The Korean Communications
in Statistics Vol. 8, No. 1, 2001
pp. 15-27

Nonparametric Estimation in Regression Model[1)]

Sang Moon Han[2)]

Abstract

One proposal is made for constructing nonparametric estimator of slope parameters in a regression model under symmetric error distributions . This estimator is based on the use of idea of Johns for estimating the center of the symmetric distribution together with the idea of regression quantiles and regression trimmed mean. This nonparametric estimator and some other L-estimators are studied by Monte Carlo.

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

1. 서 론

회귀모형에서 로버스트적 추정법은 지난 30여년간 꾸준히 연구되어왔고 최소자승법(least squares method)의 대안으로 많은 추정량들이 제시되어 왔다. 이러한 로버스트적 추정량들은 크게 분류하여 M-, L-, R-군으로 나누어 질 수 있는데 M-추정량은 Huber (1973)의 ψ 함수를 이용한 위치모수추정법에서 자연스럽게 회귀모수추정법으로 확장될수 있으며 R-추정법은 Jureckova(1971), Jaeckel(1972), Archie(1974), Hettemansperger 와 McKean(1977)등에 의해서 회귀모형으로 확장되었다. 그리고 Hogg(1988)등은 M-,L-,R-추정법에 대한 광범위한 모의실험을 통해 기존의 추정법에 대한 로버스트 추정법의 우월성을 보였다. L-추정법은 Bickel(1973)에 의해서 처음으로 제시되었고 Bickel의 추정량은 점근적인 좋은 성질에도 불구하고 복잡한 형태를 취하고 있고 계산하기에 매우 복잡하였다. 이에 따라 Konker 와 Bassett (1978,1982)는 예비적합(preliminary fit)에 의한 잔차의 순서통계량에 의하지 않는 L-추정량을 제시하였다. 그들은 소위 특정한 check함수상에서의 M-추정량으로 회귀분위수를 정의하는 방법을 취하고 이의 점근적성질이 위치모수의 위치분위수와 비슷하다는 것을 발견하였다. 그리고 Ruppert 와 Carroll(1980) 은 Konker 와 Bassett의 아이디어를 확장하여 절사회귀추정량을 제시하였으며 이 추정량의 점근적 성질은 위치모수에서의 절사추정량과 비슷하고 이 추정량은 디자인 행렬의 재모수화(reparameterization)에 대해서도 불변(invariance)인 좋은 성질을 가지고 있다는 것을 발견하였다.

---

1) This research is supported by research professor fund of Seoul City University in 1999

2) Professor, Department of Computer Science and Statistics, University of Seoul, JeonnongDong 90, Dongdaemoonku, Seoul.
E-mail: smhan@uoscc.uos.ac.kr

그리고 이 추정량은 Barrodales 과 Roberts(1974), Armstrong 과 Kung(1978)등 다수의 저자에 의해 제안된 표준적인 L_1 -알고리즘의 약간의 수정에 의해 쉽게 계산되어진다는 장점을 가지고 있다. 아마도 이 추정량이 L-추정량으로 처음으로 실용적이고 유용한 추정량일 것이다. 그 이후 De Jongh과 De Wet(1985) 은 Jaeckel(1971)의 위치모수추정법을 Ruppert 와 Carroll 의 절사회귀추정법을 사용하여 회귀모형에 확장하였다. 본 논문에서는 대칭분포하에서 위치모수 추정에 있어서의 Johns(1974)의 아이디어를 활용하여 대칭오차분포하에서의 회귀기울기를 추정하는 추정법을 제시하고자 한다.

그러면 본 논문에서 이용되는 회귀분위수추정량과 절사회귀추정량에 대해 간단히 언급하기로 하자. 먼저 다음과 같은 표준적인 회귀모형을 가정하자.

$$\underline{y}=X\underline{\beta}+\underline{z} \tag{1.1}$$

단, $\underline{y}=(y_1,\cdots,y_n)'$, X 행렬은 $n\times p$ 인 기지의 행렬이고, i 번째 행벡터는 $\underline{x_i}'$, $\underline{\beta}=(\beta_0,\beta_1,\cdots,\beta_{p-1})'$ 는 미지인 모수벡터라고 하고 $\underline{z}=(z_1,\cdots,z_n)'$ 에서 각각의 좌표확률변수는 영(zero)에 대해 대칭이며, 서로 독립이고 동일한 미지인 분포함수 F 와 확률밀도함수 f 를 가진다고 하자 .

회귀분위수의 아이디어의 근간은 위치모수에 있어서의 일반적인 θ -차 표본 백분위수(sample quantiles)는 다음과 같은 check 함수를 가지는 M-추정량에 의해 구할 수 있다는 데 있다.

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

단, $\theta\in(0,1)$. 그리고 (1.2)식을 적용하여 $K(\theta)$ 를 θ-차 회귀분위수라고 하면 이 $K(\theta)$ 는 (1.3)식을 만족하는 값이 된다.

$$\min_{\underline{b}\in R^p}\sum_{i=1}^{n}\rho_\theta(y_i-\underline{x_i}'\underline{b}) \tag{1.3}$$

이 회귀분위수를 사용하여, Koenker 와 Bassett(1980) 다음과 같은 절사회귀추정량을 제시하였다. $0<p_1<p_2<1$ 인 값에 대해, $K(p_1)$, $K(p_2)$ 를 각각 p_1-차 , p_2-차 회귀분위수라고 하자. 이때

$$a_i=\begin{cases}1 & , \ \underline{x_i}'K(p_1)\leq y_i\leq \underline{x_i}'K(p_2)\\ 0 & , \text{ 다른 경우}\end{cases} \tag{1.4}$$

라고 하면 절사회귀추정량 $L(\underline{p})$ 는 $a_i=1$ 인 관측치만을 사용한 최소제곱추정량이다. 즉,

$$L(\underline{p})=(X'AX)^{-}X'A\underline{y} \tag{1.5}$$

이고 $A=diag(a_i)$, X 디자인 행렬이고 x_i' ,$i=1,\cdots,n$ 는 X의 행벡터이며 $p=(p_1,p_2)'$이다. 만약 X 가 절편을 포함하고 $n\to\infty$일때 $n^{-1}X'X\to Q$ 이며 Q 는 양정치행렬(positive definite matrix)이고 ξ_1 , ξ_2 는 각각 기저분포 F 의 p_1-차 와 p_2-차 분위수이며 $p=(p_1, p_2)$ 라고 할 때 Ruppert 와 Carroll(1980)는 몇가지 조건하에서 다음의 관계식이 성립함을 보였다.

$$\sqrt{n}(L(p)-\beta)=n^{-\frac{1}{2}}Q^{-1}\sum_{i=1}^{n}\{x_i'(\psi(z_i)-E(\psi(z_i))+\delta(p)\}+o_p(1)$$

단,

$$\begin{aligned}\psi(x) &= \xi_1/(p_2-p_1) \quad , \quad x<\xi_1\\ &= x/(p_2-p_1) \quad , \quad \xi_1\le x\le\xi_2\\ &= \xi_2/(p_2-p_1) \quad , \quad x>\xi_2\end{aligned} \tag{1.6}$$

$$\sqrt{n}(L(p)-\beta-\delta(p)) \xrightarrow{D} N(0,\sigma^2(p,F)Q^{-1})$$

단, $\xrightarrow{D}$ 는 분포적 수렴을 의미한다.

$\delta(p)=((p_2-p_1)^{-1}\int_{\xi_1}^{\xi_2}xdF(x),0,\cdots,0))'$ 이고 $\sigma^2(p,F)$는 위치모수에 있어서 p_1-차와 p_2-차 절사평균의 점근적 분산을 의미한다. 그리고 이 결과들은 본 논문에서 제안한 추정량의 점근적 성질을 구할 때 사용되어 질 것이다.

2. 제안된 추정량의 점근적 성질

본절을 시작하기 전에 본 논문에서 제시된 정리나 따름정리등에 부과되는 몇 가지 기호나 가정들을 소개하겠다. (1.1)식에서 사용된 y, X, 그리고 z 등은 표본수 n에 종속하지만 이것을 암묵적으로 표시하지 않겠다. $e=(1,0,\cdots,,0)'$ 는 $(p\times1)$벡터이고 , I_p 는 $(p\times p)$ 단위행렬이라 하자. $0<p_i<1/2$에 대해 $\xi_i=F^{-1}(p_i)$ 로 표시하고 지저분포의 영에 대한 대칭성에 의해 $-\xi_i=F^{-1}(1-p_i)$ 로 표시하며, $N_p(\mu,\Sigma)$ 는 평균벡터가 μ 이고 공분산 행렬이 Σ 인 p-변량 정규분포라고 하자. 그리고 본 논문에서는 Ruppert 와 Carroll(1980)의 논문에서 결과를 인용하여 제안된 추정량의 점근적성질을 다루기 때문에 상기논문에서의 조건들을 디자인 행렬과 예비적합추정량에 대하여 다음과 같은 가정을 한다.

A1. $x_i=(x_{i1},x_{i2},\cdots,x_{ip})'$ 를 행렬 X 의 i-번째 행이라고 하고 $x_{i1}=1$, $i=1,2,\cdots,n$ 이며 $\sum_{i=1}^{n}x_{ij}=0, j=2,3,\cdots,p$ 이다.

A2. $\lim_{n\to\infty}(\max_{j\le p,\, i\le n} n^{-1/2}|x_{ij}|)=0$ 이다.

A3. $\lim_{n\to\infty} n^{-1}(X'X)=Q$ 를 만족시키는 양정치행렬 Q 가 존재한다.

A4. $\hat{\beta}_0$ 를 예비 적합 추정량이라고 할때, 적당한 상수 c에 대해 $\sqrt{n}(\hat{\beta}_0-\beta-c\underline{e})=O_p(1)$ 이다.

우리들의 제안된 추정량은 다음과 같다.

$0<p_0<p_1<\cdots<p_k=1/2$ 라고 하자. 그리고, $K(p_i)$ 를 p_i-차 회귀분위수라고 하자. 그러면 $i=1,2,\cdots,n$ 에 대해

$$
\begin{aligned}
a_1&=\begin{cases}1 & , \underline{x}_i'K(p_0)<y_i\le \underline{x}_i'K(p_1) \text{ 혹은 } \underline{x}_i'K(1-p_1)<y_i\le \underline{x}_i'K(1-p_0)\\ 0 & ,\text{다른 경우}\end{cases}\\
a_2&=\begin{cases}1 & , \underline{x}_i'K(p_1)<y_i\le \underline{x}_i'K(p_2) \text{ 혹은 } \underline{x}_i'K(1-p_2)<y_i\le \underline{x}_i'K(1-p_1)\\ 0 & ,\text{다른 경우}\end{cases}\\
&\vdots\\
a_k&=\begin{cases}1 & , \underline{x}_i'K(p_{k-1})<y_i< \underline{x}_i'K(1-p_{k-1})\\ 0 & ,\text{다른 경우}\end{cases}
\end{aligned}
\tag{2.1}
$$

라고 정의하고 $L_1, L_2, \cdots, L_k$ 를 각각 $a_1=1, a_2=1,\cdots, a_k=1$에 대응하는 관측치로만 구성된 최소제곱추정량이라고 하자. 즉,

$$
\begin{aligned}
L_1&=(X'A_1X)^{-}X'A_1\underline{y},\\
L_2&=(X'A_2X)^{-}X'A_2\underline{y},\\
&\vdots\\
L_k&=(X'A_kX)^{-}X'A_k\underline{y}
\end{aligned}
\tag{2.2}
$$

단 $A_i=diag(a_i)$, $i=1,2,\cdots,k$ 이고 X 는 행벡터 $\underline{x}_i'$, $i=1,2,\cdots,n$ 로 구성된 행렬이다. 그러면 우리들의 추정량은 다음과 같은 형태를 가진다.

$$B_k=w_1L_1+w_2L_2+\cdots+w_kL_k,\quad \sum_{i=1}^{k}w_i=1\ .$$

따라서, 기울기 추정량은 $L_1, L_2, \cdots, L_k$ 로 부터 기울기 부분을 제거하여 얻어질 수 있다 . 그러면 기울기 부분을 제거한 $(p-1)$ 차원 추정량을 $L_1^*, L_2^*, \cdots, L_k^*$라고 하자. 따라서 우리들의 기울기 추정량은 다음과 같은 형태를 가진다.

$$S_k=w_1L_1^*+w_2L_2^*+\cdots+w_kL_k^*,\quad \sum_{i=1}^{k}w_i=1 \tag{2.3}$$

데이터의 수가 100개 이하인 실제적인 응용에서는 k=2 인 경우가 충분하며, 이 경우 기저분포

가 정규분포와 유사하면 w_1과 w_2 에 비슷한 가중치를 줄 것이며, 기저분포가 정규분포보다 두꺼운 꼬리부분을 가지면 w_1에 w_2 보다 더 적은 가중치를 줄 것이다. 차후에 대칭오차 분포하의 위치모수의 추정에서의 Johns의 추정량과 유사한 가중치 $w_1, w_2, \cdots, w_k$ 를 부과하는 방법을 제시하겠다. 먼저 이를 위한 필요한 정리를 간단히 제시한다.

정리 2.1 k는 고정되어 있고, $0<p_0<p_1<\cdots<p_k=1/2$ 이며, $p_1-p_0=p_2-p_1=\cdots=p_k-p_{k-1}=q_k$ 라고 하면,

$$\sqrt{n}\,[B_k-\underline{\beta}] \xrightarrow{D} N_p(0, (\underline{w}'\Sigma\underline{w})Q^{-1})$$

이다. 단, $\underline{w}=(w_1, w_2, \cdots, w_k)'$, $i=1,2,\cdots,k$ 이고 $i\le j\le k$에 대해 $\Sigma=(\sigma_{ij})_{k\times k}$ 는

$$\begin{aligned}\sigma_{ii} &= 1/2q_k^{-2}\left\{\xi_j\int_{\xi_{i-1}}^{\xi_i}F(x)dx-\int_{\xi_{i-1}}^{\xi_i}xF(x)dx\right\}\\ \sigma_{ij} &= \sigma_{ji}=1/2q_k^{-2}\left\{\xi_j\int_{\xi_{i-1}}^{\xi_i}F(x)dx-\xi_{j-1}\int_{\xi_{i-1}}^{\xi_i}F(x)dx\right\}\end{aligned} \tag{2.4}$$

이다.

증명. $j=1,2,\cdots,k-1$ 에 대해 L_j 는 다음과 같이 표시된다.

$$L_j=[X'(A-B)X]^{-}X'(A-B)\underline{y}$$

단, $i=1,2,\cdots,n$ 에 대해, $A=diag(a_i)$, $B=diag(b_i)$ 이고

$$a_i=\begin{cases}1 & , \quad \underline{x}_i'K(p_{j-1})<y_i<\underline{x}_i'K(1-p_{j-1}),\\ 0 & , \quad \text{다른경우}\end{cases}$$

$$b_i=\begin{cases}1 & , \quad \underline{x}_i'K(p_j)<y_i<\underline{x}_i'K(1-p_j),\\ 0 & , \quad \text{다른경우}\end{cases}$$

이다. 그리고 A, B의 정의와 Ruppert 와 Carroll(1978)의 정리 3을 적용하면.

$$\begin{aligned}&n^{-1/2}X'(A-B)[\underline{y}-(A-B)X\underline{\beta}]\\ &=n^{-1/2}X'A(\underline{y}-AX\underline{\beta})-n^{-1/2}X'B(\underline{y}-BX\underline{\beta})\\ &=n^{-1/2}\left\{(1-2p_{j-1})\sum_{i=1}^{n}x_i\psi_{j1}(z_i)-(1-2p_j)\sum_{i=1}^{n}x_i\psi_{j2}(z_i)\right\}+o_p(1)\end{aligned}$$

단, ψ_{j1}, ψ_{j2} 는

$$\begin{aligned}
\psi_{j1}(x) &= \xi_{j-1}/(1-2p_{j-1}), && x < \xi_{j-1} \\
&= x/(1-2p_{j-1}), && \xi_{j-1} \le x \le -\xi_{j-1} \\
&= -\xi_{j-1}/(1-2p_{j-1}), && x > -\xi_{j-1}, \\
\psi_{j2}(x) &= \xi_j/(1-2p_j), && x < \xi_j \\
&= x/(1-2p_j), && \xi_j \le x \le -\xi_j \\
&= -\xi_j/(1-2p_j), && x > -\xi_j
\end{aligned}$$

그리고,

$$n[X'(A-B)X]^{-} = \frac{1}{2q_k} Q^{-1} + o_p(1)$$

위의 결과에 의해 $j=1,2,\cdots,k-1$에 대해,

$$n^{1/2}(L_j - \beta) = n^{-1/2} Q^{-1} \sum_{i=1}^{n} x_i \Psi_j(z_i) + o_p(1) \tag{2.5}$$

단, $j=1,2,\cdots,k-1$에 대해,

$$\Psi_j(x) = \frac{1}{2q_k}[(1-2p_{j-1})\psi_{j1}(x) - (1-2p_j)\psi_{j2}(x)]$$

그리고

$$\begin{aligned}
2q_k\Psi_j(x) &= \xi_{j-1} - \xi_j \quad , && x < \xi_{j-1} \\
&= x - \xi_j \quad , && \xi_{j-1} \le x \le \xi_j \\
&= 0 \quad , && \xi_j \le x \le -\xi_{j-1} \\
&= x - \xi_j \quad , && -\xi_j \le x \le -\xi_{j-1} \\
&= \xi_j - \xi_{j-1} \quad , && x > -\xi_{j-1}
\end{aligned}$$

$j=k$ 에 대해, (1.6) 식의 결과를 이용하면

$$n^{1/2}(L_k - \beta) = n^{-1/2} Q^{-1} \sum_{i=1}^{n} x_i \Psi_k(z_i) + o_p(1) \tag{2.6}$$

단,

$$\begin{aligned}
2q_k\Psi_k(x) &= \xi_k \quad , && x < \xi_k \\
&= x \quad , && \xi_k < x < -\xi_k \\
&= -\xi_k \quad , && x > -\xi_k
\end{aligned}$$

만약 $\Psi_k^*(z_i)=\sum_{r=1}^{k} w_r\Psi_r(z_i)$ 이라고 하면 정리 2.1의 추정량의 점근적 p-변량 정규분포를 증명하기위해 임의의 상수벡터 $\underline{c}\in R^p$에 대해, $W_n=\underline{c}'Q^{-1}n^{-1/2}\sum_{i=1}^{k}\underline{x}_i\Psi_k^*(z_i)$ 이 점근적 일변량 정규분포를 따르면 된다.

$W_{in}=\underline{c}'Q^{-1}n^{-1/2}\,\underline{x}_i\Psi_k^*(z_i)$ 라고 하자. 그러면,

$$E(W_{in})=0$$
$$Var(W_{in})=\underline{c}'Q^{-1}n^{-1}\,\underline{x}_i Var\{\Psi_k^*(z_i)\}\,\underline{x}_i'Q^{-1}\underline{c}$$

따라서, $E(W_n)=0$, $n\to\infty$일때 $Var(W_n)\to Var\{\Psi_k^*(z_i)\}\underline{c}'Q^{-1}\underline{c}=(\underline{w}'\Sigma\underline{w})\underline{c}'Q^{-1}\underline{c}$ 이고 Σ는 $(\Psi_1(z),\Psi_2(z),\cdots,\Psi_k(z))$의 공분산 행렬이다 . Lindeberg의 C.L.T.에 의해 임의의 $\varepsilon>0$에 대해,

$$\frac{1}{Var(W_n)}\sum_{i=1}^{k}\int_{\{|W_{in}|>\varepsilon\sqrt{Var(W_n)}\}} W_{in}^2\,dF_{in}\to 0 \text{ as } n\to\infty \tag{2.7}$$

임을 보이면 충분하다. 단, F_{in} 은 W_{in}의 분포함수이다 . 적당한 양의 수 M에 대해, $|\Psi_k^*(z)|\le M$ 이므로 (2.7)은 다음과 같은 부등식이 성립한다.

$$\frac{M^2}{Var(W_n)}\sum_{i=1}^{k}\int_{\{\max|W_{in}|>\varepsilon\sqrt{Var(W_n)}\}} n^{-1}\underline{c}'Q^{-1}\,\underline{x}_i\,\underline{x}_i'Q^{-1}\underline{c}\,dF_{in}$$
$$\le \frac{M^2}{Var(W_n)}\sum_{i=1}^{k}\underline{c}'Q^{-1}(n^{-1}\,\underline{x}_i\,\underline{x}_i')Q^{-1}\underline{c}\;P(\max|W_{in}|>\varepsilon\sqrt{VarW_n})$$

그리고 조건 A3에 의해

$$P(\max|W_{in}|>\varepsilon\sqrt{VarW_n})$$
$$=P(\max n^{-1/2}\,|c'Q^{-1}\,\underline{x}_i\Psi_k^*(z_i)|>\varepsilon\sqrt{VarW_n})\to 0 \text{ as } n\to\infty \tag{2.8}$$

위의 식 (2.8)에 의해 $n\to\infty$일 때 (2.7)의 오른편은 영으로 수렴한다. 따라서, $\sqrt{n}[B_k-\beta]\xrightarrow{D} N_p(0,(\underline{w}'\Sigma\underline{w})Q^{-1})$이 성립하고 Σ 의 요소는 $\Psi_i(z)$, $i=1,2,\cdots,k$ 의 분산-공분산에 의해 결정된다 . 그리고 (2.5)와 (2.6)의 점근적 표현식에 의한 간단한 계산에 의해 $\Psi_i(z)$의 분산인 σ_{ii} 와 $\Psi_i(z)$와 $\Psi_j(z)$의 공분산인 σ_{ij} 의 값은 정리 2.1에서 주어진 것과 같음을 알 수 있다. 그리고 기울기 모수의 점근적성질은 정리 2.1로 부터 간단히

$\sqrt{n}(S_k - \underline{\beta}_0) \xrightarrow{D} N_{p-1}(0, (\underline{w}'\Sigma\underline{w})Q_0^{-1})$임을 알수 있다. 단, 이때 $\underline{\beta}_0$는 $\underline{\beta}$의 첫 번째 요소를 제외한 벡터이며, Q_0^{-1}는 Q^{-1}의 첫 번째행과 첫 번째열을 제외한 행렬이다.

다음은 S_k 의 점근적 분산을 최소로 하는 가중치 벡터 $\underline{w}$ 를 결정하는 문제를 생각하자. 그러나 일반적으로 S_k의 점근적 분산을 추정하는 것은 매우 힘들고 이 경우 대칭분포하에서의 위치모수 추정법을 제시한 Johns(1974)의 아이디어를 응용하여 Σ 의 근사적 공분산을 최소로 하는 가중치 벡터 $\underline{w}$ 의 값을 결정하는 문제로 환원하여 보자. $i=1,2,\cdots,k$, 에 대해 다음의 근사값을 생각하자.

$$\sigma_{ii} \sim 1/2q_k^{-2}b_id_i^2 \qquad (2.9)$$
$$\sigma_{ij} \sim 1/2q_k^{-2}b_id_id_j$$

단, $d_i=\xi_i-\xi_{i-1}$, $b_i=p_{i-1}+1/2q_k$. 그러면 Σ 의 근사적 공분산은

$$B=1/2q_k^{-2}D\begin{bmatrix} b_1 & b_1 & \cdots & b_1 \\ b_1 & b_2 & \cdots & b_2 \\ \vdots & \vdots & & \vdots \\ b_1 & b_2 & \cdots & b_k \end{bmatrix}D$$

이며 $D = diag(d_1, d_2, \cdots, d_k)$이다. 그리고 다음의 식을 만족하는 가중치 벡터 $\underline{w}$ 를 구해보자.

$$\min_{\underline{w}'\underline{1}=1} (\underline{w}'B\underline{w})Q_0^{-1}$$

단, $\underline{1}=(1,1,\cdots,1)'$. Q_0^{-1} 는 고정된 값이므로, $\underline{w}'\underline{1}=1$인 조건아래에서 $\underline{w}'B\underline{w}$ 를 최소로 하는 가중치 벡터의 값을 결정하면 된다. 그리고 이 문제는 간단한 Lagrange 승수법을 적용하여 해결할 수 있으며 그 해는 $\underline{w}=(\underline{1}'B^{-1}\underline{1})^{-1}B^{-1}\underline{1}$ 로 표시된다. 그리고 이 해를 이용하여,

$$e_1=2/d_1(b_2/b_1d_1-1/d_2)q_k,$$
$$e_i=1/d_i(2/d_i-1/d_{i-1}-1/d_{i+1})q_k, \qquad i=2,3,\cdots,k-1,$$
$$e_k=2/d_k(-1/d_k-1/d_{k-1})q_k,$$

로 정의하면, 다음과 같은 구체적인 해의 형태를 얻을 수가 있다. 즉,

$$w_i=e_i/\left\{\sum_{i=1}^{k}e_i\right\}, \qquad i=1,2,\cdots,k.$$

만약 예비적합에 의한 잔차로써 d_i 의 일치추정량(consistent estimator) $\hat{d}_i$로 치환하여 구성된 e_i의일치추정량 e_i^* 를 얻었다고 하면 추정된 가중치 를 이용한 우리들의 최종적인 추정량의 형

태는 다음과 같다.

$$\widehat{S_k} = \omega_1^* L_0(\underline{p_1}) + \omega_2^* L_0(\underline{p_2}) + \cdots + \omega_k^* L_0(\underline{p_k}) \ . \tag{2.10}$$

단, $\omega_i^* = e_i^* / \left\{ \sum_{i=1}^{k} e_i^* \right\}$, $(i=1,2,\cdots,k)$. 마지막으로 $\sqrt{n}\ (\widehat{S_k} - \underline{\beta_0})$ 와 $\sqrt{n}\ (S_k - \underline{\beta_0})$ 이 동일한 점근적 분포를 따른다는 것을 보여주자. 이를 위해 다음 정리 2.2는 $i=1,2,\cdots,k$ 에 대해 가중치 ω_i 의 일치추정량 ω_i^* 를 예비적합에 따른 잔차로부터 얻기위해 매우 유용한 정리이다.

정리 2.2 $0<p_0<p_1<\cdots,<p_k=1/2$ 라 하고 $\widehat{\xi_i}$ 를 가정 A4 를 만족시키는 예비적합 추정량 $\widehat{\underline{\beta_0}}$ 로부터 얻은 np_i 번째 순서잔차량이라고 하자. 그러면, $i=1,2,\cdots,k$ 에 대해 $\widehat{d_i}$ d_i 의 일치 추정량이다 .

증명. Ruppert 와 Carroll (1980)의 예비정리 1을 사용하여 , $i=1,2,\cdots,k$ 에 대해

$$\begin{aligned}
\sqrt{n}(\widehat{\xi_i} - \xi_i) &= [f(\xi_i)]^{-1} n^{-1/2} \sum_{i=1}^{n} \psi_{p_i}(z_i - \xi_i) - e' n^{1/2}(\widehat{\underline{\beta_0}} - \underline{\beta}) + O_p(1) \\
\sqrt{n}(\widehat{\xi_{i-1}} - \xi_{i-1}) &= [f(\xi_{i-1})]^{-1} n^{-1/2} \sum_{i=1}^{n} \psi_{p_{i-1}}(z_i - \xi_{i-1}) - e' n^{1/2}(\widehat{\underline{\beta_0}} - \underline{\beta}) + O_p(1)
\end{aligned} \tag{2.11}$$

을 얻는다. 단 $\psi_\theta(x) = \theta - I(x<0)$. 따라서 위의 두 개의 식을 빼면

$$\begin{aligned}
\sqrt{n}(\widehat{d_i} - d_i) = \ & [f(\xi_i)]^{-1} n^{-1/2} \sum_{i=1}^{n} \psi_{p_i}(z_i - \xi_i) - \\
& [f(\xi_{i-1})]^{-1} n^{-1/2} \sum_{i=1}^{n} \psi_{p_{i-1}}(z_i - \xi_{i-1}) + O_p(1)
\end{aligned} \tag{2.12}$$

C.L.T.에 의해 (2.12)식의 오른편의 처음 두 개의 항은 유한한 분산을 가지는 점근적 정규분표를 따른다. 따라서, $i=1,2,\cdots,k$ 에 대해 $\widehat{d_i} - d_i = o_p(1)$.

정리 2.2 에 의해 회귀모형에서 기저오차분포의 백분위수의 차이에 대한 일치추정량을 얻었고 이를 이용하여 (2.10)에서 주어진 $\widehat{S_k}$ 의 일치추정량을 구성할 수가 있다.

따름정리 2.1 $\sqrt{n}(S_k - \underline{\beta_0})$ 와 $\sqrt{n}(\widehat{S_k} - \underline{\beta_0})$ 는 동일한 극한 분포를 가진다.

증명. S_k 와 $\widehat{S_k}$ 의 구성에 의해 ,

$$\sqrt{n}\,[(S_k - \underline{\beta_0}) - (\widehat{S_k} - \underline{\beta_0})] = \sum_{i=1}^{k} (\omega_i - \omega_i^*) \sqrt{n} L_0(p_i) \tag{2.13}$$

를 얻을 수가 있고, $\omega_i^* \xrightarrow{p} \omega_i$ 와 $\sqrt{n}L_0(p_i)$ 는 분포상 유계(bounded in distribution)이므로, $\sqrt{n}(S_k - \underline{\beta_0})$ 와 $\sqrt{n}(\widehat{S_k} - \underline{\beta_0})$ 는 동일한 극한분포를 가진다.

3. 모의 실험 및 결론

간단한 모의실험을 단순회귀모형상에서 기울기의 추정으로 시도하였다. 이 경우 우리들이 제시한 추정량 $\widehat{S_k}$ 는 대칭기저오차분포상에서 100개 이하의 표본수에 대해 $k=2$ 인 경우 기존의 L-추정량에 비해 비교적 잘 작동하였고, $k=3$, $k=4$등 k 값이 크짐에 따라 우리들의 추정량은 지나친 적응성(overadaptation)에 의해 잘 작동되지 않았다. 따라서 우리들의 모의실험은 $k=2$인 경우에 한정하였고 이 경우 우리들의 추정량을 JH 라고 불렀다. 우리들의 JH 추정량에서 이 추정량을 구성하기위한 예비적합추정량으로 다양한 대칭기저오차분포상에서 비교적 로버스트한 추정량인 최소절대값추정량(least absolute value estimator)인 L_1 ,즉 $K(0.5)$를 사용하였다.

$p_0=0.05$를 사용하였고, $p_1-p_0=p_2-p_1=q_2=0.225$ 를 사용하였다. 즉 $b_1=p_0+0.5q_2$, $b_2=p_1+0.5q_2$라고 놓고, $1\le i\le j\le 2$ 에 대해 $d_1=\xi_1-\xi_0$ 과 $d_2=\xi_2-\xi_1$ 로 놓자. 그러면 $k=2$ 인 경우의 공분산 행렬 Σ 의 요소의 근사값은 $\sigma_{11}\sim 1/2q_2^{-2}b_1d_1^2$, $\sigma_{12}\sim 1/2q_2^{-2}b_1d_1d_2$, $\sigma_{22}\sim 1/2q_2^{-2}b_2d_2^2$ 의 형태를 가진다. 만약 $V_1=1/2b_1d_1^2$, $V_2=1/2b_2d_2^2$ 라고 하면,

$$(q_2)^{-2}\underline{\omega}'\begin{pmatrix} V_1 & C \\ C & V_2 \end{pmatrix}\underline{\omega}Q_0^{-1}, \quad \underline{\omega}'\underline{1}=1 \tag{3.1}$$

이고 (3.1)의 해는 $\omega_1=(V_2-C)/(V_1-C+V_2)$ 과 $\omega_2=(V_2-C)/(V_1-C+V_2)$ 로 표시된다. 예비적합추정치로부터 구해진 잔차순서통계량으로부터 d_1과 d_2의 일치추정치인 $\widehat{d_1}$ 와 $\widehat{d_2}$를 구하고 이에 따라 가중치 벡터의 일치추정량인 $\widehat{\underline{\omega}}=(\widehat{\omega_1},\widehat{\omega_2})'$ 를 구한다. 그리고 기저오차분포의 대칭성에 따라 $\widehat{d_1}$ 와 $\widehat{d_2}$ 의 일치추정량은 다음과 같은 형태를 사용하였다. 즉

$$\widehat{d_i}=\{(\widehat{\xi_i}-\widehat{\xi_{i-1}})+(\widehat{\xi_{i-1}}-\widehat{\xi_i})\}/2 \quad , i=1,2$$

단 $i=0,1$에 대해 , $\widehat{\xi_{i-1}}$ 은 p_{i-1} 차 그리고 $\widehat{\xi_{i-1}}$ 는 $(1-p_{i-1})$ 차 잔차순서통계량이다. 이와 같은 방법에 따라 우리들의 추정량 은 다음과 같은 형태를 가진다. 만약 LS_1 은 $\underline{x_i}'K(0.05)\le y_i\le \underline{x_i}'K(0.275)$ 혹은 $\underline{x_i}'K(0.725)\le y_i\le \underline{x_i}'K(0.95)$ 에 속하는 관측치 y_i 들만으로 구해진 기울기에 대한 최소제곱추정량이고, LS_2 는 $\underline{x_i}'K(0.275)\le y_i\le \underline{x_i}'K(0.725)$ 에 속하는 관측치 y_i 들 만으로 구성된 기울기에 대한 최소제곱추정량이라고 할 때

$$JH=\frac{\widehat{w_1}}{(\widehat{w_1}+\widehat{w_2})}LS_1+\frac{\widehat{w_2}}{(\widehat{w_1}+\widehat{w_2})}LS_2 \tag{3.2}$$

로 주어진다. 그리고 디자인 행렬 X 의 첫 번째 열은 1의 값으로 채웠고, $i=1,2,\cdots,n$ 에 대해, 두 번째 열의 (x_i) 는 정규득점함수의 역함수인 $\Phi^{-1}(i/n+1)$ 를 사용하였다. 단 여기서 Φ^{-1} 는 누적표준정규분포의 역함수이다 그리고 이 값들은 $\sum x_i^2=1$가 되도록 표준화하였다 . 본 논문의 모의실험에서는 6개의 대칭기저오차분포에 대해 5개의 $L-$추정량을 사용하였고, 표본의 개수는 80개로 제한하였다. 6개의 대칭기저오차분포중에 2개는 가벼운 꼬리부분을 가지는 분포(normal 과 slate) 그리고 2개는 약간 무거운 꼬리부분을 가지는 분포(slacu 와 10% contamination), 나머지 2개는 중심부분이 아주 뾰족하거나 무거운 꼬리를 가지는 분포(double exponential 과 Cauchy)로 선택하였다. 모의실험에서 다루는 분포는 $Z=Y/X$로부터 생성된다. 이 경우 Y , X 는 서로 독립이고 Y 는 표준정규분포이다.

Normal (NOR): $X=1$

Slate(TE): $X=U^{1/10}$

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

10% Contamination(CON): $X=\begin{cases}1 & ,\text{확률 } 0.9\\ 1/3 & ,\text{확률 } 0.1\end{cases}$

Double expontial(DE): $X=1/W^{1/2}$, 단 W 는 $\chi^2(2)$.

Cauchy(CA): $X=|V|$, 단 V 는 $N(0,1)$.

그리고 이 모의실험에서는 제안된 JH추정량, 최소제곱추정량(L.S.), 최소절대값추정량(L_1)그리고 10% 및 20% 절사회귀추정량이 포함되었다. 이와 같은 조건하에서, 단순회귀모형인 $Y_i=\beta_0+\beta_1X_i+Z_i$ 에서 $\beta_0=\beta_1=0$로 놓고 β_1 의 2000번의 반복에 의해 추정 평균제곱오차(MSE)를 계산하였다.(표 1 참조). JH 추정량은 제시된 6개의 분포군에 대해 비교적 잘 작동하였다. 그러나 10% 와 20% 절사회귀추정량은 약간 무거운 꼬리를 가지는 분포군(Slacu와 10% Contamination)에서 특히 잘 작동되었다. 최소제곱추정량은 가벼운 꼬리를 가지는 분포군(Normal 또는 Slate)에서는 매우 잘 작동하였지만 뾰족한 중심부분을 가진 분포(D.E) 또는 약간 무거운 꼬리를 가진 분포군에서는 잘 작동되지 않았고, 아주 무거운 꼬리를 가진 분포군(Cauchy)에서는 전혀 작동하지 않았다. L_1 추정량은 뾰족한 중심부분을 가진 분포 또는 아주 무거운 꼬리를 가진 분포를 제외한 분포군에서는 잘 작동되지 않았다. 이미 잘 알려진 바와 같이, 가벼운 꼬리를 가진 분포군외의 분포군에 대해 제시된 추정량들은 최소제곱추정량보다 비교적 잘 작동되었다. 본 논문에서 강조하고 싶은 것은 JH추정량이 기존에 잘 알려진 절사 회귀추정량에 못지 않게 작동되며 특히 무거운 꼬리를 가진 대칭기저오차분포에 대해 매우 잘 작동하는 점이다.

각 셀이 첫 번째 값은 10^3(추정량의 MSE), 그리고 표준오차는 주어진 값의 약 3 % 내지 4

% 이다. 두 번째 값은 주어진 대칭기저오차분포에서 최소의 MSE를 가지는 추정량에 대한 상대 효율을 % 로 표시한 것이다.

[표1] MSE와 추정량들의 상대 효율 (n=80)

	NOR	TE	CU	CON	DE	CA
LS	1002 100	1253 100	2929 74.9	1767 75.5	2002 69.3	--- 0
L1	1535 65.3	1865 67.2	2806 78.2	1776 75.1	1392 96.1	2786 92.9
TRIM 10%	1068 93.8	1305 96.0	2194 100	1334 100	1793 87.1	5894 43.9
TRIM 20%	1138 88.0	1392 90.0	2205 99.5	1351 98.7	1536 90.4	3642 71.1
JH	1124 89.1	1388 93.6	2307 95.1	1443 92.4	1388 100	2589 100

References

[1] Adichie, J. N. (1974), "Rank score comparision of several regression parameters." *Annals of Statistics* **2**, 396-402.

[2] Armstrong,R.D. and Kung, M.T. (1978), "Algorithm AS 132. Least absolute value estimates fo a simple linear regression problem." *Applied Statistics 27*, 363-366

[3] Barrodale, I., and Roberts, F. D. K. (1974), "Solution of an overdetermined system of equations in the L_1 Norm." *Communications of the Association for Computing Machinery* **17**, 407-415.

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

[5] Hettmansperger, T. P., and Mckean J. W. (1977), "A robust alternative based on ranks to least squares in analyzing linear models" *Technometrics* **19**, 274-284.

[6] Hogg, R. V. (1979), "Statistical Robustness : One view of its use in applications today." *The American Statistician* **33**, 108-115.

[7] Hogg, R. V.,Bril, G. K., S. M., Han, L., Yuh (1988), "An Argument for Adaptive Robust Estimation." Probability and Statistics, EssaysinHonor of Graybill, F.A., North Holland, 135-148.

[8] Huber, P. J. (1973), "Robust regression : Asymptotics, conjectures and Monte Carlo." *Annals of statistics* **1**, 799-821

[9] Huber, P. J. (1981), "Robust Statistics", John Wiley.

[10] Jaeckel, L. A. (1972), "Estimating regression coefficient by minimizing dispersion of the residuals." *Annals of Mathematical Statistics* **43**, 149-1458.

[11] Johns, M. V. (1974), "Nonparametric estimation of location." *Journal of the American Statistical Association* **69**, 453-460.

[12] Jureckova, J. (1971), "Nonparametric estimation of regression coefficients." *Annals of Mathematical Statistics* **42**, 1328-1338

[13] Koenker, R. Bassett, C. (1982), "Robust tests for heterosedasticity based on regression quantiles." *Econometrica* **50**, 43-61

[14] P. J. de Jongh, T. de Wet (1985), "A Monte Carlo comparision of regression trimmed means." *Communications in Statistics (Theory and Methods)* **10**, 2457-2472

[15] Ruppert, D., Carroll, R. J. (1978), "Robust regression by trimmed least squares estimation." *Institute of Statistics Mimeo Series 1186*, Univ. of North Carolina at Chapel Hill, Dept. of Statistics.

[16] Ruppert, D., Carroll, R. J. (1980), "Trimmed least squares estimation in the linear model." *Journal of the American Statical Association* **75**, 828-838.