碩士學位論文

# 분위수 회귀분석 추정의 모수적 비모수적 방법 비교

韓國外國語大學校 大學院

統 計 學 科

禹 松 濟

碩士學位論文

# 분위수 회귀분석 추정의 모수적 비모수적 방법 비교

Comparison of parametric and nonparametric quantile regression method

指導 姜 奇 勳 敎授

이 論文을 碩士學位請求論文으로 提出합니다.

2011年 月 日

韓國外國語大學校 大學院

統 計 學 科

禹 松 濟

이 論文을 禹松濟의 碩士學位 論文으로 認定함.

2011年　　月　　日

審査委員　＿＿＿＿＿＿＿＿　(인)

審査委員　＿＿＿＿＿＿＿＿　(인)

審査委員　＿＿＿＿＿＿＿＿　(인)

韓國外國語大學校 大學院

# 요 약

어떠한 확률변수의 분포를 자세히 알고 싶을 때나 대칭이 아닌 한쪽으로 치우쳐진 분포를 분석할 경우에는 회귀분석보다는 분위수 회귀분석을 고려하는 것이 일반적이다. 분위수 회귀분석 역시 회귀분석과 마찬가지로 모수적, 비모수적 추정 방법이 있는데 그 두 모형간의 비교에 대해서 연구한 논문은 아직까지 없었다. 본 논문에서는 모수적 모형으로 추정한 분위수 회귀분석 모형과 비모수적으로 추정한 분위수 회귀분석 모형의 성능의 차이를 비교하고자 한다. 이를 위해 다양한 모의실험을 시행하였고 실제자료를 이용한 분석도 실시하였다. 모형에 따라 성능이 다른데 결과적으로 복잡한 데이터일 경우에는 비모수적으로 추정한 분위수 회귀분석모형이 더 좋은 결과를 보였다.

**주요용어** : 분위수 회귀분석, 비모수적 커널 평활, 이중 커널 평활

# 목 차

# <표> 목 차



# [그 림] 목 차

# 1. 서론

일반적으로 회귀분석 함수는 변량 $X$가 주어졌을 때 반응변수인 $Y$의 평균이 어떻게 변하는가에 대해서 보여준다. 가장 간단한 단순 선형회귀분석의 경우 $X$와 $Y$의 관계는 선형으로 가정하며 $X$가 주어졌을 때 $Y$의 평균적인 값을 추정하는 것을 목표로 한다. 하지만 $X$가 주어졌을 때 $Y$의 평균이 아닌 중앙값을 추정하고 싶을 경우도 있을 것이다.

중앙값은 데이터를 크기의 순서대로 나열하였을 때 중앙에 위치한 값으로 정의되는데 총 데이터의 수 $n$이 홀수일 경우에는 $(n+1)/2$번째 크기의 값이고, $n$이 짝수일 경우에는 $n/2$번째와 $(n+2)/2$번째 크기의 값의 산술평균이 중앙값이 된다. 일반적으로 중앙값을 추정하면 그 값을 $m$이라고 두는데, $m$은 그 값보다 큰 값과 작은 값으로 데이터를 두 개로 나누게 된다. 만약 어떤 확률변수 $Y$의 중앙값이 $m$이라면 $P(Y \leq m) = P(Y \geq m) = 1/2$으로 동일하게 된다. 이 의미는 확률변수 $Y$의 누적분포함수 $F(y) = P(Y \leq y)$에서 $F(m) = 1/2$이라는 의미가 된다. 만약 사람들의 연봉과 같은 데이터가 있으면 그 모양은 정규분포처럼 어떤 값을 중심으로 좌우로 대칭인 형태가 아니라 돈을 많이 받는 쪽, 혹은 돈을 적게 받는 쪽으로 치우쳐진 형태를 보이는 것이 일반적이다. 분포가 대칭형태가 아니라 어느 한 쪽으로 치우쳐진 자료를 분석할 때는 평균보다 중앙값이 더욱 더 많은 정보를 주게 된다. 또한 어떠한 경우에는 $F(y) = 1/4$이나 $F(y) = 3/4$이 되는 $y$값에 더 흥미를 가지고 그 값이 더욱 많은 정보를 주게 될 때도 있을 것이다. $F(y) = 1/4$이나 $F(y) = 3/4$이 되는 $y$값을 각각 25% 분위수와 75%분위수라고 한다. 이것을 일

반화 시키면 $0 < p < 1$일 때 $100p\%$ 분위수는 $F(y) = p$가 되는 $y$값을 의미한다. 나이($X$)에 따른 어린이의 키($Y$)에 관한 자료를 분석하는 경우에는 나이별 아이들의 평균키 보다는 어떤 아이의 키가 그 나이에서 어느 정도에 위치하는지에 대해 더욱 관심이 있을 수 있다. 이런 경우 나이가 주어졌을 때 키의 분위수를 추정하게 되는데 이것이 바로 분위수 회귀분석이다.

본 논문의 순서는 다음과 같다. 2장에서는 분위수 회귀분석의 자세한 소개와 함께 모수적, 비모수적 분위수 회귀분석 방법을 소개한다. 3장에서는 여러 가지 모형에 대한 모의실험과 실제자료를 통해 추정법의 성능을 비교하고, 4장에서는 간단한 결론을 제시하고자 한다.

# 2. 분위수 회귀분석

## 2.1. 분위수 회귀분석의 소개

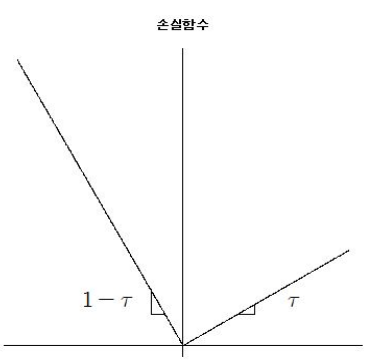
서론에서 간략히 소개 했지만 분위수 회귀분석은 독립변수 $X$가 주어졌을 때 반응변수 $Y$의 분위수를 추정하는 회귀분석 방법이고 이것을 모형식으로 표현하면 식 (2.1)과 같다.

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

식 (2.1)의 모형식에서 $\boldsymbol{\beta}(\tau)$의 추정을 어떻게 할 것인가? 전통적인 회귀분석에서의 $\beta$의 추정은 $\sum_{i=1}^{n}(y_i - \boldsymbol{x}^T\boldsymbol{\beta})^2$을 최소로 하는 문제를 풀면 되었다. 분위수 회귀분석에서 이 문제를 해결하기 위해 우선 다음과 같은 손실함수를 생각해 보자.

$$\rho_\tau(u) = u(\tau - I(u<0)),\ 0<\tau<1 \tag{2.2}$$

이 때 함수 $\rho_\tau(\cdot)$을 그림으로 그리면 [그림 2.1]과 같다.




[그림 2.1] 손실함수 $\rho_\tau(\cdot)$에 대한 그래프

어떤 확률변수 $X$와 $X$를 추정한 값 $\hat{x}$이 있을 때 식 (2.2)의 손실함수을 이용하여 $(X-\hat{x})$의 기댓값을 최소로 만드는 $\hat{x}$값을 찾아보자.

$$E\rho_\tau(X-\hat{x}) = (\tau-1)\int_{-\infty}^{\hat{x}}(x-\hat{x})dF(x) + \tau\int_{\hat{x}}^{\infty}(x-\hat{x})dF(x) \tag{2.3}$$

만약 미분과 적분의 순서를 바꿀 수 있다면 식 (2.2)를 $\hat{x}$으로 미분하여 그 값을 0으로 만들면 식 (2.4)를 얻을 수 있다.

$$0 = (1-\tau)\int_{-\infty}^{\hat{x}} dF(x) - \tau\int_{\hat{x}}^{\infty} dF(x) = F(\hat{x}) - \tau \tag{2.4}$$

확률변수 $X$가 연속형이라고 가정하면 $X$의 누적분포함수인 $F(\cdot)$는 일대일 함수이고 그러면 식 (2.3)을 최소화하는 해는 유일하게 $\hat{x} = F^{-1}(\tau)$로 나온다. 즉, $\tau$번째 분위수는 식 (2.2)에 나와 있는 손실함수의 평균을 최소로 하는 값을 찾으면 된다.

이러한 사실과 회귀분석 방법에서 모형을 추정하는 방법을 이용하면 분위수 회귀분석에서도 모형을 추정할 수 있다. 만약 확률변수 $\{y_1, y_2, \cdots, y_n\}$이 주어졌고, $u(x) = x^2$일 때 $y_i$의 평균은 식 (2.5)를 만족하는 $\mu$가 된다.

$$\min_{\mu \in R} \sum_{i=1}^{n} u(y_i - \mu) \tag{2.5}$$

회귀분석에서는 손실함수가 제곱함수이고, 만약 여기서 상수 $\mu$를 모수적 함수인 $\mu(x, \beta)$로 둔다면 식은 (2.6)으로 바뀌게 된다.

$$\min_{\beta \in R^p} \sum_{i=1}^{n} u(y_i - \mu(x_i, \beta)) \tag{2.6}$$

식 (2.6)을 풀게 되면 조건부 기댓값의 함수인 $E(Y|X=x)$를 얻을 수 있다.

분위수 회귀분석은 회귀분석의 아이디어와 같은 방법을 이용한다. $\{y_1, y_2, \cdots, y_n\}$

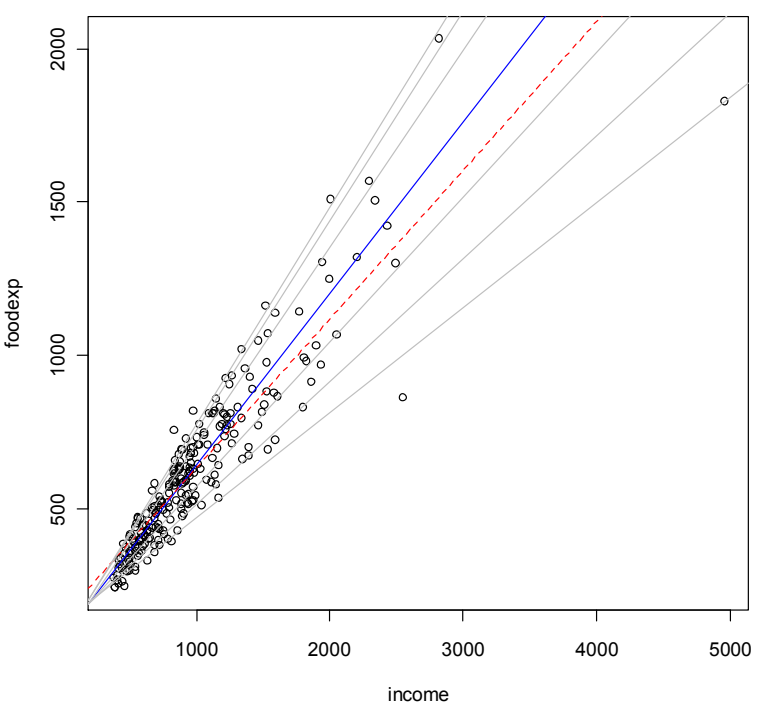
의 $\tau$번째 분위수를 구하고 싶으면 식 (2.7)을 풀면 된다.

$$\min_{\xi \in R} \sum \rho_\tau (y_i - \xi) \tag{2.7}$$

만약 조건부 중앙값을 얻고 싶다면 식 (2.7)에서 상수였던 $\xi$ 대신에 모수적 함수인 $\xi(x_i, \beta)$를 넣고, $\tau = 1/2$으로 설정해 놓으면 된다. 일반적으로 다른 분위수의 조건부 함수를 얻고 싶으면 다음을 풀면 된다.

$$\min_{\beta \in R^p} \sum_{i=1}^{n} \rho_\tau (y_i - \xi(x_i, \beta)) \tag{2.8}$$

[그림 2.2]는 Engel data로 얻은 간단한 분위수 회귀의 예이다. Engel data는 Koenker. R.과 Bassett. G(1982)가 발표한 자료로 수입에 따른 음식지출에 대한 데이터이다. 그림에서 중간의 점선은 조건부 평균의 함수이고, 실선은 아래로부터 {0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95}의 조건부 분위수를 추정한 결과이다.




[그림 2.2] Engel data에 대한 산점도 및 분위수 회귀모형의 적합

## 2.2 비모수적 분위수 회귀분석

### 2.2.1 국소선형 추정방법

비모수적 분위수 회귀분석 방법에 대해 말하기 앞서 비모수적 회귀분석 방법에 대해서 언급하겠다. 전통적인 모수적 모형의 모형식은 식 (2.9)와 같다.

$$y_i = f(\boldsymbol{\beta}, \boldsymbol{x_i'}) + \epsilon_i \tag{2.9}$$

여기서 $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)'$, $\boldsymbol{x_i'} = (x_{i1}, x_{i2}, \cdots, x_{ik})$이고, $\epsilon_i$는 각각 독립으로 평균이 0, 분산이 $\sigma^2$인 정규분포를 가정한다. 모수적 모형에서는 독립변수 $X$와 반응변수 $Y$의 관계를 가정하여 모형을 추정한다. 가정된 모형이 맞을 경우 모수적 모형은 강력한 성능을 보이지만 가정이 맞지 않을 경우 좋지 않은 성능을 보인다. 대다수의 경우 $X$와 $Y$의 관계를 잘 모르기 때문에 모형에 대한 가정을 하지 않는 것이 비모수적 방법이다. 일반적인 비모수적 모형의 모형식은 식 (2.10)과 같다.

$$\begin{aligned} y_i &= f(\boldsymbol{x_i'}) + \epsilon_i \\ &= f(x_{i1}, x_{i2}, \cdots, x_{ik}) + \epsilon_i \end{aligned} \tag{2.10}$$

이 때의 추정 대상은 $f(\cdot)$가 된다. 많은 비모수적 회귀분석에서는 $f(\cdot)$를 완만하고 미분가능한 함수라고 가정한다. $f(\cdot)$를 추정하는 비모수적 방법에는 여러 가지 방법이 있는데 국소선형 추정량(Local linear estimator)은 국소다항 추정량(Local polynomial estimator)에 포함되는 추정량이므로 국소다항 추정법을 소개하겠다.

분위수 회귀에 앞서 회귀분석에서의 국소다항 추정법을 소개하고자 한다. 추정하고자 하는 함수 $E(Y|X=x) = m(x)$를 $a$지점에서 테일러 전개를 하면

$$m(x) = m(a) + (x-a)m'(a) + (x-a)^2 m''(a)/2! + \cdots + (x-a)^n m^{(n)}(a)/n! + \cdots$$

이다. 이 때 $m(a) = \beta_0$, $m'(a) = \beta_1$, $m''(a)/2! = \beta_2$, $m^{(n)}(a)/n! = \beta_n$이라하면 $m(x)$를 $\beta$를 이용하여 근사시킬 수 있다.

$$m(x) \approx \beta_0 + \beta_1(x-a) + \beta_2(x-a)^2 + \cdots + \beta_n(x-a)^n \tag{2.11}$$

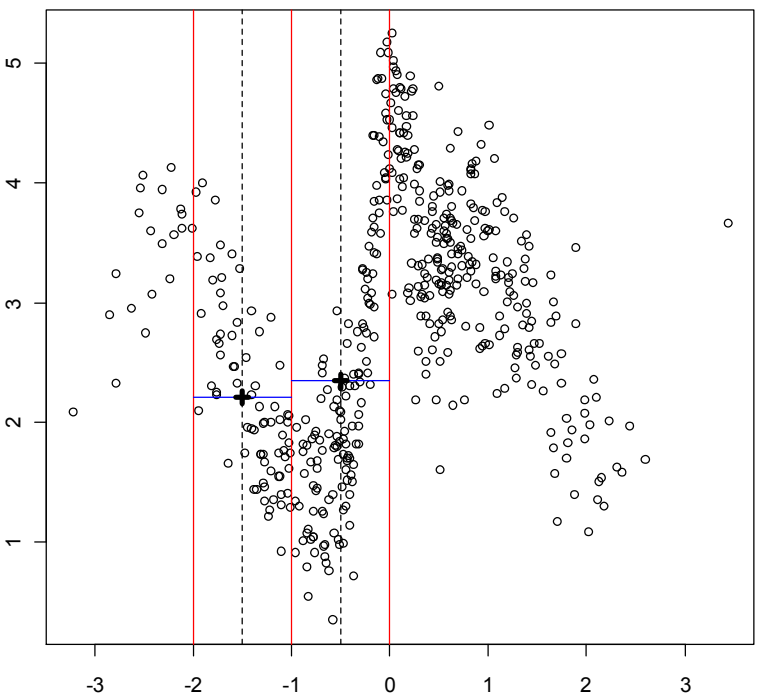
$\hat{m}(x) = \underset{f(x)}{\operatorname{argmin}} \sum_{i=1}^{n}(y_i - f(x_i))^2$ 로 구할 수 있는데 $\hat{m}(a)$를 추정하기 위해서 국소적으로 $\underset{\beta_0}{\operatorname{argmin}} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1(x-a) - \cdots \beta_n(x-a)^n)^2$을 구하는 것이 국소다항 추정법의 아이디어이다. 이 때 국소상수(Local constant) 추정은 $m(x)$를 $a$ 지점에서 첫 번째 항까지 테일러 전개를 한 $m(x) \approx m(a)$로 추정하는 것이고, 국소선형 추정은 $m(x)$를 $a$ 지점에서 두 번째 항까지 테일러 전개를 한 $m(x) \approx m(a) + (x-a)m'(a)$로 추정하는 것이다. 예를 들어 국소상수 같은 경우는 국소적으로 $m(x) \approx m(a)$로 추정을 하는 것인데, 국소적 전에 먼저 전체적으로 생각을 한다면 자료 $(y_i, x_i)$ $i = 1, \cdots, n$이 주어졌고 $m(x) = E(Y \mid X = x)$를 추정하고자 할 경우 $m(x) = \beta_0$라는 하나의 상수 값이라고 가정한다면 그 추정은 식 (2.12)와 같을 것이다.

$$\hat{m}(x) = \underset{\beta_0}{\operatorname{argmin}} \sum_{i=1}^{n}(y_i - \beta_0)^2 = \bar{y} \tag{2.12}$$

$x$값에 상관없이 모든 지점에서 $m(x) = \bar{y}$ 값이 될 것이다. 하지만 어떤 지점 $a$를 중심으로 그 주변의 데이터만을 이용해서 $m(x) = \beta_0$가 되는 값을 찾은 뒤 각 지점들을 연결하는 것을 식으로 표현하면 식 (2.13)과 같다.

$$\hat{m}(a) = \hat{E}(Y \mid X = a) = \underset{\beta_0}{\operatorname{argmin}} \sum_{i=1}^{n}(Y_i - \beta_0)^2 I_{(-h,h)}(X_i - a) \tag{2.13}$$

여기서 $I_A(x)=\begin{cases}1, & \text{if} \quad x\in A\\ 0, & \text{if} \quad x\notin A\end{cases}$ 이다.




[그림 2.3] $Y=2.5+\sin(2X)+e^{-16X^2}+0.5\epsilon$ 함수의 국소적 평균

[그림 2.3]은 국소상수 추정법의 기본적인 아이디어를 나타낸 그림으로 평활량 $h=0.5$, $X\sim N(0,1)$ 그리고 $\epsilon\sim N(0,1)$일 경우 $Y=2.5+\sin(2X)+e^{-16X^2}+0.5\epsilon$에서 난수 500개를 생성하여 $a=-1.5$와 $a=-0.5$ 지점의 국소적 평균값을 구한 그림이다.

하지만 여기서 $I_A(\cdot)$함수는 0 또는 1의 값만 갖는 계단형 함수로 미분이 불가능한 함수이다. 그리고 $a$지점과 $h$만큼 떨어져 있는 모든 데이터들에게 똑같은 가중치를 준다. $m(a)$값을 추정하는데 있어서 $a$지점에서 가까이 있는 데이터들에게는 더 높은 가중치를 주고, $a$지점에서 멀리 떨어져 있는 데이터들에겐 적은 가중치

를 주는 것이 더 합당할 것이다. 그래서 나온 양수이고, 대칭이고, 단봉형의 함수가 $K$(kernel)이다. 커널함수는 보통 대칭인 확률분포 함수를 쓰기도 하는데, 대표적인 종류로는 표준정규분포 함수, 에파니치코프(Epanechnikov) 커널 함수 등이 있다. 커널함수에서 중요한 것은 평활량인 $h$를 정하는 것인데 이것은 2.2.3에서 소개하겠다.

커널함수를 이용하여 국소선형을 추정하면 식 (2.13)은 아래와 같이 변한다.

$$\widehat{m}(a) = \hat{E}(Y \mid X = a) = \underset{\beta_0}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 (X_i - a))^2 I_{(-h,\,h)}(X_i - a) \tag{2.14}$$

$$= \underset{\beta_0}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 (X_i - a))^2 I_{(-1,1)}\left(\frac{X_i - a}{h}\right)$$

$$\Rightarrow \underset{\beta_0}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 (X_i - a))^2 \frac{1}{h} K\left(\frac{X_i - a}{h}\right)$$

$K_h(X_i - a) = K\{\ (X_i - a)/h\}/h$일 때 국소다항의 일반적인 식은 식 (2.15)와 같다.

$$\underset{\alpha}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \alpha - \beta_1 (X_i - a) - \beta_2 (X_i - a)^2 - \cdots \beta_p (X_i - a)^p)^2 K_h(X_i - a) \tag{2.15}$$

분위수 회귀분석에서 국소다항 추정법은 회귀분석일 때와 거의 동일하다. 회귀분석에서는 손실함수를 제곱함수를 이용하여 손실함수를 최소로하는 추정값을 구하였다. 손실함수를 제곱함수가 아닌 $\rho_\tau(\cdot)$를 사용하면 분위수 회귀분석에의 국소다항 추정법이 된다. 즉, 분위수 회귀에서의 국소다항 추정법의 표현식은 식 (2.16)과 같다.

$$\underset{\beta_0}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau (Y_i - \beta_0 - \beta_1 (X_i - a) - \beta_2 (X_i - a)^2 - \cdots \beta_p (X_i - a)^p) K_h(X_i - a) \tag{2.16}$$

식 (2.16)을 최소로 하는 $\beta_0$가 바로 국소다항 방법으로 추정한 $\widehat{Q_Y}(\tau \mid X = a)$가 되

고, $p=1$로 두고 식 (2.16)을 최소로 하는 $\beta_0$를 찾으면 그것이 바로 국소선형 추정량이 된다.

### 2.2.2 이중 커널 추정방법

이중 커널 추정방법은 커널 가중치를 $x$에 대해서 뿐만이 아니라 $y$에 대해서도 커널 가중치를 주는 것이다. $x$의 커널 함수 $K$가 아닌 새로운 대칭이면서 확률분포인 커널함수 $W$가 있고, $W$의 누적분포함수인 $\Omega$가 있다고 하자. 새로운 커널함수의 평활량이 $h_2$라 하고, $W_{h_2}(Y_j-u)= W\{(Y_j-u)/h_2\}/h_2$이라고 할 때, 식 (2.17)이 성립한다.

$$\int_{-\infty}^{y} W_{h_2}(Y_j-u)du= \Omega\left(\frac{y-Y_j}{h_2}\right) \tag{2.17}$$

Yu와 Jones(1998)에 의하면 만약 $h_2\to 0$이라면 식 (2.18)이 성립한다고 알려졌다.

$$E\left\{\Omega\left(\frac{y-Y}{h_2}\right)\mid X=x\right\}\approx F(y\mid x) \tag{2.18}$$

식 (2.18)의 근사값을 다시 테일러전개로 $X=a$에서 1차항까지 근사하면

$$E\left\{\Omega\left(\frac{y-Y}{h_2}\right)\mid X=x\right\}\approx F(y\mid x)\approx F(y|a)+\frac{\partial F(y\mid a)}{\partial a}(z-a)\equiv \alpha+\beta(x-a) \tag{2.19}$$

가 된다. 이 경우 $\tilde{\alpha}=\tilde{F}_{h_1,h_2}(y|x)$가 되고 이것을 국소선형 추정법에 그대로 이용하면 식 (2.20)이 된다.

$$\underset{\alpha}{\operatorname{argmin}}\sum_{i=1}^{n}\left(\Omega\left(\frac{y-Y_i}{h_2}\right)-\alpha-\beta(X_i-a)\right)^2\frac{1}{h_1}K\left(\frac{X_i-a}{h_1}\right) \tag{2.20}$$

이렇게 추정되는 $\tilde{\alpha}=\tilde{F}_{h_1,h_2}(y|x)$를 명확하게 쓰면 식 (2.21)이다.

$$\tilde{F}_{h_1,h_2}(y|x)=\frac{1}{\sum_j w_j(x;h_1)}\sum_j w_j(x;h_1)\Omega\left(\frac{\tilde{y}-Y_j}{h_2}\right) \tag{2.21}$$

$$w_j(x;h_1)=K\left(\frac{x-X_j}{h_1}\right)[S_{n,2}-(x-X_j)S_{n,1}],\ S_{n,l}=\sum_{i=1}^{n}K\left(\frac{x-X_j}{h_1}\right)(x-X_i)^l,\ l=1,2,\cdots$$

이중 커널 방법을 이용한 $y$의 조건부 $\tau$번째 분위수는 식 (2.21)의 역함수이다.

$$\tilde{y}=\tilde{q}_\tau(x)=\tilde{F}^{-1}_{h_1,h_2}(\tau|x) \tag{2.22}$$

Yu와 Jones(1998)에 의하면 이중커널 방법을 사용하면 단일 커널을 이용한 local linear 방법보다 잔차 제곱의 성질이 더 좋으며, 단일 커널의 문제점인 분위수가 교차되는 현상도 방지할 수 있다고 한다.

### 2.2.3 평활량 $h$의 선택

평활량 $h$에 따라 추정된 함수가 매우 민감하게 반응하므로 $h$를 추정하는 것은 매우 중요한 일이다. Yu 와 Jones(1998)는 분위수 회귀분석에 쓰이는 단일, 이중 커널에서 자동적으로 $h$를 찾아내 선택하는 방법을 제시하였다. 그들이 제시한 방법대로 조건부 분위수에 대한 $h$를 찾을려면 먼저 조건부 평균에 대한 $h$를 찾는 것을 우선으로 하고 있다. 그들의 제안대로 본 연구에서는 조건부 평균에 대한 평활량인 $h_{DPI}$를 사용하여 조건부 분위수의 $h$를 추정하였다. $x$축 방향의 커널함수에 대한 평활량 $\hat{h}_1(p)$를 얻기 위해 Yu 와 Jones는 다음의 추정량을 제안하였다.

$$\hat{h}_1(p)=\hat{h}_{DPI}\left(\frac{p(1-p)}{\phi(\Phi^{-1}(p))^2}\right)^{\frac{1}{5}} \tag{2.23}$$

여기서 $\phi(\cdot)$과 $\Phi(\cdot)$는 각각 표준정규분포의 확률밀도함수와 누적확률분포함수이다. 이중 커널에서 사용되는 $y$축 방향의 평활량 $\hat{h}_2(p)$를 얻기 위해 Yu와 Jones는 $\hat{h}_1(p)$를 이용하는 것을 제안했다. 만약 $\hat{h}_1(0.5) < 1$ 이라면 식 (2.24)로 추정하고, 그렇지 않으면 식 (2.25)와 같이 추정한다.

$$\hat{h}_2(p) = \max\left(\frac{\hat{h}_1^5(0.5)}{\hat{h}_1^3(p)}, \frac{\hat{h}_1(p)}{10}\right) \tag{2.24}$$

$$\hat{h}_2(p) = \frac{\hat{h}_1^4(0.5)}{\hat{h}_1^3(p)} \tag{2.25}$$

# 3. 성능 비교 및 평가

## 3.1. 모의실험

모의실험은 공개언어인 R을 이용하여 하였는데 모수적 분위수 회귀함수의 적합은 R의 패키지 중 하나인 quantreg 패키지의 rq 함수를 이용하여 하였으며, 비모수적 분위수 함수 중 local linear 방법으로 분위수 회귀함수를 적합시킨 것은 마찬가지로 quantreg 패키지의 lprq 함수를 사용하여 하였다. 이중커널 방법으로의 분위수 회귀함수 적합은 R 프로그램으로 작성하였다. 모의실험을 하기 위해 참 분위수 함수의 모양이 단순한 형태부터 복잡한 형태까지 총 4개의 모형으로 데이터를 생성하였다. 모의실험에 사용한 모형은

모형1. $Y_i = X_i^2 + \epsilon_i,\ \epsilon_i \sim N(0,1),\ X_i \sim U(-3,3)$

모형2. $Y_i = 10 - 2(X_i - 5)/5 + b_i,\ b_i = (X_i/5+1)\{(e^{\epsilon_i} - e^{0.02})/e^{0.04}(e^{0.04}-1)\},$

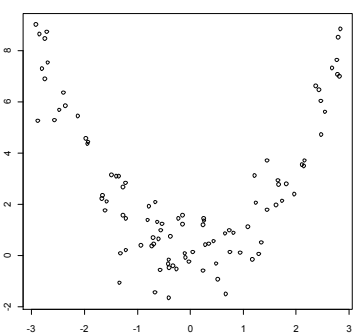
$\epsilon_i \sim N(0,0.2),\ X_i \sim U(-0.5, 5.5)$

모형3. $Y_i = \sin(\pi X_i) + \epsilon_i,\ \epsilon_i \sim N(0, e^{2\pi X_i}),\ X_i \sim U(0,1)$

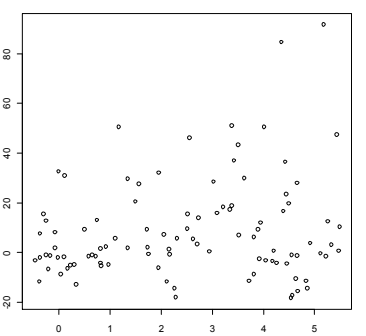
모형4. $Y_i = 2.5 + \sin(2X_i) + 2e^{-16X_i^2} + 0.5\epsilon_i,\ \epsilon_i \sim N(0,1),\ X_i \sim N(0,1)$이다.

모수적 분위수 회귀모형을 적합시키려면 먼저 자료를 보고 그 자료에서 독립변수 $X$와 반응변수 $Y$가 어떤 관계인지 유추해서 그 모형으로 적합시켜야 한다. 선택한 모형들은 독립변수와 반응변수의 관계를 알기 쉬운 자료부터 어떤 관계인지

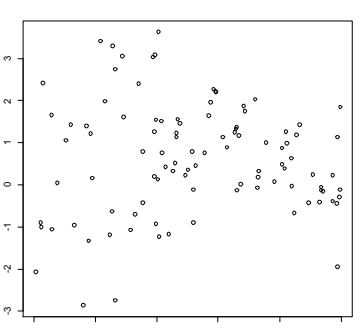
알기 어려운 자료까지 단계별로 네 가지를 골랐다. 모형1은 직관적으로 독립변수와 반응변수의 관계가 2차 함수관계를 가지고 있다고 여겨지는 자료이며 모형4는 어떤 관계가 있는지 알기 어려운 자료이다. 모형2와 모형3은 그 중간인데 [그림 3.1]은 각 모형에서 100개의 데이터를 생성하여 그린 산점도이며, [그림3.2]는 각 모형의 데이터와 참 분위수 함수를 함께 그린 것이다.



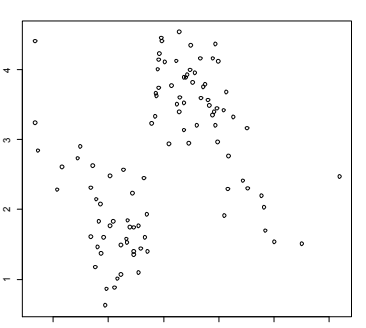
모형1



모형2

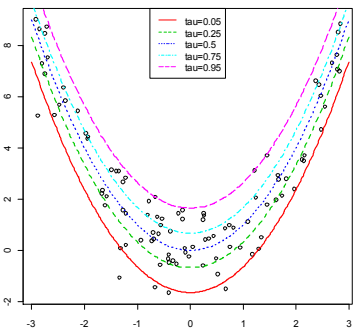


모형3

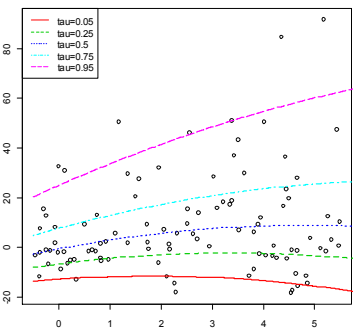


모형4

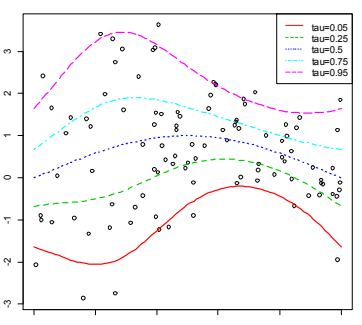
[그림3.1] 각 모형에 대한 산점도



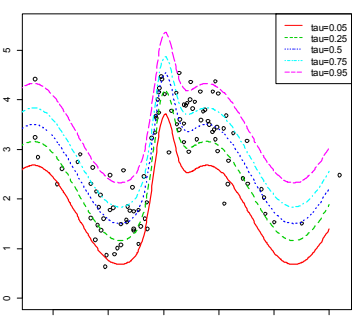

모형1




모형2




모형3




모형4

[그림3.2] 각 모형에 대한 산점도와 참 $\tau$ 분위수 함수

$\tau = (0.05, 0.25, 0.5, 0.75, 0.95)$

모수적 모형의 선택은 데이터의 산점도를 보고 하였으며 모형1에서는 2차함수, 모형 2,3,4에는 각각 log함수, 3차함수, sin함수를 적합시켰다. 데이터의 수가 1000개 이고, 반복수가 100번일 경우 모수적 모형과 비모수적 모형(국소선형, 이중커널)을 비교한 결과는 <표3.1>과 같다. 표에 나온 값들은 모의실험으로 구한

추정량에 100을 곱한 값이며, 괄호안의 숫자는 표준오차를 의미한다.

<표 3.1> 모수적, 비모수적 분위수 추정방법의 MISE 비교

| 모형 | $p$ | 모수적 모형 | | | 비모수적 모형 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | 국소선형 | | | 이중 커널 | | |
| | | Ivar | $Ibias^2$ | MISE | Ivar | $Ibias^2$ | MISE | Ivar | $Ibias^2$ | MISE |
| 모형 1 | 0.05 | 8.6278 | 0.1444 | 8.7722 (0.717) | 22.002 | 3.1317 | 8.7722 (1.121) | 19.191 | 145.51 | 164.70 (2.941) |
| | 0.25 | 3.1087 | 0.0194 | 3.1280 (0.302) | 9.6270 | 1.9393 | 11.566 (0.428) | 10.865 | 15.204 | 26.096 (0.726) |
| | 0.5 | 2.9600 | 0.0027 | 2.9627 (0.245) | 8.3571 | 1.7008 | 10.058 (0.447) | 9.8260 | 1.9381 | 11.764 (0.416) |
| | 0.75 | 3.4956 | 0.0042 | 3.4998 (0.311) | 9.4216 | 1.8843 | 11.306 (0.524) | 11.905 | 31.774 | 43.678 (1.124) |
| | 0.95 | 8.7997 | 0.0532 | 8.8529 (0.766) | 19.003 | 3.8614 | 22.865 (1.207) | 25.554 | 266.13 | 291.68 (5.013) |
| 모형 2 | 0.05 | 2.8806 | 0.6762 | 3.5568 (0.320) | 3.9719 | 0.2573 | 4.2292 (0.447) | 3.9702 | 0.3000 | 4.2703 (0.414) |
| | 0.25 | 3.0326 | 0.5862 | 3.6188 (0.310) | 5.2149 | 0.1662 | 5.3812 (0.448) | 5.1059 | 0.3079 | 5.4138 (0.432) |
| | 0.5 | 3.8928 | 0.5543 | 4.4471 (0.375) | 7.5270 | 0.0689 | 7.5959 (0.645) | 7.6511 | 0.1270 | 7.7781 (0.675) |
| | 0.75 | 8.8312 | 0.7111 | 9.5423 (0.818) | 14.862 | 0.0898 | 14.952 (1.166) | 15.190 | 0.1277 | 15.318 (1.182) |
| | 0.95 | 50.351 | 1.1430 | 51.484 (4.838) | 75.536 | 0.5192 | 76.055 (6.606) | 75.250 | 1.7584 | 77.008 (6.212) |
| 모형 3 | 0.05 | 2.2668 | 1.1003 | 3.3671 (0.201) | 2.8601 | 1.1408 | 4.0009 (0.250) | 3.0175 | 1.6318 | 4.6493 (0.241) |
| | 0.25 | 0.9318 | 0.2574 | 1.1892 (0.086) | 1.3866 | 0.1186 | 1.5052 (0.115) | 1.3491 | 0.1056 | 1.4547 (0.110) |
| | 0.5 | 0.7096 | 0.0348 | 0.7445 (0.064) | 1.0829 | 0.0370 | 1.1199 (0.090) | 1.0724 | 0.0566 | 1.1290 (0.092) |
| | 0.75 | 0.7547 | 0.0899 | 0.8446 (0.071) | 1.1762 | 0.1144 | 1.2906 (0.087) | 1.1708 | 0.1647 | 1.3356 (0.083) |
| | 0.95 | 1.8109 | 0.6785 | 2.4893 (0.015) | 2.5433 | 0.7178 | 3.2611 (0.201) | 2.6682 | 0.8697 | 3.5380 (0.178) |
| 모형 4 | 0.05 | 21.773 | 783.90 | 805.68 (15.68) | 140.02 | 44.968 | 184.99 (58.97) | 37.412 | 23.964 | 61.375 (2.324) |
| | 0.25 | 9.2735 | 1153.2 | 1162.5 (16.35) | 429.29 | 7.0965 | 436.38 (243.7) | 32.902 | 3.0417 | 35.916 (1.439) |
| | 0.5 | 7.9600 | 1156.2 | 1164.2 (14.94) | 146.39 | 2.3533 | 148.75 (56.88) | 30.256 | 0.6722 | 30.929 (1.363) |
| | 0.75 | 12.587 | 940.55 | 953.14 (14.08) | 417.25 | 10.463 | 427.72 (263.6) | 31.026 | 3.2295 | 34.255 (1.717) |
| | 0.95 | 34.907 | 1126.5 | 1161.4 (14.82) | 319.14 | 55.435 | 374.57 (250.0) | 36.896 | 24.502 | 61.372 (2.993) |

참 분위수의 함수의 형태에 따라 모수적 방법과 비모수적 방법의 성능이 달라졌

다. 참 분위수 함수의 형태가 구불구불한 경우일수록 비모수적으로 추정한 분위수 회귀모형의 MISE가 더 작으며 또한 모수적으로 분위수 회귀모형을 추정할 경우에는 목표로 하는 분위수에 따라 모수적 모형을 다르게 하는 것이 모든 분위수를 하나의 모형으로 추정하는 것 보다 더 성능이 좋았다. 가정한 모수적 모형이 자료에 잘 맞지 않을 경우에는 모수적 모형의 MISE가 비모수적 분위수 회귀모형의 MISE 보다 월등히 크게 나타나는 결과가 있었다.

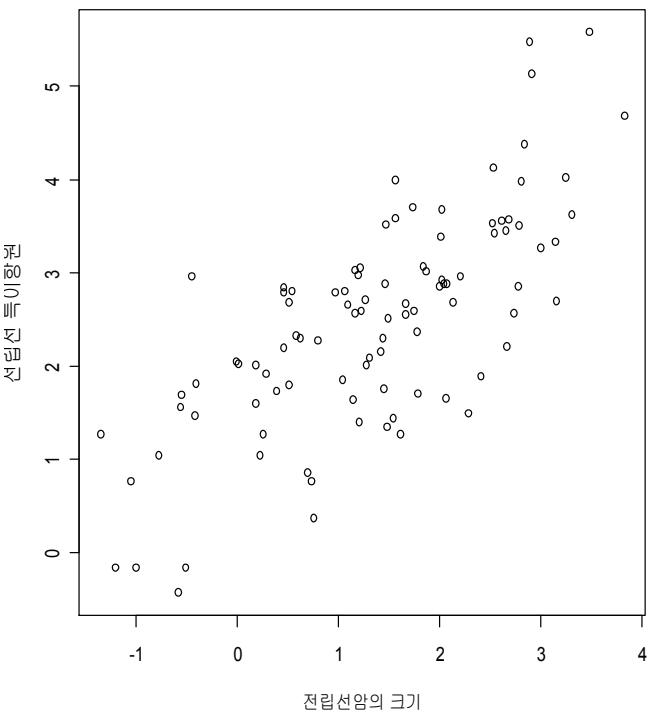
## 3.2. 실제자료 분석

### 3.2.1 전립선암 자료

두 가지 실제자료를 분석하였는데 먼저 Stamey등(1989)이 발표한 전립선암 자료를 가지고 실제자료 분석을 하였다. 이 자료는 전립선 절제수술을 통지 받은 환자들을 대상으로 수집한 자료이며, 전립선암의 크기, 전립선의 무게 등 8가지 항목을 조사하여 어느 항목이 전립선 특이항원(prostate specific antigen)에 영향을 주는가를 분석할 목적으로 모여진 자료이다. 전립선 특이항원은 전립선의 상피세포에서 합성되는 단백분해 효소로 전립선 이외의 조직에서는 거의 발현되지 않아 전립선암의 선별에 이용되는 유용한 종양표지자이다. 또한 전립선암의 선별 검사뿐만 아니라 수술 후 재발 판정에도 유용하게 이용할 수 있다. 하지만 전립선 특이항원은 전립선 조직에는 특이적이지만 종양에는 특이적이지 않아 전립선 비대증, 전립선염, 전립선 경색 등에서도 증가할 수 있다.

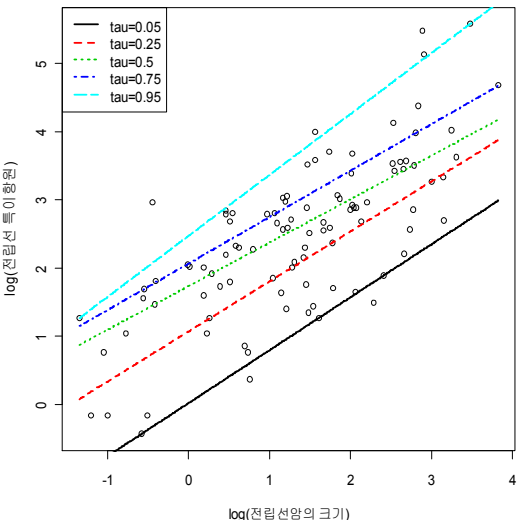
이 논문에서는 특히 전립선암의 크기와 전립선 특이항원의 관계에 대해 분위수

회귀분석을 하였는데 [그림3.3]은 전립선암의 크기 값에 로그를 취한 값과 전립선 특이항원 값에 로그를 취한 값의 산점도이며, 단위는 ng/mL이다.

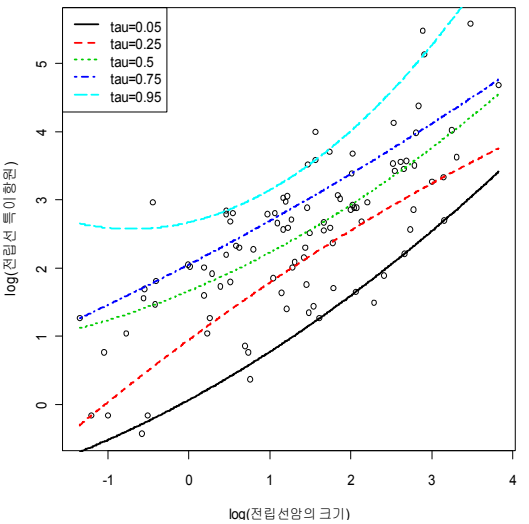



[그림3.3] 전립선암의 크기와 전립선 특이항원의 산점도

산점도를 보면 전립선암의 크기가 커질수록 전립선 특이항원이 높아진다는 것 확인할 수 있다. 산점도를 보고 대체적으로 전립선암의 크기와 전립선 특이항원의 관계가 선형관계라고 생각하여 모수적으로는 분위수 회귀직선을 추정하였고 높은 분위수에서는 2차 함수 관계가 있다고도 판단되어 2차 함수식을 추정해보았다. 비모수적으론 국소선형 방법과 이중커널 방법으로 모형을 추정해 보았다.



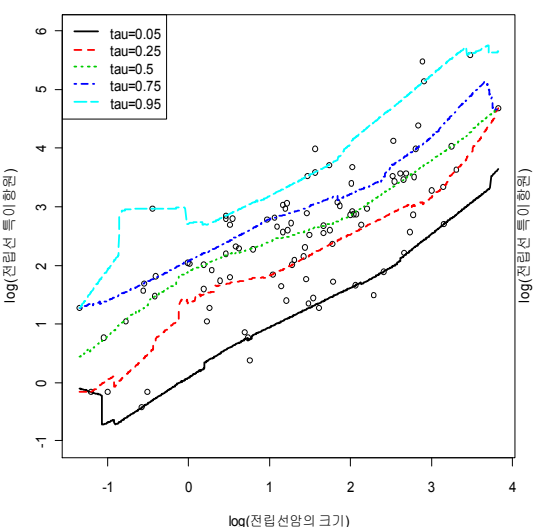

$Q_Y(\tau|X=x)=\beta_0(\tau)+\beta_1(\tau)x$



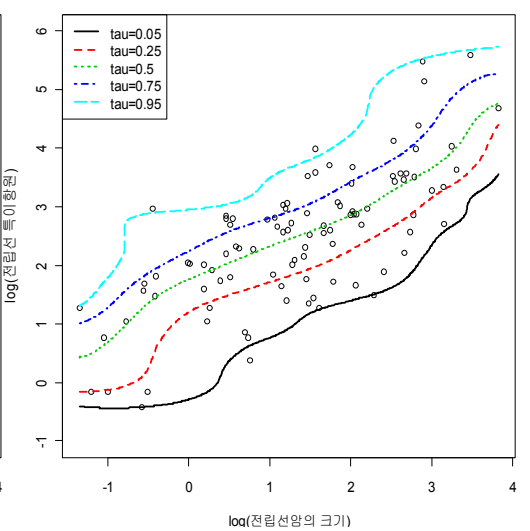

$Q_Y(\tau|X=x)=\beta_0(\tau)+\beta_1(\tau)x+\beta_2(\tau)x^2$

[그림3.4] 전립선암의 크기와 전립선 특이항원의 모수적 분위수 회귀모형

$\tau=(0.05, 0.25, 0.5, 0.75, 0.95)$




국소선형방법




이중커널방법

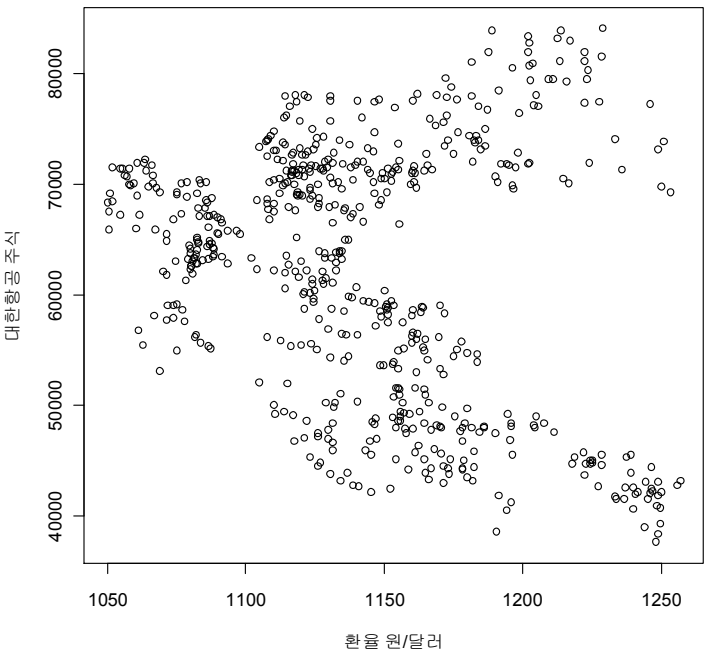
[그림3.5] 전립선암의 크기와 전립선 특이항원의 비모수적 분위수 회귀모형

$\tau=(0.05, 0.25, 0.5, 0.75, 0.95)$

모수적 모형에서는 전립선암의 크기가 커질수록 전립선 특이항원이 높아진다고 나왔다. 특히 0.95같이 높은 분위수에서는 직선의 기울기가 더 올라가서 전립선암의 크기가 클 경우 전립선 특이항원의 조건부 분포가 치우쳐진 형태로 추정되었다. 비모수적 모형에서는 국소선형 추정방법에서는 전립선암의 크기가 작을 경우 분위수가 교차되는 문제점이 발견되었고, 역시나 이중커널 방법보다는 굴곡진 모양이 나왔다. 전립선암의 크기가 클 경우의 전립선 특이항원의 조건부 분포는 모수적 모형과 다르게 나왔는데 국소선형 모형에서는 0.25, 0.5, 0.75의 세 분위수가 모여있는 모양을 보아 분포의 모양이 양 끝에 봉우리가 두 개 있는 형태의 모양으로 나왔고, 이중커널에서는 0.25, 0.5, 0.75의 세 분위수가 모여있지 않아 국소선형 모형과는 다르게 나왔다. 전립선암의 크기에 따라 전립선 특이항원이 무한대로 높아지진 않는다고 판단되어 이번 자료에서는 모수적 모형보다 비모수적 모형을 이용하는 것이 더 좋다고 짐작된다.

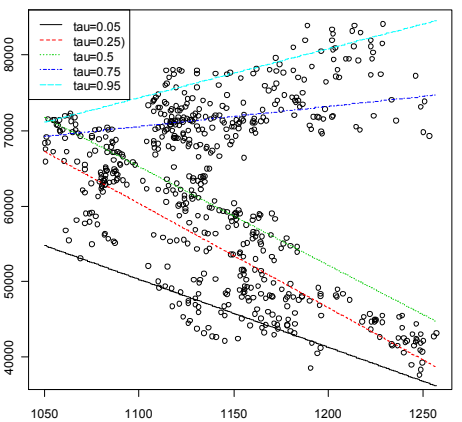
### 3.2.2 대한한공 주식자료

두번째 실제자료는 2009년 7월 22일부터 2011년 12월 12일까지 총 300일간의 원/달러 환율과 환율에 영향을 많이 받는 업종 중 하나인 대한항공의 주식이다. 먼저 자료의 산점도를 살펴보면 [그림3.6]과 같다.

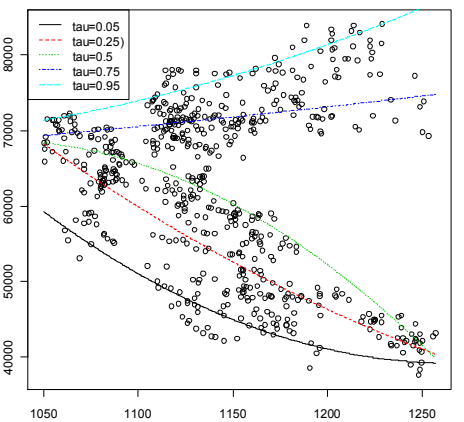



[그림3.6] 환율과 대한항공 주식의 산점도

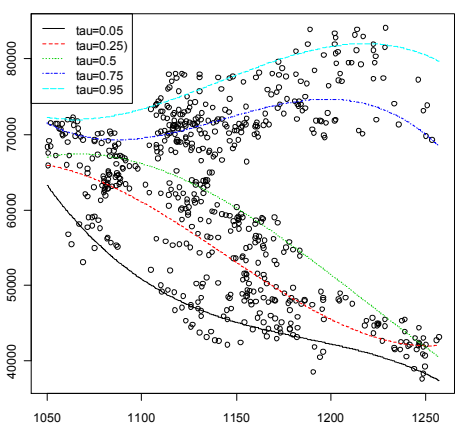
산점도를 보면 환율이 올라가면 여행사인 대한항공의 주식은 하락하거나 상승하는 둘 중 하나의 경우로 나온다. 산점도를 보고 어떤 모형이 적합할지 알 수 없어서 환율 원/달러(= $X$)와 대한항공 주식(= $Y$)의 관계를 1차식부터 5차식까지 전부 적합 시켜 보았고, 비모수적으론 국소선형 방법과 이중커널 방법으로 모형을 추정하였다.



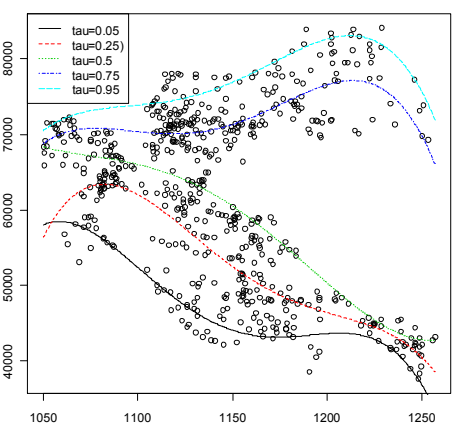

$$Q_Y(\tau|X=x)=\beta_0(\tau)+\beta_1(\tau)x$$



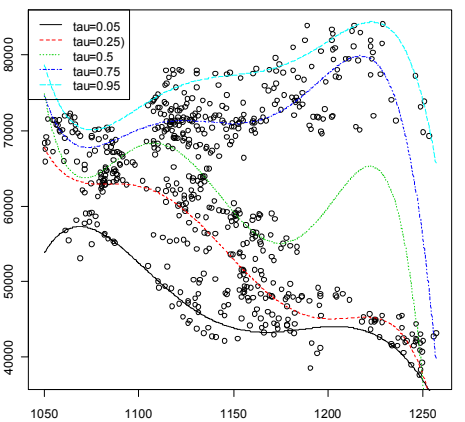

$$Q_Y(\tau|X=x)=\beta_0(\tau)+\beta_1(\tau)x+\beta_2(\tau)x^2$$




$$Q_Y(\tau|X=x)=\beta_0(\tau)+\beta_1(\tau)x+\beta_2(\tau)x^2+\beta_3(\tau)x^3$$



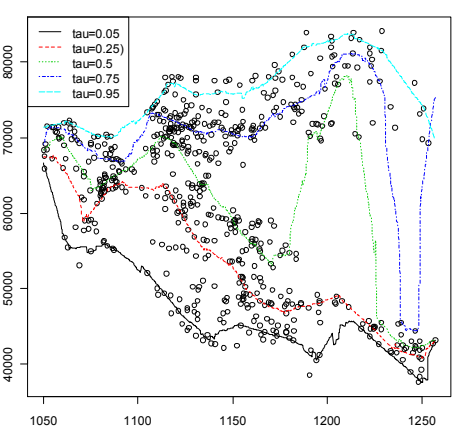

$$Q_Y(\tau|X=x)=\beta_0(\tau)+\beta_1(\tau)x+\beta_2(\tau)x^2+\beta_3(\tau)x^3+\beta_4(\tau)x^4$$



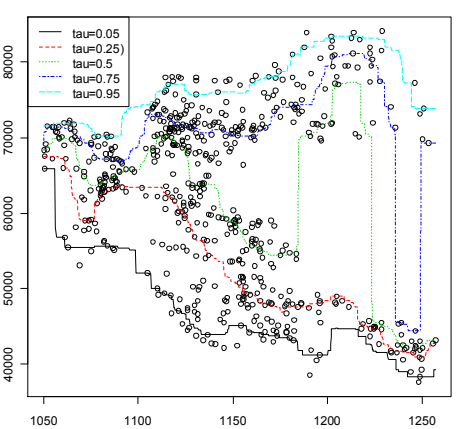

$$Q_Y(\tau|X=x)=\beta_0(\tau)+\beta_1(\tau)x+\beta_2(\tau)x^2+\beta_3(\tau)x^3+\beta_4(\tau)x^4+\beta_5(\tau)x^5$$

[그림3.7] 환율과 대한항공 주식의 모수적 분위수 회귀모형

$\tau=(0.05, 0.25, 0.5, 0.75, 0.95)$




국소선형방법




이중커널방법

[그림3.8] 환율과 대한항공 주식의 비모수적 분위수 회귀모형

$\tau=(0.05, 0.25, 0.5, 0.75, 0.95)$

모수적 모형에서는 전부 분위수가 교차되게 나왔다. 이는 분위수의 성질을 생각하면 나와서는 안 되는 결과이기 때문에 가정한 모형이 자료에 적합하지 않다는 것을 의미한다. 비모수적 분위수 회귀모형은 분위수가 교차되어 나오지 않았지만 모수적 분위수 회귀모형 보다는 거친 형태로 추정되었다.

# 4. 결론

Koenker와 Bassett(1978)이 소개한 분위수 회귀모형은 확률변수들 사이에 확률적인 관계구조를 포함한 함수 모형을 좀 더 완벽하게 추정하도록 제공한다. 본 논문에서는 분위수 회귀모형을 추정하는 방법을 살펴보고 비교를 해보았다. 그로인해 먼저 모수적 추정방법의 자세한 설명과 비모수적 추정방법 중 특히 국소선형방법과 이중 커널 방법에 대해 설명을 하였고 모의실험을 통해 여러 가지 상황에서의 모수적, 비모수적 방법의 성능을 비교하였다. 비교는 평균제곱오차를 적분한 MISE라는 값으로 하였으며 모의실험을 통해 얻은 추정량의 표준오차 역시 제공하였다. 모의실험 결과 참 분위수 함수의 형태가 간단한 경우에는 모수적으로 추정한 분위수 회귀모형이 더 좋고, 참 분위수 함수의 형태가 복잡해질수록 비모수적으로 추정한 분위수 회귀모형의 성능이 더 좋았다. 모수적으로 분위수 회귀모형을 추정할 경우에는 분위수에 따라 모수적 모형을 다르게 하는 것도 좋은데 그것은 분위수에 따라 데이터의 형태가 다르기 때문이다. 가정한 모수적 모형이 자료에 잘 맞지 않을 경우에는 모수적 모형의 성능이 비모수적 분위수 회귀모형에 비해 크게 떨어졌다. 그래서 분위수 회귀분석을 할 경우에 다음과 같은 순서로 분석하기를 제안한다. 첫 번째로 데이터의 산점도를 그려서 전체적인 데이터의 형태를 살펴보고, 설명변수 $X=x$ 주변의 종속변수 $Y$ 데이터를 살펴보며 원하는 분위수 함수의 형태를 대략적으로 결정한다. 그 형태가 간단한 형태면 모수적 모형으로 분위수 회귀모형을 적합시키고, 대략적인 형태가 복잡한 형태라면 비모수적 분위수 회귀 모형을 적합시키면 보다 효율적인 분위수 회귀분석이 이루어질 것 이라고 생각된다.

# 참고문헌

Koenker, R. and Bassett, G. (1978) "Regression Quantiles," *Econometrica*, 46, 33-50.

Koenker, R. and Bassett, G. (1982) "Robust Tests of Heteroscedasticity based on Regression Quantiles," *Econometrica*, 50, 43-61.

Koenker, R. (2004) "Quantile Regression," *Cambridge University Press.*

Li, Y., Graubard, B. I. and Korn, E. L. (2009) "Application of nonparametric quantile regression to body mass index percentile curves from survey data," *Statistics in Medicine,* 29, 558-572.

Ruppert, D., Sheather, S. J. and Wand, M. P. (1995) "An effective bandwidth selector for local least squares regression," *Journal of the American Statistical Association,* 90, 1257-1270.

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," *The Statistician,* 52, 331-350.