The Korean Journal of Applied Statistics (2012) DOI: http://dx.doi.org/10.5351/KJAS.2012.25.5.793
25(5), 793–802

Nonparametric Estimation using Regression Quantiles in a Regression Model

Sang Moon Han[1] · Byoung Cheol Jung[2]

[1]Department of Statistics, University of Seoul; [2]Department of Statistics, University of Seoul

(Received August 11, 2012; Revised September 15, 2012; Accepted September 21, 2012)

Abstract

One proposal is made to construct a nonparametric estimator of slope parameters in a regression model under symmetric error distributions. This estimator is based on the use of the idea of minimizing approximate variance of a proposed estimator using regression quantiles. This nonparametric estimator and some other L-estimators are studied and compared with well known M-estimators through a simulation study.

Keywords: Regression quantile, regression trimmed mean, L-estimator.

1. 서론

회귀모형에서 로버스트 추정법은 지난 40여년간 꾸준히 연구되어왔고 최소제곱법(least squares method)의 대안으로 많은 추정량들이 제시되어 왔다. 이러한 로버스트적 추정법들은 크게 분류하여 M-, L-, R-군 등으로 나누어 질 수 있다.

R-추정법은 M-추정법이나 L-추정법에 비해 많이 사용되지 않으나 Adichie (1974) 및 Hettmansperger와 McKean (1977) 등에 의해 회귀 모형상에서의 추정법이 연구되었고, M-추정법은 Huber (1973)의 ψ 함수를 이용한 위치모수추정법에서 자연스럽게 회귀 모수추정법으로 확장될 수 있다. 특히 Han (2003)은 Hogg (1983)의 선택 통계량(selector statistics)을 사용한 위치 모형에서의 모수 추정법을 회귀 모형에서 선택통계량을 사용한 M-추정법으로 확장하였다. L-추정법은 Bickel (1973)에 의해서 처음으로 제시되었다. Bickel의 추정량은 점근적인 좋은 성질에도 불구하고 복잡한 형태를 취해 계산하기에 매우 복잡하였다. 이에 따라 Koenker와 Bassett (1978, 1982)는 예비적합(preliminary fit)에 의한 잔차의 순서통계량에 의하지 않는 L-추정량 형태의 회귀분위수(regression quantile)를 제안하였다. 그들은 소위 특정한 check 함수상에서의 M-추정량으로 회귀분위수를 정의하는 방법을 취하고 이의 점근적 성질이 위치모수의 위치분위수와 비슷하다는 것을 발견하였다. Ruppert와 Carroll (1980)은 Koenker와 Bassett의 아이디어를 확장한 절사회귀추정량을 제시하였다. 그들은 절사회귀추정량의 점근적 성질이 위치모수에서의 절사추정량과 비슷하고, 이 추정량은 디자인 행렬의 재모수화(reparameterization)에 대해서도 불변(invariance)인 좋은 성질을 가지고 있다는 것을 발견하였다. 특히 이 추정량은 Barrodale과 Roberts (1974) 및 Portnoy와 Koenker (1997) 등 다수의 저자에 의해

This work was supported by the 2011 Research Fund of the University of Seoul.
[1]Corresponding author: Professor, Department of Statistics, University of Seoul, 163 Siripdaero, Dongdaemun-gu, Seoul 130-743, Korea. E-mail: smhan@uos.ac.kr

제안된 표준적인 L-알고리즘에 대한 약간의 수정에 의해 쉽게 계산되어진다는 장점을 가지고 있다. 아마도 이 추정량이 L-추정량으로 처음으로 실용적이고 유용한 추정량일 것이다. L-추정량에 대한 이후 연구로 De Jongh과 De Wet (1985)은 Jaeckel (1971)의 위치 모수 추정법을 Ruppert와 Carroll의 절사회귀추정법을 사용하여 회귀모형에 확장하였다. Hogg 등 (1988) 등은 M-, L-, R-추정법에 대한 광범위한 모의실험을 통해 기존의 추정법에 대한 로버스트 추정법의 우월성을 보였다. 이와 같은 연구이외에도 다수의 연구자들이 회귀 백분위수에 대하여 연구하였다. 보다 자세한 회귀 백분위수에 대한 문헌은 Koenker 등 (2005)을 참조하기 바란다.

본 논문에서는 Johns (1974)의 위치 모수의 추정법에서 가중치를 주는 방법과 Ruppert와 Carroll (1980)의 결과를 응용해서 회귀 백분위수를 이용한 회귀계수 추정 문제를 고려하고자 한다. 먼저 본 논문에서 이용되는 회귀분위수추정량에 대해 간단히 언급하기로 하자. 다음과 같은 표준적인 회귀모형을 가정하자.

$$\mathbf{y} = X\boldsymbol{\beta} + \mathbf{z}, \tag{1.1}$$

여기서 $\mathbf{y} = (y_1, \ldots, y_n)'$인 $n \times 1$ 종속변수 벡터이고, X는 $n \times p$인 알려진 설명변수 행렬을 나타내고 $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})'$는 $p \times 1$ 회귀계수 벡터를 나타낸다. 아울러 $\mathbf{z} = (z_1, \ldots, z_n)'$는 $n \times 1$ 오차항 벡터를 나타내며, 이의 각 확률변수는 영(zero)에 대하여 대칭이며, 서로 독립이고 동일한 미지인 분포함수 F와 확률밀도함수 f를 가진다고 가정하자. 회귀분위수 아이디어의 근간은 위치모수의 표본 백분위수를 회귀모형으로 자연스럽게 확장한 것이다. 먼저 $0 < \theta < 1$에 대하여 $\Psi_\theta(x) = \theta - I(x < 0)$라고 하자. 여기서 $I(\cdot)$는 지시함수를 나타낸다. 이때 $\rho_\theta(x) = x\Psi_\theta(x)$라 정의하면 위치모수에 대한 일반적인 θ-차 표본 백분위수는 다음과 같은 check 함수를 가지는 M-추정량에 의해 구할 수 있다.

$$\rho_\theta(x) = \begin{cases} \theta x, & x \geq 0, \\ (\theta - 1)x, & x < 0. \end{cases} \tag{1.2}$$

식 (1.2)를 회귀모형에 적용하여 계산한 $K(\theta)$를 θ-차 회귀분위수라 정의하면 $K(\theta)$는 다음의 식을 만족하는 값이 된다.

$$\min_{b \in R^p} \sum_{j=1}^{n} \rho_\theta \left(y_j - \mathbf{x}_j' \boldsymbol{b} \right). \tag{1.3}$$

2. 제안된 추정량의 점근적 성질

먼저 본 절을 시작하기 전에 본 논문에서 제시된 정리나 따름정리 등에 부과되는 몇 가지 기호나 가정들을 소개하고자 한다. $\mathbf{e} = (1, 0, \ldots, 0)'$는 첫 번째 항만 1이고 나머지는 모두 0으로 이루어진 $p \times 1$ 벡터이고, I_p는 $p \times p$ 단위행렬이라 하자. $0 < p_i < 1/2$에 대하여 $\eta(p_i) = F^{-1}(p_i)$로 표시하고, 기저분포의 0에 대한 대칭성에 의해 $-\eta(p_i) = F^{-1}(1 - p_i)$로 표시하고자 한다. 아울러 $N_p(\boldsymbol{\mu}, \Sigma)$는 평균벡터가 $\boldsymbol{\mu}$이고 공분산행렬이 Σ인 p-변량 정규분포라 정의하자. 본 논문에서는 $0 < p_0 < p_1 < \cdots < p_k = 1/2$이라고 놓고 $n \to \infty$일 때, $n^{-1}X'X \to Q$, Q는 양정치행렬이라고 하자. 그리고 $b_i = p_i + 0.5r_i$, $r_i = p_i - p_{i-1}\,(i = 1, 2, \ldots, k)$로 놓고 $K(b_i)$를 $100b_i$ 회귀 분위수라 하자. 그러면 $K_i = (K(b_i) + (K(1 - b_i))/2$로 정의할 때 본 논문에서 제안하는 추정량은 다음과 같은 형태를 갖는다.

$$B_k = w_1K_1 + w_2K_2 + \cdots + w_kK_k, \quad \sum_{i=1}^{k} w_i = 1. \tag{2.1}$$

식 (2.1)에서 $K_1^*, \ldots, K_k^*$를 $K_1, K_2, \ldots, K_k$의 추정량이라 했을 때, 본 논문에서 제안하는 회귀모수 $\boldsymbol{\beta}$에 대한 추정량은 다음과 같이 주어진다.

$$S_k = w_1 K_1^* + w_2 K_2^* + \cdots + w_k K_k^*, \quad \sum_{i=1}^{k} w_i = 1. \tag{2.2}$$

이때 가중치 벡터 $\boldsymbol{w} = (w_1, \ldots, w_k)'$는 S_k의 점근적 분산이 최소가 되도록 선택할 수 있다. 가중치의 결정문제는 본 장의 후반부에서 자세히 다룰 것이다. k의 선택문제에 있어서 실제적인 모의실험에서 데이터의 수가 100개 이하인 경우는 $k = 2$인 경우가 충분하며, 데이터의 수가 커지는 경우 2보다 큰 k도 고려할 수 있다. 실제 본 연구의 모의실험 결과 $n = 160$인 경우 $k = 3$에서 $k = 2$보다 더 좋은 결과를 얻었다. 가중치의 경우, $k = 2$인 경우 기저분포가 정규분포와 유사하면 w_1과 w_2에 비슷한 가중치를 줄 것이며, 기저분포가 정규분포보다 두꺼운 꼬리부분을 가지면 w_1에 w_2보다 더 적은 가중치를 줄 것이다.

이제 본 논문에서 제안하는 추정량의 점근적 성질 파악을 위해 필요한 정리를 제시하고자 한다. 몇 가지 가정하에서 k개의 회귀분위수 $K(q_i), i = 1, 2, \ldots, k$이 주어질 때 각 추정량을 $K^*(q_i), i = 1, 2, \ldots, k$이라고 하면 본 논문의 정리 2.1을 위해 필요한 다음과 같은 보조정리를 얻을 수 있다.

보조정리 2.1 $i \le j \le k$에 대하여 $\Omega = (\Omega_{ij})_{k\times k}$는 $0 < q_1 < \cdots < q_k < 1$에 대하여 $\sqrt{n}(K^*(q_1) - K(q_1), \ldots, K^*(q_k) - K(q_k)) \xrightarrow{d} N_{kp}(0, \Omega \otimes Q^{-1})$이다. 여기서 $\Omega_{ij} = q_i(1-q_j)/f(\eta(q_i))f(\eta(q_j)), 1 \le i \le j \le k$, $K(q_i) = \boldsymbol{\beta} + \eta(q_i)\mathbf{e}$이고, $\otimes$는 크로네커 곱을 나타내며, $\xrightarrow{d}$는 분포적 수렴(CONVERGENCE IN DISTRIBUTION)을 의미한다.

증명: DasGupta (2008) 정리 7.6 참조. □

보조정리 2.2 $i \le j \le k$에 대하여 $\sqrt{n}(K_1^* - \boldsymbol{\beta}, \ldots, K_k^* - \boldsymbol{\beta}) \xrightarrow{d} N_{kp}(0, \Pi \otimes Q^{-1})$이다. 여기서 $\Pi = (\pi_{ij})_{k\times k}$이며, $0 < b_1 < \cdots < b_k < 1/2$에 대하여 $\pi_{ii} = (b_i/2)f^2(\eta(b_i))$이고, $\pi_{ij} = (b_i/2)f(\eta(b_i))f(\eta(b_j))$이다.

증명: 본 논문의 정의에서 $K_i^* = (K^*(b_i) + K^*(1-b_i))/2$이고 $K_i = (K(b_i) + K(1-b_i))/2 = (\boldsymbol{\beta} + \eta(b_i) + \boldsymbol{\beta} + \eta(1-b_i))/2 = \boldsymbol{\beta}$가 된다. K_i^*가 두 개의 회귀백분위수 $K^*(b_i)$와 $K^*(1-b_i)$의 선형결합이므로 보조정리 2.1을 적용하면 $\sqrt{n}(K_1^* - \boldsymbol{\beta}, \ldots, K_k^* - \boldsymbol{\beta})$는 점근적으로 다변량 정규분포를 따른다. 따라서 분산-공분산 행렬의 요소만 결정하면 되므로 $\sqrt{n}(K_i^* - \boldsymbol{\beta})$의 분산과 $i \ne j$에 대하여 $(\sqrt{n}(K_i^* - \boldsymbol{\beta}), \sqrt{n}(K_j^* - \boldsymbol{\beta}))$의 공분산 형태만 결정하면 된다. □

보조정리 2.1에서 $k = 2$인 경우를 적용하면 점근적 결합분포(Asymptotic joint distribution)는 다음과 같이 구해진다.

$$\sqrt{n}\left(K^*(b_i) - K(b_i), K^*(1-b_i) - K(1-b_i)\right) \xrightarrow{d} N_{2p}\left(0, \Omega \otimes Q^{-1}\right),$$

여기서

$$\Omega = \begin{pmatrix} b_i(1-b_i)/f^2(\eta(b_i)) & b_i^2/f^2(\eta(b_i)) \\ b_i^2/f^2(\eta(b_i)) & b_i(1-b_i)/f^2(\eta(b_i)) \end{pmatrix}.$$

윗 식을 통해 $\sqrt{n}(K^*(b_i) - K(b_i)) \xrightarrow{d} N_p(0, b_i(1-b_i)/f^2(\eta(b_i))Q^{-1})$이고 $\sqrt{n}(K^*(1-b_i) - K(1-b_i)) \xrightarrow{d} N_p(0, b_i(1-b_i)/f^2(\eta(b_i))Q^{-1})$이므로 간단한 계산에 의해서 $\sqrt{n}(K_i^* - \boldsymbol{\beta}) \xrightarrow{d} N_p(0, (b_i/2)f^2(\eta(b_i))Q^{-1})$임을 알 수 있고, 이를 통해 π_{ii}를 얻을 수 있다. 더불어 보조정리 2.1을 이용한 간단한 계산과정으로 공분산은 $\mathrm{Cov}(\sqrt{n}(K_i^* - \boldsymbol{\beta}), \sqrt{n}(K_j^* - \boldsymbol{\beta})) \to (b_i/2)f(\eta(b_i))f(\eta(b_j))\ Q^{-1}$가 되며, 이를 통해 π_{ij}를 얻을 수 있다.

정리 2.1 k가 임의의 고정된 값이라 하고, $0 < p_0 < p_1 < \cdots < p_k = 1/2$라고 하자. $q_i = p_i + 0.5r_i$, $r_i = p_i - p_{i-1}$, $i = 1, 2, \ldots, k$일 때 $\sqrt{n}(S_k - \boldsymbol{\beta}) \xrightarrow{d} N_p(0, w'\Pi w Q^{-1})$이다.

증명: 식 (2.2)로부터 $S_k = w_1 K_1^* + w_2 K_2^* + \cdots + w_k K_k^*$이므로 $S_k - \boldsymbol{\beta} = w_1(K_1^* - \boldsymbol{\beta}) + w_2(K_2^* - \boldsymbol{\beta}) + \cdots + w_k(K_k^* - \boldsymbol{\beta})$이 된다. 보조정리 2.2로부터 $\sqrt{n}(S_k - \boldsymbol{\beta}) \xrightarrow{d} N_p(0, w'\Pi w Q^{-1})$임을 쉽게 확인할 수 있다. □

이제 S_k의 점근적 분산을 최소로 하는 가중치 벡터 w를 결정하는 문제를 생각하자. 일반적으로 S_k의 점근적 분산은 기저분포가 미지이기 때문에 적은 수의 표본으로 추정하는 것은 매우 힘들다. 그러므로 본 논문에서는 이의 근사를 생각하고자 한다. 먼저 $i = 1, 2, \ldots, k$에 대하여 $p_2 - p_1 = p_3 - p_2 = \cdots = p_k - p_{k-1} = q_k$라고 하고 다음의 근사를 생각하자.

$$f(\eta(b_i)) \sim \frac{q_k}{d_i},$$

여기서 $d_i = \eta(b_i) - \eta(b_{i-1})$이고 $b_i = p_{i-1} + 0.5q_k$이다. 위의 근사식은 히스토그램을 이용한 간단한 근사이고, 이를 이용하여 π_{ii}와 π_{ij}의 근사를 다음과 같이 생각할 수 있다.

$$\pi_{ii} \sim \frac{1}{2} r_k^{-2} b_i d_i^2, \qquad \pi_{ij} \sim \frac{1}{2} r_k^{-2} b_i d_i d_j. \tag{2.3}$$

식 (2.3)을 이용하면 Π의 근사적 공분산은 다음과 같이 구해진다.

$$S = \frac{1}{2} q_k^{-2} D \begin{bmatrix} b_1 & b_1 & \cdots & b_1 \\ b_1 & b_2 & \cdots & b_2 \\ \vdots & \vdots & \ddots & \vdots \\ b_1 & b_2 & \cdots & b_k \end{bmatrix} D, \tag{2.4}$$

여기서 $D = \mathrm{diag}(d_1, d_2, \ldots, d_k)$이다. 그러므로 본 논문에서 구하고자 하는 가중치 벡터 w는 식 (2.4)에 나타난 S를 이용하여 다음과 같은 식을 만족하는 값을 찾으면 된다.

$$\min_{w'i=1} (w'Sw) Q^{-1}, \tag{2.5}$$

여기서 $i = (1, 1, \ldots, 1)'$는 모든 원소가 1로 이루어진 $k \times 1$벡터이다. 이때 Q^{-1}는 고정된 값이므로 $w'i = 1$인 조건하에서 $w'Sw$를 최소로하는 가중치 벡터의 값을 결정하면 된다. 이 문제는 간단한 라그랑쥐(Lagrange) 승수법을 적용하여 해결할 수 있으며, 그 해는 $w = (i'S^{-1}i)^{-1}S^{-1}i$와 같이 구해진다. 이러한 해를 이용하면 가중치에 대하여 다음과 같은 구체적인 해의 형태를 얻을 수 있다.

$$w_i = \frac{\delta_i}{\sum_{i=1}^{k} \delta_i}, \quad i = 1, 2, \ldots, k, \tag{2.6}$$

여기서 δ_i는 다음과 같이 구해진다.

$$\begin{aligned} \delta_1 &= \left(\frac{2b_2}{b_1}\right)\left(\frac{1}{d_1^2} - \frac{1}{d_1 d_2}\right) \\ \delta_i &= \frac{1}{d_i}\left(\frac{2}{d_i} - \frac{1}{d_{i-1}} - \frac{1}{d_{i+1}}\right), \quad i = 2, 3, \ldots, k-1 \\ \delta_k &= \frac{1}{d_k^2} - \frac{1}{d_{k-1} d_k}. \end{aligned} \tag{2.7}$$

만약 예비적합에 의한 잔차로서 d_i의 일치추정량 $\hat{d}_i$로 치환하여 구성된 δ_i의 일치추정량 δ_i^*를 얻었다고 하면 추정된 가중치를 이용한 본 논문의 최종적인 추정량의 형태는 다음과 같이 얻어진다.

$$\hat{S}_k = w_1^* K_1^* + w_2^* K_2^* + \cdots + w_k^* K_k^*, \tag{2.8}$$

여기서 $w_i^* = \delta_i^* / \sum_{i=1}^k \delta_i^*$, $i = 1, 2, \ldots, k$이다.

마지막으로 $\sqrt{n}(\hat{S}_k - \boldsymbol{\beta})$와 $\sqrt{n}(S_k - \boldsymbol{\beta})$가 동일한 점근적 분포를 갖는다는 것을 보여주고자 한다. 이를 위해 다음 정리 2.2는 가중치 w_i의 일치추정량 w_i^*를 예비적합에 따른 잔차로 부터 얻기 위해 매우 유용한 정리이다.

정리 2.2 $0 < p_0 < p_1 < \cdots < p_k = 1/2$라 하고 $\hat{\eta}(p_i)$를 $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathbf{0}} - \boldsymbol{\beta} - c\mathbf{e}) = O_p(1)$을 만족시키는 예비적합 추정량 $\hat{\boldsymbol{\beta}}_{\mathbf{0}}$로부터 얻은 np_i번째 순서 잔차량이라고 하자. 그러면 $i = 1, 2, \ldots, k$에 대하여 $\hat{d}_i$는 d_i의 일치추정량이다. 단 c는 상수이다.

증명: Rupport와 Carroll (1980)의 보조정리 1을 사용하면 $i = 1, 2, \ldots, k$에 대하여 다음과 같은 결과를 얻을 수 있다.

$$\sqrt{n}\left(\hat{\eta}(p_i) - \eta(p_i)\right) = \left[f(\eta(p_i)\right]^{-1} n^{-\frac{1}{2}} \sum_{j=1}^{n} \psi_{p_i}\left(z_j - \eta(p_i)\right) - e' n^{\frac{1}{2}} \left(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}\right) + O_p(1)$$

$$\sqrt{n}(\hat{\eta}(p_{i-1}) - \eta(p_{i-1})) = [f(\eta(p_{i-1})]^{-1} n^{-\frac{1}{2}} \sum_{j=1}^{n} \psi_{p_{i-1}}(z_j - \eta(p_{i-1})) - e' n^{\frac{1}{2}} \left(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}\right) + O_p(1), \tag{2.9}$$

여기서 $\psi_\theta(x) = \theta - I(x < 0)$이고 $I(\cdot)$은 지시함수를 나타낸다. 그러므로 식 (2.9)에 나타난 두 식을 빼면 다음과 같은 결과를 얻을 수 있다.

$$\sqrt{n}\left(\hat{d}_i - d_i\right) = [f(\eta(p_i)]^{-1} n^{-\frac{1}{2}} \sum_{j=1}^{n} \psi_{p_i}(z_j - \eta(p_i)) - [f(\eta(p_{i-1})]^{-1} n^{-\frac{1}{2}} \sum_{j=1}^{n} \psi_{p_{i-1}}(z_j - \eta(p_{i-1})).$$
$$+ O_p(1). \tag{2.10}$$

중심극한정리에 의하여 식 (2.10)의 오른쪽의 처음 두 개의 항은 유한한 분산을 갖는 점근적 정규분포를 따른다. 그러므로 모든 $i = 1, 2, \ldots, k$에 대하여 $\hat{d}_i - d_i = o_p(1)$을 만족하게 된다.

정리 2.2에 의해 회귀모형에서 기저 오차분포의 백분위수 차이에 대한 일치추정량을 얻을 수 있으며, 이를 이용하여 식 (2.8)에 주어진 $\hat{S}_k$의 일치추정량을 구성할 수가 있다. □

따름정리 2.1. $\sqrt{n}(\hat{S}_k - \boldsymbol{\beta})$와 $\sqrt{n}(S_k - \boldsymbol{\beta})$는 동일한 극한분포를 가진다.

증명: $\hat{S}_k$와 S_k의 구성에 의하여 다음과 같은 결과를 얻는다.

$$\sqrt{n}\left[\left(\hat{S}_k - \boldsymbol{\beta}\right) - \left(S_k - \boldsymbol{\beta}\right)\right] = \sum_{i=1}^{k} \left(w_i^* - w_i\right) \sqrt{n}\left(K_i^* - \boldsymbol{\beta}\right). \tag{2.11}$$

식 (2.11)에서 $w_i^* \xrightarrow{p} w_i$와 $\sqrt{n}(K_i^* - \boldsymbol{\beta})$는 분포상 유계(bounded in distribution)이므로 $\sqrt{n}(\hat{S}_k - \boldsymbol{\beta})$와 $\sqrt{n}(S_k - \boldsymbol{\beta})$는 동일한 극한분포를 갖는다. 단 $\xrightarrow{p}$는 확률적 수렴(convergence in probability)을 의미한다. □

3. 모의실험

3.1. 모의실험 디자인 및 추정량

본 연구에서 제안한 추정량의 효율성을 비교하기 위하여 모의실험을 실시하였다. 모의실험은 단순회귀모형 $y_j = \beta_0 + \beta_1 x_j + \epsilon_j \, (j = 1, \ldots, n)$을 사용하였다. β_0와 β_1의 참 값은 모두 0으로 가정하였으며, 설명변수 X의 첫 번째 열은 모두 1이고 두 번째 열은 $x_j = \Phi^{-1}(j/n+1)$을 이용하였다. 이때 $\Phi^{-1}(\cdot)$는 표준정규분포의 CDF의 역함수를 나타낸다. 아울러 x_j들은 $\sum x_j^2 = 1$이 되도록 표준화하였다.

본 논문에서 제안하는 추정량은 $k = 2$인 경우와 $k = 3$인 경우 등 2개의 추정량이다. 먼저 $k = 2$인 경우의 추정량은 다음과 같다.

$$\mathrm{JH}(2) = \frac{\hat{w}_1}{\hat{w}_1 + \hat{w}_2} K_1^* + \frac{\hat{w}_1}{\hat{w}_2 + \hat{w}_2} K_2^*. \tag{3.1}$$

식 (3.1)에 나타난 추정량에서 $p_0 = 0.05$와 $p_1 - p_0 = p_2 - p_1 = q_2 = 0.225$를 사용하였다. 즉, $b_1 = p_0 + 0.5q_2$, $b_2 = p_1 + 0.5q_2$라 놓고 $d_1 = \eta(p_1) - \eta(p_0)$, $d_2 = \eta(p_2) - \eta(p_1)$로 놓자. 그러면 $k-2$인 경우 분산-공분산 행렬 Π의 구성요소 각각의 근사값은 $\pi_{11} \sim 1/2q_2^{-2} b_1 d_1^2$, $\pi_{12} \sim 1/2q_2^{-2} b_1 d_1 d_2$ 및 $\pi_{22} \sim 1/2q_2^{-2} b_2 d_2^2$의 형태를 갖는다. 만약 $V_1 = 1/2 b_1 d_1^2$, $V_2 = 1/2 b_2 d_2^2$, $C = 1/2 b_1 d_1 d_2$라고 하면 본 논문에서 제안하는 JH(2)추정량의 분산의 근사값은 다음과 같이 구해진다.

$$V\left(\mathrm{JH}(2)\right) = q_2^{-2} w' \begin{bmatrix} V_1 & C \\ C & V_2 \end{bmatrix} w Q_0^{-1}, \quad w' i = 1. \tag{3.2}$$

아울러 식 (3.2)의 분산을 최소로 하는 해는 $w_1 = (V_2 - C)/(V_1 + V_2 - 2C)$과 $w_2 = (V_1 - C)/(V_1 + V_2 - 2C)$임을 쉽게 확인할 수 있다. 회귀 예비적합 추정치(L_1)로부터 얻어진 잔차를 순서통계량하여 d_1과 d_2의 일치추정량 $\hat{d}_1$과 $\hat{d}_2$를 구하고 이로부터 가중치 벡터의 일치추정량 $\hat{w} = (\hat{w}_1, \hat{w}_2)'$을 구할 수 있다. 이때 기저분포의 대칭성의 가정에 의해 d_1과 d_2의 일치추정량은 다음의 형태를 취한다.

$$\hat{d}_i = \frac{(\hat{\eta}(p_i) - \hat{\eta}(p_{i-1})) + (\hat{\eta}(1 - p_{i-1}) - \hat{\eta}(1 - p_i))}{2}, \quad i = 1, 2, \tag{3.3}$$

여기서 $\hat{\eta}(p_i)$와 $\hat{\eta}(1 - p_i)$는 각각 잔차의 $100p_i$와 $100(1 - p_i)$ 백분위수이다. 마지막으로 JH(2) 추정량에 사용되는 회귀 백분위 추정량으로는 $K_1^* = [K^*(0.1625) + K^*(0.8375)]/2$와 $K_2^* = [K^*(0.3875) + K^*(0.6125)]/2$를 사용하였다. 이와 같은 과정에서 JH(2) 추정량을 얻고, JH(2)추정량에서 기울기 부분만을 고려한 것이 본 논문에서 제안하는 회귀모형의 기울기에 대한 추정량이 된다.

$k = 3$인 경우의 회귀계수 추정량 JH(3)도 JH(2)를 구하는 것과 비슷한 방법으로 $p_0 = 0.05, p_1 = 0.20, p_2 = 0.35, p_3 = 0.5$를 사용해서 쉽게 추정할 수 있다. 이에 대한 자세한 과정은 지면관계상 생략하였다.

본 논문에서 제안한 추정량과 효율성을 비교하기 위하여 최소제곱 추정량(LS), 대표적인 L-추정량과 M-추정량을 사용하였다. 먼저 L-추정량의 경우 최소절대값추정량(L_1), 10% 및 20% 절사회귀(regression trimmed mean) 추정량 $T(10)$과 $T(20)$ 등 3개의 추정량을 사용하였으며, M-추정량의 경우 Huber, Tukey 및 Huber-Collins의 회귀 M-추정량을 각각 사용하였다. 여기에서는 단순회귀모형에서 Huber와 Tukey의 회귀 M-추정량에 대하여 간단히 언급하고자 한다. 먼저 회귀 최소절대값 추정량인 L_1을 회귀모수에 대한 예비추정량이라 하고, 그 추정량으로 부터 얻은 잔차를 각각 $r_j = y_j - \hat{\beta}_0 = \hat{\beta}_1 x_j$이라 하자. 이 경우 $\mathrm{MAD} = \mathrm{median}_j\{|r_j - \mathrm{med}_l(r_j)|\}$라 하고, $s = \mathrm{MAD}/0.6745$라 하면 회귀계수에 대한 M-추정량의 형태는 다음의 가중최소제곱 방정식의 해가 된

다 (Holland와 Welsch, 1977; Hogg 등, 1988).

$$\sum_{j=1}^{n} w_j (y_j - \beta_0 - \beta_1 x_j)^2 = 0. \tag{3.4}$$

이때 조절상수 k를 갖는 Huber의 M-추정량과 Tukey의 M-추정량은 다음과 같은 형태를 갖는다.

$$\tilde{\beta}_1 = \left[\sum_{j=1}^{n} w_j \sum_{j=1}^{n} X_j Y_j - \sum_{j=1}^{n} w_j X_j \sum_{j=1}^{n} w_j Y_j\right] \Big/ \left[\sum_{j=1}^{n} w_j X_j^2 - \left(\sum_{j=1}^{n} w_j X_j\right)^2\right]. \tag{3.5}$$

식 (3.5)에 나타난 M-추정량에서 Huber 형태는 가중치 w_j를 다음과 같이 정하여 얻을 수 있다.

$$w_j = \begin{cases} 1, & \text{if } |r_j| \le ks, \\ ks, & \text{otherwise.} \end{cases} \tag{3.6}$$

식 (3.6)에 나타난 가중치를 사용한 M-추정량을 $H(k)$라 정의한다. Tukey 형태의 M-추정량은 다음과 같은 가중치 w_j를 사용하여 얻어진다.

$$w_j = \begin{cases} \{1 - (r_j/ks)^2\}^2, & \text{if } |r_j| \le ks, \\ 0, & \text{otherwise.} \end{cases} \tag{3.7}$$

식 (3.7)에 나타난 가중치를 사용한 M-추정량을 $T(k)$라 정의한다. Huber (1981)에 의하면 기저 오차분포의 형태가 $f(x)$일 때, 회귀 M-추정량의 점근적 분산은 $E(\phi)^2/(E((\phi'))^2 * X'X)^{-1}$와 같은 형태를 갖는다고 알려져 있다. 여기서 $\phi = \log \rho(x)$, $\rho = -\log f(x)$이다. 본 연구의 모의실험에서 사용된 추정량은 $H(1.25)$ 및 $T(4.82)$이다. 아울러 Huber-Collins형태의 회귀 M-추정량에는 $p = 1.277$, $x_1 = 1.344$ 및 $r = 4$를 사용하였다 (Hampel 등, 1986). 이와 같은 M-추정량은 HC라 표기하도록 한다.

모의실험에 사용된 오차분포는 아래의 6개를 사용하였다. 그리고 $X = Z/V$ 형태의 확률변수에서 Z, V는 서로 독립이고 Z는 표준정규분포의 확률변수이고, V는 다음과 같다.

(1) Normal (NOR): $V = 1$.

(2) Slate (TE): $V = U^{1/10}$, 여기서 U는 균일분포(0, 1)을 따르는 확률변수.

(3) Slacu (CU): $V = U^{1/3}$.

(4) Contaminated Normal (CON): $V = \begin{cases} 1, & \text{with prob. } 0.9, \\ 1/3, & \text{with prob. } 0.1. \end{cases}$

(5) Double Exponential (DE): $V = 1/\sqrt{W}$, 여기서 W는 자유도 2인 카이제곱 확률변수.

(6) Cauchy (CA): $V = |Z'|$, 여기서 Z'는 표준정규분포를 따르는 확률변수.

이와 같이 모의실험에서 사용된 6개의 대칭 오차분포 중에서 2개는 가벼운 꼬리부분을 가지는 분포(NOR, TE)이고, 2개는 약간 무거운 꼬리부분을 갖는 분포(CU, CON)이며, 1개는 중심부분이 아주 뾰족한 분포(DE)이다. 나머지 하나는 아주 꼬리부분이 아주 무거운 분포(CA)이다.

모의실험에 사용된 표본수는 $n = 40$, 80 및 160이었으며, 모든 실험조합에서 총 10,000번의 반복을 실시하였다. 각 반복에서는 본 논문에서 제안한 2개의 L-추정량, 최소제곱추정량, 대표적인 L-추정량 3개(L_1, $T(10)$, $T(20)$) 및 대표적인 3개의 M-추정량($H(1.25)$, $T(4.82)$, HC)을 이용하여 회귀계수를 추정하였다. 10,000번의 반복을 통하여 얻어진 각 추정량들을 이용하여 평균제곱오차(Mean Square Error; MSE)를 계산하여 본 논문에서 제안하는 추정량의 효율성을 파악하였다.

Table 3.1. MSE of various estimators of slope parameter[a]

Sample size	Estimator	Distribution					
		NOR	TE	CU	CON	DE	CA
$N = 40$	LS	9923*	12491*	30342	17914	20067	-[b]
	L_1	15630	19356	27905	18436	13662	28276
	$H(1.25)$	10772	13555	20689	13441	13343	32956
	$T(4.82)$	10682	13513	20683*	13298	13662	27486
	HC	11251	14197	21100	13523	13699	26073
	$T(10)$	10448	13657	21686	13421	13768	53879
	$T(20)$	11254	13465	21887	13254*	15324	30114
	JH(2)	11447	14312	23683	14531	12693*	22483*
	JH(3)	11765	15454	24746	14846	12887	23272
$N = 80$	LS	10080*	12504*	30824	17840	19386	-
	L_1	15825	18920	28472	17605	13291	28573
	$H(1.25)$	10957	13274	21331	13075	13247	34341
	$T(4.82)$	10882	13183	21252*	12870*	13130	28518
	HC	11427	13810	21674	13152	13179	26942
	$T(10)$	10686	13051	21947	13347	15127	53211
	$T(20)$	11383	13924	22053	13511	13285	29843
	JH(2)	11523	14082	23564	14435	12356	22807*
	JH(3)	11187	14554	23941	14356	12254*	23475
$N = 160$	LS	10064*	12643*	30045	17789	20110	-
	L_1	15536	18533	28043	18001	13118	28433
	$H(1.25)$	10886	13345	20953	13228	13328	33289
	$T(4.82)$	10598	13365	20896*	12779*	13328	27984
	HC	11289	14021	21335	13345	13621	26554
	$T(10)$	10735	13265	21454	13365	15088	52930
	$T(20)$	11552	13547	21787	13433	13287	28944
	JH(2)	11488	14065	23169	14781	12272	22768*
	JH(3)	11044	13207	22786	14031	12054*	22876

[a] MSE값은 실제 얻어진 MSE값을 10,000배 한 것이다.

[b] CA의 경우 LS추정량의 MSE가 너무 큰 값이 나오므로 −로 표시하였다.

∗ 제안된 추정량들 중에서 MSE 값이 가장 작은 추정량을 ∗로 표시하였다.

3.2. 모의실험 결과

다음 Table 3.1은 모의실험 결과 얻어진 표본수에 따른 각 추정량의 MSE를 나타낸다. Table 3.1의 결과를 요약하면 다음과 같은 결과를 얻을 수 있다.

최소제곱추정량(LS)은 기저 오차분포가 가벼운 꼬리(NOR, TE)일 때 최상의 효율을 가졌다. 그러나 그 외의 분포 군에서는 다른 추정량들에 비해 효율이 현격히 떨어지며, 특히 극단적인 무거운 꼬리를 가지는 경우(CA) 다른 추정량에 비하여 0에 가까운 효율을 가졌다.

M-추정량인 $H(1.25)$, $T(4.82)$, HC 중에서 $T(4.82)$가 전반적으로 가장 효율이 좋은 추정량임을 알 수 있다. 반면 HC 추정량은 극단적으로 무거운 꼬리를 가진 분포의 경우 외에는 $H(1.25)$, $T(4.82)$에 비해 효율이 떨어짐을 알 수 있다. 특히 $H(1.25)$와 $T(4.82)$는 뾰족한 중심 부분을 가지는 이중지수분포(DE)나 무거운 꼬리를 가지는 코쉬 분포(CA) 외의 분포 군에 대해 본 논문에서 제안한 L-추정량인 JH(2)나 JH(3) 보다 약간 좋은 효율을 보여주었다. L-추정량에 속하는 20% 절사회귀추정량인 $T(20)$

추정량은 모든 기저 오차 분포 군에서 무난한 효율을 보여주었다. 반면 기저 오차분포가 뾰족한 중심부분을 가지는 DE 나 무거운 꼬리를 가지는 CA분포에서는 같은 L-추정량인 JH(2)나 JH(3) 보다 비효율적인 추정량임을 알 수 있다. 최소절대값 추정량(L_1)은 DE처럼 극단적으로 뾰족한 중심을 가진 분포나 CA처럼 극단적으로 무거운 꼬리를 가지는 분포 외에는 다른 모든 M-추정량에 비해 훨씬 못한 효율성을 보이는 것으로 나타났다.

본 논문에서 제안한 JH(2) 추정량은 표본수 $N = 40$에서는 JH(3)보다 좋은 효율을 보여주었다. 표본수 $N = 80$에서는 두 추정량간 뚜렷한 차이를 보이지 않다가 표본수가 $N = 160$일 때 극단적으로 무거운 꼬리를 가지는 코쉬 분포(CA) 외에서는 JH(3)보다 못한 효율을 보여주었다. 즉 표본 수가 증가함에 따라 JH(2)나 JH(3) 추정량의 효율도 좋아졌으나 JH(3) 추정량이 JH(2)보다 더 효율적인 추정량으로 나타났다.

4. 결론 및 제언

본 연구에서는 선형회귀모형에서 회귀분위수를 이용한 L-추정법을 제안하였다. 이는 회귀분위수를 이용하여 근사분산을 최소화하는 추정량을 제안한 것이다. 모의실험 결과 본 논문에서 제안한 L-추정량 JH(2)와 JH(3)는 무거운 꼬리를 가진 분포에 대해서 기존의 L-추정량이나 M-추정량에 비하여 효율적인 것으로 나타나 기존의 M-추정법에 대한 하나의 대안으로 사용될 수 있을 것으로 생각된다. 향후 기저 분포에 대한 정보를 제공하는 선택통계량을 이용한 적합한 추정량의 선택에 대한 좀 더 정교한 연구가 이루어진다면 좀 더 효율적인 L-추정법을 제안할 수 있을 것으로 생각한다.

References

Adichie, J. N. (1974). Rank score comparison of several regression parameter, *Annals of Statistics*, **2**, 396–402.

Barrodale, I. and Roberts, F. D. K. (1974). Solution of an overdetermined system of equations in the L norm, *Communications of the Association for Computing Machinery*, **17**, 407–415.

Bickel, P. J. (1973). On some analogues to linear combinations of ordered statistics in the linear model, *Annals of Statistics*, **1**, 597–616.

DasGupta, P. J. (2008). *Asymptotic Theory of Statistics and Probability*, Springer.

De Jongh, P. J. and De Wet, T. (1985). A Monte Carlo comparison of regression trimmed means, *Communications in Statistics - Theory and Methods*, **10**, 2457–2472.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. M. and Stahel, W. A. (1986). *Robust Statistics, the Approach Based on Influence Function*, John Wiley.

Han, S. M. (2003). Adaptive M-estimation in regression model, *The Korean Communications in Statistics*, **10**, 859–871.

Hettmansperger, T. P. and McKean, J. W. (1977). A robust alternative based on ranks to least squares in analyzing linear models, *Technometircs*, **19**, 275–284.

Hogg, R. V. (1983). On adaptive statistical inference, *Communications in Statistics - Theory and Methods*, **11**, 2531–2542.

Hogg, R. V., Bril, G. K., Han, S. M. and Yuh, L. (1988). An argument for adaptive Robust estimation, *Probability and Statistics, Essays in Honor of Graybill, F. A.*, North Holland, 135–148.

Holland, P. W. and Welsch, R. E. (1977). Robust regression using iteratively reweighted least-squares, *Communications in Statistics - Theory and Methods*, **6**, 813–827.

Huber, P. J. (1973). Robust regression: Asymptotics, Conjectures and Monte Carlo, *Annals of statistics*, **1**, 799–821.

Huber, P. J. (1981). *Robust Statistics*, John Wiley.

Jaeckel, L. A. (1971). Robust estimates of location: Symmetry and asymmetric contamination, *The Annals of Mathematical Statistics*, **42**, 1020–1034.

Johns, M. V. (1974). Nonparametric estimation of location, *Journal of the American Statistical Association*, **69**, 453–460.

Koenker, R. and Bassett, C. (1978). Regression quantiles, *Econometrica*, **46**, 33–50.

Koenker, R. and Bassett, C. (1982). Robust tests for Heterosedasticity based on regression quantiles, *Econometrica*, **50**, 43–61.

Koenker, R., Hammond, P. and Holly, A. (2005). *Quantile Regression*, Cambridge University Press.

Portnoy, S. and Koenker, R. (1997). The Gaussian Hare and the Laplacian Tortoise: Computability of squared-error versus absolute-error estimators, *Statistical Science*, **12**, 299–300.

Ruppert, D. and Carroll, R. J. (1980). Trimmed least squares estimation in the linear model, *Journal of the American Statical Association*, **75**, 828–838.