Journal of the Korean Data & Information Science Society
2013, **24**(2), 211–222

http://dx.doi.org/10.7465/jkdi.2013.24.2.211
한국데이터정보과학회지

조건부 분위수의 중도절단을 고려한 비모수적 추정[†]

김은영[1] · 최혜미[2]

[12]전북대학교 통계학과

접수 2013년 1월 9일, 수정 2013년 2월 1일, 게재확정 2013년 2월 18일

요 약

중도절단된 자료가 있을 경우 조건부 분위수함수를 비모수적으로 추정하는 문제에 대하여 다루고 있다. 역함수에 근거한 방법인 Yu와 Jones (1998)에 의해 제안된 중복커널기법 추정량과 Lee 등 (2006)의 국소로지스틱기법 추정량을 중도절단된 자료가 있는 경우로 수정하여 새롭게 제안하고, 이들을 기존의 Koenker와 Bassett (1978)의 점검함수에 근거한 커널평활 추정량들과 모의실험을 통해 비교해 보았다. 모의실험을 통하여 역함수에 근거한 추정량들은 조건부 분포가 대칭인 모형에서, 점검함수기법 추정량들은 한쪽으로 치우친 분포인 경우에 조건부 분위수를 대체로 더 잘 추정하고 있음을 알 수 있었다.

주요용어: 국소선형기법, 조건부 분위수, 중도절단.

1. 서론

대부분의 회귀모형을 통한 연구는 공변량 $X = x$가 주어져 있을 때 반응변수 Y의 조건부 분포 $F(y|x)$의 평균 $E(Y|X = x) = m(x)$에 관하여 이루어지고 있지만, 조건부 중앙값 등과 같은 Y의 조건부 분위수

$$q(x;p) = F^{-1}(p|x) = \inf\{y : F(y|x) \geq p\}, \quad 0 < p < 1$$

의 추정에 대한 관심 또한 높아지고 있으며, 다양한 분야에서 이용되고 있다. 예를 들어 병원에서는 환자의 상태가 정상범위에 속하는지 여부를 판단하기 위하여 참조차트 (reference chart)를 많이 사용하고 있는데, 이는 환자의 기초정보를 공변량으로 하는 조건부 분위수에 해당하는 것이다. 또한, 임금과 소득을 연구하는 표준 분석 방법 도구로 조건부 분위수를 사용하여, 서로 다른 소비 그룹에 속하는 개인별 차이를 인식하거나, 집단의 구성원간의 임금분포를 통해 적절한 세금을 징수하고 사회정책을 세우고 있다. 이외에도 금융분야에서 VaR 추정을 하는 경우나 환경·기후 현상모형을 세우는 경우에 조건부 극단 분위수 (conditional extreme quantile)를 이용하고 있다.

이와 같이 여러 분야에 응용되고 있는 조건부 분위수의 다양한 추정량들이 제안되어왔다 (Yu 등, 2003; Koenker, 2005). 이들 추정량은 크게 조건부 분포함수 $F(\cdot|x)$를 먼저 추정한 후 역함수를 취하는 방법과 조건부 분위수 $q(x;p)$가

$$\underset{a}{\operatorname{argmin}}\, E\{\rho_p(Y - a)|X = x\}$$

[†] 이 논문은 2012년도 정부 (교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임 (No. 2012R1A1A3010532).
[1] (561-756) 전북 전주시 덕진구 덕진동 1가 664-14, 전북대학교 통계학과, 석사과정.
[2] 교신저자: (561-756) 전북 전주시 덕진구 덕진동 1가 664-14, 전북대학교 통계학과 (응용통계연구소), 부교수. E-mail: hchoi@jbnu.ac.kr

와 동일함을 이용하여 Koenker와 Bassett (1978)가 제안한 추정법으로 나눌 수 있다. 여기에서 $\rho_p(z) = z(p - I(z < 0))$이고 점검함수 ("check" function)라고 불리며, $I(z < 0)$는 $z < 0$이면 1, 아니면 0의 값을 가지는 지시함수를 나타내고 있다. 이 두 가지 접근법을 편의상 각각 "역함수에 근거한 추정"과 "점검함수에 근거한 추정"이라고 하겠다.

본 논문에서는 중도절단자료가 있는 경우의 조건부 분위수의 비모수적 추정량에 대하여 논의하고 있다. 중도절단은 생존자료에서 많이 발생하며, 생존시간의 조건부 분위수모형은 가속수명모형이나 Cox 비례위험모형의 대안으로 사용되어지기도 한다. 최근 들어 고정 또는 임의중도절단을 고려한 조건부 분위수의 (준)모수적 추정량들이 많이 제안되고 있다 (Chernozhukov와 Hong, 2002; Bang과 Tsiatis, 2002; Portnoy, 2003). 그러나 본 논문에서는 분포 가정에 민감하지 않은 비모수적 추정량 중에서 특히 커널 평활의 국소선형기법 (local linear technique)을 이용한 추정량을 고려하고 있다. 이는 다른 대표적인 비모수적 함수 추정 기법인 스플라인 (spline)을 이용한 방법에 비해서 점근적 이론 성질이 잘 규명되어 있어서 적합도에 대한 평가가 더 용이하여 더욱 각광을 받고 있다. 또한 R에서는 기본적으로 제공되는 함수 외에도 다양한 패키지들에서 국소선형회귀기법을 적용할 수 있게 하고 있다.

점검함수에 국소선형기법을 적용한 추정량으로는 Gannoun 등 (2007)과 Ghouch와 Van Keilegom (2009)을 들 수 있다. 이 두 추정량은 중도절단이 없이 완전히 관측된 경우에 서로 일치한다. 다른 한편으로 역함수기법으로 유도된 추정량으로는 Yu와 Jones (1998)와 Lee 등 (2006)의 추정량을 들 수 있는데, 전자는 중복커널 국소선형법 ("double-kernel" local linear approach)을 후자는 국소로지스틱 회귀방법 (local logistic regression)을 적용하고 있다. 그러나 이들은 완전하게 관측된 자료의 경우에 제안된 추정량이어서, 본 논문에서는 누적한계추정량 (product-limit)을 사용한 가중치 부여를 통하여 중도절단을 고려하도록 수정한 추정량을 제안하였다. 여기에서 누적한계 (product-limit) 추정량은 Kaplan과 Meier에 의해 제안된 생존함수의 대표적인 비모수적 추정량으로 다음과 같이 정의된다.

$$\hat{S}(t) = \prod_{i:y_{(i)} \le t} \left(1 - \frac{1}{n-i+1}\right)^{\delta_{(i)}} \tag{1.1}$$

이때 $y_{(i)}$는 관측된 시간 y_i들의 순서통계량이고, $\delta_{(i)}$는 $y_{(i)}$에 대응하는 중도절단 여부를 나타내는 δ(중도절단자료이면 0, 아니면 1)이다. 만약 추정량 $\hat{S}(t)$에서 $\delta_{(i)}$대신에 $1-\delta_{(i)}$를 대입하면, $\hat{S}(t)$는 중도절단시간의 생존함수 추정량이 된다. 누적한계추정량은 효율적인 추정방법으로 비모수적 최대가능도 추정량으로 알려져 있으며, 생존분석에서 널리 응용되고 있다.

본 논문의 구성은 다음과 같다. 2절에서는 역함수와 점검함수에 근거한 조건부 분위수 비모수적 추정량들의 소개와 중도절단을 고려하여 수정된 추정량을 제안하고, 3절에서 이들 추정량을 모의실험을 통해 비교하고 있다. 끝으로 4절에서는 결론 및 앞으로의 연구방향에 관하여 논의한다.

2. 조건부 분위수 비모수적 추정량들

임의중도절단자료를 확률변수를 사용하여 나타내면 다음과 같다. 생존시간 Y와 중도절단시간 C에 대하여 $Z = \min(Y, C)$와 $\delta = I(Y < C)$라고 하고 공변량을 X라고 할 때, $(Z_1, X_1, \delta_1), \ldots, (Z_n, X_n, \delta_n)$는 서로 독립인 임의 중도절단이 있는 관측변수들을 나타낸다. 본 논문에서는 Park과 Kim (2011)에서와 같이 우측임의중도절단 (right random censoring)만을 고려하고 있다. 임의중도절단의 경우 일반적으로 생존시간 T와 중도절단시간 C가 서로 독립이라는 가정 한다. 또한 중도절단시간 C와 공변량 X도 서로 독립이라고 가정하였다. 이는 중도절단이 공변량과 관계없이 일어나는 경우를 반영하고 있다. 공변량 $X = x$가 주어져 있을 때 생존시간 Y의 조건부 분포함수를 $F(\cdot|x)$로 나타내고, 공변량 X의 확률밀도함수를 $f_X(\cdot)$, 중도 절단시간의 분포함수를 $G(\cdot)$로 나타내었다.

2.1. 역함수에 근거한 추정량

2.1.1. 중복커널방법

중복커널방법은 $I(Y_i \le y)$의 평활된 변형을 사용한 국소선형회귀 기법으로 조건부 확률밀도함수 추정을 위해 Fan 등 (1996)이 제안한 방법이다. 이를 Yu와 Jones (1998)가 조건부 분포함수 추정을 위해 적용하고, Kim 등 (2010)이 중도절단을 고려한 추정량으로 확장하였다.

본 논문에서는 국소선형회귀기법 적용을 위해 사용한 커널을 K, 반응변수 Y의 평활에 사용할 커널을 L이라 하고, 각 커널의 띠폭 (bandwidth)을 h와 l로 나타냈으며, 두 커널 모두 대칭인 확률밀도함수라고 가정한다. 여기에서 커널 L의 분포함수를 Ω로 나타내었다. 즉

$$\Omega\left(\frac{y-Y}{l}\right) = \int_{-\infty}^{t} \frac{1}{l} L\left(\frac{Y-u}{l}\right) du$$

이다. 여기에서 띠폭 l이 작아지면

$$E\left[\Omega\left(\frac{y-Y}{l}\right)\middle| X=x\right] \approx F(y|x) = E[I(Y \le y)|X=x]$$

임을 주목하고, Yu와 Jones (1998)는 조건부 분포함수 추정량으로

$$\underset{a,b}{\operatorname{argmin}} \sum_{i=1}^{n} \left(\Omega\left(\frac{y-Y_i}{l}\right) - a - b(X_i - x)\right)^2 K\left(\frac{x-X_i}{h}\right) \tag{2.1}$$

을 만족하는 $\hat{a}$를 제안하였다. 그러나 본 논문에서 고려하는 중도절단자료의 경우에는 Y_i 대신에 $Z_i = \min(Y_i, C_i)$와 $\delta_i = I(Y_i < C_i)$만을 관측하게 되어, 식 (2.1)의 수정이 필요하다. 중도절단시간 C가 (X,Y)와 독립이므로, C의 생존함수를 $\bar{G}(\cdot) = 1 - G(\cdot)$로 확률밀도함수를 $g(\cdot)$로 나타낼 때

$$\begin{aligned} E\left[\frac{\delta}{\bar{G}(Z)}\Omega\left(\frac{y-Z}{l}\right)\middle| X=x\right] &= \int\int_{t<c} \Omega\left(\frac{y-t}{l}\right)\frac{1}{\bar{G}(t)} f(t|x)g(c)dcdt \\ &= \int \Omega\left(\frac{y-t}{l}\right) f(t|x)dt \\ &= E\left[\Omega\left(\frac{y-Y}{l}\right)\middle| X=x\right] \end{aligned}$$

라는 관계가 성립한다. Kim 등 (2010)에서는 조건부 분포함수의 중복커널추정량으로 이 관계를 이용하여, (2.1)을 수정하고 $\bar{G}$을 누적한계추정량으로 대체한

$$\underset{a,b}{\operatorname{argmin}} \sum_{i=1}^{n} \left(\frac{\delta_i}{\hat{\bar{G}}(Z_i)}\Omega\left(\frac{y-Z_i}{l}\right) - a - b(X_i - x)\right)^2 K\left(\frac{x-X_i}{h}\right)$$

을 만족하는 명확한 해 $(\hat{a}, \hat{b})$를 구하고, 조건부 분포함수의 추정량 $\hat{F}_K(y|x) \equiv \hat{a}$, 즉

$$\frac{\sum_{i=1}^{n} w_{h,i}\frac{\delta_i}{\hat{\bar{G}}(Z_i)}\Omega\left(\frac{y-Z_i}{l}\right)}{\sum_{i=1}^{n} w_{h,i}}$$

을 제안하였다. 여기에서 $s_j(x) = \sum (x-X_i)^j L((x-X_i)/l),\ j = 0,1,2$ 라고 할 때

$$w_{h,i} = L\left(\frac{x-X_i}{l}\right)\{s_2(x) - (x-X_i)s_1(x)\}$$

을 나타낸다. 참고로 $\hat{b}$는 Y의 x에서의 조건부 분포함수의 도함수값 추정을 위해 사용될 수 있다. Kim 등 (2010)은 위험률함수 추정문제를 다룬 연구로 조건부 분위수 추정에 대한 논의는 하지 않았다. 본 논문에서는 추정된 조건부 분포함수에 대하여

$$\hat{F}_K(\hat{q}_{YJ}(x;p)) = p$$

을 만족하는 역함수기법에 근거한 조건부 분위수 추정량 $\hat{q}_{YJ}(x:p)$을 제안하고 있다.

2.1.2. 국소로지스틱회귀법

공변량 $X = x$일 때 $I(Y \le y)$는 성공확률이 $F(y|x)$인 베르누이 분포를 따르므로, Lee 등 (2006)은 $I(Y_i \le y)$의 조건부 로그가능도함수가

$$\sum_{i=1}^{n} \left\{ I(Y_i \le y) \log\left(\frac{F(y|X_i)}{1-F(y|X_i)}\right) + \log(1-F(y|X_i)) \right\} \tag{2.2}$$

임을 이용하여, 다음과 같이 국소조건부 로그가능도함수에 커널 가중치를 부여한

$$\sum_{i=1}^{n} \left\{ I(Y_i \le y)(a+b(X_i-x)) - \log(1+e^{a+b(X_i-x)}) \right\} K\left(\frac{x-X_i}{h}\right) \tag{2.3}$$

을 최소로 하는 조건부 분포함수 추정량을 제안하였다. 식 (2.3)은 $\text{logit}(F(y|u))$를 $a+b(u-x)$로 국소적으로 선형화하여, 식 (2.2)의 $F(y|X_i)$에

$$\frac{\exp(a+b(X_i-x))}{1+\exp(a+b(X_i-x))}$$

을 대입하여 얻어진 것이다. 중도절단자료를 포함하는 경우 적절한 수정이 필요한데, 본 논문에서는

$$\begin{aligned} E\left[\frac{\delta}{\bar{G}(Z)} I(Z \le y) \middle| X = x\right] &= \int\int_{t<c} \frac{I(t \le y)}{\bar{G}(t)} f(t|x) g(c) dc dt \\ &= \int I(t \le y) f(t|x) dt \\ &= E\{I(Y \le y)|X = x\} = F(y|x) \end{aligned} \tag{2.4}$$

임을 주목하고, (2.3)의 $I(Y_i \le y)$에 $I(Z_i \le y)\frac{\delta_i}{\hat{\bar{G}}(Z_i)}$를 대입한

$$\sum_{i=1}^{n} \left\{ I(Z_i \le y) \frac{\delta_i}{\hat{\bar{G}}(Z_i)} (a+b(X_i-x)) - \log(1+e^{a+b(X_i-x)}) \right\} K\left(\frac{x-X_i}{h}\right) \tag{2.5}$$

을 최소로 하는 $\hat{a}$를 이용하여 조건부 분포함수 추정량

$$\hat{F}_L(y|x) = \frac{\exp(\hat{a})}{1+\exp(\hat{a})}$$

을 구하고, 중도절단을 고려한 조건부 분위수 추정량을

$$\hat{q}_{LGL}(x;p) = \hat{F}_L^{-1}(p|x)$$

제안하였다.

국소로지스틱기법을 적용하여 추정된 조건부 분포함수가 항상 0과 1사이에 있다고 보장할 수 있다는 장점이 있다. 중복커널방법과 국소로지스틱회귀법을 적용한 조건부 분포함수 추정량은 모든 y에 대해서 단조성을 가지고 있지는 않아서 역함수를 취할 때 문제가 발생할 수 있으나, $p > .5$인 경우는 $\hat{q}(x:p) = \sup\{y : \hat{F}(y|x) \le p\}$로 $p \le .5$인 경우는 $\hat{q}(x:p) = \inf\{y : \hat{F}(y|x) \ge p\}$로 추정하여 문제를 해결할 수 있었다. 중복평활기법은 평활을 이중으로 하게 되어 계산량이 많은 단점이 있으나, 단일 커널방법보다 추정량의 더 평활이 잘 되고, 중도 절단이 없는 경우 이론적으로도 적분제곱오차의 평균(MISE; mean integrated squared error)이 더 좋다고 알려져 있다 (Yu와 Jones, 1998).

2.2. 점검함수에 근거한 추정량

Gannoun 등 (2007)과 Ghouch와 Van Keilegom (2009)은 중도 절단이 있는 자료의 경우에 점검함수 $\rho_p(z) = z(p - I(z < 0))$에 국소선형기법을 적용하여 조건부 분위수 추정량을 제안하였다. 이 두 추정량은 서로 다른 방법으로 중도절단을 고려하고 있지만, 중도절단이 없는 경우에는 두 추정량은

$$\begin{aligned}(\hat{a}, \hat{b}) &= \underset{a,b}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_p(Y_i - a - b(X_i - x)) K\left(\frac{x - X_i}{h}\right) \\ &= \underset{a,b}{\operatorname{argmin}} \sum_{i=1}^{n} [Y_i - a - b(X_i - x)]\,[p - I(Y_i < a + b(X_i - x))]\, K\left(\frac{x - X_i}{h}\right) \end{aligned} \tag{2.6}$$

을 만족하는 $\hat{a}$로 동일하다. 이때 $\hat{b}$는 x에서의 조건부 분위수 도함수값 추정량에 해당한다.

관계식 (2.4)을 이용하여 제안된 Ghouch와 Van Keilegom (2009)의 추정식에 본 논문의 중도절단시간이 공변량과 독립이라는 가정을 반영하면

$$\sum_{i=1}^{n} [Z_i - a - b(X_i - x)] \left[p - \frac{\delta_i}{\hat{\bar{G}}(Z_i)} I(Z_i < a + b(X_i - x))\right] K\left(\frac{x - X_i}{h}\right) \tag{2.7}$$

과 같다. 이를 최소로 하는 $\hat{a}$를 조건부 분위수 추정량 $\hat{q}_{CH1}(x;p)$라고 나타내기로 한다.

Gannoun 등 (2007)은 Cai (2003)가 생존시간의 조건부 평균함수 $m(x) = E(Y|X = x)$를 추정하기 위해 도입한 가중치 $w_i = \frac{\delta_i}{n\hat{\bar{G}}(Z_i)}$를 사용하여 식 (2.6)을

$$(\hat{a}, \hat{b}) = \underset{a,b}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_p(Z_i - a - b(X_i - x)) \frac{\delta_i}{\hat{\bar{G}}(Z_i)} K\left(\frac{x - X_i}{h}\right) \tag{2.8}$$

으로 수정하였다. 이를 만족하는 $\hat{a}$를 조건부 분위수 추정량 $\hat{q}_{CH2}(x;p)$라고 나타내기로 한다.

점검함수기법은 조건부 분포함수의 추정단계 없이 바로 조건부 분위수를 구할 수 있으나, 점검함수 ρ_p가 미분가능하지 않아 정확한 해를 거의 찾을 수 없으며, 수치적으로도 불안정한 현상이 발생할 수 있다.

3. 모의실험을 통한 비교

본 논문에서 제안한 중도절단자료가 있는 경우 역함수에 근거한 중복커널추정량 $\hat{q}_{YJ}$와 국소로지스틱 추정량 $\hat{q}_{LGL}$의 소표본에서의 성질을 점검함수에 근거한 Ghouch와 Van Keilegom (2009)의 $\hat{q}_{CH1}$과 Gannoun 등 (2007)의 $\hat{q}_{CH2}$와 모의실험을 통하여 비교해 보았다. Yu와 Jones (1998)에서 중도절단이 없을 경우 $\hat{q}_{YJ}$와 $\hat{q}_{CH1}$ 또는 $\hat{q}_{CH2}$ (참고: 중도절단자료가 없을 때 $\hat{q}_{CH1}$=$\hat{q}_{CH2}$)의 비교를 위해 사용한 네 개의 모형을 본 논문에서도 그대로 사용하였다.

1. 거의 직선에 가까운 분위수, 이분산 모형 (heteroscedastic):

$$Y = \sin(0.75X) + 1 + 0.3\sqrt{(\sin(0.75X) + 1)}E$$
$$X \sim N(0, 0.0625)), \quad E \sim N(0, 1)$$

2. 부드러운 곡선형 분위수, 등분산 (homoscedastic) 모형:

$$Y = 2.5 + \sin(2X) + 2\exp(-16X^2) + 0.5E$$
$$X \sim N(0, 1), \quad E \sim N(0, 1)$$

3. 단순한 분위수, 한쪽으로 치우친 (skew) 모형:

$$Y = 2 + 2\cos(X) + \exp(-4X^2) + E$$
$$X \sim N(0, 1), \quad E \sim Exp(1)$$

4. 단순한 분위수, 이분산 모형 (heteroscedastic):

$$Y = 2 + X + \exp(-X)(E - \log 2.6)$$
$$X \sim U[0, 5], \quad E \sim Exp(1)$$

위의 모든 모형에서 변수 X는 E와 독립이다. 중도절단시간 C는 각 모형마다 주어진 E의 분포와 같은 종류의 분포를 따르며 (즉 모형1과 2: 정규분포, 모형 3과 4: 지수분포), 평균과 분산을 조절하여 중도절단비율이 0% (C0), 약 20% (C2)와 약 30% (C3)인 모형에서 표본을 생성하였다. 표본 크기 $n = 100$으로 하고, 각 모형마다 세 분위수(p=0.1, p=0.5, p=0.9)를 계산하는 실험을 100번 반복 하였다. 모의실험 결과는 ISE (integrated squared error)에 의해 평가하였다.

$$ISE_j = \int_{\mathcal{I}} [\hat{q}^{(j)}(x; p) - q(x; p)]^2 f_X(x) dx, \quad j = 1, \ldots, 100$$

이때 $q(x; p)$는 모형의 조건부 p번째 분위수이며, $\hat{q}^{(j)}(x; p)$는 j번째 표본에서 얻은 추정치이다. 그리고, 적분 범위 $\mathcal{I}$는 생성된 자료의 거의 모두를 포함하여 계산하도록 모형마다 서로 다른 범위: (1) [-.5,.5], (2) [-2,2], (3) [-2,2], (4) [0,5]를 사용하였다.

커널 K로는 정규확률밀도함수, L로는 $U(-1, 1)$의 밀도함수를 사용하였다. 커널 K로 정규확률밀도함수 대신에 Epanechnikov 커널을 사용하여 모의실험을 수행해 추정량의 성질을 비교해 보기도 했는데, 정규확률밀도함수 커널과 비슷하였다. 커널평활의 경우 띠폭 h의 선택이 매우 중요한데 (Huh, 2012), 본 논문에서는 각 표본마다 띠폭을 변화시켜가면서 구한 조건부 분위수 추정량의 ISE가 가장 작은 띠폭을 선택하였다. 중복커널추정량의 경우 반응변수의 평활을 위해 사용된 커널 L의 띠폭(l)도 변화시켜가며 최적의 띠폭 (h, l)을 찾아야 하겠지만, 계산 량이 너무 많아지는 것을 피하고, Lee 등 (2006)은 l이 h보다 빠르게 0으로 수렴하는 것을 가정하고 있으므로, 각 h별로 $l = h/10, h/5, h/2$ 세 가지 경우에 추정치를 계산하고 최소로 하는 ISE를 구하여 비교하였다.

추정량 $\hat{q}_{LGL}$, $\hat{q}_{CH1}$과 $\hat{q}_{CH2}$은 $\hat{q}_{YJ}$와 다르게 추정식의 최소값 $\hat{a}$를 명확하게 찾을 수 없으므로 `R` 소프트웨어의 `optim`이라는 함수를 사용하여 수치해석적으로 구하였다. 함수 `optim`은 일반적인 제약조건이 있는 경우에는 적당하지 않으나, 다양한 알고리즘으로 비선형 함수의 최적의 해를 찾을 수 있어서 통계학에서 최대 가능도 계산 등에 많이 사용되고 있다. 각 모형별로 얻어진 100개의 ISE의 중앙값을 Table 3.1로 정리하였다. 그리고 Figure 3.1-3.4와 같이 모형 별로 상자그림을 그려 보았는데, 그림에서 YJ는

Table 3.1 Median ISEs of $\hat{q}_{YJ}$, $\hat{q}_{LGL}$, $\hat{q}_{CH1}$ and $\hat{q}_{CH2}$ for each of Models 1-4

Model	Censoring rate	p	$\hat{q}_{YJ}$	$\hat{q}_{LGL}$	$\hat{q}_{CH1}$	$\hat{q}_{CH2}$
Model 1	C0	0.1	0.0024	0.0033	0.0034	0.0034
		0.5	0.0012	0.0025	0.0016	0.0016
		0.9	0.0038	0.0035	0.0022	0.0022
	C2	0.1	0.0027	0.0033	0.0033	0.0044
		0.5	0.0018	0.0031	0.0027	0.0053
		0.9	0.0071	0.0067	0.0099	0.0079
	C3	0.1	0.0035	0.0041	0.0036	0.0077
		0.5	0.0022	0.0039	0.0027	0.0094
		0.9	0.0092	0.0106	0.0265	0.0209
Model 2	C0	0.1	0.1669	0.1075	0.1285	0.1285
		0.5	0.0581	0.0616	0.0636	0.0636
		0.9	0.1371	0.0724	0.0765	0.0765
	C2	0.1	0.1729	0.1059	0.1314	0.1947
		0.5	0.0662	0.0671	0.0787	0.1133
		0.9	0.2483	0.1610	0.7120	0.1500
	C3	0.1	0.1792	0.1160	0.1359	0.1924
		0.5	0.0773	0.0801	0.0894	0.1086
		0.9	0.3839	0.2261	0.8072	0.1364
Model 3	C0	0.1	0.0348	0.0263	0.0155	0.0154
		0.5	0.0413	0.0452	0.0408	0.0408
		0.9	0.1679	0.2242	0.1727	0.1726
	C2	0.1	0.0385	0.0306	0.0191	0.0243
		0.5	0.0605	0.0664	0.0505	0.0508
		0.9	0.4194	0.4796	0.3897	0.2789
	C3	0.1	0.0445	0.0347	0.0209	0.0376
		0.5	0.0776	0.0784	0.0545	0.0750
		0.9	0.6804	0.6796	0.7909	0.4308
Model 4	C0	0.1	0.0605	0.0053	0.0005	0.0005
		0.5	0.0065	0.0087	0.0007	0.0007
		0.9	0.0763	0.0222	0.0100	0.0100
	C2	0.1	0.0606	0.0083	0.0005	0.0007
		0.5	0.0099	0.0130	0.0007	0.0008
		0.9	0.3206	0.0838	0.0261	0.0137
	C3	0.1	0.0601	0.0103	0.0008	0.0011
		0.5	0.0161	0.0190	0.0010	0.0011
		0.9	0.7235	0.1547	0.0338	0.0263

$\hat{q}_{YJ}$, LGL은 $\hat{q}_{LGL}$, CH2는 $\hat{q}_{CH2}$를 나타내고 있다. $\hat{q}_{CH1}$는 수치적으로 불안정하여 이상치 (중도절단자료의 경우 각 모형별로 10^{10}보다 큰 값들이 10%이상)들로 인해 상자그림를 그리지 않았다. 점검함수에 직접 중도절단자료의 영향을 반영하는 것 ($\hat{q}_{CH1}$)은 원래 미분 불가능한 함수인 점검함수를 수치적인 해 구하기에 더욱 어려운 함수로 만들어서 커널 가중치에 반영하는 것 ($\hat{q}_{CH2}$)보다 추정 효율을 떨어뜨림을 짐작할 수 있다. 모의실험 결과를 요약하면 다음과 같다.

(i) 분위수별로 볼 때 $p = 0.5$인 경우가 가장 잘 추정되었고, $p = 0.9$인 경우 가장 잘 추정되지 않았다. 그 이유는 모형에서 우측임의중도절단만을 고려한 영향으로 보인다.

(ii) 같은 모형 내에서 중도절단비율이 증가할 수록 추정이 잘 안 됨을 알 수 있다. 중도절단이 있는 경우에 $p = 0.9$는 네 개의 추정량 모두 추정의 정도가 상당히 떨어지며 특히 $\hat{q}_{CH1}$는 심각하게 좋지 않음을 볼 수 있다.

(iii) 역함수에 근거한 추정량은 조건부 분포함수가 대칭인 모형 1과 2에서 이 더 좋은 성질을 보이고 있다. 특히 모형 1에서는 $\hat{q}_{YJ}$, 모형 2에서는 $\hat{q}_{LGL}$가 좋은 추정을 보이고 있으나, 그 차이는 크지 않고 한 쪽으로 치우친 모형 3과 4에서는 $\hat{q}_{LGL}$의 ISE 중앙값이 $\hat{q}_{YJ}$보다 대체로 더 작음을 볼 수 있다.

(iv) 점검함수에 근거한 추정량 $\hat{q}_{CH2}$는 네 모형 모두에서 좋은 추정을 보이고 있다. 특히 조건부 분포함수가 한 쪽으로 치우친 모형 3과 4에서 특히 더 좋은 성질을 보이고 있으며, $p = 0.1, 0.5$인 경우 $\hat{q}_{CH1}$과 $\hat{q}_{CH2}$의 ISE 중앙값만을 비교하면 거의 비슷하게 잘 추정하고 있으나, $\hat{q}_{CH1}$는 수치적으로 너무 불안정함을 볼 수 있었다.

(v) 잘 추정되고 있지 않는 $p = 0.9$ 분위수 경우를 제외한 $p = 0.1$과 $p = 0.5$ 경우, 모형 1과 2는 $\hat{q}_{LGL}$가 모형 3과 4는 $\hat{q}_{CH2}$가 가장 좋은 성질을 보이고 있다.

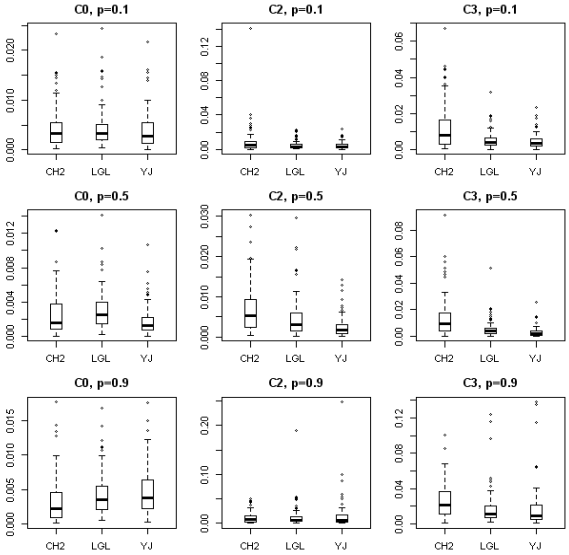

Figure 3.1 The boxplots of ISEs for Model 1

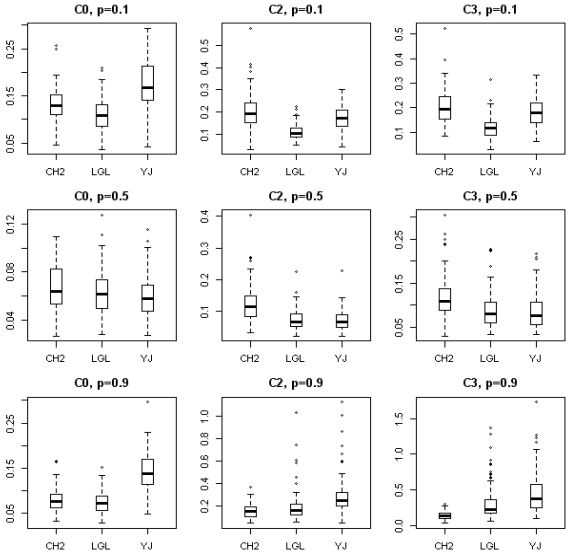

Figure 3.2 The boxplots of ISEs for Model 2

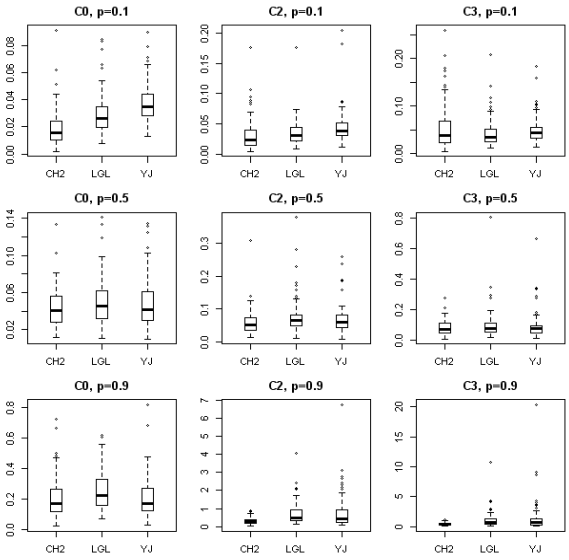

Figure 3.3 The boxplots of ISEs for Model 3

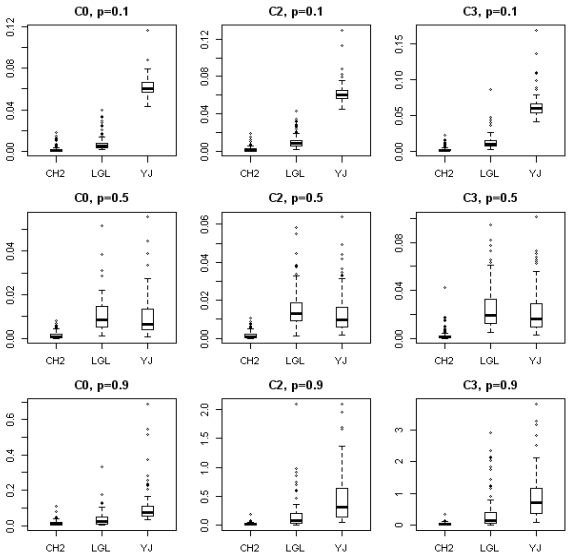

Figure 3.4 The boxplots of ISEs for Model 4

4. 결론

중도절단자료의 경우 공변량이 주어져 있을 때 반응변수의 조건부 분위수함수 $q(x;p), 0 < p < 1$ 비모수적으로 추정하는 방법에 관심을 가지고, 조건부 분포함수를 추정한 후 역함수를 취한 추정량들의 수

정된 형태를 제안하고 기존의 점검함수에 근거한 추정량과 모의실험을 통해 소표본적 성질을 비교하였다. 모의실험 결과를 정리한 Table 3.1에서 볼 수 있듯이 중앙값 측면으로 볼 때는 대칭인 분포 모형인 모형 1과 2에서는 역함수에 근거한 추정이 한 쪽으로 치우친 분포인 모형 3과 4에서는 점검함수에 근거한 추정이 더 좋은 성질을 보이고 있다. $\hat{q}_{YJ}$는 $\hat{F}_K$의 역함수로 커널가중치를 곱한 제곱함수의 최소값에 해당하여 조건부 분포가 지수분포인 모형 3과 4의 경우보다 정규분포인 경우 모형 1과 2의 경우에서 더 잘 추정하고 있는 것으로 보인다. 점검함수 $\rho_p(x) = x(p - I(x < 0))$은 제곱함수에 비하여 큰 값의 영향에 더 강건한 (robust)함수이므로 $\rho_p(\cdot)$에 근거한 추정량이 모형 3과 4에서 $\hat{q}_{YJ}$보다 더 좋은 성질을 보임을 짐작할 수 있다. 그런데 ISE의 산포도까지 고려해보면, 추정이 잘 이루어지지 않은 $p = 0.9$ 분위수를 제외하고 논의할 때 역함수에 근거한 국소로지스틱 추정량과 점검함수에 근거한 Gannoun 등 (2007)의 추정량이 조건 분위수 추정을 위해 더 추천할 만하다. 득히 네 추정량 중 $p = 0.9$과 C3에서 가장 잘 추정하고 있는 점검함수에 근거한 $\hat{q}_{CH2}$이 중도절단의 영향이 가장 작은 것으로 보인다.

본 논문에서는 R의 기본 함수인 `optim`를 이용하여 $\hat{q}_{CH1}$를 구하였는데, 수치적으로 매우 불안정한 면이 있었다. 다른 추정량들과 보다 공정한 비교를 위해서는 $\hat{q}_{CH1}$만을 위한 계산 알고리즘 개발이 선행되어야 할 것이다. 모의실험을 통하여 네 가지 추정량의 소표본적 성질을 비교해 보았는데, 이들의 대표본적 성질을 이론적으로 밝히고 비교해 보아야 할 것이며, 모의실험에서 네 추정량 모두 중도절단이 있는 경우에 $p = 0.9$ 분위수 추정의 정도가 모형 1을 제외하고는 심각하게 떨어지고 있는 점을 주목하여 그 원인분석과 적절한 대안을 생각해 보아야 할 것이다. 그리고 제안된 추정방법을 실제 자료에 적용시켜 분석해 볼 필요도 있다. 이때 최적의 띠폭은 교차타당성 (cross-validation) 방법으로 선택할 수 있을 것이다. 커널평활을 이용한 비모수적 추정에서 띠폭은 매우 중요한 역할을 하고 있는데, 중도절단자료가 없는 경우에는 Yu와 Jones (1998)과 Lee 등 (2006)에서 띠폭 선택 방법에 대한 논의가 있지만, 중도절단된 자료가 있는 경우에는 이에 대한 연구가 미미하다. 따라서, 본 논문에서 고려한 추정량의 최적의 띠폭 결정방법에 대한 이론적인 연구 또한 이루어져야 할 것이다. 또한 추정의 편리성을 위해 공변량과 중도절단 변수가 독립이라고 가정하였지만, 실제로는 이 가정을 만족하기가 쉽지 않기 때문에 서로 의존하는 경우를 고려한 추정량을 제시하여야 할 것이다.

참고문헌

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

Cai, Z. (2003). Wighted local linear approach to censored nonparametric regression. In *Recent Advances and Trends in Nonparametric Statistics*, edited by M. G. Akritas and D. M. Politis, Elsevier, 217–231.

Chernozhukov, V. and Hong, H. (2002). Three-step censored quantile regression and extramarital affairs. *Journal of the American Statistical Association*, **97**, 872–882.

Fan, J., Yao, Q. and Tong, H. (1996). Estimation of conditional densities and sensitivity measures. *Biometrika*, **83**, 189–206.

Gannoun, A., Saracco, J. and Yu, K. (2007). Comparison of kernel estimators of conditional distribution and quantile regression under censoring. *Statistical Modelling*, **7**, 329–344.

Ghouch, A. E. and Keilegom, I. V. (2009). Local linear quantile regression with dependent censored data. *Statistica Sinica*, **19**, 1621–1640.

Huh, J. (2012). Bandwidth selection for discontinuity point estimation in density. *Journal of the Korean Data & Information Science Society*, **23**, 79–87.

Kim, C., Oh, M., Yang, S. and Choi, H. (2010). A local linear estimation of conditional hazard function in censored data. *Journal of the Korean Statistical Society*, **39**, 347–355.

Koenker, R. and Bassett, G. S. (1978). Regression quantiles. *Econometrica*, **46**, 33–50.

Koenker, R. (2005). *Quantile regression*, Economic Society Monographs 38, Cambridge University Press, Cambridge.

Lee, Y. K., Lee, E. R. and Park, B. U. (2006). Conditional quantile estimation by local logical regression. *Nonparametric Statistics*, **18**, 357–373.

Park, H. and Kim, J. S. (2011). An estimation of the treatment effect for the right censored data. *Journal of the Korean & Information Science Society*, **22**, 537–547.

Portnoy, S. (2003). Censored regression quantiles. *Journal of the American Statistical Association*, **98**, 1001–1012.

Susarla, V., Tsai, W. Y. and Van Ryzin, J. (1984). A Buckley-James type estimator for the mean with censored data. *Biometrika*, **71**, 624–625.

Yu, K. and Jones, M. C. (1998). Local linear quantile regression. *Journal of the American Statistical Association*, **93**, 228–237.

Yu, K., Lu, Z. and Stander, J. (2003). Quantile regression: Applications and current research areas. *Journal of the Royal Statistical Society B*, **52**, 331–350.

Journal of the Korean Data & Information Science Society
2013, **24**(2), 211–222

http://dx.doi.org/10.7465/jkdi.2013.24.2.211
한국데이터정보과학회지

Nonparametric estimation of conditional quantile with censored data[†]

Eun-Young Kim[1] · Hyemi Choi[2]

[12]Department of Statistics, Chonbuk National University

Received 9 January 2013, revised 1 February 2013, accepted 18 February 2013

Abstract

We consider the problem of nonparametrically estimating the conditional quantile function from censored data and propose new estimators here. They are based on local logistic regression technique of Lee *et al.* (2006) and "double-kernel" technique of Yu and Jones (1998) respectively, which are modified versions under random censoring. We compare those with two existing estimators based on a local linear fits using the check function approach. The comparison is done by a simulation study.

Keywords: Censoring, conditional quantile, local linear fit.

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 2012R1A1A3010532).

[1] Graduate student, Department of Statistics, Chonbuk National University, Jeonbuk 561-756, Korea.
[2] Corresponding author: Associate professor, Department of Statistics (Institute of Applied Statistics), Chonbuk National University, Jeonbuk 561-756, Korea. E-mail: hchoi@jbnu.ac.kr