The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.7.1361
29(7), 1361–1371

Penalized quantile regression tree

Jaeoh Kim[a] · HyungJun Cho[a] · Sungwan Bang[b,1]

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

(Received September 2, 2016; Revised October 27, 2016; Accepted October 31, 2016)

Abstract

Quantile regression provides a variety of useful statistical information to examine how covariates influence the conditional quantile functions of a response variable. However, traditional quantile regression (which assume a linear model) is not appropriate when the relationship between the response and the covariates is a nonlinear. It is also necessary to conduct variable selection for high dimensional data or strongly correlated covariates. In this paper, we propose a penalized quantile regression tree model. The split rule of the proposed method is based on residual analysis, which has a negligible bias to select a split variable and reasonable computational cost. A simulation study and real data analysis are presented to demonstrate the satisfactory performance and usefulness of the proposed method.

Keywords: decision tree, penalized regression, quantile regression

1. 서론

설명변수의 함수로서 반응변수의 조건부 평균(conditional mean)을 추정하는 최소제곱법(ordinary least squares; OLS)은 여전히 통계학의 중심에 있다. 그러나 설명변수가 반응변수에 대한 분포의 형태(shape)나 척도(scale) 및 위치(location)에 어떻게 영향을 미치는지에 대하여 보다 정교한 요구가 지속되고 있다. 이러한 요구에 대해 Koenker와 Bassett (1978)은 반응변수의 조건부 분포에 대하여 많은 정보를 제공하는 분위수 회귀모형(quantile regression; QR)을 제안하였으며, 이러한 분위수 회귀모형은 유용성과 강건성을 바탕으로 경제(economics), 생물통계(biostatistics) 및 생존분석(survival analysis) 등 다양한 분야에서 활용되고 있다 (Koenker, 2005).

일반적인 분위수 회귀모형은 선형(linear)형태를 가정하며 때로는 모형의 유연성을 위하여 이러한 선형의 가정을 완화할 필요가 있다. 비선형 분위수 회귀모형의 추정을 위해 다양한 비모수적(nonparametric) 추정법들이 제안되었다. Yu와 Jones (1998)과 Hallin 등 (2009)이 제안한 국소 선형분위수 회귀(local linear quantile regression)방법, Koenker와 Mizera (2004)의 분위수 평활 스플라인(quantile smoothing splines)을 활용한 방법, Li 등 (2007)과 Liu와 Wu (2011)의 커널함수(kernel function)를

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (NRF-2015R1C1A1A02036473) for S. Bang and by the Ministry of Education (NRF-2015R1D1A1A09058602) for H. Cho.
[1]Corresponding author: Department of Mathematics, Korea Military Academy, 574, Hwarang-ro, Nowon-gu, Seoul 01805, Korea. E-mail: wan1365@gmail.com

적용한 방법 등이 있다. 특히 Chaudhuri와 Loh (2002)는 나무구조 모형(tree-structured model)에 분위수 회귀모형을 적용한 분위수 회귀나무모형(quantile regression tree; QRT) 방법을 제안하였다.

QRT는 나무모형이 갖는 우수한 해석(interpretation)과 높은 예측(prediction)의 장점을 고스란히 갖는다. 또한 Chaudhuri와 Loh (2002)는 QRT에서 각 노드별로 적합하는 조각별 다항식(piecewise polynomial) 분위수 회귀모형의 점근적(asymptotic) 성질을 보여 나무구조 모형의 우수성을 입증하였다. 그러나 QRT는 모형구축 단계에서 매우 중요한 적합변수의 선택(fitting variable selection)이 이루어지지 않는 단점이 있다. 특히 적합변수의 선택은 변수의 수가 많은 고차원 자료(high dimensional data)나 설명변수들간 상관관계(correlation)가 높은 자료의 경우에서 그 중요성이 더욱 증대된다. 전통적인 최량부분집합(best subset)을 이용한 변수선택 방법은 변수의 수가 증가함에 따라 계산상 어려움이 있을 수 있고, 불연속성(discreteness)으로 인하여 극단적인 변수선택을 할 가능성이 높다 (Breiman, 1995; Fan과 Li, 2001). 단계적 변수선택 방법(stepwise variable selection)은 국소최적해(local optimal solution)를 선택할 수 있는 문제, 높은 변동성(variability)과 변수선택 단계에서 확률적 오차(stochastic error)를 무시하는 문제를 갖는다 (Shen과 Ye, 2002). 이러한 문제에 대해 벌점화 회귀(penalized regression)모형은 변수선택과 회귀계수의 축소추정을 동시에 수행하여 좋은 해결방법이 된다. L_1 벌칙항(penalty term)으로 알려진 least absolute shrinkage and selection operator(LASSO) (Tibshirani, 1996)와 비볼록 벌점(nonconvex penalty)에 기반한 smoothly clipped absolute deviation(SCAD) (Fan과 Li, 2001), minimax concave penalty(MCP) (Zhang, 2010) 외에 다양한 벌점화 방법이 개발되고 있다. 분위수 회귀모형에 대한 벌점화 방법은 Koenker (2004)의 LASSO 형태 벌칙항 이후 다양한 연구가 있었다. 특히 분위수 회귀모형에 SCAD와 adaptive-LASSO 벌칙항을 적용한 방법론은 oracle property를 만족하는 것으로 알려져 있다 (Wu와 Liu, 2009).

본 연구에서는 분위수 회귀모형의 선형가정을 완화하고 보다 정교한 해석과 우수한 예측을 기대하기 위해 벌점화 분위수 회귀나무모형(penalized quantile regression tree; PQRT)을 제안한다. PQRT는 높은 예측력과 설명변수간 관계 규명의 통계적 학습(statistical learning) 측면에서 기존 방법보다 많은 부분 충족할 수 있다. 이는 나무모형의 각 노드 t에서 유의한 설명변수를 식별하여 적합하기 때문에 가능하다. 본 논문에서는 PQRT의 벌칙항으로 분위수 회귀모형에서 Oracle Porperty를 만족하는 SCAD 벌칙항을 사용하였다.

본 논문의 구성은 다음과 같다. 제 2절에서 분위수 회귀모형(QR)과 분위수 회귀나무모형(QRT)에 대해 소개하고, 제 3절에서 QRT의 제한사항을 보완한 PQRT를 각 노드별 적합식과 나무모형의 분할 및 정지규칙을 중심으로 설명한다. 제 4절에서 모의 실험과 실증 예제를 보이며 제 5절에서 결론을 제시한다. 본 연구에서 QRT 및 PQRT는 모두 R 3.3.1을 이용하여 구현하였다.

2. 분위수 회귀나무모형(QRT) 소개

2.1. 분위수 회귀모형

p차원 설명변수 $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$와 1차원 반응변수 $y_i \in R$로 이루어진 표본의 크기가 n인 자료 $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$가 주어졌다고 하자. 반응변수 y에 대한 $100\tau\%$ 조건부 분위수 함수(conditional quantile function of $y|\boldsymbol{x}$) $q_\tau(y|\boldsymbol{x})$는

$$P(Y \leq q_\tau(\boldsymbol{X})|\boldsymbol{X} = \boldsymbol{x}) = \tau, \quad \text{for } 0 < \tau < 1 \tag{2.1}$$

와 같이 정의된다. 이때 선형 분위수 함수 $q_\tau(\boldsymbol{x}) = \mathbf{x}^T\boldsymbol{\beta}_\tau$는 체크 손실함수(check loss function)

$\rho_\tau(u) = u(\tau - I(u < 0))$를 이용하여

$$\min_{\beta_\tau} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_\tau) \tag{2.2}$$

을 통해 추정되며 (Koenker과 Bassett, 1978), 여기서 $\mathbf{x} = (1, \boldsymbol{x}^T)^T$이며 $\boldsymbol{\beta}_\tau = (\beta_{0,\tau}, \beta_{1,\tau}, \ldots, \beta_{p,\tau})^T$는 $100\tau\%$ 분위수 함수의 회귀계수 벡터이다. 분위수 회귀모형은 절대 손실함수(absolute loss function)를 일반화한 체크 손실함수를 사용하므로 이상점(outlier)이 존재하거나 꼬리가 두꺼운 오차항의 경우에도 추정의 강건성과 효율성이 좋은 것으로 알려져 있다. 그러나 선형 분위수 회귀모형은 반응변수와 설명변수가 비선형의 관계를 갖는 자료의 분석에는 사용이 매우 제한되므로 비모수적 방법이 필요하다.

2.2. 분위수 회귀나무모형

Chaudhuri와 Loh (2002)가 제안한 분위수 회귀나무모형은 비선형자료에 대한 효과적인 해석과 우수한 예측 능력을 동시에 갖춘 방법으로 평가된다. 식 (2.1)은 자연스럽게 임의의 노드(arbitrary node) t에서의 조건부 분위수 함수 $q_\tau(\boldsymbol{x}, t)$

$$q_\tau(\boldsymbol{x}, t) = \mathbf{x}^T \boldsymbol{\beta}_\tau(t) = \beta_{0,\tau}(t) + x_1\beta_{1,\tau}(t) + \cdots + x_p\beta_{p,\tau}(t), \quad \text{for } t = 1, 2, \ldots, \widetilde{T} \tag{2.3}$$

로 확장될 수 있으며, 여기서 $\boldsymbol{\beta}_\tau(t)$는 노드 t에서의 회귀계수 벡터이고 $\widetilde{T}$은 나무모형에서 최종노드(terminal node)의 총 개수를 나타낸다. 특정 노드 t에서 비선형적 관계식에 의해 2개의 선형 분위수 함수로 결합되어 있어 분할변수(split variable) Z_s와 분할점 C를 이용하여 노드 t의 자료를 이질적 부분집합 $t_L(Z_s \le C)$와 $t_R(Z_s > C)$로 분할할 수 있다면 노드 t에서의 비선형 함수는 t_L과 t_R에서 각각 선형 분위수 함수로 추정된다. 노드 t에서 잔차(residual)는

$$e_{i,\tau}(t) = y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_\tau(t), \quad \text{for } t = 1, 2, \ldots, \widetilde{T}_n \tag{2.4}$$

와 같이 계산되고, 노드 t에서 불순도(impurity) 함수 $R(t)$는

$$R_\tau(t) = \sum_{i \in t} \rho_\tau(e_{i,\tau}(t)), \quad \text{for } t = 1, 2, \ldots, \widetilde{T}_n \tag{2.5}$$

와 같이 정의된다. 식 (2.4)와 (2.5)를 통해 불순도가 작을 수록 모형의 적합이 더 좋을 것임을 직관적으로 알 수 있다. 일반적으로 회귀나무모형은 가능한 모든 분할점과 분할집합에 대해 불순도의 감소량 $\Delta(t) = R(t) - [R(t_L) + R(t_R)]$를 모두 계산하여 이 중 가장 큰 $\Delta(t)$에 해당하는 분할변수와 분할점(또는 분할집합)을 선택한다. 이러한 회귀나무모형의 알고리즘에는 Breiman 등 (1984)의 classification and regression tree(CART), Quinlan (1993)의 C4.5 등이 있으나 두 알고리즘 모두 계산시간이 과도하게 소요되고 분할변수 선택편향(selection bias)의 문제가 있다 (Loh, 2002; Hothorn 등, 2006; Eo와 Cho, 2014). 특히 이 문제는 Breiman 등 (1984)의 2.5절과 11.8절에서도 분할 가능점이 많은 경우 분할변수 선택의 편향이 발생한다는 점과 "end cuts preference", "middle cut preference"로 자세히 다루바 있다. 반면 Generalized, Unbiased, Interaction Detection and Estimation(GUIDE) 알고리즘 (Loh, 2002, 2009)은 잔차(residual)를 분석하여 분할변수를 먼저 선정하고 선정된 분할변수에 대해 분할점(또는 분할집합)을 결정하는 방법으로 계산시간 및 분할변수 선택편향의 문제를 극복한 방법으로 평가된다. 이러한 이유로 Chaudhuri와 Loh (2002)는 GUIDE 알고리즘을 QRT에 적용하였다.

3. 벌점화 분위수 회귀나무모형(PQRT)

3.1. 적합식

회귀나무모형의 노드 t에서 해석력과 정확도를 향상하기 위해 불필요한 잡음(noise) 변수를 제거하고 유의한 적합변수(fitting variable)만을 선택하는 문제는 중요한 사안이다 (Kim과 Xing, 2012; Chang, 2014). 이를 위해 Kim 등 (2007)은 전통적인 회귀나무모형의 노드 t에서 단계적 변수선택 방법을 적용하여 소수의 적합변수만을 가지는 나무모형을 구축하는 방법을 제안하였다. 그러나 적합변수간 상관관계가 높거나 고차원 문제의 경우 이 방법은 부적합할 수 있다. 또한 Kim과 Xing (2012)은 CART 알고리즘에 대해 LASSO 벌칙항을 추가한 모형을 제안하였으나, CART 알고리즘이 갖는 분할변수 선택편향의 문제와 계산시간이 과도하게 소요될 수 있는 문제를 내재한다.

따라서 본 연구에서는 GUIDE 알고리즘을 기반으로 하는 QRT의 적합식에 벌칙항을 추가하여 축소추정을 하는 모형 (3.1)

$$\min_{\beta_\tau(t)} \sum_{i\in t} \rho_\tau \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_\tau(t) \right) + n \sum_{j=1}^{p} p_\lambda \left(|\beta_{j,\tau}| \right) \tag{3.1}$$

을 나무모형의 각 노드 t의 적합식으로 제안한다. 분위수 회귀모형의 벌점화 추정법에는 LASSO 형태의 벌칙항을 부여한 Koenker (2004), adaptive-LASSO와 SCAD 형태의 벌칙항을 적용한 Wu와 Liu (2009)의 연구가 있으며 본 연구에서는 분위수 회귀모형에서 oracle property를 만족하는 SCAD 벌칙항을 적용하였다.

3.2. 분할 및 정지규칙

PQRT 분할규칙의 특징은 Loh (2002)에서 제안한 것과 같이 먼저 분할변수를 선택한 후에 분할점(또는 분할집합)을 선택하는 것이다. 분할변수를 먼저 선택하는 이유는 분할변수 선택간 발생하는 선택편향의 가능성을 최대한 낮추기 위함이다. 이 때 분할후보변수와 반응변수를 적합한 결과를 바탕으로 잔차분석을 하는데 이것은 변수간 상호작용을 반영할 수 있다. 다시 말하면 분할후보변수가 반응변수와 어떤 관계를 갖는지 통계적으로 확인하여 가장 유의한 관계를 갖는 변수를 분할변수로 선택하게 된다. 다음은 상세한 분할변수 선택 알고리즘이다.

1단계 식 (2.4)를 이용하여 잔차를 계산.

2단계 분할표(contingency table) 작성

(2-1) 잔차의 부호에 따라 구분하여 행(row)으로 설정,

(2-2) 분할후보변수가 범주형이면 각 범주를 열(column)로 설정하고, 연속형이라면 노드 t에서의 표본 개수에 따라 3 또는 4분위수로 나눈 기준을 열로 설정.

3단계 각 분할후보변수에 대해 분할표를 이용하여 카이제곱검정(chi-square test) 통계량을 계산.

4단계 분할후보변수별 자유도(degree of freedom)를 일치시키기 위하여 Wilson Hilferty 변환 시행.

5단계 4단계 결과를 비교하여 가장 작은 유의확률을 갖는 분할후보변수 선택.

이렇게 선택된 분할변수에 대해 가능한 모든 점(또는 집합)의 불순도를 계산하여 불순도 감소량이 최대가 되는 점(또는 집합)을 분할점(또는 분할집합)으로 선택한다.

나무모형의 과적합(overfitting)과 과소적합(underfitting)을 줄이기 위해 나무모형의 크기를 결정하는 것은 중요한 문제이다. PQRT는 최종노드의 표본수, 나무모형의 깊이와 카이제곱 통계량으로부터 계산

Table 4.1. Estimated selection probabilities for split variable in Model (4.1)

Variable	QRT				PQRT			
	Average	τ			Average	τ		
		0.25	0.5	0.75		0.25	0.5	0.75
Z_1	0.758	0.630	0.870	0.775	0.792	0.730	0.885	0.760
Z_2	0.057	0.075	0.035	0.060	0.060	0.075	0.040	0.065
Z_3	0.065	0.100	0.030	0.065	0.062	0.065	0.030	0.090
Z_4	0.065	0.095	0.040	0.060	0.040	0.070	0.015	0.035
Z_5	0.055	0.100	0.025	0.040	0.047	0.060	0.030	0.050

되는 유의확률의 최소값을 제한하여 나무모형의 성장을 정지한다. 또한 이러한 직접적인 기준으로 나무모형의 크기를 제한하는 경우 발생하는 과적합과 과소적합 문제를 방지하기 위해 Eo와 Cho (2014)가 제안한 M-단계 정지규칙을 추가로 적용한다. 이는 사전 가지치기(pre-pruning)와 사후 가지치기(post-pruning)을 결합한 효과적인 방법이다.

4. 모의 실험과 실증 예제

본 절에서는 나무모형의 최종 노드에서 적합되는 적합변수와 분할변수 선택에 대해 나무모형이 1회 분할되는 경우에 대하여 QRT와 PQRT의 성능을 비교하였다. 또한 두 종류의 오차항을 가정하여 최종 노드가 3개인 나무모형에 대해 두 방법의 예측력을 비교하였다. 실증 예제는 보스턴 집값 자료(https://archive.ics.uci.edu/ml/datasets/Housing)를 이용하여 PQRT의 우수한 성능을 확인하였다. PQRT의 경우 조율모수(tuning parameter)는 R의 'rqPen' 패키지를 이용하여 결정하였다.

4.1. 분할변수 및 적합변수 선택에 대한 모의실험

$X_j\,(j = 1,\ldots,8)$를 적합변수로 두고, 5개의 분할변수 $Z_s\,(s = 1,\ldots,5)$를 고려하자. 적합변수 간 상관관계를 위해 $(X_1,\ldots,X_8)$은 평균이 0이며 공분산행렬(covariance matrix)이 $0.5^{|a-b|}$ $(a = 1,\ldots,8;\ b = 1,\ldots,8)$인 다변량 정규분포(multivariate normal distribution)를 따른다. 여기서 $Z_1 \sim C_2$, $Z_2 \sim C_{12}$, $Z_3 \sim N(0,1)$, $Z_4 \sim \text{Exp}(1)$, $Z_5 \sim \text{DU}(1,2,\ldots,6)$이며 C_L은 순서가 없는 $1,2,\ldots,L$에 대해 동일한 확률 $1/L$을 갖는 분포이며 $\text{Exp}(\lambda)$는 모수(rate parameter)가 λ인 지수분포(exponential distribution)이고 $\text{DU}(1,2,\ldots,U)$는 집합 $\{1,2,\ldots,U\}$에 대한 이산균등분포(discrete uniform distribution)를 나타낸다. 특히 $\text{DU}(1,2,\ldots,U)$는 순서가 있는 변수에 대한 이산균등분포인 점에서 C_L과 차이가 있다. C_L에서 $L = 12$인 경우 가능한 분할 집합의 수는 $2^{(12-1)} - 1$개인 반면, $\text{DU}(1,2,\ldots,12)$에서는 $12 - 1$개가 가능한 분할 집합의 수가 된다.

모형 (4.1)을 이용하여 반응변수를 독립적으로 100개 생성하며 이때 오차항 ε은 표준정규분포를 따른다.

$$Y = \begin{cases} 3X_1 + 1.5X_5 + \varepsilon, & \text{if } Z_1 = 1, \\ 1 + 4X_1 + 2X_5 + \varepsilon, & \text{otherwise.} \end{cases} \tag{4.1}$$

모형 (4.1)에서 적합변수는 X_1과 X_5이며, 분할변수는 Z_1이다. 그러므로 분할변수 Z_1을 기준으로 분할되고 X_1과 X_5에만 회귀계수를 추정한다면 이상적이다. Table 4.1은 이 과정을 독립적으로 200회 반복한 QRT와 PQRT의 분할변수 선택 결과를 나타낸다. Table 4.1을 통해 분할변수 선택시 PQRT가 QRT에 비해 전반적으로 우수한 선택확률을 갖는 것을 알 수 있다. $\tau = 0.75$에서처럼 일부 분위수에서

Table 4.2. Estimated selection probabilities for fitting variable in Model (4.1)

Method	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Oracle	1.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
QRT	1.000	0.997	0.994	0.998	1.000	0.996	0.997	0.999
PQRT	1.000	0.313	0.328	0.315	1.000	0.318	0.328	0.316

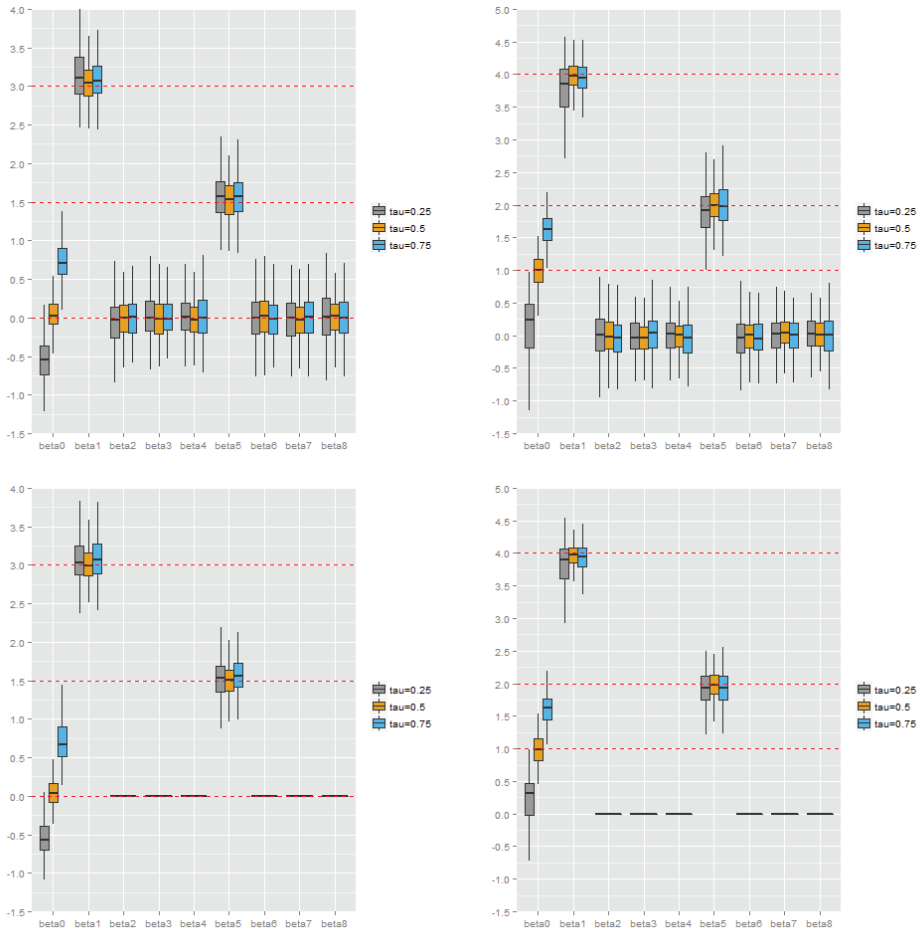

Figure 4.1. Boxplots of estimated parameter values at left and right nodes in Model (4.1). Upper left and right are estimated parameter values at left and right nodes using QRT, respectively. Below are using PQRT. Red line represents true value.

PQRT의 성능이 다소 감소하기도 하였으나, 모의실험에서 고려된 분위수 이외의 다른 분위수에 대하여 대부분 PQRT가 더 정확한 분할변수를 선택하는 것을 확인할 수 있었으며, 지면관계로 인하여 이러한 결과는 생략하였다.

Table 4.2는 각각의 적합변수들이 모형에 선택된 비율로서 여기서 1은 적합변수로 200회 모두 선택된 것을 의미한다. QRT는 정보가 없는 적합변수, 즉 잡음변수를 거의 모든 경우 모형에 포함하지만 PQRT는 잡음변수를 대략 30% 정도만을 모형에 포함하였다. 축소추정(shrinkage estimation)의 정도

Table 4.3. Mean absolute errors for Model (4.2)

Error dist.	Method	Average	τ		
			0.25	0.5	0.75
$N(0,1)$	QRT	0.7686	0.8074	0.7216	0.7768
		(0.0147)	(0.0165)	(0.0132)	(0.0145)
	PQRT	0.6196	0.6461	0.5812	0.6314
		(0.0174)	(0.0171)	(0.0175)	(0.0178)
DE(0, 1)	QRT	0.8717	1.0703	0.7790	0.7659
		(0.0200)	(0.0237)	(0.0158)	(0.0204)
	PQRT	0.6802	0.8339	0.6127	0.5941
		(0.0198)	(0.0227)	(0.0196)	(0.0171)

The numbers in parentheses are standard errors.

를 보다 더 정확히 알기 위해 Figure 4.1과 같이 회귀계수별 추정값을 상자그림(boxplot)으로 확인하였다. Figure 4.1에서 가로축은 적합변수에 대한 계수값 $\beta_1, \ldots, \beta_8$을 나타낸다. Table 4.2와 Figure 4.1의 결과로서 PQRT가 QRT에 비해 더 정확한 적합을 하고 있음을 확인할 수 있다.

4.2. 예측력에 대한 모의실험

4.1절에서 설정한 동일한 적합변수와 분할변수에 대해 모형 (4.2)를 이용하여 예측력에 대한 모의실험을 진행하였다. 이때 오차항은 표준정규분포와 이중지수분포(double exponential distribution; DE)를 가정하였으며, 크기가 10,000인 검정자료(test data)를 독립적으로 추가 생성하여 평균절대오차(mean absolute error; MAE)를 계산하고 이를 통해 예측력을 비교하였다. 이 과정은 100회 독립적으로 반복 시행하였다.

$$Y = \begin{cases} 0.5 + 0.5X_1 + \varepsilon, & \text{if } Z_1 = 1 \text{ and } Z_3 \geq 0, \\ -0.5 - 0.5X_1 + \varepsilon, & \text{if } Z_1 = 2, \\ 1 + X_1 + \varepsilon, & \text{otherwise.} \end{cases} \tag{4.2}$$

예측력 비교를 위한 MAE는

$$\text{MAE}(\tau_k) = \frac{1}{10000} \sum_{r=1}^{\tilde{T}} \sum_{i \in \tilde{t}_r} \left| q_{\tau_k}\left(\boldsymbol{x_i}, \tilde{t}_r\right) - \hat{q}_{\tau_k}\left(\boldsymbol{x_i}, \tilde{t}_r\right) \right| \tag{4.3}$$

와 같이 정의된다. Table 4.3은 QRT와 PQRT의 예측력을 비교한 것으로 시험한 모든 분위수와 오차항 분포에서 PQRT가 더 작은 MAE를 보인다. 이러한 결과는 4.1절에서 보인 분할 및 적합변수를 더 적절하게 선택한 결과로 설명할 수 있다. 특히 PQRT는 일부 분위수에서 분할변수 선택의 성능이 다소 감소하였으나, QRT에 비해 노드 t에서 더 정확한 적합을 함으로써 QRT보다 우수한 예측력을 보였다. 이러한 PQRT의 우수한 예측력은 모의실험에서 고려된 분위수 이외의 다른 분위수에서도 확인할 수 있었다.

4.3. 실증 예제: 보스턴 주택가격 자료를 중심으로

본 절에서는 앞에서 설명한 QRT와 PQRT의 성능을 실증 예제를 통해 비교하고자 한다. 분석에 사용된 자료는 보스턴 집값 자료로 506개의 관측치에 대해 주택가격(CMEDV)에 로그를 취한 것을 반응변수로

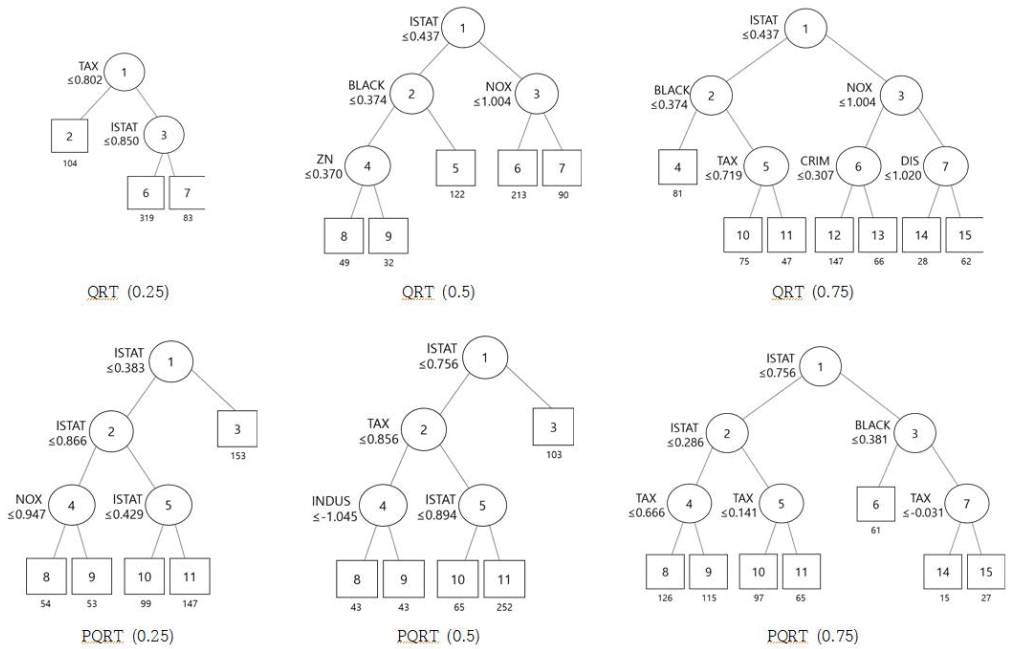

Figure 4.2. Estimated tree models using QRT and PQRT at three quantile levels.

두고 분할변수는 10개(crime rate; CRIM, proportion of area zoned with large lots; ZN, proportion of non-retail business acres per town(INDUS), nitric oxides concentration; NOX, proportion of owner-occupied units built prior to 1940; AGE, weighted distances to five Boston employment centers; DIS, property tax rate; TAX, pupil-teacher ratio by town; PTRATIO, black population proportion town; BLACK, and lower status population proportion; LSTAT)를 설정하였다. 적합변수는 분할변수 10개와 독립적으로 표준정규분포를 따르는 확률변수 20개를 합하여 총 30개를 사용하였다.

Figure 4.2는 QRT와 PQRT 방법을 이용하여 506개 전체 관측치를 대상으로 3가지 분위수에 대한 회귀나무모형을 추정한 결과이다. QRT는 주택가격이 낮을 때 TAX를 가장 중요한 변수로 선정한 반면, PQRT는 LSTAT을 선정하였다. 주택가격 분포의 중위수(median)에 대해 QRT와 PQRT 모두 LSTAT을 우선 선정하였으나, 이후 QRT는 BLACK과 NOX를 선정한 반면 PQRT는 TAX를 선정하였다. 높은 주택가격에 대해서도 나무모형의 깊이가 깊어짐에 따라 상이한 결과를 보였다. Figure 4.3은 0.5 분위수에서 추정된 나무모형에 대해 최종노드의 회귀계수 추정값을 상자그림으로 나타낸다. Figure 4.3의 세로축은 추정된 계수값을 나타내며 가로축은 위에서 열거한 분할변수 10개와 독립적으로 표준정규분포를 따르는 확률변수 20개를 순차적으로 숫자로 표시한 것이다. Figure 4.3에서 볼 수 있듯이 PQRT가 전반적으로 축소추정을 하고 있음을 알 수 있으며, 특히 PQRT는 잡음을 위하여 추가한 표준정규 확률변수 20개를 모형에서 제거함으로써 보다 간결하고 정확한 모형을 추정하는 것을 알수 있다. 다른 분위수에서도 유사한 결과를 보이며 본 논문에서는 지면 절약을 위해 0.5 분위수만 제시하였다.

QRT와 PQRT의 예측력을 비교하기 위하여 506개의 관측치를 350개 훈련자료와 156개 검정자료로 구분하여 체크 손실함수를 이용하여 오차를 확인하였다. 이 과정은 독립적으로 100회 반복하였으며 그 결과는 Table 4.4와 같다. 앞선 모의실험 결과와 동일하게 PQRT의 예측력이 QRT보다 좋음을 알 수 있다.

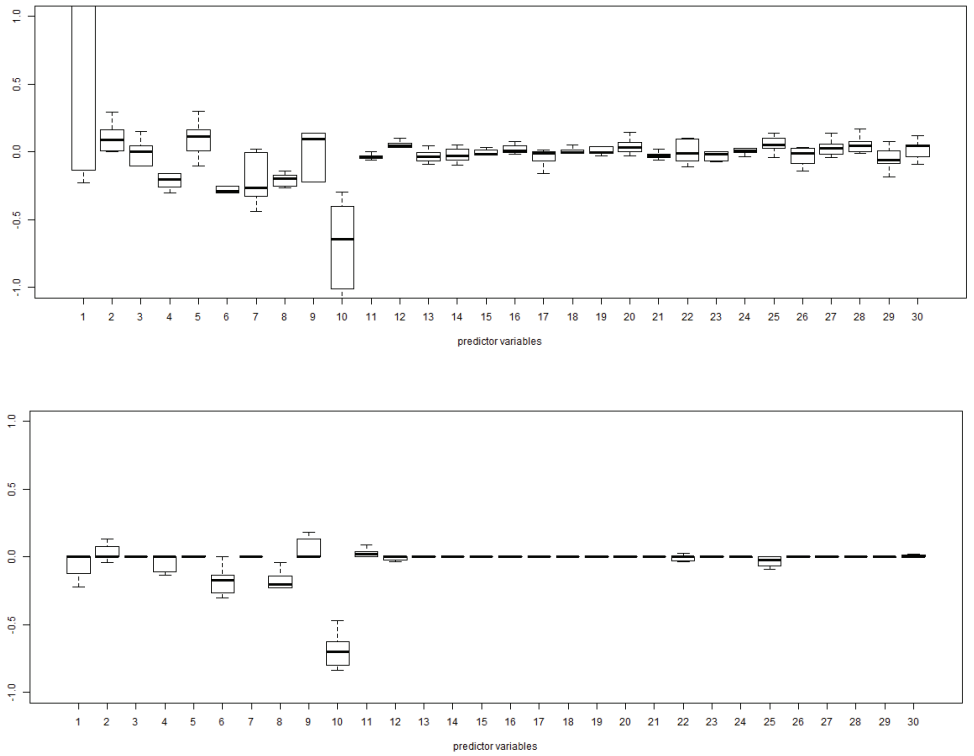

Figure 4.3. Boxplots of estimated parameter values at terminal nodes for corrected Boston house price data. Upper is estimated parameters using QRT at quantile level 0.5. Below is using PQRT at same quantile level.

Table 4.4. Prediction accuracy of the corrected Boston house price data.

Method	Average	τ		
		0.25	0.5	0.75
QRT	0.2481	0.2212	0.3086	0.2145
	(0.0154)	(0.0072)	(0.0311)	(0.0079)
PQRT	0.1787	0.1682	0.1962	0.1716
	(0.0019)	(0.0021)	(0.0015)	(0.0021)

The numbers in parentheses are standard errors.

5. 결론

분위수 회귀모형이 매우 유용한 통계적 도구로 이용되지만, 설명변수와 반응변수의 함수관계를 선형으로 가정하는 것은 때때로 부정확하며 왜곡된 결과를 초래하는 원인이 될 수 있다. 설명변수와 반응변수의 함수관계를 비선형으로 가정할 때 주로 비모수적 방법을 고려하며 이 중 나무모형은 우수한 해석과 예측이 가능한 모형으로 알려져 있다. 한편 데이터의 형태가 다양해지고 차원이 증가함에 따라 벌점화 회귀모형의 필요성은 어느 때보다 더욱 높다고 할 수 있다. 본 연구에서는 데이터의 형태가 고차원 또는 설명변수간 상관관계가 높은 경우 등에 유용한 벌점화 추정법을 분위수 회귀나무모형에 적용하는

PQRT 방법을 제안하였다. 모의실험을 통하여 제안한 PQRT가 적합변수와 분할변수의 선택에서 우수한 성능을 보이는 것을 확인하였으며, 상이한 오차항의 분포에 대해 예측 오차가 작음을 보였다. 나아가 실증예제에 적용하여 기존 QRT보다 우수한 성능을 입증하였다. 본 연구에서는 비선형모형의 비모수적 추정법인 QRT의 제한사항을 보완한 PQRT에 초점을 두어 linear quantile regression(LQR), penalized LQR(PLQR) 등과 같은 선형모형을 가정하는 추정법은 성능의 비교 대상에서 고려하지 않았다.

References

Breiman, L. (1995). Better subset regression using the nonnegative garrote, *Technometrics*, **37**, 373–384.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*, Wadsworth, Belmont.

Chang, Y. J. (2014). Multi-step quantile regression tree, *Journal of Statistical Computation and Simulation*, **84**, 663–682.

Chaudhuri, P. and Loh, W. Y. (2002). Nonparametric estimation of conditional quantiles using quantile regression trees, *Bernoulli*, **8**, 561–576.

Eo, S. H. and Cho, H. (2014). Tree-structured mixed-effects regression modeling for longitudinal data, *Journal of Computational and Graphical Statistics*, **23**, 740–760.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, **96**, 1348–1360.

Hallin, M., Lu, Z., and Yu, K. (2009). Local linear spatial quantile regression, *Bernoulli*, **15**, 659–686.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning a conditional inference framework, *Journal of Computational and Graphical Statistics*, **15**, 651–674.

Kim, H., Loh, W. Y., Shih, Y. S., and Chaudhuri, P. (2007). Visualizable and interpretable regression models with good prediction power, *IIE Transactions*, **39**, 565–579.

Kim, S. and Xing, E. P. (2012). Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eQTL mapping, *The Annals of Applied Statistics*, **6**, 1095–1117.

Koenker, R. (2004). Quantile regression for longitudinal data, *Journal of Multivariate Analysis*, **91**, 74–89.

Koenker, R. (2005). *Quantile Regression*. Cambridge university press, New York.

Koenker, R. and Bassett, Jr, G. (1978). Regression quantiles, *Econometrica: Journal of the Econometric Society*, **46**, 33–50.

Koenker R. and Mizera, I. (2004). Penalized triograms: total variation regularization for bivariate smoothing, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **66**, 145–163.

Li, Y., Liu, Y., and Zhu, J. (2007). Quantile regression in reproducing kernel Hilbert spaces. *Journal of the American Statistical Association*, **102**, 255–268.

Liu, Y. and Wu, Y. (2011). Simultaneous multiple non-crossing quantile regression estimation using kernel constraints, *Journal of Nonparametric Statistics*, **23**, 415–437.

Loh, W. Y. (2002). Regression trees with unbiased variable selection and interaction detection, *Statistica Sinica*, **12**, 361–386.

Loh, W. Y. (2009). Improving the precision of classification trees, *Annals of Applied Statistics*, **3**, 1710–1737.

Quinlan, J. R. (1993). *C4.5: Programming for Machine Learning*, Morgan Kaufmann Publishers Inc., San Francisco.

Shen, X. and Ye, J. (2002). Adaptive model selection, *Journal of the American Statistical Association*, **97**, 210–221.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**, 267–288.

Wu, Y. and Liu, Y. (2009). Variable selection in quantile regression, *Statistica Sinica*, **19**, 801–817.

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

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty, *The Annals of Statistics*, **38**, 894–942.

벌점화 분위수 회귀나무모형에 대한 연구

김재오[a] · 조형준[a] · 방성완[b,1]

[a]고려대학교 통계학과, [b]육군사관학교 수학과

(2016년 9월 2일 접수, 2016년 10월 27일 수정, 2016년 10월 31일 채택)

요 약

분위수 회귀모형은 설명변수가 반응변수의 조건부 분위수 함수에 어떻게 관계되는지 탐색함으로서 많은 유용한 정보를 제공한다. 그러나 설명변수와 반응변수가 비선형 관계를 갖는다면 선형형태를 가정하는 전통적인 분위수 회귀모형은 적합하지 않다. 또한 고차원 자료 또는 설명변수간 상관관계가 높은 자료에 대해서 변수선택의 방법이 필요하다. 이러한 이유로 본 연구에서는 벌점화 분위수 회귀나무모형을 제안하였다. 한편 제안한 방법의 분할규칙은 과도한 계산시간과 분할변수 선택편향 문제를 극복한 잔차 분석을 기반으로 하였다. 본 연구에서는 모의실험과 실증 예제를 통해 제안한 방법의 우수한 성능과 유용성을 확인하였다.

주요용어: 의사결정나무, 분위수 회귀 모형, 벌점화 회귀 모형

본 연구는 2015년 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2015R1C1A1A02036473, NRF-2015R1D1A1A09058602).

[1]교신저자: (01805) 서울특별시 노원구 화랑로 574, 육군사관학교 수학과. E-mail: wan1365@gmail.com