yellow curve-the smoother based on the observed data in (a)

$48)_+$)과 (1)을 통해 변환된 데이터에 (2)를 적합하여 추정한 (빨강색) 회귀선 그리고 변환하지 않은 데이터(Figure2-(a))에 (2)를 적합하여 추정한 (노랑색) 회귀선을 비교하여 나타낸다. Simulation study에서는 변환하지 않은 데이터를 사용한 경우에 변환한 데이터를 사용한 경우보다 실제 회귀선을 과소 추정하는 것이 확인되었으나 스탠포드 심장 이식 데이터에서는 데이터 변환 여부에 따른 차이가 크지 않았다.

References

Jonathan Buckley and Ian James. Linear regression with censored data. *Biometrika*, 66(3):429–436, 1979.

William S Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836, 1979.

Jianqing Fan. Design-adaptive nonparametric regression. *Journal of the American statistical Association*, 87(420):998–1004, 1992.

Jianqing Fan and Irene Gijbels. Censored regression: local linear approximations and their applications. *Journal of the American Statistical Association*, 89 (426):560–570, 1994.

Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

Charles J Stone. Consistent nonparametric regression. *The annals of statistics*, pages 595–620, 1977.

Geoffrey S Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.