The Korean Communications
in Statistics Vol. 10 No. 3, 2003
pp. 859-871

Adaptive M-estimation in Regression Model

Sang Moon Han[1)]

Abstract

In this paper we introduce some adaptive M-estimators using selector statistics to estimate the slope of regression model under the symmetric and continuous underlying error distributions. This selector statistics is based on the residuals after the preliminary fit L_1 (least absolute estimator) and the idea of Hogg(1983) and Hogg et. al. (1988) who used averages of some order statistics to discriminate underlying symmetric distributions in the location model. If we use L_1 as a preliminary fit to get residuals, we find the asymptotic distribution of sample quantiles of residual are slightly different from that of sample quantiles in the location model. If we use the functions of sample quantiles of residuals as selector statistics , we find the suitable quantile points of residual based on maximizing the asymptotic distance index to discriminate distributions under consideration. In Monte Carlo study, this adaptive M-estimation method using selector statistics works pretty good in wide range of underlying error distributions.

Keywords : regression, adaptive M-estimators, selector statistics

1. 서론

위치모형과 회귀모형에서의 로버스트적 추정법은 1960년대 부터 지난 40여년간 꾸준히 연구되어왔고 위치모수의 추정에 있어서의 표본평균과 회귀모수의 추정에 있어서의 최소자승법(least squares method)의 대안으로 기저 오차분포의 형태에 둔감한 많은 로버스트적 추정량들이 제시되어 왔다. 그리고 정규가설을 이용한 표본평균이나 최소자승법이 정규가설이 무너질 경우 거의 무용한 추정량이 될 수도 있다는 문제점은 Huber(1964,1973)에 의해 많이 해결되었다. Huber (1964)는 ψ 함수를 이용하여 위치모수에 대한 M-추정법을 제안하였고, Huber(1973)에 의해 자연스럽게 회귀모수추정법으로 확장되었으며 현재 가장 많이 사용되는 로버스트적 추정방법으로 알려져 있다. 그리고 Hogg(1988)등은 광범위한 모의실험을 통해 위치모수 및 회귀모수에 대한 기존의 추정법에 대한 로버스트 추정법의 우월성을 보였다. Huber의 방법과 다른
각도로 Hogg(1967,1983)는 대칭인 위치모수의 추정에 있어서 기저 오차분포에 대한 형태를 선택통계량에 의해 파악하고, 이에 알맞은 추정량을 배분하는 적응 추정법을 제시하였다.

1) Professor, Dept. of Statistics, Univ. of Seoul, 130-743 Seoul
E-mail : smhan@ uoscc.uos.ac.kr

예컨대 선택통계량에 의해 오차분포가 정규분포 보다 이중지수분포에 가깝다는 사실을 알고 있다면 표본평균보다 중앙값을 중심에 대한 추정량으로 사용하는 것이 효율적일 것이다. 이러한 아이디어를 사용한 추정량들이 프린스턴 로버스트 추정법 연구(1972)에서 가치를 인정 받았다. 그러나 그가 사용한 선택통계량은 표본 첨도(sample kurtosis)를 이용한 것으로 점근적 수렴속도가 느린 단점을 지니고 있어 이후 Hogg등(1975), Hogg(1983), Hogg등(1988)에 의해 순서 통계량들의 평균들의 함수형태의 선택통계량들을 이용하여 이러한 단점을 보완하여 왔다. 그리고 Hogg(1983)가 제안한 선택통계량들을 이용하고 Hogg 등(1988)에 의해 제안된 두 분포의 거리를 최대로 하는 선택통계량의 구성방법을 Han(2002)이 위치모수의 추정에 적용하였다. 본 논문에서는 이와 유사한 방법을 회귀모수 추정에 확장하여, L_1에 의한 예비적합에 의해 구성된 잔차(residuals)의 백분위수(quantile)의 함수들로 구성된 선택통계량에 적용하여 그 통계량의 점근적 성질 및 이와 같이 제안된 선택통계량에 의한 회귀 기울기에 대한 적합 M-추정량들을 제시하려한다.

2. 선택통계량의 제안

먼저 본 논문에서 이용되는 회귀모형과 최소절대값 추정량(L_1)을 간단히 언급하기로 하자. 다음과 같은 표준적인 회귀모형을 가정하자.

$$\underline{y} = X\underline{\beta} + \underline{z} \tag{2.1}$$

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

$$\rho_\theta(x) = \begin{cases} \theta x & , \ x \geq 0 \\ (\theta - 1)x & , \ x < 0 \end{cases} \tag{2.2}$$

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.

$$\min_{\underline{b} \in R^p} \sum_{i=1}^{n} \rho_\theta(y_i - \underline{x_i}'\underline{b}) \tag{2.3}$$

본 논문에서는 잔차를 얻기위한 예비추정량으로 L_1 즉 $K(1/2)$ 를 이용할 것이다.

이 절에서 이용될 정리를 간단히 소개하고 이것을 이용하여 제안하게 될 선택통계량들의 점근적 성질을 규명하고자 한다. 먼저 본 논문에서 제안될 추정량들의 성질을 규명하기위해 몇가지 가정과 모형에 대해 제시하고자 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(positive definite matrix) 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.

본 논문에서 잔차를 얻기위한 예비추정량으로 최소절대값 추정량(least absolute value estimator: L_1 추정량)으로 하는 이유는 기저오차분포에 영향을 덜 받는 추정량이기 때문이다. n개의 데이터를 모형(2.1)에 L_1 추정량을 적합시킬 때 잔차의 $100p$ 백분위수를 r_p 라고 하고 $\eta(p)=F^{-1}(p)$라 하면 가정 A1-A4가 만족될 때 다음과 같은 정리와 보조정리들이 성립한다.

정리 2.1 $n\to\infty$ 이고, $0<p<1$ 에 대해 $\hat{\beta}_0$ 가 회귀모수에 대한 예비추정량일 때

$$n^{1/2}(r_p-\eta(p))=f^{-1}(\eta(p))n^{-1/2}\sum_{i=1}^{n}[\psi_p(Z_i-\eta(p))]-\underline{e}'n^{1/2}(\hat{\beta}_0-\beta)+o_p(1)$$

이다. 단 $\underline{e}=(1,0,0,\cdots,0)'$ 인 열 벡터이다.

증명 Ruppert와 Carroll(1980) 보조정리1 참조.

따름정리 2.2 $n\to\infty$ 이고, $0<p<1$ 에 대해

$$n^{1/2}(r_p-\eta(p))=n^{-1/2}\sum_{i=1}^{n}[f^{-1}(\eta(p))\psi_p(Z_i-\eta(p))-f^{-1}(0)(1/2-I(Z_i<0)]+o_p(1)$$

이다. 단 $I(\cdot)$는 지시함수(indicator function)이고, $\psi_p(Z_i-\eta(p))=p-I(Z_i<\eta(p))$ 이다.

증명 $\hat{\beta}_0$ 이 회귀모수에 대한 L_1 추정량이라 하면, Ruppert 와 Carroll(1980)은

$$n^{1/2}(\hat{\beta}_0-\beta)=n^{-1/2}\sum_{i=1}^{n}Q^{-1}\underline{x}_i(f(0))^{-1}(1/2-I(Z_i<0))+o_p(1) \tag{2.4}$$

임을 보여주었다. 단 $\underline{x}_i'$ 는 디자인 행렬 X 의 i 번째 행벡터이다. $\underline{0}$ 를 $(p-1)\times 1$ 인 열 벡터라고 하면 가정 1과 가정 3에 의해, 적당한 행렬 Q^* 에 대해

$$Q=\begin{bmatrix} 1 & \underline{0}' \\ \underline{0} & Q^* \end{bmatrix}, \quad Q^{-1}=\begin{bmatrix} 1 & \underline{0}' \\ \underline{0} & Q^{*-1} \end{bmatrix}$$

가 성립함을 알 수 있다. 따라서 (2.4)식을 정리 2.1에 있는 식에 대입하고, $\underline{e}'Q^{-1}\underline{x}_i=1$ 임을 이용하면 따름정리 2.2의 결과를 얻는다.

정리 2.3 $n\to\infty$ 이고, $0<p<1$일 때, $n^{1/2}(r_p-\eta(p))$ 는 정규분포

$$N(\ 0,\ f^{-2}(\eta(p))p(1-p)-pf^{-1}(\eta(p))f^{-1}(0)+f^{-2}(0)/4\ ) \tag{2.5}$$

를 따른다.

증명 위의 따름정리 2.2에서 $f^{-1}(\eta(p))\ \psi_p(Z_i-\eta(p))-f^{-1}(0)(1/2-I(Z_i<0)$ 는 각각 i.i.d.이므로 중심극한정리를 적용하면 정규분포를 따른다. 따라서 평균과 분산만 결정하면 된다. 이때 평균이 0임은 자명하고, 분산은

$$\begin{aligned} &E[f^{-1}(\eta(p))\ \psi_p(Z_i-\eta(p))-f^{-1}(0)(1/2-I(Z_i<0)]^2 = \\ &f^{-2}(\eta(p))E(p-I(Z_i<\eta(p)))^2+f^{-2}(0)E(1/2-I(Z_i<0))^2 \\ &-2f^{-1}(0)f^{-1}(\eta(p))E(p-I(Z_i<\eta(p)))(1/2-I(Z_i<0)) = \\ &f^{-2}(\eta(p))p(1-p)-pf^{-1}(0)f^{-1}(\eta(p))+f^{-2}(0)/4 \end{aligned}$$

가 된다.

정리2.4 $n\to\infty$일 때, $n^{1/2}(r_p-\eta(p))$과 $n^{1/2}(r_q-\eta(q))$ 의 공분산은 $0<p<q<1$에 대해

$$\begin{aligned} &f^{-1}(\eta(p))f^{-1}(\eta(q))p(1-q)-f^{-1}(0)f^{-1}(\eta(p))\min(p/2,(1-p)/2) \\ &-f^{-1}(0)f^{-1}(\eta(q))\min(q/2,(1-q)/2)]+f^{-2}(0)/4 \end{aligned} \tag{2.6}$$

이다. 단 $\min(a,b)=a,\ a\le b,\quad \min(a,b)=b,\ a\ge b$ 이다.

증명 $0<p<q<1/2$인 경우만 보이면 충분하다. 따름정리 2.2를 이용하고, $Z_i\ (i=1,2,,\cdot,n)$ 가 서로 독립인 확률변수임을 이용하여 공분산을 계산하여 보면,

$$\begin{aligned} &cov(n^{1/2}(r_p-\eta(p)),n^{1/2}(r_q-\eta(q))= \\ &n^{-1}E[\ \sum_{i=1}^{n}\psi_p(Z_i-\eta(p)-f^{-1}(0)(1/2-I(Z_i<0))] \\ &[\ \sum_{i=1}^{n}\psi_q(Z_i-\eta(q)-f^{-1}(0)(1/2-I(Z_i<0))]\ = \\ &n^{-1}[f^{-1}(\eta(p))f^{-1}(\eta(q))\ \sum_{i=1}^{n}E[\psi_p(Z_i-\eta(p))][\psi_q(Z_i-\eta(q))] \end{aligned} \tag{2.7}$$

$$-f^{-1}(0)f^{-1}(\eta(p))\sum_{i=1}^{n}E[\psi_p(Z_i-\eta(p))][1/2-I(Z_i<0)] \tag{2.8}$$

$$-f^{-1}(0)f^{-1}(\eta(q))\sum_{i=1}^{n}E[\psi_q(Z_i-\eta(q))][1/2-I(Z_i<0)] \tag{2.9}$$

$$+f^{-2}(0)(\eta(q))\sum_{i=1}^{n}E[1/2-I(Z_i<0)]^2] \tag{2.10}$$

을 얻는다.

(2.7)식의 기대치항을 계산하면,

$$E[\psi_p(Z_i-\eta(p))][\psi_q(Z_i-\eta(q))]=E[p-I(Z_i<\eta(p))][q-I(Z_i<\eta(q))]$$
$$=pq-pq-pq+p=p(1-q)$$

이다. 마찬가지 방법으로 (2.8), (2.9), (2.10)식의 기대치항을 계산하여 정리하면 (2.6)식을 얻을 수 있다.

<u>언급사항</u> : 위에서 언급된 정리 2.3 과 정리 2.4 에서 보여주는 것은 잔차를 얻기위한 예비추정으로 사용된 L_1 추정 이후의 잔차의 백분위수의 분산과 공분산이 대칭인 위치모형에서의 표본백분위수의 분산과 공분산과 차이가 있다는 점이다. 평균은 모두 0으로 동일하나 분산과 공분산은 (2.5)식과 (2.6)식에서 각각 L_1 추정에 의한 효과인 $-f^{-1}(0)f^{-1}(\eta(p))p+f^{-2}(0)/4$ 와 $-f^{-1}(0)f^{-1}(\eta(p))\min(p/2,(1-p)/2)\quad -f^{-1}(0)f^{-1}(\eta(q))\min(q/2,(1-q)/2)]+f^{-2}(0)/4$ 이 첨가된다는 사실이 흥미롭다.

본 논문에서 제안하는 선택통계량은 Hogg(1983)가 제안한 순서통계량의 일부 표본들의 평균을 이용한 선택통계량을 잔차의 표본 백분위수를 이용하여 다음과 같이 제안한다.

$$H_2=\frac{r_{1-\beta}-r_{\beta}}{r_{1-\gamma}-r_{\gamma}},\quad ,\ H_3=\frac{r_{1-\alpha}-r_{\alpha}}{r_{1-\beta}-r_{\beta}},\quad H=H_2+H_3 \tag{2.12}$$

단, 여기서 $\alpha<\beta<\gamma<.5$ 이고 $r_{\alpha}, r_{\beta}, r_{\gamma}$ 는 각각 잔차의 $100\alpha, 100\beta, 100\gamma$ 백분위수이다. 이때, 기저오차분포의 밀도함수가 연속이며 대칭인 f(x)의 형태를 가질 때 제안된 선택통계량인 H_2 와 H_3 는 점근적인 분포는 다음과 같다.

<u>정리 2.3</u> H_2 와 H_3가 (2.12)과 같이 정의되고, $\eta(\alpha),\eta(\beta),\eta(\gamma)$ 각각 기저분포의 제 $100\alpha, 100\beta, 100\gamma$ 백분위수일 때, $n\to\infty$이면

$$\sqrt{n}(H_2-\mu_{H_2})\to N(0,\ \boldsymbol{a}'A\,\boldsymbol{a}) \tag{2.13}$$

$$\sqrt{n}(H_3-\mu_{H_3})\to N(0,\ \boldsymbol{b}'B\,\boldsymbol{b}) \tag{2.14}$$

이다.

단 $\mu_{H_2}=\frac{\eta(1-\beta)-\eta(\beta)}{\eta(1-\gamma)-\eta(\gamma)}, \mu_{H_3}=\frac{\eta(1-\alpha)-\eta(\alpha)}{\eta(1-\beta)-\eta(\beta)}$ 이고 열벡터 $\boldsymbol{a}$는 H_2를 각각

$r_{1-\beta}, r_\beta, r_{1-\gamma}, r_\gamma$ 에 대해 편미분하여 $\eta(1-\beta), \eta(\beta), \eta(1-\gamma), \eta(\gamma)$ 에서 계산한 열 벡터이고, A 는 $r_{1-\beta}, r_\beta, r_{1-\gamma}, r_\gamma$ 인 4개의 잔차 백분위수가 구성하는 분산-공분산 행렬이다. $\boldsymbol{b}$ 와 B 도 마찬가지로 정의된다.

증명: 먼저 (2.13)식을 증명하기 위해 $r_{1-\beta}, r_\beta, r_{1-\gamma}, r_\gamma$ 를 각각 X_1, X_2, X_3, X_4 라고 하고 $\eta(1-\beta), \eta(\beta), \eta(1-\gamma), \eta(\gamma)$ 를 각각 $\mu_1, \mu_2, \mu_3, \mu_4$ 라고 하자. $H_2(\boldsymbol{X}) = H(X_1, X_2, X_3, X_4)$ 라 놓으면 $H(\boldsymbol{X}) = (X_1 - X_2)/(X_3 - X_4)$ 이고, 이 식을 $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4)'$ 근방에서 다변량 Taylor 전개하면,

$$H(\boldsymbol{X}) - H(\boldsymbol{\mu}) = \sum_{i=1}^{4}(X_i - \mu_i)\partial H/\partial X_i \mid_{\boldsymbol{X}=\boldsymbol{\mu}} + \frac{1}{2!}\left\{\sum_{i=1}^{4}(X_i - \mu_i)^2 \partial^2 H/\partial X_i^2 \mid_{\boldsymbol{X}=\boldsymbol{\mu}} + 2\sum_{i<j}\sum(X_i - \mu_i)(X_j - \mu_j)\partial^2 H/\partial X_i \partial X_j \mid_{\boldsymbol{X}=\boldsymbol{\mu}}\right\} + \cdots$$

그런데 $i = 1, 2, 3, 4$ 에 대해 $\sqrt{n}(X_i - \mu_i)$ 은 (2.5)과 같은 점근적으로 정규분포를 따르고, $\sqrt{n}(X_i - \mu_i)^2 = \sqrt{n}(X_i - \mu_i)(X_i - \mu_i)$ 에서 $X_i - \mu_i \xrightarrow{p} 0$이므로, $\sqrt{n}(X_i - \mu_i)^2 \xrightarrow{p} 0$ 이다. 마찬가지로, $\sqrt{n}(X_i - \mu_i)(X_i - \mu_i) \xrightarrow{p} 0$. 그리고 3차항 이상의 항에 대해서도 모두 $\xrightarrow{p} 0$이다. 따라서, $\sqrt{n}(H(\boldsymbol{X}) - H(\boldsymbol{\mu}))$ 은 점근적으로 $\sqrt{n}\sum_{i=1}^{4}(X_i - \mu_i)\partial H/\partial X_i \mid_{\boldsymbol{X}=\boldsymbol{\mu}}$ 와 동일한 분포를 따른다. 즉 $\sqrt{n}\sum_{i=1}^{4}(X_i - \mu_i)\partial H/\partial X_i \mid_{\boldsymbol{X}=\boldsymbol{\mu}}$ 는 정리 2.3과 정리 2.4에 의해 점근적으로 평균이 0이고 분산이 $\boldsymbol{a}'A\boldsymbol{a}$ 임을 쉽게 확인 할 수 있어므로, (2.2)식은 증명된다. 마찬가지 방법으로 (2.14)식도 증명된다.

다음으로 본 논문에서 최종적인 선택통계량으로 사용될 $H = H_2 + H_3$ 통계량의 점근적 분포는 다음과 같다.

따름정리2.5 $H = H_2 + H_3$ 이고 $\eta(\alpha), \eta(\beta), \eta(\gamma)$ 가 각각 오차분포의 $100\alpha, 100\beta, 100\gamma$ 백분위수일 때, $n \to \infty$이면

$$\sqrt{n}(H - \mu_H) \to N(0,\ \boldsymbol{c}'C\boldsymbol{c}) \ , \tag{2.15}$$

이다.

단, 여기서 $\mu_H = \frac{\eta(1-\beta) - \eta(\beta)}{\eta(1-\gamma) - \eta(\gamma)} + \frac{\eta(1-\alpha) - \eta(\alpha)}{\eta(1-\beta) - \eta(\beta)}$ 이고 열벡터 $\boldsymbol{c}$는 선택통계량 H를 각각 잔차의 표본백분위수인 $r_{1-\alpha}, r_\alpha, r_{1-\beta}, r_\beta, r_{1-\gamma}, r_\gamma$ 에 대해 편미분하여 $\eta(1-\alpha), \eta(\alpha), \eta(1-\beta), \eta(\beta), \eta(1-\gamma), \eta(\gamma)$ 에서 계산한 열 벡터이고, C는 6개의 잔차의 백분위수인 $r_{1-\alpha}, r_\alpha, r_{1-\beta}, r_\beta, r_{1-\gamma}, r_\gamma$ 가 구성하는 분산-공분산 행렬이다.

증명 $r_{1-\alpha}$, r_{α}, $r_{1-\beta}$, r_{β}, $r_{1-\gamma}$, r_{γ} 를 각각 X_1, X_2, X_3, X_4, X_5, X_6 라고 하고 $\eta(1-\alpha), \eta(\alpha), \eta(1-\beta), \eta(\beta), \eta(1-\gamma), \eta(\gamma)$ 를 각각 $\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6$ 라고 하자. 그리고 $H(\boldsymbol{X}) = H(X_1, X_2, X_3, X_4, X_5, X_6)$ 라 놓고 $H(\boldsymbol{X})$ 를 $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6)'$ 근방에서 다변량 Taylor 전개하면, 정리2.3의 증명과 동일한 방법으로 증명된다.

여기서 제안하는 선택통계량은 Han (2002)의 위치모수의 추정에 있어서의 통계량 H와 마찬가지로 다양한 기저오차분포들을 잘 구분(discrimate)하여 주도록 α, β 와 γ 의 값을 결정해 주는 것이다. 많은 기저오차분포에 대해 모두 고려 할 수는 없으므로, 가벼운 꼬리를 가진 NOR 과 뾰족한 중심을 가진 DE, 그리고 극단적으로 무거운 꼬리를 가진 CA분포에 대해 이 세 개의 분포를 동시에 잘 구분하는 α, β 와 γ 의 값을 정하고자 한다. 이를 위해 $\sqrt{n}(H-\mu_H)$ 가 각각의 오차분포 NOR, DE, CA에서 서로 다른 점근적 정규분포를 따르기 때문에 각각의 평균과 분산을 각각 $(0, \sigma^2_{NOR}), (0, \sigma^2_{DE}), (0, \sigma^2_{CA})$ 라고 하자. 이 점근적 분포들은 이것들이 많이 떨어져 있을수록 선택통계량 H 에 의해 잘 구분이 될 것이다. 여기서 두 개의 분포의 점근적 거리를 다음과 같이 정의하자. 예컨대 오차분포가 NOR과 DE 일 때, 이것들의 거리를 다음과 같이 정의하자.

$$nD^2 = (\mu_{NOR} - \mu_{DE})^2 \left(\frac{1}{2}(\sigma^2_{NOR} + \sigma^2_{DE})\right)^{-1} \tag{2.16}$$

식 (2.4)는 동일한 분산-공분산 행렬 를 가지는 두 개의 다변량 분포 $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), (\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ 사이의 Mahalanobs 거리가 $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ 로 되고(1981, Mardia), 이것의 변형으로 분산-공분산 행렬이 $\boldsymbol{\Sigma}_1$ 과 $\boldsymbol{\Sigma}_2$ 로 같지 않을 때. Nakanish와 Sato(1985) 및 Hogg 등(1988)은 다음과 같이 두 개의 다변량 분포의 거리를 정의하였다.

$$nD^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \left(\frac{1}{2}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)\right)^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{2.17}$$

(2.16)식은 (2.17)식의 두 개의 분포에 대한 점근적 거리의 1차원적인 표현이다. (2.16)식을 이용하여 선택통계량 H 에 의한 (NOR,DE), (DE,CA), (NOR,CA) 사이의 점근적 거리를 각 d_1, d_2, d_3 라고 하고 $0 < \alpha < \beta < \gamma < 0.5$에 대해 $\alpha = 0.01$, $\beta = 0.02$ 그리고 $\gamma = 0.03$ 부터 시작하여 0.01씩 값을 증가시켜 각 분포들간의 거리를 <표 2.1>에서 보는 바와 같이 구성하면 d_1, d_2, d_3 의 값을 동시에 최대가 되게 만드는 α, β, γ 의 값이 없다. 따라서 $d_1 + d_2 + d_3$ 의 값을 최대로 하는 $\alpha = 0.02$, $\beta = 0.06$, $\gamma = 0.28$ 의 값을 최종적인 값으로 하여 선택통계량을 결정하였다. 즉

$$H_2 = \frac{r_{0.94} - r_{0.06}}{r_{0.72} - r_{0.28}}, \quad H_3 = \frac{r_{0.98} - r_{0.02}}{r_{0.94} - r_{0.06}}. \quad H = H_2 + H_3$$

<표 2.1> 선택통계량 H 를 사용한 두 분포들간의 점근적 거리

α 값	$\alpha=0.01$									$\alpha=0.02$								
β 값	$\beta=0.04$			$\beta=0.05$			$\beta=0.06$			$\beta=0.04$			$\beta=0.05$			$\beta=0.06$		
점근거리 / γ 값	d_1	d_2	d_3	d_1	d_2	d_3	d_1	d_2	d_3	d_1	d_2	d_3	d_1	d_2	d_3	d_1	d_2	d_3
0.22	628	719	1050	625	723	1120	620	703	1147	559	736	1092	564	776	1218	565	780	1279
0.24	644	720	1063	642	724	1137	638	702	1168	578	738	1108	582	777	1236	584	782	1301
0.26	655	717	1071	653	721	1147	650	699	1181	588	736	1114	594	775	1248	597	778	1316
0.28	658	711	1075	658	713	1152	655	690	1185	593	731	1123	600	768	1254	604	770	1324
0.30	655	702	1074	656	701	1151	655	677	1186	592	723	1130	600	757	1255	605	758	1325
0.32	646	690	1070	648	686	1145	648	659	1171	585	711	1128	594	743	1251	600	741	1318

3. 추정량의 제안

본 논문에서 제안할 추정량은 Huber와 Tukey의 회귀 M-추정량의 조절상수(tuning constant)를 선택통계량 H 의 값에 따라 결정하는 적응 M-추정량이다. 4절의 모의실험에서 단순회귀모형을 가정하고 회귀기울기에 대한 다양한 추정량의 효율을 생각할 것이므로 여기서 단순회귀모형에서의 Huber와 Tukey의 회귀 M-추정량에 대해 간단히 언급하겠다. 먼저 L_1 을 회귀모수에 대한 예비추정량이라 하고 이것의잔차를 각각 $r_i=y_i-\hat{\beta}_0-\hat{\beta}_1x_i,\quad (i=1,2,\cdots,n)$ 이라고 하고, $MAD=\ median_i\{|r_i-med_j(r_i)|\}$ 으로 그리고 $s=MAD/.6745$ 라 하면, 회귀모수에 대한 M-추정량의 형태는 다음의 가중최소제곱 방정식의 해가 된다는 것은 잘 알려져 있다.(1977 Holland와 Welsh, 1988 Hogg 등).

$$\sum_{i=1}^{n} w_i(y_i-\beta_0-\beta_1x_i)^2=0 \tag{3.1}$$

이때 조절상수 k를 가진 Huber의 M-추정량 와 Tukey의 M-추정량 $T(k)$는 각각

$$H(k)=\ [\ \sum_{i=1}^{n} w_i \sum_{i=1}^{n} X_iY_i-\ \sum_{i=1}^{n} w_iX_i \sum_{i=1}^{n} w_iY_i]/[\ \sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_iX_i^2-(\ \sum_{i=1}^{n} w_iX_i)^2]$$

단 $w_i=1,\ |r_i|\le ks$ 인 경우

$=ks/|r_i|,$ 다른 경우

$$T(k)=\left[\sum_{i=1}^{n} w_i \sum_{i=1}^{n} X_iY_i-\sum_{i=1}^{n} w_iX_i \sum_{i=1}^{n} w_iY_i\right]/\left[\sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_iX_i^2-\left(\sum_{i=1}^{n} w_iX_i\right)^2\right]$$

단 $w_i=\{1-(r_i/ks)^2\}^2$, $|r_i|\leq ks$ 인 경우

$=0$, 다른 경우

와 같이 정의된다. 그리고 기저 오차분포의 형태가 f(x) 일 때, 회귀 M-추정량의 점근적 분산은 $E(\psi)^2/(E(\psi'))^2(X\cdot X)^{-1}$ 와 같은 형태를 가진다는 사실은 잘 알려져 있다.(Huber(1981)). 단 $\psi=\log\rho(x)$, $\rho=-\log f(x)$ 이다. 단순회귀모형에서 $\sum_{i=1}^{n} x_i^2=1$인 경우 <표 3.1>과 <표 3.2>는 위의 공식을 사용하여 조절상수의 값에 따른 회귀기울기에 대한 M-추정량의 점근적 분산을 각각 구한 것이다. 상기 표에서 보는 바와 같이 NOR처럼 가벼운 꼬리를 가진 분포에서는 Huber와 Tukey의 M-추정량 모두 조절상수의 값이 크질수록 점근적 분산이 작아지는 경향이 있고, DE나 CA처럼 뾰족한 중심부분을 가진 분포나 극단적으로 무거운 꼬리를 가진 분포에서는 조절 상수의 값이 작아질수록 점근적 분산이 작아지는 경향이 있다. 그리고 확률 0.9로 표준정규분포를 따르고, 확률 0.1로 평균이 0, 분산이 9인 분포를 따르는 오염분포(CON), 자유도 3인 t-분포(T(3))처럼 중간정도의 두께를 가진 분포에서는 일정한 규칙이 없다는 사실을 알수 있다. 그리고 (2.6)에서 제안된 선택통계량 H 를 사용한 추정량을 제안하기위해 n=40개의 표본을 사용하여 4000번의 반복에 의한 모의실험에서 H 값의 변화에 따른 빈도수를 가벼운 꼬리를 가진 오차분포(NOR), 중간정도의 꼬리를 가진 오차분포(CON), 뾰족한 가운데 부분을 가진 오차분포(DE) 및 무거운 꼬리를 가진 분포(CA)에 대해 모의실험하였다. <표 3.3>에서 보는 바와 같이 NOR처럼 가벼운 꼬리나 중간정도의 두께를 가진 오차분포하에서는 선택 통계량 H 의 값이 작아지는 경향이 있고, DE나 CA처럼 뾰족한 중심부분을 가진 분포나 극단적으로 무거운 꼬리를 가진 분포에서는 H 의 값이 커지는 경향이 있다. 이와 같은 모의실험의 결과 본 논문에서 제안하는 두가지 형태의 적응 M-추정량은 다음과 같다.

$$AH=\begin{cases} H(1.5), & \text{if } H\leq 4.5 \\ H(1.2), & \text{if } 4.5< H< 9.0 \\ H(0.8), & \text{if } H> 9.0 \end{cases} \tag{3.2}$$

$$AT=\begin{cases} B(5.5), & \text{if } H\leq 4.5 \\ B(4.8), & \text{if } 4.5< H< 9.0 \\ B(4.0), & \text{if } H\geq 9.0 \end{cases} \tag{3.3}$$

<표 3.1> Huber M-추정량의 점근적 분산

	H(.5)	H(1.0)	H(1.5)	H(2.0)
NOR	1.2625	1.1073	1.0371	1.0140
CON	1.4789	1.3330	1.2959	1.3237
T(3)	1.5695	1.5207	1.5824	1.6862
DE	1.1652	1.3226	1.4653	1.5888
CA	2.1553	2.5465	2.9927	3.5208

<표 3.2> Tukey M-추정량의 점근적 분산

	B(3)	B(6)	B(9)	B(12)
NOR	1.2930	1.0160	1.0040	1.0013
CON	1.5156	1.2780	1.3760	1.4802
T(3)	1.7098	1.5904	1.7489	1.8799
DE	1.4037	1.4946	1.6383	1.7531
CA	2.2245	2.5891	3.2365	3.9847

<표 3.3> H 값의 변화에 따른 빈도수

H값 / 분포	4.5<	4.5-5.0	5.0-5.5	5.5-6.0	6.0-6.5	6.5-7.0	7.0-7.5	7.5-8.0	8.0-8.5	8.5-9.0	9.0>
NOR	2839	705	306	96	42	9	1	0	1	0	1
CON	1292	907	710	520	302	161	65	23	9	6	5
DE	225	333	492	546	476	456	353	247	178	194	500
CA	4	16	27	48	69	103	120	147	160	176	3132

4. 모의실험

우리의 모의실험에서는 단순회귀모형 $y_i=\beta_0+\beta_1 x_i+\varepsilon_i,\ i=1,2,\cdots,n$ 을 가정하고, $\beta_0=\beta_1=0$ 을 놓고, 이때 회귀기울기를 추정하여 평균제곱오차(M.S.E.)를 비교하여 추정량의 효율을 비교할 것이다. 디자인 행렬 X의 첫 번째 열을 모두 1을 그리고 두 번째열은 $x_i=\Phi^{-1}(i/n+1)$ 을 이용하였다. 여기서 $\Phi^{-1}(\cdot)$ 는 표준정규분포의 역함수이다.

모의실험에 사용된 분포는 아래의 7개를 사용하였다. 그리고 $S=Y/Z$ 형태의 확률변수에서 Y, Z 는 서로 독립이고 Y 는 표준정규분포의 확률변수이고, Z 는 다음과 같다.

(1) Normal(NOR) : $Z=1$

(2) Slate (TE) : $Z=U^{\frac{1}{10}}$, 여기서 U 는 표준균일분포를 따르는 확률변수.

(3) Slacu (CU) : $Z=U^{\frac{1}{3}}$

(4) Slash (SH) : $Z = U^{\frac{1}{3}}$

(5) Coutaminated (Con) : $Z = \begin{cases} 1 & , \text{확률 } 0.9 \\ 1/3 & , \text{확률 } 0.1 \end{cases}$

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

(7) Cauchy (CA) : $Z = |V|$, 여기서 V는 표준정규분포를 따르는 확률변수.

그리고 제안되는 추정량은 모두 7개로 구성되어 있는데 최소제곱추정량(LS), 최소절대값추정량(L_1), Huber의 H(1.25), Tukey의 T(4.82), p=1.277, x_1=1.344, r=4를 적용한Huber-Collins의 회귀 M-추정량 HC (Hampel 등(1986) 참조)와 위의 (3.2)과 (3.3)에서 정의한 선택통계량을 사용한 두 개의 적응 회귀 M-추정량 AH와 AT로 되어있다. 모의실험의 표본수와 반복횟수는 각각 N=20, 40, 80이고 10000번이고, 평균제곱오차(MSE)* 10000의 크기로 <표 4.1>에 결과가 주어져있다. 그리고 모의실험에서 사용된 측도(scale)은 s=MAD/.6745를 사용하였다. 위의 모의실험 결과를 요약하면 다음과 같다.

(1) 표본수가 N=20, 40, 80으로 변함에 따라 회귀기울기 추정량의 효율에는 큰 변동이 없음을 알 수 있고,적응 M-추정량인 AH와 AT는 광범위한 분포군에 대해 좋은 효율을 가짐을 알수 있다. 특히 AH와 AT는 각각 비적응 M-추정량인 H(1.25)와 T(4.82)에 비해 NOR,TE등 가벼운 꼬리를 가진 분포나 SH, CA등 극단적으로 무거운 꼬리를 가진 분포에 대해서는 상당히 우월하나, CU, CON, DE 처럼 중간정도의 두께의 꼬리를 가진 분포나 극단적으로 뾰족한 중심부를 가진 분포에 대해서는 비슷한 효율을 알수 있다. 즉 선택통계량의 사용에 의해 중간정도의 꼬리의 두께를 가진 분포에 대해 효율의 감소가 거의 없이 가벼운 꼬리나 극단적으로 무거운 꼬리를 가진 분포에 대해서는 많은 효율의 증가를 가져다 줌을 알 수 있다.

(2) 비적응 추정량인 H(1.25), T(4.82), HC중에 T(4.82)가 전반적으로 가장 효율이 좋은 추정량임을 알 수 있다. 그리고, HC추정량은 극단적으로 무거운 꼬리를 가진 분포의 경우외에는 H(1.25), T(4.82)에 비해 효율이 떨어짐을 알 수 있다.

(3) AH는 AT에 비해 H(1.25)에 비해 CU, CON, DE등 중간정도의 꼬리를 가진 분포나 중심부분이 뾰족한 분포에 대해서는 효율이 약간 좋으나, NOR, TE등 가벼운 꼬리를 가진 분포군에 대해서는 효율이 약간 떨어지고, 특히 무거운 꼬리를 가진 분포군에서는 효율이 떨어짐을 알수있다.

(4) 잘 알려진 사실이지만, 표본평균(Mean)은 다른 모든 M-추정량에 비해 가벼운 꼬리를 가진 분포군의 위치모수 추정에는 나은 효율을 가지나, 다른 경우에는 못한 효율을 가지며, 특히 무거운 꼬리를 가진 분포군의 위치모수 추정에는 아주 좋지못한 효율을 가진다. 그리고 표본중앙값(Med)은 DE처럼 극단적으로 뾰족한 중심을 가진 분포나, SH나 CA처럼 극단적으로 무거운 꼬리를 가지는 분포외에는 다른 모든 M-추정량에 비해 훨씬 못한 효율을 가진다. 결론적으로 위의 모의실험 결과 선택통계량을 이용한 M-추정법은 기존의 M-추정법에 비해 선택통계량의 사용에 의해 중간정도의 꼬리의 두께를 가진 분포에 대해 효율의 감소가 거의 없이 가벼운 꼬리나 극단적

으로 무거운 꼬리를 가진 분포에 대해서는 많은 효율의 증가를 가져다 줌을 알 수 있다. 좀더 많은 연구가 이루어져야 하겠지만, 기존의 M-추정법의 하나의 대안으로 사용될 수 있는 추정법임을 분명히 확인 할 수 있겠다. 앞으로 연구가 더 이루어져아 하는 부분은 선택통계량에 대한 좀더 정교한 연구가 이루어져야 하겠고, 선택통계량을 이용한 비대칭인 오차분포에 대한 회귀 모수 추정법의 연구가 이루어져야 하겠다.

< 표 4.1> 추정량들의 평균제곱오차

추정량 \ 분포	표본수	NOR	TE	CU	CON	DE	SH	CA
LS	N=20	9916*	12688*	29465	17914	19676	-	-
	N=40	9923*	12491*	30342	18083	20067	-	-
	N=80	10080*	12504*	30824	17840	19386	-	-
L_1	N=20	15522	19459	28349	18392	13645	67175	28276
	N=40	15630	19356	27905	18436	13662	68289	28573
	N=80	15825	18920	28472	17665	13291	67296	29692
H(1.25)	N=20	10851	13746	21145	13411	13343	68089	33447
	N=40	10772	13555	20689	13441	13662	68924	32956
	N=80	10957	13274	21331	13075	13130	68270	34341
T(4.82)	N=20	10774	13668	21110*	13177*	13555	58671	28220
	N=40	10682	13513	20683	13298*	13826	59125	27486
	N=80	10882	13183	21252*	12870*	13351	58564	28518
HC	N=20	11305	14376	21546	13425	13497	56023	26638
	N=40	11251	14197	21100	13523	13699	56661	26073
	N=80	11427	13810	21674	13152	13179	55708	26942
AH	N=20	10573	13509	21180	13414	13301*	61498	28574
	N=40	10503	13262	20714	13477	13607*	62485	28147
	N=80	10694	12986	21309	13074	13089*	61694	29497
AT	N=20	10389	13207	21166	13283	13683	55720*	25989*
	N=40	10269	13020	20638*	13433	13917	56371*	25564*
	N=80	10497	12801	21285	12917	13493	55386*	26427*

1) MSE값은 실제 얻어진 MSE값을 10000배 한 것이다.

2) 표본평균의 MSE가 SH와 CA의 경우 너무 큰 값이 나오므로 - 로 표시하였다.

3) 제안된 추정량들 중에 MSE값이 가장 작은 추정량을 *로 표시하였다

참고문헌

[1] Andrews, D.F. , Bickel, P.J. , Hampel, F.R. , Huber P.J. , Roger W.H. , Tukey, J.W. (1972). Robust Estimation of Location: Survey and Advances. Princeton Univ. Press, Princeton, NJ.

[2] Hogg, R.V. (1967). Some oberservations on robust estimation, J. Amer. Stat. Assoc. 62, 1179-1186.

[3] Hogg, R.V., Fisher D.M., Randles R.H. (1975). A two-sample adaptive distribution-free test. J. Amer. Stat. Assoc. 70, 656-661.

[4] Hogg, R.V., (1983). On adaptive statistical inference, Comm. Statist. 11, 2531-2542.

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

[6] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.M., Stahel, W.A.(1986). Robust Statistics, the Approach Based on Influence Functions. John Wiley and Sons.

[7] Han, S.M. (2002). Adaptive M-estimation using Selector Statistics in Location Model. 한국통계학회 논문집 제 9권 제 2호. 325-335

[8] Holland, P.W., Welsh, R.E. (1977). Robust Regression using Iteratively Reweighted Least Squares, Comm. Statist. A6, 813-827.

[9] Huber, P.J. , (1964). Robust estimation of location parameter, Ann. Math. Stat. 35, 73-101

[10] Huber, P.J.,(1972). Robust statistics: A review. Ann. Math. Stat. 43, 1041-1967.

[11] Huber, P.J.,(1981). Robust statistics. John Wiley and Sons.

[12] Nakanish, H., Sato, Y. (1985). The performance of the linear and quadratic discriminant functions for three types of non-normal distributions, Comm. Statist. 14, 1181-1200.

[13] Ruppert, D., Carroll, R. J. (1980). Trimmed Least Squares Estimation in the Linear Model, J. Amer. Stat. Assoc. 75, 828-838.

[2003년 8월 접수, 2003년 11월 채택]