The Korean Journal of Applied Statistics (2020) DOI: https://doi.org/10.5351/KJAS.2020.36.6.753
33(6), 753–762

Analysis of AI interview data using unified non-crossing multiple quantile regression tree model

Jaeoh Kim[a] · Sungwan Bang[b,1]

[a]Center for Army Analysis and Simulation, ROK Army HQs;
[b]Department of Mathematics, Korea Military Academy

(Received August 4, 2020; Revised September 21, 2020; Accepted September 21, 2020)

Abstract

With an increasing interest in integrating artificial intelligence (AI) into interview processes, the Republic of Korea (ROK) army is trying to lead and analyze AI-powered interview platform. This study is to analyze the AI interview data using a unified non-crossing multiple quantile tree (UNQRT) model. Compared to the UNQRT, the existing models, such as quantile regression and quantile regression tree model (QRT), are inadequate for the analysis of AI interview data. Specially, the linearity assumption of the quantile regression is overly strong for the aforementioned application. While the QRT model seems to be applicable by relaxing the linearity assumption, it suffers from crossing problems among estimated quantile functions and leads to an uninterpretable model. The UNQRT circumvents the crossing problem of quantile functions by simultaneously estimating multiple quantile functions with a non-crossing constraint and is robust from extreme quantiles. Furthermore, the single tree construction from the UNQRT leads to an interpretable model compared to the QRT model. In this study, by using the UNQRT, we explored the relationship between the results of the Army AI interview system and the existing personnel data to derive meaningful results.

Keywords: quantile regression, quantile regression tree, unified non-crossing multiple quantile regression tree, AI interview

1. 서론

분위수 회귀(quantile regression)는 조건부 평균을 추정하는 ordinary least squares (OLS)에 비해 이 상치(outlier)에 상대적으로 영향을 덜 받는 강건한 방법이다 (Koenker와 Bassett, 1978). 또한 연구 자가 관심있는 상위 또는 하위 분위수별로 설명변수와 반응변수간 인과관계를 탐색할 수 있는 점에서 매우 유용하다. 분위수 회귀는 생존분석(survival analysis) 자료에 대한 모형 (Portnoy, 2003; Luo 등, 2013; Sun 등, 2016), 경시적 자료(longitudinal data)를 분석하기 위한 모형 (Wang과 Fygenson, 2009; Farcomeni, 2012) 등과 같이 다양한 분야로 연구가 이루어졌으며, 본 연구에서는 선형가정(linearity assumption)을 완화한 회귀나무모형(regression tree model)으로의 확장에 주목한다.

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NO. 2020R1F1A1A01065107).
[1]Corresponding author: Department of Mathematics, Korea Military Academy, 574, Hwarang-ro, Nowon-gu, Seoul 01805, Korea. E-mail: wan1365@gmail.com

의사결정나무(decison tree)는 분류(classification)와 회귀(regression)에 모두 가능한 방법으로 불순도(impurity)가 감소하도록 나무구조의 모형을 성장시키며 적절한 규칙으로 가지치기(pruning)를 하는 방법이다. 이 방법은 단순한 구조로 해석이 용이하며 자료에 대한 훌륭한 시각화(visualization)을 제공하여 높은 설명력(interpretation)을 갖는다. 또한 선형성(linearity), 정규성(normality) 및 등분산성(homoskedasticity) 가정이 필요하지 않은 비모수적 방법이다.

분위수 회귀에 대해 Chaudhuri와 Loh (2002)는 의사결정나무 구조로 확장한 방법(quantile regression tree; QRT)을 제안하였다. 이 방법은 나무구조의 장점을 활용하여 분위수 함수를 추정하는 장점이 있다. 그러나 분위수별 교차하는 문제, 극단 분위수에서 불안정한 결과를 갖는 문제, 분위수별로 서로 다른 나무모형을 구성하여 종종 해석이 난해한 문제 등을 내재한다.

이후 Kim 등 (2019)은 연구자가 관심있는 여러 분위수 함수에 대해 나무구조를 동시에 추정하는 방법을 제안하였다. Unified non-crossing multiple quantile regression tree model (UNQRT)은 비교차(non-crossing) 제약식을 부여한 상태로 하나의 나무구조에서 비교차 다중 조건부 분위수 함수를 동시에 추정하는 방법이다. UNQRT는 비교차 제약식을 활용한 다중 조건부 분위수 함수를 추정함으로써 극단 분위수에서 불안정한 결과를 갖는 문제와 분위수 함수별 교차하는 문제를 효과적으로 해결한다. 또한 연구자가 관심있는 분위수에 대해 하나로 통합된 나무모형을 제시함으로서 의사결정나무의 해석력을 극대화할 수 있는 방법이다.

한편 대한민국 육군은 미래 첨단과학기술군을 이끌어갈 우수인재 선발을 위해 인공지능(artificial intelligence; AI) 면접체계를 시범적으로 적용하고 있다. AI 면접체계는 최근 채용 절차의 공정성과 효율성을 향상하기 위하여 많은 기업에서 도입하고 있으며 국방부 '4차 산업혁명 스마트 국방혁신'의 세부사업 중 하나로 육군이 선도적으로 추진하고 있다. 육군은 2022년부터 간부선발 전 과정에 AI 면접체계를 도입하는 것을 목표로 추진 중이며 장교 및 부사관 장기복무 선발, 육사생도 선발 및 전문학위 위탁교육 선발 등에 시범적용하며 체계를 보완하고 정확도를 검증하고 있다.

본 연구의 목적은 육군이 AI 면접체계를 전면 도입하기 위해 19년 부사관 장기복무 선발시 시범적용한 자료를 UNQRT를 활용하여 분석하는 것이다. 이 자료는 AI 면접체계를 통해 대상자별로 산출된 종합점수와 병과, 인사평가, 교육성적 등 대상자별 인사자료의 일부로 구성되어 있다. 이러한 자료에 대해 본 연구에서는 AI 면접체계 결과의 고득점자 및 저득점자 등과 기존 인사자료로 판단할 수 있는 개인별 특성간의 인과관계를 분위수 함수를 활용하여 규명하고자 한다. 이를 위해 비교차 다중 조건부 분위수 함수를 동시 추정하여 하나의 통합된 나무구조로 나타내는 UNQRT를 활용한다. UNQRT는 우수한 예측력과 더불어 극단 분위수에서도 비교적 안정적인 결과를 보여주며, 분위수를 모두 통합한 하나의 나무구조로 자료를 표현하여 해석력이 뛰어난 점에서 본 자료 분석에 충분히 타당하다고 할 수 있다.

본 연구의 구성은 다음과 같다. 제 2절에서 QR, QRT 및 UNQRT에 대해 설명하고, 제 3절에서 육군의 AI 면접 관련 자료에 대해 소개한다. 이어서 제 4절에서 이를 분석한 결과를 제시한다. 마지막 제 5절에서 결론 및 향후 연구를 제안한다.

2. 모형 소개

2.1. 분위수 회귀모형

p차원 설명변수 $\boldsymbol{x} = (x_1, \ldots, x_p)^T$와 반응변수 $y \in R$로 이루어진 표본의 크기가 n인 자료 $\{x_i, y_i\}_{i=1}^n$을 가정하자. 이때 $0 < \tau_1 < \tau_2 < \cdots < \tau_K < 1$에 대해 반응변수 Y의 조건부 분위수 함수(conditional quantile

function) $q_\tau(y|\boldsymbol{x})$는

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

과 같다. Koenker과 Bassett (1978)는 체크 손실함수(check loss function) $\rho_\tau(u) = u(\tau - I(u<0))$를 이용하여 선형 분위수 함수 $q_\tau(\boldsymbol{x}) = \mathbf{x}^T\boldsymbol{\beta}_\tau$를

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

과 같이 추정하는 것을 제안하였으며, 여기서 $\mathbf{x} = (1, \boldsymbol{x}^T)^T$이며 $\beta_\tau = (\beta_{0,\tau}, \beta_{1,\tau}, \ldots, \beta_{p,\tau})^T$는 $100\tau\%$ 분위수 함수의 회귀계수 벡터이다.

2.2. 분위수 회귀나무모형

식 (2.1)의 조건부 분위수 함수는 회귀나무모형의 임의의 노드 t에서 각각의 조건부 분위수 함수 $q_\tau(\boldsymbol{x}, t)$

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

로 확장될 수 있다. 이때 $\boldsymbol{\beta}_\tau(t)$는 노드 t에서의 회귀계수 벡터이고 $\tilde{T}$은 회귀나무모형에서 최종노드(terminal node)의 총 개수이다.

Chaudhuri와 Loh (2002)는 분위수 회귀나무모형 구축을 위해 조건부 분위수 함수의 조각별 다항 추정(piecewise polynomial estimate)을 제안하였다. 이 방법은 Breiman 등 (1984)이 제안한 조각별 상수 중위수 회귀나무(piecewise constant median regression tree)에 대비하여 다음과 같은 장점이 있다. 먼저, Breiman 등 (1984)의 방법은 종종 나무모형의 크기가 매우 커져 해석에 어려움이 있는 반면, 조각별 다항 추정은 각 노드에 적합되는 다항식을 조절함으로서 나무모형의 크기를 변경할 수 있다 (Chaudhuri와 Loh, 2002). 둘째, 상수항을 적합하는 조각별 상수 중위수 회귀나무에 비해 다항식을 적합함으로서 추정의 정확도를 높일 수 있다.

Chang (2016)은 고차원 대용량 자료에 대한 분위수 회귀 모형의 변수 선택의 해결 방법으로 회귀나무 모형을 적용하였다. 이를 통해 각 분위수 별로 소수의 주요 변수를 선택함으로서 차원을 적절하게 축소하는 효과와 함께 해석력을 향상하였다.

2.3. 비교차 다중 분위수 회귀나무모형

Kim 등 (2019)은 일반적인 분위수 회귀모형의 선형 가정을 완화한 상태로 설명변수가 반응변수의 조건부 분위수 함수에 어떻게 관계되는지 탐색하기 위하여 통합 비교차 다중 분위수 회귀나무모형을 제안하였다. 이 방법은 조건부 분위수 함수의 조각별 다항 추정시 발생하는 분위수별 교차하는 문제, 극단 분위수에서 불안정한 결과를 갖는 문제, 분위수별로 서로 다른 나무모형을 구성하여 해석이 난해한 문제를 효과적으로 해결한다. 또한 UNQRT의 계산 알고리즘에서는 Breiman 등 (1984)의 classification and regression tree (CART) 알고리즘에서 종종 발생하는 과도한 계산비용과 분할 가능점(또는 분할집합)이 많은 경우 분할변수 선택의 편향이 발생하는 것에 대해 generalized, unbiased, interaction detection and estimation (GUIDE) 알고리즘 (Loh, 2002)을 적용하여 개선하였다. GUIDE 알고리즘은 잔차와 설명변수간 관계를 통계적 방법

을 통해 검정함으로서 분할변수를 먼저 선택하고, 선택된 분할변수에 대해서만 분할점(또는 분할집합)을 선택하는 방법이다.

Kim 등 (2019)의 UNQRT 알고리즘을 요약하면 다음과 같다.

2.3.1. **적합식** UNQRT는 나무모형의 각 노드 t에서 다중 분위수 함수를 동시에 추정하는 다음과 같은

$$\min_{\beta_{\tau_1}(t),\cdots,\beta_{\tau_K}(t)} \sum_{k=1}^{K} w(k) \sum_{i\in t} \rho_{\tau_k}\left(y_i - x_i^T \beta_{\tau_k}(t)\right), \tag{2.4}$$
$$\text{subject to } x^T\left(\boldsymbol{\beta}_{\tau_{k+1}}(t) - \boldsymbol{\beta}_{\tau_k}(t)\right) \geq 0 \quad \text{for } k = 1,2,\dots,K-1 \text{ and } \forall \mathbf{x} \in [0,1]^p ,$$

최적화 모형을 제안하였다. 식 (2.4)에서는 분위수별 비교차 제약식을 부여하고, $100\tau_k\%$ 분위수의 가중치 $w(k)$를 목적함수에 포함하였다. $\phi(\cdot)$과 $\Phi(\cdot)$을 표준정규분포의 확률밀도함수와 누적분포함수로 정의할 때 $E[\rho_\tau(\varepsilon - \Phi^{-1}(\tau))] = \phi(\Phi^{-1}(\tau))$ 이므로 체크 함수의 기댓값은 중위수(median)에서 가장 크고 극단 분위수로 갈 수록 감소하게 된다 (Liu와 Wu, 2011). Kim 등 (2019)은 $w(k) = 1/\phi(\Phi^{-1}(\tau))$를 부여하여 극단 분위수를 무시하고, 중위수 부근의 분위수에 지나치게 의존하는 문제를 해결하였다.

식 (2.4)에서 비교차 제약식은 $2^p(K-1)$개가 되어 효율적인 계산 알고리즘이 요구된다. UNQRT에서는 Bondell 등 (2010)이 제안한 슬랙변수(slack variable) $\alpha_{j,k}^{+}(t) \geq 0$와 $\alpha_{j,k}^{-}(t) \geq 0$를 이용한 재모수화(reparameterization)를 통해 다음과 같은

$$\begin{aligned} \arg\min_{\beta_{\tau_k}} \min_{(t)} & \sum_{k=1}^{K} w(k) \sum_{i\in t} \rho_{\tau_k}\left(y_i - \beta_{0,\tau_k}(t) - \sum_{j=1}^{p} x_{ij}\beta_{j,\tau_k}(t)\right) \\ \text{subject to} \quad & \beta_{j,\tau_{k+1}}(t) - \beta_{j,\tau_k}(t) = \alpha_{j,k}^{+}(t) - \alpha_{j,k}^{-}(t) \\ & \beta_{0,\tau_{k+1}}(t) - \beta_{0,\tau_k}(t) - \sum_{j=1}^{p} \alpha_{j,k}^{-}(t) \geq 0 \\ & \alpha_{j,k}^{+}(t) \geq 0 \quad \text{and} \quad \alpha_{j,k}^{-}(t) \geq 0, \quad \text{for } j=1,\dots,p \text{ and } k=1,\dots,K-1. \end{aligned} \tag{2.5}$$

최적화 모형을 각 노드 t의 적합식으로 활용하였다.

2.3.2. **분할변수 선택** UNQRT에서는 CART와 같이 분할변수의 모든 분할점(또는 분할집합)에 대해 적합하는 것이 아니라, 먼저 분할변수를 선택한 뒤 선택된 분할변수에 대해서만 분할점(또는 분할집합)을 선택하는 2단계 알고리즘을 적용한다. 먼저 분할변수 Z의 선택을 위해 다음과 같은

$$i(t) = \sum_{k=1}^{K} w(k) \sum_{i\in t} \rho_{\tau_k}\left(e_{i,\tau_k}(t)\right). \tag{2.6}$$

노드 t에서 불순도 함수 $i(t)$를 정의한다. 이때 $e_{i,\tau_k}(t) = y_i - \mathbf{x}_i^T\widehat{\boldsymbol{\beta}}_{\tau_k}(t)$로 노드 t에서 k번째 분위수 τ_k의 i번째 잔차(residual)이다. UNQRT에서는 이 잔차와 분할 후보변수간 무작위성(randomness)을 측정하여 가장 무작위성이 낮은 변수를 분할변수로 선택한다. 이를 위하여 각 분위수 τ_k에 대해 잔차를 부호로 구분하

고 분할 후보변수를 범주 또는 적절한 구간 L (사분위수를 기준으로 하여 통상 4)로 구분하여 $2 \times L$ 분할표(contingency table)를 구성한다. 식 (2.5)에서 비교차 다중 분위수 함수를 동시에 추정한 뒤 그 잔차를 분위수별 분할표에 활용하므로 이를 하나의 분할표로 통합하여 기존의 극단 분위수에서의 불안정한 문제를 극복한다 (Kim 등, 2019). 분할 후보변수별로 구성되는 분할표에 대해 명목(nominal) 변수간 관계(association)를 측정하는 불확실성 계수(uncertainty coefficient); (Theil, 1970)로 무작위성을 측정하여 가장 높은 값을 갖는 변수를 최종 분할변수로 선택한다. Chaudhuri와 Loh (2002)에서 분위수별로 분할변수를 선택하는 것은 종종 해석의 어려움을 주는 반면, UNQRT는 합리적인 방법으로 모든 분위수를 통합하여 하나의 분할변수를 선택하므로 해석이 있어 장점을 갖는다. 분할점(또는 분할집합)을 선택하는 알고리즘은 Kim 등 (2019)에 상세히 기술되어 있으므로 생략한다.

2.3.3. 나무크기 결정 나무모형의 크기를 결정하는 것은 과적합(overfitting) 및 과소적합(underfitting) 문제와 연관되어 중요하다. UNQRT는 사전-가지치기(pre-pruning)와 사후-가지치기(post-pruning) 알고리즘을 동시에 적용한다. 먼저 사전-가지치기 알고리즘으로 불확실성 계수, 나무의 크기, 최종 노드의 표본개수의 사전 정해진 하한값에 도달하면 나무모형의 성장을 정지한다. 다음 사후-가지치기를 위해 M-단계 알고리즘을 적용한다. M-단계 알고리즘은 M개 하위 노드의 불순도 감소량이 유의한 수준에 도달하는지 탐색하여 나무모형의 성장 여부를 결정하는 것이다. UNQRT는 사전-가지치기와 사후-가지치기 알고리즘을 동시에 적용하여 과적합과 과소적합 문제를 경감한다.

3. AI 면접체계 및 관련 분석자료

AI 면접체계는 지원자의 표정, 음성, 맥박, 안면 변화 등 생리적 반응을 기반으로 하는 외적 요소와 기존 조직내 고성과자의 데이터와 지원자의 질의응답 결과를 분석한 내적 요소를 종합분석하여 지원자의 선발 여부에 대한 의사결정을 지원하는 것이다. AI 면접은 자기소개 및 지원동기 등 기본질문, 개인성향을 판단하는 탐색질문, 특정 상황에 대한 대응능력을 확인하는 상황질문, 간단한 뇌과학 게임과 개인별 맞춤형 심층질문 순으로 진행하며 대략 1시간 정도 소요되는 것이 일반적이다. 많은 기업과 정부기관에서 선발과정에서 면접에 작지 않은 어려움을 호소하고 있다. 특히, 대면면접은 면접관의 주관성에 따라 평가점수의 큰 차이가 발생할 가능성이 높기 때문에 같은 기준으로 선발기관이 추구하는 인재상에 부합하는 선발을 하길 원할 것이다. 선발의 공정성과 효율성을 증대하기 위해 ’19년도 약 80여개 기관이 AI 면접을 도입하였다.

육군도 ‘4차 산업혁명 스마트 국방혁신’을 구현하기 위한 과제로 AI 면접체계 도입을 정책적으로 결정하였다. 육군이 시범적으로 적용하고자 하는 AI 면접체계는 ‘신뢰’, ‘전략’, ‘관계’, ‘실행’, ‘가치’, ‘조직적합’, ‘호감도’의 7개 세부 역량을 측정하며 종합점수를 제시한다. 이 체계는 장교 및 부사관의 장기복무 선발, 전문학위 교육선발, 초임 부사관 선발 등에 적용하고자 준비 중이며, 본 연구에서는 부사관 장기복무 선발을 위해 시범적용한 573명의 자료를 분석한다. 분석의 목적은 육군이 도입하고자 하는 AI 면접체계와 육군이 관리하는 개인별 인사자료간의 관계를 분위수별로 탐색하는 것이다. 이를 위해 본 연구에서는 AI 면접체계의 log(종합점수)를 반응변수로 설정한다. 이것은 장기복무 선발을 최종 판단할 수 있는 중요한 변수이다. 설명변수는 육군이 관리하는 개인별 인사자료의 일부로 범주형 변수인 ‘병과(BRC)’, ‘세부 군사특기(MOS)’와 ‘추천여부(RCD)’이며, 연속형 변수인 ‘인사평가(PE)’, ‘교육성적(EDU)’, ‘체력(PHY)’, ‘자기개발(SFD)’와 ‘대면면접 결과(FIR)’이다. 육군의 병과는 다양하게 구성되어 있으나, 본 연구에서는 ‘전투병과(C)’, ‘전투지원병과(CS)’, ‘전투근무지원병과(CSS)’로 구분하였다. 군사특기는 전투임무 ‘1’, 전투

지원임무 '2', 행정임무 '3', 기타임무 '4'로 분류하였다. 인사평가는 근무간 년 2회 상급자로부터 평가받은 인사고과를 점수화 한 것이다. 교육성적, 체력, 자기개발은 육군에서 공통적으로 관리하는 요소를 기준으로 점수화하여 분석에 활용하였다. 이 분석을 위해 적용하는 UNQRT는 관심있는 분위수에 대해 동시추정한 비교차 다중 조건부 분위수 함수와 하나의 나무모형을 제공하여 높은 예측력과 우수한 해석력을 기대할 수 있다.

4. 분석 결과

본 절에서는 제 3절에서 소개한 자료에 대한 분석 결과를 제시한다. 분석을 위한 적합변수는 '대면면접 결과(FIR)'이며 나무모형을 구성하기 위한 분할변수는 '병과(BRC)', '세부 군사특기(MOS)', '인사평가(PE)', '교육성적(EDU)', '추천여부(RCD)', '체력(PHY)'과 '자기개발(SFD)'이다. 먼저 QRT가 추정한 조건부 분위수 함수와 UNQRT의 동시추정 비교차 다중 조건부 분위수 함수를 비교하고, 두 방법을 통해 구축한 나무모형의 결과를 살펴본다. 이어서 10-겹 교차검증(10-fold cross-validation)방식으로 모형의 성능을 평가한다. 본 논문에서는 Kim 등 (2019)이 'supplemetary materials'에서 UNQRT 알고리즘을 구현하여 제공하는 프로그램을 분석에 활용하였다.

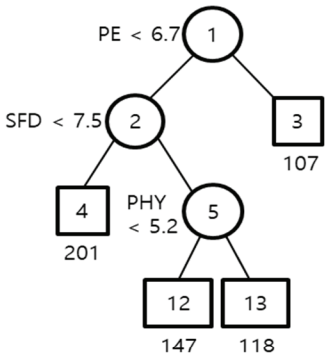
4.1. 분위수 함수 추정 및 나무모형 구축 결과

Figure 4.1은 QRT와 UNQRT로 나무모형을 구축한 결과이다. QRT는 각 분위수별로 서로 상이한 나무모형을 구축한 것을 확인할 수 있다. 특히, 0.1 분위수에서는 '인사평가(PE)', '자기개발(SFD)'과 '체력(PHY)'이 중요한 변수지만, 다른 분위수에서는 이 변수들이 분할변수로 선택되지 않았다. 이러한 현상은 0.3과 0.7 분위수의 분할변수에서도 동일하게 발생하였다. QRT를 통해 분위수별로 구축된 나무모형의 해석은 각 분위수별로 선택된 분할변수가 매우 상이하여 해석이 용이하지 않다. 더군다나 일반적으로 병과가 매우 중요한 변수임에도 불구하고 0.1 분위수에서는 이를 분할변수로 사용하지 않았으며 0.9분위수에서도 7번 노드에 이르러서야 사용하는 등 해석이 극단 분위수에서 불안정함을 보였다. 반면, UNQRT는 비교차 다중 조건부 분위수 함수를 동시에 추정하여 하나의 통합된 나무모형을 제시한다. UNQRT에서는 '병과'로 먼저 분할한 뒤 전투병과의 경우 '추천여부', '인사평가'와 '자기개발'로 전투지원 및 전투근무지원병과의 경우 '추천여부', '교육성적'을 분할변수로 선택한다. 군 인사분야 전문가 의견에 따르면 병과는 가장 지배적인 변수로 볼 수 있으며, 지휘관에 의한 추천 여부는 병과에 관계없이 장기복무 선발에 중요한 요소이다. 이러한 사항을 고려할 때 UNQRT는 타당한 해석을 제공했다고 할 수 있다. 또한 장기복무 선발에 있어 전투병과와 비전투병과의 추천 이후 어떠한 변수가 중요한 영향을 미치는지 알 수 있는 점도 매우 흥미롭다.

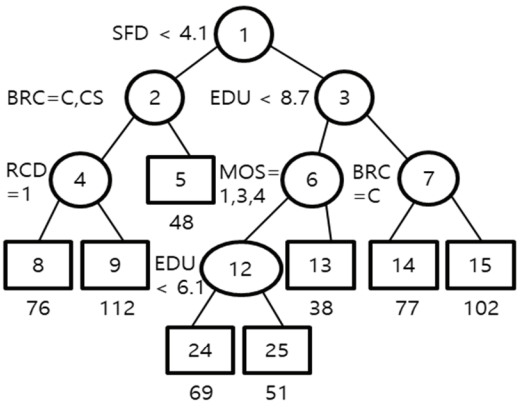
Figure 4.2는 5개 분위수(0.1, 0.3, 0.5, 0.7, 0.9) 함수에 대해 QRT와 UNQRT의 추정된 결과이다. 지면을 절약하기 위해 최종 노드에서 적합된 결과는 생략한다. QRT의 경우 분위수별로 교차하는 현상이 다수 발생함을 알 수 있으며 이것은 분위수 함수의 가정을 위배하는 것이다. 반면, UNQRT는 분위수별 비교차된 추정 결과를 보인다.

4.2. 예측력 비교

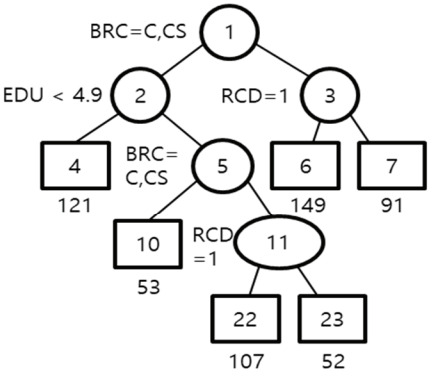
QRT와 UNQRT의 성능을 비교하기 위해 주어진 자료에 대해 5-겹 교차검증을 실시한다. 이때 시험



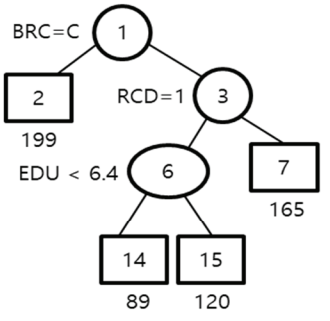
QRT at τ=0.1



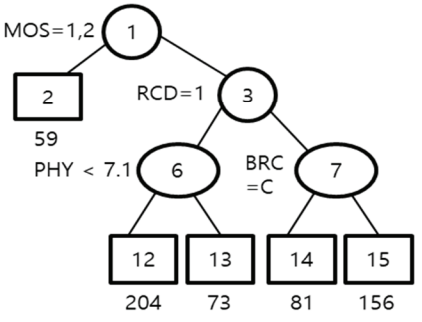
QRT at τ=0.3



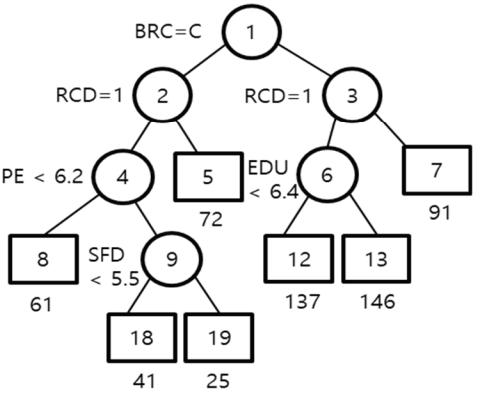
QRT at τ=0.5



QRT at τ=0.7

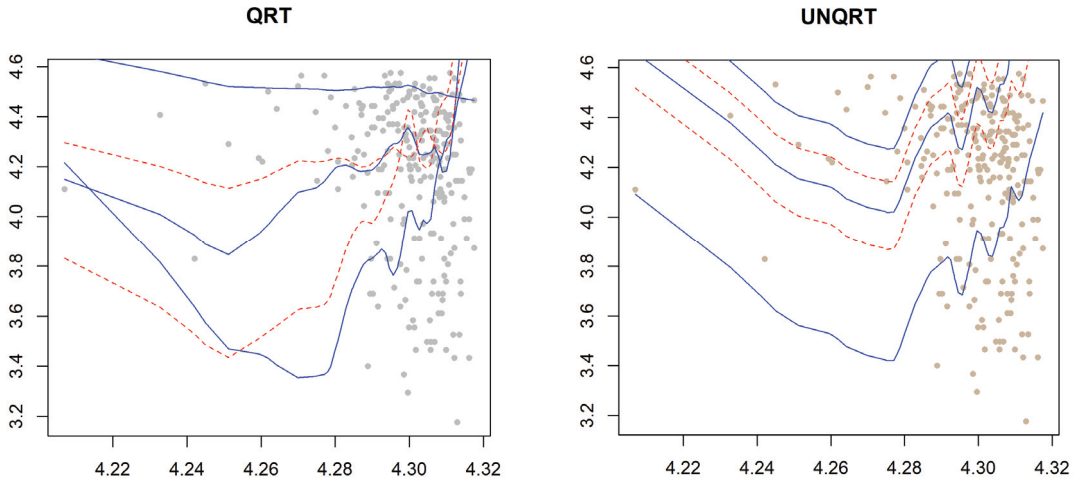


QRT at τ=0.9



UNQRT at τ=0.1, 0.3, 0.5, 0.7, 0.9

Figure 4.1. The tree models from QRT and UNQRT for the AI interview data. The numbers within circles or squares are node numbers and the numbers beneath each terminal node are the node sample sizes. The observation goes to the left if the condition is satis?ed at the intermediate nodes; otherwise, it goes to the right.



(a) Estimated quantile functions by QRT (b) Estimated quantile functions by UNQRT

Figure 4.2. Fitted quantile functions of the log result of AI interview against log FIR for the AI interview data. The locally weighted scatterplot smoothing is used for effective visualization. The left and right panels show the QRT and UNQRT estimates, respectively. The blue solid lines represent the estimated quantile functions at $\tau = 0.1$, 0.5, and 0.9, and the dashed red lines represent the estimated quantile functions at $\tau = 0.3$ and 0.7.

데이터에 대해 다음과 같은

$$i(t) = \sum_{k=1}^{K} w(k) \sum_{i \in t} \rho_{\tau_k}\left(e_{i,\tau_k}(t)\right). \tag{4.1}$$

평균예측오차(mean prediction error)를 측정한다.

Table 4.1은 각 방법에 대해 분위수별 평균예측오차를 독립적으로 100번 반복한 결과이다. UNQRT는 모든 분위수에서 QRT보다 우수한 성능을 보였으며, 특히 극단인 0.1 분위수에서 상대적으로 QRT에 비해 월등히 우수하였다.

5. 결론 및 향후연구

조건부 평균함수를 추정하는 OLS에 대비하여 조건부 분위수 함수를 추정함으로서 분위수별로 설명변수와 반응변수가 인과관계를 탐색 할 수 있는 분위수 회귀모형은 매우 유용한 통계적 방법임이 분명하다. 그러나 자료의 형태가 다양해지고 연구자가 자료로부터 더 많은 정보를 얻고자 함에 따라 이러한 전통적인 분위수 회귀는 적절하지 않은 경우가 종종 발생한다. 특히 선형성에 대한 가정은 많은 경우에 지나치게 강한 가정으로 작용하여 적절한 자료 분석을 하지 못할 수 있다. 본 연구에서는 나무모형과 결합된 분위수 회귀 방법인 UNQRT를 소개하고 이를 실제 자료에 적용하였다. UNQRT는 기존 분위수 회귀를 나무모형으로 결합한 방법에 대비하여 많은 장점을 가진다. UNQRT는 비교차 제약식을 부여한 상태로 다중 분위수 함수를 동시에 추정함으로서 분위수 함수의 교차 문제를 해결하며, 극단 분위수에서 안정된 결과를 기대할 수 있고, 하나의 통합된 나무모형을 제시하여 우수한 해석력이 있다.

Table 4.1. Prediction accuracy of the AI interview data of the ROK army

Method	Average	τ				
		0.1	0.3	0.5	0.7	0.9
QRT	3.035	4.225	3.776	2.169	2.339	2.664
	(0.441)	(0.525)	(0.477)	(0.390)	(0.407)	(0.428)
UNQRT	1.267	1.667	1.553	1.002	1.118	0.995
	(0.115)	(0.129)	(0.133)	(0.098)	(0.105)	(0.090)

The numbers in parentheses are standard errors.

본 연구에서 다룬 실제 자료는 육군이 AI 면접체계 도입을 위해 시범 적용하고 있는 자료의 일부로 4차 산업혁명을 대비하는 차원에서 국방부와 육군이 선도적으로 추진하는 핵심 사업에 관한 것이다. UNQRT를 활용하여 AI 면접체계의 결과와 기존 인사자료간 관계를 충분히 탐색하여 의미있는 다양한 결과를 도출하였다. 또한 QRT에 대비하여 우수한 예측력과 해석력을 입증하였다. AI 면접체계 결과가 선발 여부임을 고려할 때 득점 결과의 분위수별로 기존 육군이 유지하는 인사적 특성과 관계를 알 수 있었다. 이러한 결과는 AI 면접체계의 도입을 위한 정책적 보완사항을 도출하기 위한 기초 자료로 의미가 있다.

References

Bondell, H. D., Reich, B. J., and Wang, H. (2010). Noncrossing quantile regression curve estimation, *Biometrika*, **97**, 825–838.

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

Chang, Y. (2016). Variable selection with quantile regression tree, *The Korean Journal of Applied Statistics*, **29**, 1095–1106.

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

Farcomeni, A. (2012). Quantile regression for longitudinal data based on latent Markov subject-specific parameters, *Statistics and Computing*, **22**, 141–152.

Kim, J., Cho H., and Bang, S. (2019). Unified noncrossing multiple quantile regressions tree, *Journal of Computational and Graphical Statistics*, **28**, 454–465.

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

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.

Luo, X., Huang, C. Y., and Wang, L. (2013). Quantile regression for recurrent gap time data, *Biometrics*, **69**, 375–385.

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

Sun, X., Peng, L., Huang, Y., and Lai, H. J. (2016). Generalizing quantile regression for counting processes with applications to recurrent events, *Journal of the American Statistical Association*, **111**, 145–156.

Theil, H. (1970). On the estimation of relationships involving qualitative variables, *American Journal of Sociology*, **76**, 103–154.

Wang, H. J. and Fygenson, M. (2009). Inference for censored quantile regression models in longitudinal studies, *The Annals of Statistics*, **37**, 756–781.

통합 비교차 다중 분위수회귀나무 모형을 활용한 AI 면접체계 자료 분석

김재오[a] · 방성완[b,1]

[a]육군본부 분석평가단, [b]육군사관학교 수학과

(2020년 8월 4일 접수, 2020년 9월 21일 수정, 2020년 9월 21일 채택)

요 약

본 연구는 대한민국 육군이 선도적으로 도입하고자 노력하고 있는 AI 면접체계의 자료를 통합 비교차 다중 분위수 회귀나무 모형(unified non-crossing multiple quantile tree; UNQRT)을 활용하여 분석한 것이다. 분위수 회귀가 일반적인 선형회귀에 비하여 많은 장점을 가지지만, 선형성 가정은 여전히 많은 현실 문제해결에 있어 지나치게 강한 가정이다. 선형성을 완화한 모형의 하나인 기존 나무모형 기반의 분위수 회귀는 추정된 분위수 함수별로 교차하는 문제와 분위수별로 나무모형을 제시하여 해석력을 저하시키는 문제가 있다. 통합 비교차 다중 분위수회귀나무 모형은 비교차 제약식을 부여한 상태로 다중 분위수 함수를 동시에 추정함으로서 분위수 함수의 교차 문제를 해결하며, 극단 분위수에서 안정된 결과를 기대할 수 있고, 하나의 통합된 나무모형을 제시하여 우수한 해석력이 있다. 본 연구에서는 통합 비교차 다중 분위수회귀나무 모형을 활용하여 육군 AI 면접체계의 결과와 기존 인사자료간 관계를 충분히 탐색하여 의미있는 다양한 결과를 도출하였다.

주요용어: 분위수 회귀모형, 분위수 회귀나무모형, 비교차 분위수 회귀나무모형, 인공지능 면접체계

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NO. 2020R1F1A1A01065107).

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