The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.6.1095
29(6), 1095–1106

Variable selection with quantile regression tree

Youngjae Chang[a,1]

[a]Department of Information Statistics, Korea National Open University

(Received August 22, 2016; Revised October 8, 2016; Accepted October 8, 2016)

Abstract

The quantile regression method proposed by Koenker *et al.* (1978) focuses on conditional quantiles given by independent variables, and analyzes the relationship between response variable and independent variables at the given quantile. Considering the linear programming used for the estimation of quantile regression coefficients, the model fitting job might be difficult when large data are introduced for analysis. Therefore, dimension reduction (or variable selection) could be a good solution for the quantile regression of large data sets. Regression tree methods are applied to a variable selection for quantile regression in this paper. Real data of Korea Baseball Organization (KBO) players are analyzed following the variable selection approach based on the regression tree. Analysis result shows that a few important variables are selected, which are also meaningful for the given quantiles of salary data of the baseball players.

Keywords: quantile regression, regression tree, variable selection

1. 서론

Koenker과 Bassett (1978)에 의해 제안 된 분위수 회귀분석법은 독립변수들이 주어졌을 때, 종속변수의 조건부 분위수에 초점을 맞추어 독립변수들과 종속변수의 해당 특정 분위수와의 관계를 분석하는 방법이다. 전형적인 통상최소제곱(ordinary least squares) 추정방법이 독립변수가 주어졌을 때, 종속변수의 평균의 움직임에 주목한 방법이라는 점에서 선형회귀모형에서 적용되었던 방법론을 분위수 회귀모형에 직접 적용하기는 어렵다. 따라서, 기존의 선형모형의 틀을 벗어나 다양한 측면에서 모형의 개선을 위한 알고리즘이 제안되었다. 한편 고차원 대용량 자료의 경우에는 차원 축소의 문제, 조금 더 폭을 좁혀 생각해보면 변수선택의 문제를 통해 의사 결정에 영향을 미치는 주요 요인들을 파악하거나 적절한 규모의 모형을 적합하는 과정이 중요하며 결과적으로 이러한 변수선택은 모형의 예측력을 제고하는 데 유용한 것으로 나타난다 (Chang, 2014). 분위수 회귀 모형의 경우에도 이러한 점을 감안하여 변수선택에 관한 논의가 꾸준히 이루어져 왔다. 본 논문에서는 분위수 회귀분석의 변수선택의 문제를 보다 직관적이고 간단하게 해결하기 위한 방법으로서 회귀나무 모형을 원용하였다. 고차원 자료 분석의 관점에서 분위수 회귀나무 모형 방법을 구현해 보고 실제 자료를 분석해 보았다. 본 논문의 구성은 다음과 같다. 2장에서 분위수 회귀모형과 회귀나무의 기본적인 알고리즘을 개괄한 뒤, 3장에서 분위수 회귀나무의 변수선택 방법을 고찰하고 이 변수선택 방법론을 이용하여 실제 데이터를 분석하여 보았으며, 마지막으로 4장에서는 결론 및 향후 연구과제에 대해 간략히 정리하였다.

This research was supported by Korea National Open University Research Fund.
[1]Department of Information Statistics, Korea National Open University, 86, Daehak-ro, Jongno-gu, Seoul 03087, Korea. E-mail: yjchang@knou.ac.kr

2. 알고리즘 개괄

2.1. 분위수 회귀(quantile regression)

본 절에서는 분위수 회귀모형을 간략히 정리해 보기로 한다. 종속변수 Y와 독립변수 X가 있다고 가정하고 X는 d차원의 변수라고 하자. 이 경우 α 백분위수, Q_α는 식 (2.1)과 같이 나타낼 수 있다.

$$Q_\alpha(X = x) = \inf\{y : F(y|X = x) \geq \alpha\}. \tag{2.1}$$

조건부 분포함수 $F(y|X = x)$는 식 (2.2)처럼 표현할 수 있다.

$$F(y|X = x) = P(Y \leq y|X = x). \tag{2.2}$$

이상의 정의 하에서 분위수 회귀분석은 통상적인 회귀분석 문제에서와 마찬가지로 손실함수(loss function)를 최소화하는 회귀계수를 찾는 과정이라고 할 수 있다. 다만, 통상최소제곱 추정의 경우와는 달리 식 (2.3)과 같은 특별한 손실함수를 사용하게 된다.

$$\rho_\alpha(u) = u(\alpha - I(u < 0)). \tag{2.3}$$

손실함수 $\rho_\alpha(u)$의 특징을 꼽자면 선형 손실함수로서 α 값에 따라 비대칭적인 모습을 지닌다는 것이다. Figure 2.1은 분위수 값에 따라 손실함수가 어떻게 달라지는 지를 나타내고 있다. $\alpha = 0.5$는 중위값을 의미하며 이 경우에는 통상최소제곱 추정의 경우인 Figure 2.2와 유사하지만, 나머지 경우에는 비대칭적인 손실함수의 형태이다. 이러한 손실함수를 바탕으로 $Y = X'\beta$와 같은 선형모형의 회귀계수 추정은 식 (2.4)와 같이 주어진 $X = x$ 값에 대하여 손실함수의 기댓값을 최소로하는 조건부 분위수를 찾는 과정으로 요약할 수 있다.

$$Q_\alpha(X = x) = \arg\min_{\beta \in R^d} E(\rho_\alpha(Y - x'\beta)). \tag{2.4}$$

회귀모형을 손실함수와 연관지어 살펴보면 회귀계수 추정 원리를 조금 더 명확하게 파악할 수 있다. 예를 들어 Figure 2.1의 $\alpha = 0.3$과 같은 경우, u의 값을 $Y - x'\beta$로 대체하여 생각해 보면 이 값이 0보다 작을 때, 즉 회귀직선 보다 아래 위치한 관측치가 있을 때, 이에 대해 상대적으로 더 큰 손실함수 값이 부여된다고 볼 수 있다. 따라서 이러한 손실함수 값을 최소화하기 위해서는 가급적 회귀직선이 낮은 쪽에 위치하여야 한다. 이러한 형태의 직선 중 손실함수의 값을 최소화시키는 회귀계수를 추정함으로써 $\alpha = 0.3$ 해당 분위수 회귀모형을 적합하게 되는 것이다. $\alpha = 0.7$의 경우에도 손실함수 모양이 비대칭인 모습을 지니지만 $\alpha = 0.3$의 예와는 정 반대의 과정을 거쳐 회귀직선이 높은 쪽에 위치하게 되는 계수 값을 추정하게 되는 것이다.

이러한 이유로 분위수 회귀모형의 회귀계수 추정은 그 이론적 배경에 비해 추정 과정이 간단하지 않다. 기본적으로 최소절대편차(least absolute deviation) 손실함수에서 전형적으로 사용되는 선형계획법을 활용하게 된다. 식 (2.5)처럼 최소절대편차 추정치를 찾는 이 과정을 중위값 회귀(median regression)라고 하는데, 이는 식 (2.3)의 손실함수에서 $\alpha = 0.5$인 특별한 경우임을 쉽게 알 수 있다.

$$\arg\min_{\beta \in R^d} E(|Y - x'\beta|). \tag{2.5}$$

선형계획법을 활용하여 추정하는 데 있어서는 대체로 반복(iteration)을 통해 해를 찾는 알고리즘을 이용하게 된다. 이 과정에서도 전역 최소값(global minimum)을 찾지 못할 가능성도 있고 X의 차원 d가 매우 클 경우에는 회귀계수의 추정이 어려울 뿐만 아니라 추정 과정에서 수렴하지 않을 가능성도 증

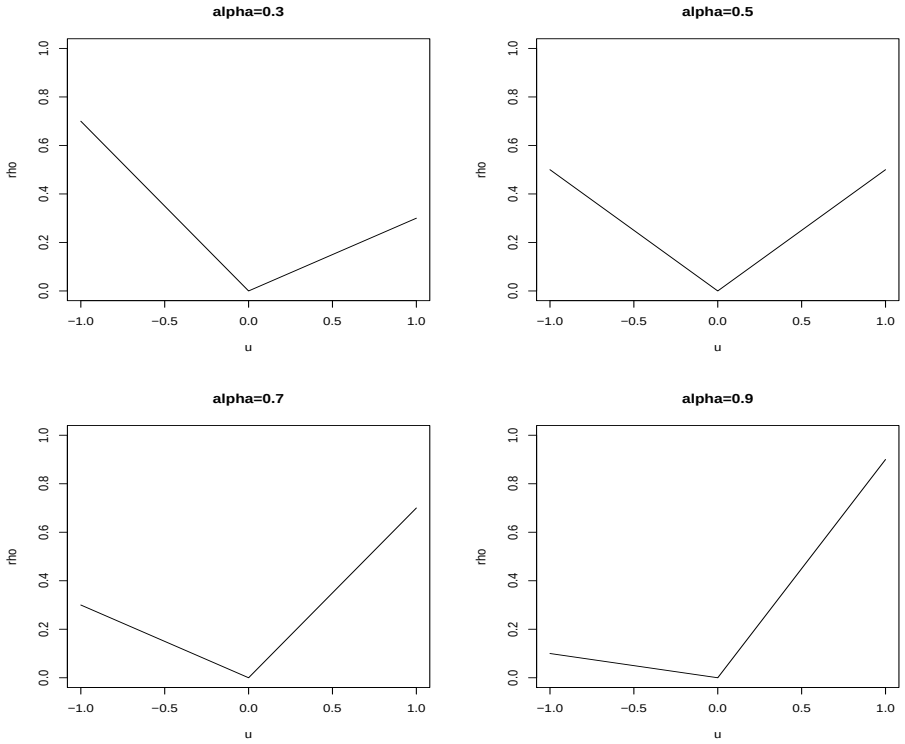

Figure 2.1. Loss function according to α.

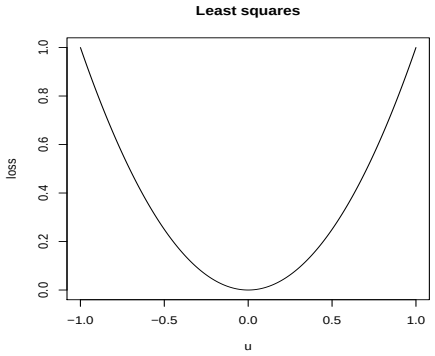

Figure 2.2. Squared error loss function.

가한다. 따라서, 고차원 자료를 다루는 분위수 회귀분석의 경우에는 더욱 주의를 요하게 된다. 이러한 측면에서 독립변수의 차원이 클 경우 이를 해결하기 위한 여러 가지 방법론들이 제기 되었다. Chang (2014)에서도 다단계 분위수 회귀나무 방법론을 이용하여 독립변수의 차원이 급증할 때에도 예측력이 저하되지 않는 알고리즘을 제안한 바 있다. 다만, 교차타당화(cross-validation)를 이용한 모형 평가 및 예측력 향상에만 국한된 것으로 변수선택이나 모형축소 측면은 고려하지 않았다. 본 논문에서는 고차원

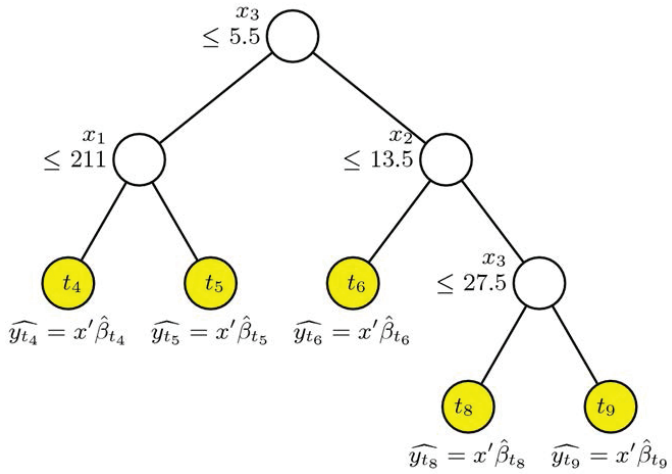

Figure 2.3. An example of piecewise linear regression tree: At each split, an observation goes to the left branch if and only if the condition is satisfied.

의 분위수 회귀모형에 있어서 변수선택과 모형 축소 방법을 살펴보기로 한다. 기존 논문에서 다루었던 변수선택의 방법 대신 회귀나무를 이용하여 직관적이고 간단하게 구현하는 방법을 이용한다.

2.2. 회귀나무(regression tree) 알고리즘

회귀나무는 데이터를 특정 기준 변수값에 따라 재귀적으로 이분할 하면서 모형을 확장하고 과다적합을 방지하기 위해 교차 타당화 방법을 통해 적정 크기의 모형을 찾는 방법이다. 조각별 선형 회귀나무란 회귀나무 모형을 구축할 때 분기가 이루어 질 자식 노드(node)에서 선형모형을 적합한 뒤 이러한 선형모형의 잔차가 최소화되는 지점을 찾아 최적 분기점으로 설정해 나가는 방법이다.

Figure 2.3은 조각별 선형 회귀나무의 예이다. 최상위 노드 또는 마디는 모든 훈련샘플(training sample)을 포함하고 있으며 여기서부터 가지가 나뉘어지며 나무가 자라게 된다. 각 노드에서는 분기변수(split variable)로 선택된 설명변수(x_1, x_2, x_3 등)의 값에 따라 가지가 나뉘어지며 이러한 이분할이 반복되는 단계 및 교차타당화를 통한 가지치기(pruning) 과정을 거쳐 최종적인 나무의 모습을 이루게 된다. Figure 2.3에서 t_4, t_5, t_6, t_8, t_9는 최종노드이며 각 최종 노드 아래에는 $\widehat{y_{t_4}} = x'\hat{\beta}_{t_4}$, $\widehat{y_{t_5}} = x'\hat{\beta}_{t_5}$, $\widehat{y_{t_6}} = x'\hat{\beta}_{t_6}$, $\widehat{y_{t_8}} = x'\hat{\beta}_{t_8}$, $\widehat{y_{t_9}} = x'\hat{\beta}_{t_9}$처럼 해당 노드에서 적합된 선형모형이 제시되어 있다.

이러한 조각별 선형 회귀나무에서 $\widehat{y_{t_4}} = x'\hat{\beta}_{t_4}$ 등과 같이 각 최종 노드에 적합되는 모형이 어떤 모형인지에 따라 구체적인 회귀나무의 형태가 정해지게 된다. Chang (2010)에서처럼 최종 노드에서 적합되는 모형이 다중 선형회귀 모형인 다중 선형 회귀나무 모형(multiple linear regression tree)인 경우, 상수항만 존재하는 상수항 모형, 최종 노드에서의 모형을 다중선형회귀 모형으로 하되 변수선택을 감안한 단계별 회귀(stepwise regression) 알고리즘을 적용한 나무모형 등 다양한 모형을 적용하여 나무를 구축할 수 있다.

Chang과 Kim (2011)에서는 간단한 시뮬레이션 연구를 통해 비선형 회귀모형 추정에 있어서 회귀나무가 적절하게 사용될 수 있음을 보였는데, 종속변수와 비선형 관계가 뚜렷한 독립변수를 분기변수로 삼아 이분할 하고 자식 노드에서의 잔차가 최소가 되는 점을 찾는 과정을 시현하였다. 이러한 관점에서 보면, 조각별 선형 회귀나무는 그 이름이 의미하는 대로 몇 개의 선형 모형의 결합으로 이루어진 회귀나무라고

이해할 수 있다. Classification And Regression Tree(CART)와 같은 전통적인 회귀나무의 자식 노드에서는 상수항 모형만 적합되는 것과는 대조적이라고 할 수 있다. 이러한 이유 때문에 대체로 일반적인 조각별 선형 회귀나무의 크기는 CART와 같은 상수항 회귀나무 모형에 비해 작게 나타난다.

이러한 조각별 선형 회귀나무의 구현이 일반화 될 수 있다면, 앞서 살펴 본 분위수 회귀에 회귀나무를 적용하는 문제도 매우 간단하게 해결될 수 있다. Loh (2002)는 회귀나무의 이러한 성질을 이용하여 비모수적인 분위수 회귀모형 추정 방법을 제안하였다. Loh가 제안한 Generalized, Unbiased, Interaction Detection and Estimation(GUIDE)는 자식 노드에서 다양한 모형적합을 가능하게 함으로써 모형 적합의 범위를 넓히는 동시에 예측력도 제고한 알고리즘이다. Loh (2002)에서는 GUIDE 알고리즘의 특징을 변수선택 편의(variable selection bias)가 거의 없으며 곡률 검정(curvature test) 단계를 통한 비선형성 포착이 용이하고, 교호효과를 고려한데다가 계산 시간이 상대적으로 빠르다는 점 등으로 꼽았다. 특히 선택편의(selection bias)는 고차원 자료의 다범주 변수로 인해 발생하는 경우가 많은데 이를 방지했다는 것은 Breiman 등 (1984)이 제안한 CART 방법론이 지니고 있던 문제점을 해결한 것으로 평가할 수 있다. 따라서, GUIDE 방법론은 일반적인 고차원 자료를 바탕으로 모형을 적합할 때 나타나는 나무모형의 예측력 저하를 상당히 완화시킬 수 있다.

2.3. 변수선택을 위한 분위수 회귀나무(quantile regression tree) 알고리즘

Chang (2010)에서 GUIDE의 조각별 선형회귀 나무 모형 중 한 형태인 단계적 선형회귀나무(stepwise linear regression tree) 모형을 이용하여 고차원 자료의 변수선택 방법을 제안한 바 있다. 그러나 분위수 회귀에는 동 방법의 직접적인 적용이 어려운 점을 감안하여 본 논문에서는 Chauduri와 Loh (2002)가 제안한 분위수 회귀 나무 알고리즘을 응용한다. 특별히 분위수 회귀의 문제에 있어서 변수선택 방법으로 상수항 모형 분위수 회귀나무를 이용하기로 한다. 다중 선형 회귀나무를 적용할 경우 분기변수로 선택되는 변수의 개수는 몇 개로 한정되지만, 최종노드에서 적합되는 모형은 전체 입력변수를 포함하는 모형으로서 결과적으로 선택된 변수만으로 모형을 구축하는 효과를 얻을 수 없기 때문이다. 즉, 상수항 분위수 회귀(constant quantile regression)모형으로 변수를 선택한 후 다중 선형 분위수 회귀나무로 모형을 적합하는 방식의 결합 알고리즘을 사용하게 되는 것이다. 이러한 결합 알고리즘을 간략히 요약하면 다음과 같다.

1. 분석하고자하는 해당 분위수 α를 지정한다.
2. 현 노드를 t라 하자. 현 노드의 데이터를 바탕으로 분위수 α에 해당하는 상수항 분위수 회귀 모형을 적합한다. 상수항 분위수 회귀모형이란, 각 노드에서 종속변수의 분위수와 다수의 입력변수들 간의 관계를 고려한 모형으로 최종 노드에서의 모형이 상수항으로만 구성된 나무모형이다.
3. 각각의 관측치에 대해서, 곡률검정(curvature test)를 이용하여 분기변수를 선택한다 곡률검정이란 종속변수와 각 독립변수 간의 관계가 비선형적인지 여부를 검정하는 과정이다. 예를 들어 식 (2.6)에서 함수 $\mu(\cdot)$가 선형일 때, 실제 데이터에 모형을 적합하였다고 가정해 보자. 가장 간단한 예로서 우리에게 익숙한 단순 선형회귀모형을 생각하면 된다.

$$y_i = \mu(x_i) + \epsilon_i, \quad i = 1, \ldots, n. \tag{2.6}$$

모형 적합 결과 산출되는 예측값 $\hat{y}_i$과 실제값 y_i와의 편차인 잔차를 수직축에, 입력변수인 x_i를 수평축에 놓고 산점도를 그린다. 이 때, 입력변수인 x_i의 각 4분위수 지점을 나누어 4개의 셀을 만들고 수직축은 0을 기준으로 잔차의 부호를 구분하는 수평선을 그어 총 8개의 셀을 만든다. 이러한 분할

표(contingency table)가 작성되면, 식 (2.7)과 같은 카이제곱 통계량을 계산할 수 있다.

$$\chi^2 = \frac{\sum(O_i - E_i)^2}{E_i}, \tag{2.7}$$

여기서 O_i는 각 셀에 위치한 관측치 수이고 E_i는 주어진 분할표의 기댓값, 즉 관측치 수의 셀당 평균값이 된다. 자유도는 $(4-1)(2-1)=3$이 된다. 만약 해당 카이제곱 통계량의 값이 크다면 이는 잔차의 분포가 치우진 것을 의미하므로 선형성에서 벗어난 것을 의미한다고 볼 수 있다. 각 입력변수별로 해당 통계량의 유의확률 p-값 비교를 통해 가장 작은 p-값을 나타내는 입력변수가 비선형성이 가장 뚜렷한 변수라고 간주하여 이 입력변수를 분기가 되어야 할 대상 변수로 선정한다. 자세한 사항은 Chang과 Kim (2011)을 참조하기 바란다. 개별 변수뿐만 아니라 각 변수 쌍에 대해 교호효과 파악을 위한 카이제곱 검정(chi-squared tests)을 실시하여 가장 작은 p-값을 나타낸 변수를 선택하게 된다.

4. 위와 같은 과정을 통해 분기변수가 결정되면 분기점(split point)을 결정하게 된다. t_L과 t_R을 각각 현 노드인 t의 좌, 우 하위 노드라고 하자.
   - 만약 X가 수치형 변수이면 t_L과 t_R의 손실함수의 합(sum of error losses)이 최소가 되게하는 분기점을 찾는다.
   - 만약 X가 범주형 변수이면 해당 변수의 분기조건은 $X \in C$와 같은 형태로 정해진다 (C는 X의 값으로 구성된 부분집합이다). 즉, 하위 노드인 t_L과 t_R에서 손실함수의 가중합이 최소가 되게하는 분기 조합을 찾게 된다.
5. 분기과정이 끝나게 되면, 교차타당화를 이용한 가지치기 단계(pruning)를 거쳐 회귀나무를 구현한다.
6. 이상의 상수항 모형 적합을 통해 구현된 회귀 나무에 나타난 변수들의 리스트를 확인하고 이 선택된 변수들로만 전체 자료를 대상으로 새로운 분위수 회귀모형을 적합한다. 이 때, 적합하는 모형은 조각별 회귀나무 모형 중 하나의 형태인 다중 선형 회귀나무 모형을 적용하게 된다.
7. 최종적으로 가지치기 단계를 거쳐 다중선형 분위수 회귀나무 모형을 얻게 된다.

3. 실증분석

본장에서는 분위수 회귀나무를 이용하여 한국 프로야구 선수들의 연봉과 성적과의 관계를 분석해 보았다. 특히 투수와 타자 등 선수들의 직전년도 성적을 독립변수로 하고 올해의 연봉을 종속변수로 하여 분위수 별로 연봉에 영향을 미치는 변수들이 차이가 있는지를 살펴보기 위해 회귀나무를 이용한 변수 선택을 실시하였다. 모든 변수를 포함한 다중 분위수 회귀 모형을 적합할 경우 너무 많은 독립변수들이 존재하여 회귀계수 추정의 불안정성 등이 발생할 수 있으므로 이러한 변수 선택 과정은 모형 적합의 용이성 면에서도 의미가 있다고 하겠다.

3.1. 데이터 설명

본 논문에서는 한국프로야구 선수들의 연봉과 직전년도 성적으로 이루어진 데이터를 분위수 회귀나무 모형을 통해 분석하였다. 2016년 개막일 로스터에 등록된 선수들을 기준으로 하되 직전년도인 2015년도 성적이 존재하는 선수들만 분석대상으로 하였다. 즉, 올해 처음 등록된 신인이나 군복무나 부상 등으로 2015년도 기록이 없는 선수들은 분석에서 제외하였다. 결과적으로 분석대상 선수들 수는 타자는

Table 3.1. Variables for KBO players

Variable for batters	Description	Variables for pitchers	Description
AVG	타율	ERA	평균자책점
G	게임	G	경기
AB	타수	W	승리
R	득점	L	패배
H	안타	SV	세이브
h2B	2루타	HLD	홀드
h3B	3루타	WPCT	승률
HR	홈런	TBF	타자수
TB	루타	IP	이닝
RBI	타점	H	피안타
SB	도루	HR	피홈런
CS	도루실패	BB	볼넷
BB	볼넷	HBP	사구
HBP	사구	SO	삼진
SO	삼진	R	실점
GDP	병살타		
SLG	장타율		
OBP	출루율		
E	실책		

141명, 투수는 85명이었다. 연봉을 기준으로 90 백분위, 50 백분위, 25 백분위 수, 즉 $\alpha = 0.9, 0.5, 0.25$ 인 경우로 나누어 각각의 회귀나무를 통해 변수를 선택하고 그 변수들만으로 분위수 회귀모형을 적합해 보았다. 낮은 분위수로 25 백분위수를 선택한 이유는 지나치게 낮은 수준의 연봉을 받는 경우를 제외하고 분석하고자 하였기 때문이다. 낮은 수준의 연봉의 경우는 대부분 신인급의 선수들로서 최저 기본연봉에 근접한 수준을 받으므로 성적에 민감하게 연동하여 연봉을 받는다고는 평가할 수 없기 때문이다. 이는 대체로 FA 계약 등과 관련된 선수들이 많이 분포되어 있는 높은 연봉수준에도 적용되는 사안이라고 볼 수 있으며 이를 감안하여 통상적인 95 백분위수 대신 90 백분위수를 분석대상으로 삼았다. 그럼에도 불구하고 이러한 분석대상 백분위수 선정은 다소 임의적인 부분이 있으므로 결과 해석상 유의점이 존재한다. 포지션별로는 투수들과 타자들의 성적을 직접적으로 비교하기 어려우므로 투수와 타자는 별도로 구분하여 분석하였다. 데이터는 한국야구위원회 홈페이지(http://www.koreabaseball.com/)에 공개된 2016년 개막일 당시의 로스터와 선수조회 기능을 이용하여 검색한 뒤 축적하였다. Table 3.1은 타자와 투수의 성적을 나타내는 주요 변수들이다. 이러한 2015년도의 성적관련 변수들을 독립변수로, 2016년 개막일 당시의 선수들의 연봉을 종속변수로 하여 분위수 회귀나무 모형을 적합해 보았다. 연봉의 단위는 천만원이며 각 변수들은 변환하지 않은 형태의 자료를 그대로 사용하였다.

3.2. 분석 결과

실증분석 결과 타자와 투수의 경우 모두 분위수 회귀 나무를 통해 독립변수의 수를 상당히 축소할 수 있는 것으로 나타났다. 먼저 타자의 분석결과를 보면, $\alpha = 0.9$인 경우 H(안타 수), h2B(2루타 수), SLG(장타율) 등의 변수가 선택되었고 $\alpha = 0.5$일 때에는 R(득점)과 HR(홈런) 등이 선택되었다 (Figure 3.1, Figure 3.2). 한편, $\alpha = 0.25$인 경우에는 RBI(타점) 및 BB(볼넷) 등의 변수가 선택되었다. 이러한 변수 선택의 결과를 통해 몇 가지 사항을 추론해 볼 수 있다 (Figure 3.3). 선택된 변수가 해당 연

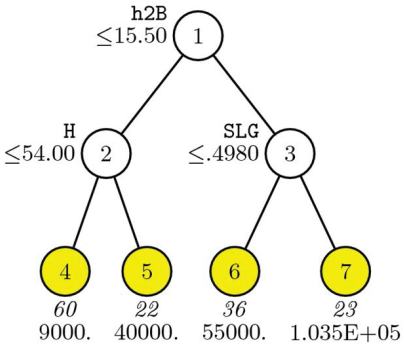

Figure 3.1. GUIDE 0.50-SE piecewise constant 0.90-quantile regression tree for predicting salary of batters.

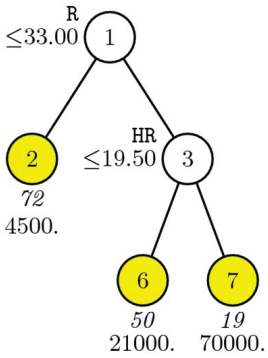

Figure 3.2. GUIDE 0.50-SE piecewise constant 0.50-quantile regression tree for predicting salary of batters.

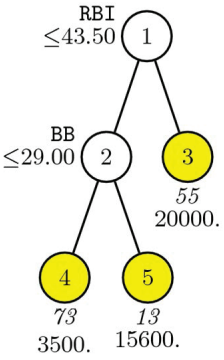

Figure 3.3. GUIDE 0.50-SE piecewise constant 0.25-quantile regression tree for predicting salary of batters.

봉 분위수를 설명하는 독립변수라는 점을 고려하면, 높은 수준의 연봉을 받는 타자들은 안타 수나 2루타 수 등 기본적인 타자의 능력에 더하여 장타율까지 연봉 수준과 관계가 있다고 볼 수 있다. 중위수 정도의 연봉을 받는 선수들은 득점과 홈런이 변수로 선택되었고 하위 25퍼센트 정도 연봉을 받는 타자들은 타점과 볼넷이 의미 있는 변수로 선택되었다. 홈런이 선택된 점을 제외하면 대체로 고연봉에 비해 상대적으로 출루가 빈번한 지가 연봉에 영향을 미치고 있는 것으로 해석할 수 있다.

Table 3.2. Quantile regression models after variable selection (batters)

Selected variables	$\alpha = 0.90$	$\alpha = 0.50$	$\alpha = 0.25$
Constant	−274.99	2,500.00	1,378.20
H	82.96	-	-
h2B	1,931.10	-	-
SLG	42,073.00	-	-
R	-	166.67	-
HR	-	1,570.70	-
RBI	-	-	177.25
BB	-	-	146.65
Sample quantile	60,000	12,000	4,500

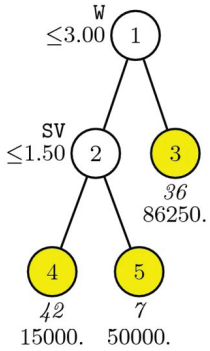

Figure 3.4. GUIDE 0.50-SE piecewise constant 0.90-quantile regression tree for predicting salary of pitchers.

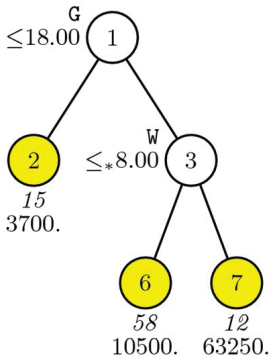

Figure 3.5. GUIDE 0.50-SE piecewise constant 0.50-quantile regression tree for predicting salary of pitchers.

이와 같은 과정을 통해 얻은 변수선택 결과를 반영하여 분위수 회귀나무 모형을 적합한 결과는 Table 3.2와 같다. 모두 나무구조가 형성되지 않고 단일 분위수 회귀모형으로 적합되었다.

한편, 투수의 경우 분위수 회귀나무를 통해 얻은 결과는 타자의 분석결과와 약간 상이하였다. $\alpha = 0.9$인 경우 W(승리 수), SV(세이브 수) 등의 변수가 선택되었고 $\alpha = 0.5$일 때에는 W(승리 수)와 G(출전 게임 수)가 선택되었다 (Figure 3.4, Figure 3.5). 한편, $\alpha = 0.25$인 경우에는 선택된 변수가 없었다. 투수의 경우에는 분위수 별로 뚜렷한 차이는 보이지 않았지만, $\alpha = 0.9$인 경우 W(승리 수)와 SV(세이

Table 3.3. Quantile regression models after variable selection (pitchers)

Selected variables	$\alpha = 0.90$	$\alpha = 0.50$	$\alpha = 0.25$
Constant	14,000	3,570.3	6,200
W	6,568.2	3,429.6	-
SV	3,751.4	-	-
G	-	8.64	-
Sample quantile	69,000	11,000	6,200

브 수)가 모두 선택된 것은 최근 선발, 중간 계투, 마무리 등 투수의 임무가 분업화되면서 각 분야에서 뛰어난 성적을 보이는 선수들이 고액 연봉을 받는 사례가 나타난 데 기인한다고 할 수 있다.

투수의 경우에도 변수선택 결과를 반영하여 분위수 회귀나무 모형을 적합하였으며 그 결과는 Table 3.3에 나타나 있다. 이 경우에도 모두 나무구조가 형성되지 않고 단일 분위수 회귀모형으로 적합되었다.

4. 결론

분위수 회귀분석법은 독립변수들이 주어졌을 때, 종속변수의 조건부 분위수에 초점을 맞추어 독립변수들과 종속변수의 해당 특정 분위수와의 관계를 분석하는 방법이다. 선형계획법 등 분위수 회귀에 있어서 추정의 방법들을 감안하면 대용량 자료 분석이 필요할 경우에는 적절한 모형 적합이 쉽지 않다는 점은 자명하다. 따라서 이러한 경우, 차원 축소의 문제가 중요한 과제로 부여된다. 본 논문에서는 이러한 변수선택의 문제를 회귀나무 방법을 이용하여 직관적이고도 이해하기 쉽게 구현해 보았다. 곡률검정에 바탕을 두고 있는 조각별 선형 회귀나무의 틀을 유지하되 분위수 회귀 방법을 응용하여 변수 선택의 문제를 해결하고자 하였다. 한국야구위원회에 등록된 선수들의 자료를 바탕으로 분위수 회귀에 관한 변수선택을 실시한 결과, 타자와 투수의 경우 모두 분위수 회귀 나무를 통해 독립변수의 수를 상당히 축소할 수 있는 것으로 나타났다. 분위수에 따라 차이는 있지만, 총 19개(타자) 또는 14개(투수) 중 2–3개의 변수만을 선택하여 모형 축소가 가능하였다. $\alpha = 0.9$인 경우 타자의 분석결과를 보면, H(안타 수), h2B(2루타 수), SLG(장타율) 등의 세 개의 변수가 선택되어 고액 연봉자의 경우 연봉에 미치는 주요 요인들로 나타났고, 투수의 경우 W(승리 수)와 SV(세이브 수)가 선택되어 최근투수의 임무가 분업화되면서 각 분야에서 뛰어난 성적을 보이는 선수들이 고액 연봉을 받고 있다는 사실을 뒷받침해 주었다. 실제 데이터를 분석하면서 주요 변수가 선택되는 등 의미 있는 결과를 얻었으나 본 연구는 다소 한계점을 지니고 있다. 회귀나무 자체의 단점이기도 하지만, 분위수 회귀나무 적용에 있어서도 수많은 변형이 가능하여 불안정성이 상존한다는 점이 제약으로 존재한다. 또한 실생활에서 접할 수 있는, 상대적으로 독립변수의 수가 많고 구성도 복잡한 실제 데이터를 분석하였으나 차원의 수가 매우 큰 고차원 자료의 경우에는 동 연구 결과의 적용 가능성을 평가할 필요가 있다. 따라서 향후 강건성을 고려하는 동시에 변수의 수가 크게 증가함에 따른 변수 선택의 성능 변화 등에 대해서도 연구를 확장하는 등 심도 있는 검토가 필요하다 하겠다.

References

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*, CRC press.

Chang, Y. (2010). The analysis of factors which affect Business Survey Index using regression trees, *The Korean Journal of Applied Statistics*, **23**, 63–71.

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

Chang, Y. and Kim, H. (2011). Tree-Structured Nonlinear Regression, *The Korean Journal of Applied Statistics*, **24**, 759–768.

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

Koenker, R. and Bassett, G. (1978). Regression Quantiles, *Journal of Econometrica*, **46**, 33–50.

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

분위수 회귀나무를 이용한 변수선택 방법 연구

장영재[a,1]

[a]한국방송통신대학교 정보통계학과

(2016년 8월 22일 접수, 2016년 10월 8일 수정, 2016년 10월 8일 채택)

요 약

Koenker 등 (1978)에 의해 제안 된 분위수 회귀분석법은 독립변수들이 주어졌을 때, 종속변수의 조건부 분위수에 초점을 맞추어 독립변수들과 종속변수의 해당 특정 분위수와의 관계를 분석하는 방법이다. 선형프로그래밍법 등을 이용한 분위수 회귀의 추정 과정을 생각해 볼 때, 고차원 대용량 자료의 경우에는 모형 적합에 어려움을 겪을 수 밖에 없다. 따라서 분위수 회귀의 문제에 있어서도 차원 축소의 문제, 조금 더 폭을 좁혀 생각해보면 변수선택의 문제를 통해 의사 결정에 영향을 미치는 주요 요인들을 파악하거나 적절한 규모의 모형을 적합하는 과정이 중요하다고 할 수 있다. 본 논문에서는 분위수 회귀의 변수선택의 문제를 보다 직관적이고 간단하게 해결하기 위한 방법으로서 회귀나무 모형을 응용하여 한국야구위원회에 등록된 선수들의 연봉과 기록 데이터를 분석해 보았다. 분석 결과, 각 분위수 별로 소수의 주요 변수가 선택되어 차원축소의 효과를 얻을 수 있었다. 또한 해당 분위수별로 선택된 변수도 해석상 의미 있는 것으로 평가할 수 있었다.

주요용어: 변수선택, 분위수 회귀, 회귀나무.

이 논문은 2015년도 한국방송통신대학교 학술연구비 지원을 받아 작성된 것임.

[1](03087) 서울시 종로구 대학로 86, 한국방송통신대학교 정보통계학과. E-mail: yjchang@knou.ac.kr