The Korean Journal of Applied Statistics (2018) DOI: http://dx.doi.org/10.5351/KJAS.2018.31.3.343
31(3), 343–352

Comparison of estimation methods for expectile regression

Jong Min Kim[a] · Kee-Hoon Kang[a,1]

[a]Department of Statistics, Hankuk University of Foreign Studies

(Received February 23, 2018; Revised April 8, 2018; Accepted April 30, 2018)

Abstract

We can use quantile regression and expectile regression analysis to estimate trends in extreme regions as well as the average trends of response variables in given explanatory variables. In this paper, we compare the performance between the parametric and nonparametric methods for expectile regression. We introduce each estimation method and analyze through various simulations and the application to real data. The nonparametric model showed better results if the model is complex and difficult to deduce the relationship between variables. The use of nonparametric methods can be recommended in terms of the difficulty of assuming a parametric model in expectile regression.

Keywords: cross-validation, nonparametric method, parametric method, P-spline, quantile regression

1. 서론

전통적인 회귀분석은 설명변수가 주어졌을 때 반응변수의 평균적인 추세가 어떻게 변하는지를 모형화한다. 가장 간단한 단순선형회귀분석의 경우에는 설명변수 X가 주어졌을 때 반응변수 Y의 조건부 평균을 선형관계로 가정하고 추정하는 것을 목표로 한다. 하지만, 경제, 금융, 의학 등의 분야에서는 X가 주어졌을 때 Y의 평균만이 아니라 Y의 극단적인 영역에서 추세를 추정하고 싶은 경우가 빈번히 발생한다. 예를 들어, 아동의 영양실조에 관해 연구한 Fenske 등 (2011), 프론티어 추정을 다룬 Schnabel과 Eilers (2009) 등을 참고할 수 있다. 이렇게 회귀분석은 설명변수의 주어진 수준에서 반응변수의 분포를 보다 면밀하게 살펴보는 방법으로 최근에 확장되어 왔다.

반응변수 Y의 꼬리 부분과 같은 극단적인 영역에서의 추세를 추정하는 방법에는 일반적으로 분위수 회귀분석(quantile regression)과 평률 회귀분석(expectile regression)이 있다. 분위수 회귀분석은 X가 주어졌을 때, Y의 조건부 분위수 탐색하기 위하여 체크함수(check function)로 L_1 거리를 이용한다. 분위수 회귀분석은 다양한 분야에 활용되었으나 Newey와 Powell (1987)은 목적함수가 미분이 불가능하고, 정규분포 같은 오차 분포의 경우에 덜 효율적이며 공분산의 계산이 어렵다는 단점을 지적하였다. 그들은 이에 대한 대안으로 L_2 거리를 체크함수로 이용하는 평률 회귀분석을 제안하였다. 평률 회귀분석은 최소제곱법을 추정에 이용할 수 있기 때문에 계산 및 이론적인 특성의 유도 등이 용이한 장점이 있

This research was supported by Hankuk University of Foreign Studies Research Fund of 2018.
[1]Corresponding author: Department of Statistics, Hankuk University of Foreign Studies, 81, Oedae-ro, Mohyeon-eup, Cheoin-gu, Yongin-si, Gyeonggi-do 17035, Korea. E-mail: khkang@hufs.ac.kr

다. 이는 분위수의 특별한 경우인 중앙값과 평률의 특별한 경우인 평균과의 성질을 비교하는 것에서도 짐작될 수 있다.

분위수 회귀의 해석은 상대적으로 쉽다고 할 수 있다. τ-분위수 회귀함수 q_τ는 데이터의 $100\tau\%$가 q_τ보다 적고 데이터의 $100(1-\tau)\%$가 q_τ보다 위에 있다는 것을 의미한다. 이에 비해 평률의 해석은 그리 직관적이지 않다. Yao와 Tong (1996)에 의하면 τ-평률 회귀함수 m_τ와 반응변수의 관측치 사이 거리의 $100\tau\%$가 m_τ 아래의 데이터에 의해 결정되고, $100(1-\tau)\%$가 m_τ 위의 부분에 의해 주어진다. 즉, 분위수 회귀분석은 관측치가 추정할 회귀함수 위 혹은 아래에 있는지에 관한 위치 정보만을 이용하고, 평률 회귀분석은 추정할 회귀함수와 관측치의 거리의 비율을 이용한다. 이 점에서 Newey와 Powell (1987)은 평률 회귀분석이 분위수 회귀분석에 비해 주어진 데이터를 더 효율적으로 사용한다고 하였다.

평률 회귀분석에 관한 최근 연구로는 Schnabel과 Eilers (2009), De Rossi와 Harvey (2009), Sobotka와 Kneib (2012), Guo와 Härdle (2012), Sobotka 등 (2013), Jiang 등 (2017), Spiegel 등 (2017), Zhao와 Zhang (2018)이 있다. 본 논문에서는 평률 회귀분석을 이용하여 주어진 설명변수의 수준에서 반응변수의 극단적인 영역의 추세를 파악함에 있어 모수적, 비모수적 추정 방법을 살펴보고 성능의 차이를 비교해 보고자 한다. 2절에서는 평률 회귀분석의 소개와 함께 모수적, 비모수적 추정 방법을 소개한다. 3절에서의 다양한 모형에 대한 모의실험과 4절에서의 실제자료 분석을 통해 방법들 간의 성능을 비교하고, 5절에서는 간략한 결론을 제시한다.

2. 본론

2.1. 평률 회귀분석의 소개

분위수 회귀분석은 설명변수 X가 주어졌을 때 분포함수가 $F_Y(\cdot)$로 주어진 반응변수 Y의 $100\tau\%$ 조건부 분위수 $Q_Y(\tau|x) = F_Y^{-1}(\tau|x)$의 추정을 목표로 한다. 전통적인 선형회귀분석에서는 조건부 기댓값인 $E(Y|\boldsymbol{x}) = \boldsymbol{x}^t\boldsymbol{\beta}$에서 $\boldsymbol{\beta}$를 추정하기 위해 최소제곱법을 이용한다. 분위수 회귀분석에서는 모수적 방법인 선형회귀분석 방법으로 추정할 때 조건부 분위수를 식 (2.1)과 같이 모형화한다.

$$Q_Y(\tau|\boldsymbol{x}) = \boldsymbol{x}^t\boldsymbol{\beta}_\tau. \tag{2.1}$$

분위수 회귀분석에서 $\boldsymbol{\beta}_\tau$의 추정은 다음의 최소화 문제를 만족하는 해로부터 구할 수 있다.

$$\min_{\boldsymbol{\beta}_\tau \in R^p} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}^t\boldsymbol{\beta}_\tau), \tag{2.2}$$

여기서 $\rho_\tau(\cdot)$는 체크함수로 $\rho_\tau(u) = u(\tau - I(u<0))$, $\tau \in (0,1)$이다. 식 (2.2)의 해는 식 (2.3)과 같이 나타내어질 수 있다.

$$\hat{\boldsymbol{\beta}}_\tau = \arg\min_{\boldsymbol{\beta}_\tau} \left\{ \sum_{i\,:\,y_i \geq \boldsymbol{x}_i^t\boldsymbol{\beta}_\tau} \tau \left|y_i - \boldsymbol{x}_i^t\boldsymbol{\beta}_\tau\right| + \sum_{i\,:\,y_i < \boldsymbol{x}_i^t\boldsymbol{\beta}_\tau} (1-\tau)\left|y_i - \boldsymbol{x}_i^t\boldsymbol{\beta}_\tau\right| \right\}. \tag{2.3}$$

식 (2.3)의 경우 목적함수가 미분이 불가능하기 때문에 선형계획법(linear programming)을 활용하여 최적화하는 것이 가능하다. 추정치를 얻는 과정에서 목적함수를 모든 점에서 미분이 가능하도록 하여 계산을 편리하게 하고, 공분산 계산의 용이성을 위해 대안으로 나온 것이 평률 회귀분석이다 (Newey와 Powell, 1987). 평률 회귀분석은 L_1 거리가 아닌 L_2 거리에 기반한 다음의 체크함수 $\tilde{\rho}_\tau(\cdot)$를 이용한다.

$$\tilde{\rho}_\tau(u) = \begin{cases} \tau u^2, & u \geq 0, \\ (1-\tau)u^2, & u < 0. \end{cases} \tag{2.4}$$

평률 회귀분석에서 모수적 방법의 예로 선형모형을 적용시키면 조건부 평률함수 $m_Y(\tau|\boldsymbol{x}) = \boldsymbol{x}^t\tilde{\boldsymbol{\beta}}_\tau$를 고려할 수 있다. 식 (2.4)의 체크함수를 이용하여 여기서 평률함수는 잔차들의 비대칭 제곱합을 최소로 하는 것을 통해 추정할 수 있다. 이를 식으로 나타내면 식 (2.5)와 같다.

$$\hat{\tilde{\boldsymbol{\beta}}}_\tau = \arg\min_{\tilde{\boldsymbol{\beta}}_\tau} \left\{ \sum_{i\,:\,y_i \geq \boldsymbol{x}_i^t\tilde{\boldsymbol{\beta}}_\tau} \tau \left(y_i - \boldsymbol{x}_i^t\tilde{\boldsymbol{\beta}}_\tau\right)^2 + \sum_{i\,:\,y_i < \boldsymbol{x}_i^t\tilde{\boldsymbol{\beta}}_\tau} (1-\tau) \left(y_i - \boldsymbol{x}_i^t\tilde{\boldsymbol{\beta}}_\tau\right)^2 \right\}. \tag{2.5}$$

분위수와 평률은 서로 다르지만 밀접한 연관이 있다. Yao와 Tong (1996)은 분위수는 분포함수에 의해 결정되지만 평률은 꼬리 부분의 기댓값에 의해 결정됨을 지적하였다. 여러 평률에 대한 회귀함수를 추정했을 때 그것들이 서로 평행하면 동질적임을 의미하고, 또한 이웃하는 낮은 평률 곡선과 높은 평률 곡선 간의 차이를 비교하면 치우침의 정도를 탐지할 수 있다고 하였다. 원칙적으로는 X가 주어졌을 때 반응변수의 조건부 분포는 평률들을 이용하여 설명될 수 있다. 보다 자세한 평률의 의미를 위해서는 Waltrup (2014), Yang과 Zou (2015)을 참고하면 된다.

2.2. 비모수적 평률 회귀분석 방법

모수 모형은 설명변수 X와 반응변수 Y간의 관계를 특정 형태를 가정하여 그에 따른 모수를 추정하는 방법이다. 이 방법은 가정된 모형이 데이터에 잘 부합할 경우 좋은 성능을 보이지만 그렇지 못할 경우 큰 문제가 생길 수 있다. 뿐만 아니라 설명변수와 반응변수 간의 평률함수 관계를 파악하기 힘든 경우가 많기 때문에 형태에 대한 가정을 하지 않고 모형을 추정하는 방법인 비모수적 접근법을 고려할 수 있다. 회귀분석에서 일반적인 비모수 모형은 식 (2.6)과 같다.

$$y_i = f(x_i) + \varepsilon_i, \tag{2.6}$$

여기서 회귀함수 f는 미분가능하고 충분히 부드러운 함수이며, $i = 1, \ldots, n$에 대해 $E(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$이라고 가정한다. f를 추정하는 방법에는 여러 가지가 있지만 비모수적으로 평률 회귀함수 $m_Y(\tau|x) = f(x)$를 추정하기 위해서는 P-스플라인(P-spline)과 비대칭 최소제곱법을 결합한 방법을 이용하고, 다변량의 경우에는 가법모형을 이용한다.

P-스플라인은 B-스플라인을 통해 소개할 수 있다. B-스플라인은 반응변수와 설명변수 사이의 관계에 대하여 매우 유연한 모델링을 가능하게 해주는 기저함수를 이용한 비모수적 방법이다. 기본적인 원리는 매듭(knot)을 통해 자료 전체를 구간으로 나누고 각 매듭과 매듭 사이의 구간을 기저 함수를 통해 추정하는 것이다. 이를 모형식으로 나타내면 식 (2.7)과 같다.

$$f(x) = \sum_{j=1}^{K} u_j B_j(x). \tag{2.7}$$

이 선형결합에서 K는 기저함수 개수를, B_j는 j번째 B-스플라인 기저함수를, u_j는 j번째 기저함수의 계수를 의미한다.

B-스플라인에 관한 내용은 De Boor (2001), Eilers와 Marx (1996)과 Fahrmeir 등 (2013)에 자세히 소개되어 있으며, 매듭을 t_j, $j = 1, \ldots, K-1$이라 할 때, 차수가 0인 B-스플라인 기저함수는 다음과 같이 정의된다.

$$B_j^0(x) = \begin{cases} 1, & t_j \leq x \leq t_j + 1\text{일 때,} \\ 0, & \text{그 밖의 경우.} \end{cases}$$

또한, 차수가 $p \geq 1$인 B-스플라인 기저함수는 $p-1$차 B-스플라인으로부터 점화식에 의해 다음과 같이 정의된다.

$$B_j^p(x) = \frac{x - t_{j-p}}{t_j - t_{j-p}} B_{j-1}^{p-1}(x) + \frac{t_{j+1} - x}{t_{j+1} - t_{j+1-p}} B_j^{p-1}(x).$$

식 (2.7)의 선형결합은 실험 설계점(design point) $x_1, \ldots, x_n$이 정해지면 설계 행렬(design matrix) $\boldsymbol{B}$와 계수 벡터 $\boldsymbol{u}$를 통해 $\boldsymbol{Bu}$로 나타 낼 수 있다. B-스플라인의 좋은 성질 중 하나는 B-스플라인의 추정된 계수가 함수 f에 대한 추정뿐만 아니라 f의 도함수도 추정할 수 있다는 것이다. 차수가 p인 B-스플라인을 사용하면 (2.8)과 같이 도함수를 추정할 수 있고, 이에 대한 증명은 Fahrmeir 등 (2013)을 참고하면 된다.

$$\frac{\partial f(x)}{\partial x} = p \sum_j \frac{\triangle u_j}{t_j - t_{j-1}} B_{j-1}^{p-1}(x), \quad \triangle u_j = u_j - u_{j-1}. \tag{2.8}$$

B-스플라인에서 기저함수의 계수 벡터 $\boldsymbol{u}$의 추정치는 식 (2.9)를 계수 벡터 $\boldsymbol{u}$에 대해 최소화함으로써 얻을 수 있다.

$$(\boldsymbol{y} - \boldsymbol{Bu})^t (\boldsymbol{y} - \boldsymbol{Bu}). \tag{2.9}$$

식 (2.9)의 계수 벡터 $\boldsymbol{u}$에 대한 최소화는 일반적인 회귀분석에서 회귀계수의 추정과 같이 정규방정식을 이용하여 다음과 같이 구할 수 있다.

$$\hat{\boldsymbol{u}} = (\boldsymbol{B}^t \boldsymbol{B})^{-1} (\boldsymbol{B}^t \boldsymbol{y}).$$

B-스플라인에서는 매듭의 수와 그 배치를 어떻게 선택하느냐에 따라 추정된 결과가 크게 영향을 받는다. 이러한 문제를 해소하기 위한 일환으로 제안된 방법이 P-스플라인이다. P-스플라인은 벌점화된(penalized) 스플라인으로 B-스플라인의 매듭의 수 결정과 이에 따른 매듭의 배치에 따라 발생할 수 있는 문제를 해소하기 위해 두 가지 절차를 이용한다. 첫 번째는 많은 매듭의 수를 전 영역에 동일한 간격으로 배치하는 것이다. 여기서 매듭의 수는 Fahrmeir 등 (2013)이 20에서 40개 사이에서 선택하는 것이 가장 효과적이라는 것을 제시하였다. 두 번째는 Eilers와 Marx (1996)가 지나치게 거친 형태로 선이 추정되는 것을 방지하기 위해 패널티 부분을 도입함으로써 선의 부드러움을 조절하였다. Eilers와 Marx (1996)는 함수의 곡률을 측정하기 위해 2차 차분(second order differences)을 제안하였고 이를 기반으로 한 벌점 부분이 식 (2.10)이다.

$$\lambda \sum_{j=3}^{K} (\triangle^2 u_j)^2, \quad \triangle^2 u_j = u_j - 2u_{j-1} + u_{j-2}, \tag{2.10}$$

여기서 λ는 조절모수를 의미하며, λ의 추정은 일반적으로 교차확인법(cross-validation)을 통해 이루어진다. 벌점 행렬을 구축하기 위해서 식 (2.11)과 같은 $(K-2) \times K$ 차원 행렬 $\boldsymbol{D}$를 이용할 수 있다.

$$\boldsymbol{D} = \begin{pmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -2 & 1 \end{pmatrix}. \tag{2.11}$$

벌점인 식 (2.10)은 행렬 $\boldsymbol{P} = \boldsymbol{D}^t\boldsymbol{D}$라 할 때, 식 (2.12)의 우변과 같이 나타낼 수 있다.

$$\lambda \sum_{j=3}^{K} (\triangle^2 u_j)^2 = \lambda \boldsymbol{u}^t \boldsymbol{P} \boldsymbol{u}. \tag{2.12}$$

P-스플라인에서 기저함수 계수 벡터 $\boldsymbol{u}$의 추정은 B-스플라인에서 추정을 위한 식 (2.9)를 최소화하는 대신 식 (2.12)가 벌점으로 추가된 식 (2.13)을 최소화하는 해를 구하면 된다.

$$(\boldsymbol{y} - \boldsymbol{B}\boldsymbol{u})^t (\boldsymbol{y} - \boldsymbol{B}\boldsymbol{u}) + \lambda \boldsymbol{u}^t \boldsymbol{P} \boldsymbol{u}. \tag{2.13}$$

식 (2.13)을 미분하고 정규방정식을 풀면 $\boldsymbol{u}$의 추정치는 식 (2.14)와 같이 구할 수 있다.

$$\hat{\boldsymbol{u}} = (\boldsymbol{B}^t\boldsymbol{B} + \lambda \boldsymbol{P})^{-1} \boldsymbol{B}^t \boldsymbol{y}. \tag{2.14}$$

비모수적 평률 회귀분석은 P-스플라인과 비대칭 제곱합을 최소로 하는 방법을 결합하여 이루어진다. 이는 식 (2.15)와 같은 목적 함수를 만들어내고 이를 추정하고자 하는 평률 τ에 대해 최소로 하는 계수 벡터 $\boldsymbol{u}_\tau$를 찾으면 τ-평률 회귀함수를 추정할 수 있다.

$$\sum_{i=1}^{n} \tilde{\rho}_\tau (y_i - Bu_i) + \lambda_\tau \boldsymbol{u}_\tau^t \boldsymbol{P}_\tau \boldsymbol{u}_\tau. \tag{2.15}$$

Waltrup (2014)은 일반적으로 비모수적으로 평률 회귀함수를 추정할 때 20개의 기저함수를 사용하고 각 기저함수는 3차함수를 이용할 것을 제안하였다.

3. 모의실험

모의실험은 오픈 소스인 `R`을 이용하여 진행하였으며 모수적, 비모수적 평률 회귀분석의 적합을 위해 `R`에 있는 `expectreg` 패키지의 `expectreg.ls` 함수를 사용하였다. 모의실험을 위한 회귀함수의 형태는 다음과 같이 총 4개의 모형을 고려하였다. 모수적 방법으로 평률 회귀모형을 적합시키기 위해서는 먼저 자료의 산점도를 살펴보고 그 자료에서 설명변수 X의 수준에서 반응변수 Y의 조건부 평률이 어떠한 형태인지 유추해서 모형을 적합시켜야 한다. 이러한 유추가 산점도로부터 쉽게 가능하지는 않기에 조건부 평균을 유추하기 쉬운 자료부터 비교적 복잡한 형태라 유추하기 어려운 자료까지를 포함하도록 설명변수와 오차항의 범위 및 분포를 달리하여 사용한 것이다.

모형 1. $Y_i = 4 + 3X_i + \varepsilon_i,\ X_i \sim U(-1, 1),\ \varepsilon_i \sim \chi^2(2)$

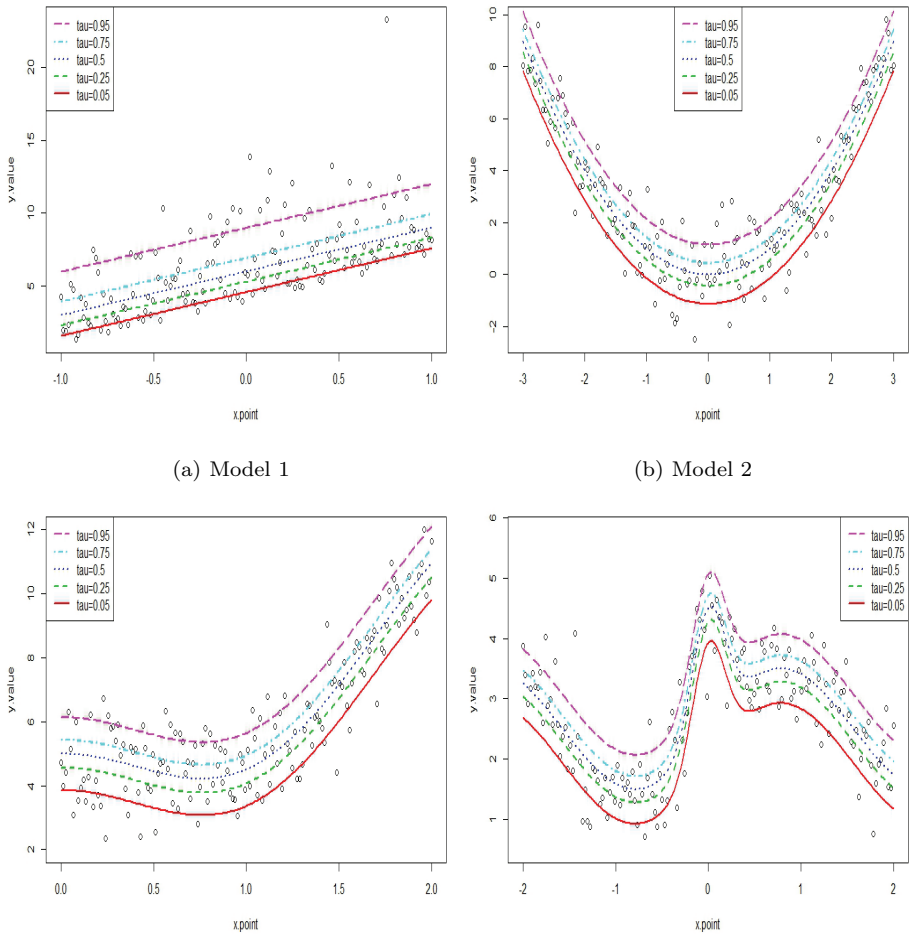
모형 2. $Y_i = X_i^2 + \varepsilon_i,\ X_i \sim U(-3, 3),\ \varepsilon_i \sim N(0, 1)$

모형 3. $Y_i = 4 + \cos(3X_i) + 1.5X_i^2 + \varepsilon_i,\ X_i \sim U(0, 2),\ \varepsilon_i \sim N(0, 1)$

모형 4. $Y_i = 2.5 + \sin(2X_i) + 2e^{-16X_i^2} + 0.5\varepsilon_i,\ X_i \sim U(-2, 2),\ \varepsilon_i \sim N(0, 1)$

모형 1과 2는 설명변수와 반응변수의 관계가 비교적 간단한 경우이며, 모형 3과 4는 어떠한 관계가 있는지 알기가 쉽지 않은 자료이다. Figure 3.1은 각각의 모형에서 100개의 데이터를 생성하여 그린 산점도와 참 평률 회귀함수를 함께 그린 것이다.

모수 모형의 선택은 자료의 산점도를 보고 결정하였으며 모형 1, 2에서는 각각 1차 함수, 2차 함수를 적합시켰고 모형 3에는 2차 함수와 지수함수를 적용시켜 보았다. 또한, 모형 4에서는 3차 함수와 sin함수를 적합시켰다. 비모수 모형의 경우 P-스플라인 방법을 통해 적합시켰으며 기저함수는 20개를 사용하



(a) Model 1 (b) Model 2

(c) Model 3 (d) Model 4

Figure 3.1. Scatter plot for each model and the true expectile regression functions.

고 각 기저함수는 3차함수를 사용하였다. 그리고 조절모수 λ는 교차확인법을 이용하여 구하였다. 데이터의 수가 500개이고, 반복수가 200번일 경우 몇 가지 평률에 대해 모수 모형과 비모수 모형을 비교한 결과는 Table 3.1과 같다. 표에 나온 값들은 각 평률에 대한 평균제곱의 차를 적분하여 구한 평균적분제곱오차(mean integrated squared error; MISE)값이며, 그 표준오차(standard error; SE)를 괄호 안에 나타내었다. 모형 3과 4의 모수적 방법에서 왼쪽 열은 각각 2차 함수와 3차 함수를 적합시킨 결과이며, 오른쪽 열은 각각 지수함수와 sin함수를 적합시킨 결과이다.

Table 3.1의 모의실험 결과 참 평률 함수의 모양에 따라 모수적 방법과 비모수적 방법의 성능이 차이를 보였다. 산점도를 통해 두 변수가 어떤 관계인지 유추할 수 있는 비교적 간단한 경우인 모형 1과 2에서는 모수적 방법이 비모수적 방법에 비해 더 좋은 성능을 보였다. 이 경우에는 모수 모형으로 참 모형을 사용한 것이므로 예상할 수 있는 결과이다. 하지만 모형 3, 4와 같이 두 변수 간의 관계를 유추하기 힘든 경우 비모수적 방법이 모수적 방법보다 더 좋은 성능을 보였다. 모형 3에서 보듯이 모수 모형의 선택

Table 3.1. MISE comparison of parametric and nonparametric expectile estimation methods

Model	τ	Parametric method		Nonparametric method
		MISE (SE)		MISE (SE)
1	0.05	0.0057 (0.0045)		0.0127 (0.0050)
	0.50	0.0149 (0.0039)		0.0424 (0.0112)
	0.95	0.4171 (0.0324)		0.8073 (0.0573)
2	0.05	0.0527 (0.0051)		0.0726 (0.0091)
	0.50	0.0102 (0.0031)		0.0294 (0.0078)
	0.95	0.0206 (0.0661)		0.0612 (0.0099)
		Quadratic fitting	Exponential fitting	
3	0.05	0.0875 (0.0014)	0.1638 (0.0018)	0.0463 (0.0019)
	0.50	0.0744 (0.0006)	0.1479 (0.0007)	0.0281 (0.0011)
	0.95	0.0874 (0.0014)	0.1659 (0.0018)	0.0482 (0.0022)
		Cubic fitting	Sine fitting	
4	0.05	1.2250 (0.0038)	1.2209 (0.0029)	0.0810 (0.0018)
	0.50	1.1450 (0.0007)	1.1443 (0.0006)	0.0466 (0.0018)
	0.95	1.9106 (0.0110)	1.9079 (0.0166)	0.0731 (0.0019)

MISE = mean integrated squared error; SE = standard error.

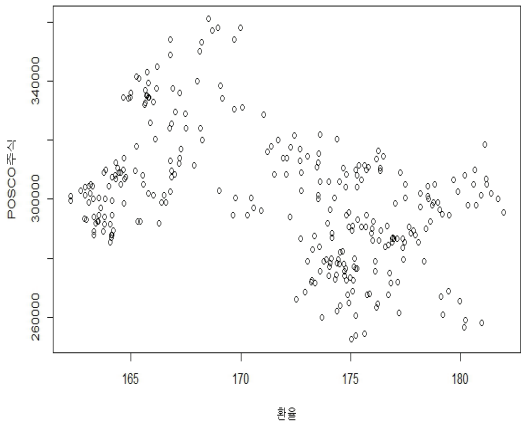


Figure 4.1. Scatter plot for RMB exchange rate and POSCO stock price.

에 따라 MISE값이 많이 달라지며 비모수 모형의 경우보다 월등히 크게 나타나는 경우가 있었다. 따라서, 모형 3, 4의 경우처럼 합리적으로 모수 모형을 유추하는 것이 쉽지 않은 경우에는 비모수 모형을 사용하는 것이 위험도 줄일 수 있고, 성능도 더 좋다고 할 수 있다.

4. 실제자료 분석

평률 회귀분석을 위한 실제자료는 2014년 1월 2일부터 2015년 3월 25일까지 총 300일간의 원/위원화 환율과 그에 영향을 많이 받는 업종 중 하나인 POSCO의 주가 자료이다 (자료 출처: http://finance.daum.net). POSCO는 해외에서도 특히 중국으로의 자동차 강판, 압연 등의 수출이 잦은 회사로 위안화 환율에 주가가 영향을 많이 받는다. 본 논문에서는 POSCO 주가(원)와 위안화 환율(원/위안화)의 관계에 대해서 평률 회귀분석을 시행하였으며 Figure 4.1은 두 변수의 산점도이다.

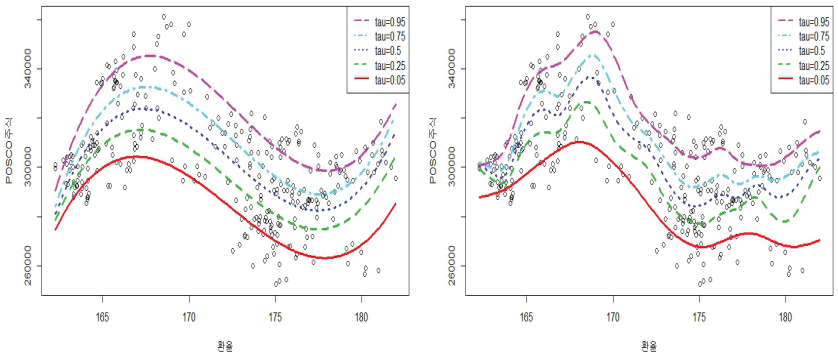


Figure 4.2. Result of fitting the parametric (left) and nonparametric (right) expectile regression for the RMB exchange rate and the POSCO stock price.

산점도를 보면 평균적으로는 환율이 올라가면 POSCO의 주가가 하락하는 것으로 보이나 두 변수간의 관계를 면밀하게 판단하기 위하여 환율(X)과 POSCO 주가(Y)의 관계를 1차식부터 4차식까지 이용하여 다항회귀분석을 시행하고 그 결과를 살펴보았다. 1차식부터 4차식까지 적합된 회귀식의 수정결정계수(adjusted R-square)는 각각 0.186, 0.189, 0.412, 그리고 0.416으로 3차식과 4차식간의 차이가 거의 없었다. 3차식 모형과 4차식 모형 모두 유의하였으나 수정결정계수, Akaike information criteria (AIC)와 Bayesian information criteria (BIC)를 모두 계산하여 모수 모형으로는 3차 함수 모형을 선택하여 평률 회귀분석을 실시하였다. 이와 같은 모형 선택의 절차를 피할 수 있다는 것이 비모수적 방법의 장점이며, 비모수 모형으로는 P-스플라인 방법으로 적합시켰다. Figure 4.2는 각각 환율과 POSCO 주가의 모수적 평률 회귀모형과 비모수적 평률 회귀모형을 적합시킨 결과이다.

POSCO가 환율에 따른 정책을 세우고 상황에 대비하기 위해서는 $\tau = 0.5$인 평균적인 추세를 기반으로 하기 보다는 Figure 4.2에서 $\tau = 0.05, 0.95$의 경우처럼 환율 변동에 따른 주가의 극단적인 추세를 기반으로 환율에 따른 정책을 세우는 것이 바람직할 것이다. 모수 모형에서는 환율이 높아지는 초기에 POSCO 주가가 높아지다가 167원 이후에는 낮아지는 추세이고 177원 이후로 급격히 상승하는 형태로 나왔다. 비모수 모형에서는 모수 모형에서와 마찬가지로 POSCO의 주가가 높아지다가 낮아지고 175원 이후로 현상 유지되다가 180원 이후에는 조금 상승하는 형태로 나왔다. 실제자료 분석에서는 참 평률함수를 알지 못하기 때문에 정확한 성능을 비교하기는 어렵지만, 조건부 평균에 해당되는 $\tau = 0.5$ 평률 회귀분석의 경우에 모수적, 비모수적 방법으로 구한 추정치의 MSE가 각각 2.795×10^8, 2.362×10^8으로 모수적 방법이 대략 18% 정도 크게 나타남을 알 수 있다. 위안화 가치가 하락하면 중국 철강업체들의 한국으로 수출이 촉진될 것이며 이에 따라 철강업체인 POSCO 주가는 하락할 것으로 예상될 수 있다. 하지만 자료 분석의 결과에 의하면 이러한 사항은 환율 변동의 구간에 따라서 다르고, 평률에 따라서도 달라질 수 있음을 확인하였다. 환율 상승에 따라 철강의 대중국 수출이 늘어날 수 있으므로 포스코 주가가 상승할 것을 기대할 수 있지만 모수적 방법의 경우처럼 그 폭이 급격하지는 않을 것으로 예상한다. 아울러, 비모수적 방법을 이용하면 환율이 165에서 167 사이, 175에서 180 사이의 구간에서처럼 주가의 평률이 국소적인 영역에서 변동하는 점도 확인할 수 있으므로, 이 자료에서는 모수적 방법보다 비모수적 방법에 의한 추세 예측에 따라 정책을 수립하는 것이 더욱 바람직하다고 판단된다.

5. 결론

본 논문에서는 자료의 극단적인 부분에서의 추세를 추정하기 위해 사용하는 평률과 평률 회귀분석에 대

해 알아보았다. 평률 회귀모형의 모수적 추정 방법과 비모수적 추정 방법, 특히 P-스플라인 방법에 대해 설명하였고 모의실험을 통해 다양한 상황에서의 두 방법의 성능을 평균제곱의 차를 적분한 MISE값을 통해 비교하였다. 모의실험 결과 참 평률 함수 모양이 간단하여 설명변수 X와 반응변수 Y의 관계를 비교적 쉽게 유추할 수 있는 경우에는 모수적 방법이 비모수적 방법보다 성능이 더 좋았다. 모수적으로 평률 회귀모형을 추정할 경우에는 평률에 따라 모수 모형의 성능에 차이를 보이는데 이는 각 평률에 따라 평률함수의 형태가 다를 수 있기 때문이다. 가정한 모수 모형이 자료에 잘 맞지 않는 경우에는 모수적 평률 회귀모형의 성능이 비모수적 평률 회귀모형에 비해 크게 떨어졌다. 따라서 평률 회귀분석을 할 경우에는 우선 산점도로부터 전체적인 평률 회귀함수의 형태를 대략적으로 파악하는 것이 우선이다. 개별 평률에 대한 관계의 유추가 쉽지 않기 때문에 이를 위해서 먼저 비모수적 방법으로 평률 회귀함수를 추정해 보는 것도 좋을 것이다. 파악된 평률 회귀함수의 형태가 간단하여 모수 모형이 적합한 경우에는 이 방법으로 평률 회귀모형을 적합시키고, 대략적인 형태가 복잡하여 설명변수와 반응변수 간의 관계를 유추하기 힘든 경우에는 비모수적 방법으로 평률 회귀모형을 적합시키는 것이 효율적인 접근법이라 할 수 있다.

References

De Boor, C. (2001). *A Practical Guide to Splines*, Springer-Verlag, New York.

De Rossi, G. and Harvey, A. (2009). Quantiles, expectiles and splines, *Journal of Econometrics*, **152**, 179–185.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties, *Statistical Science*, **11**, 89–121.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). *Regression : Models, Methods and Applications*, Springer-Verlag, Berlin Heidelberg.

Fenske, N., Kneib, T., and Hothorn, T. (2011). Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression, *Journal of the American Statistical Association*, **106**, 494–510.

Guo, M. and Härdle, W. K. (2012). Simultaneous confidence bands for expectile functions, *AStA - Advances in Statistical Analysis*, **96**, 517–541.

Jiang, C., Jiang, M., Xu, Q., and Huang, X. (2017). Expectile regression neural network model with applications, *Neurocomputing*, **247**, 73–86.

Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing, *Econometrica*, **55**, 819–847.

Schnabel S. K. and Eilers, P. H. C. (2009). An analysis of life expectancy and economic production using expectile frontier zones, *Demographic Research*, **21**, 109–134.

Sobotka, F., Kauermann, G., Waltrup, L. S., and Kneib, T. (2013). On confidence intervals for semiparametric expectile regression, *Statistics and Computing*, **23**, 135–148.

Sobotka, F. and Kneib, T. (2012). Geoadditive expectile regression, *Computational Statistics & Data Analysis*, **56**, 755–767.

Spiegel, E., Sobotka, F. and Kneib, T. (2017). Model selection in semiparametric expectile regression, *Electronic Journal of Statistics*, **11**, 3008–3038.

Waltrup, L. S. (2014). *Extensions of semiparametric expectile regression* (Ph.D. thesis), Ludwig Maximilians University Munich.

Yang, Y. and Zou, H. (2015). Nonparametric multiple expectile regression via ER-Boost, *Journal of Statistical Computation and Simulation*, **85**, 1442–1458.

Yao, Q. and Tong, H. (1996). Asymmetric least squares regression estimation: a nonparametric approach, *Journal of Nonparametric Statistics*, **6**, 273–292.

Zhao, J. and Zhang, Y. (2018). Variable selection in expectile regression, *Communications in Statistics - Theory and Methods*, **47**, 1731–1746.

평률 회귀분석을 위한 추정 방법의 비교

김종민[a] · 강기훈[a1]

[a]한국외국어대학교 통계학과

(2018년 2월 23일 접수, 2018년 4월 8일 수정, 2018년 4월 30일 채택)

요 약

설명변수가 주어졌을 때 반응변수의 평균적인 추세뿐만 아니라 극단적인 지역에서의 추세에 대해서 추정하고 싶거나 반응변수 분포의 일반적인 탐색을 위해서는 분위수 회귀분석과 평률 회귀분석을 사용할 수 있다. 본 논문에서는 평률 회귀모형의 추정을 위한 모수적 방법과 비모수적 방법의 성능을 비교하고자 한다. 이를 위해 각 추정 방법을 소개하고 여러 상황의 모의실험 및 실제자료에의 적용을 통해 비교 분석을 실시하였다. 모형에 따라 성능 차이가 있는데 자료의 형태가 복잡하여 변수 간의 관계를 유추하기 힘들 경우 비모수적으로 추정한 평률 회귀분석모형이 더욱 좋은 결과를 보였다. 일반적인 회귀분석의 경우와 달리 평률의 경우 후보가 되는 모수 모형을 상정하기 어렵다는 측면에서 볼 때, 비모수적 방법의 사용이 추천될 수 있다.

주요용어: 교차 확인법, 모수적 방법, 분위수 회귀분석, 비모수적 방법, P-스플라인

이 연구는 2018학년도 한국외국어대학교 교내학술연구비의 지원에 의하여 이루어진 것임.

[1]교신저자: (17035) 경기도 용인시 처인구 모현읍 외대로 81, 한국외국어대학교 통계학과.
E-mail: khkang@hufs.ac.kr