응용통계연구 제 **14**권 2호, *2001*년, *pp. 463–473*

회귀계수의 최소절대편차추정량의 표준편차 추정법*

이기훈[1] 정성석[2]

요 약

선형모형의 회귀계수의 L_1-추정량의 점근분포는 오차항의 중앙값에 종속되어있는데, 이 값은 잔차의 순서통계량의 함수로 추정될 수 있다. 본 논문에서는 오차항 중앙값의 추정량을 유도하는 몇 가지 방법을 소개하고 몬테칼로 실험을 통하여 가장 바람직한 추정량의 형태를 제안하였다. 또한 제안한 추정량을 이용하면 검정문제에서도 좋은 결과를 얻을 수 있음을 보였다.

주요용어: L_1-추정량, 회귀계수, 오차항의 중앙값.

1. 서론

최소절대편차(Least Absolute Deviation: LAD, L_1) 추정량은 Boscovich(1757), Laplace (1793)가 제안한 이후로, 직관적인 단순성에도 불구하고 이론의 난해함과 계산의 어려움 때문에 널리 이용되지 않았었다. 그런데 최근 들어서 컴퓨터 능력의 발달과 알고리즘의 개발, 그리고 분포이론의 정립 등으로 점차 많은 사용빈도를 보여주고 있다.

근래의 관련연구를 보면 Charnes, Cooper와 Ferguson(1955)이 LAD 방법의 해를 선형계획법(Linear Programing) 알고리즘을 이용하여 얻는 방법을 제안하였고, Birkes와 Dodge (1993)가 반복적인 수치해석적 알고리즘 해법을 제시하였다. 또한 LAD 추정량의 점근분포이론의 발달은 Basset과 Koenker (1978)가 점근분포이론을 유도한 이후부터 활발하게 진행되어, Bloomfield와 Steige(1983), Dupacovà(1987), Dupacovà와 Wets(1988) 등이 LAD 추정량의 강일치성(strong consistency)을 증명하였고, Bai, Rao와 Yin(1987)이 완화된 조건 하에서 점근분포를 유도하였다.

본 논문에서는 점근분포의 장애모수(nuisance parameter)인 LAD 추정량의 표준편차의 적절한 추정법에 관해서 관심을 갖는다. 이는 Sposito와 Tveite(1986), Birkes와 Dodge(1993) 등이 잔차의 순서통계량의 함수 형태로 제안한 바 있다. 본 논문에서는 컴퓨터 모의실험을 통하여 잔차의 순서통계량을 이용한 추정법의 특성을 살펴보고, 가장 바람직한 형태의 통계량을 제시하였다.

2장에서는 우리가 고려하는 선형회귀모형에서의 회귀계수의 LAD 추정량의 점근분포이론을 소개하였다. 3장에서는 이를 실제적으로 구간추정이나 검정 등에 이용할 수 있도록

* 본 논문은 1998년도 학술진흥재단 기초과학 연구소 지원과제(1998-015-D00048)의 지원으로 수행되었음.
1) (560-759) 전북 전주시 완산구 효자동, 전주대학교 경영학부, 부교수
E-mail: khlee@jeonju.ac.kr
2) (561-756) 전북 전주시 덕진구 덕진동, 전북대학교 수학·통계정보과학부, 부교수
E-mail: sschung@stat.chonbuk.ac.kr

오차항 중앙값의 추정값을 구하는 방법을 소개하고, 이들 추정량의 특성을 모의실험을 통하여 비교하였다. 또한 4장에서 L_1-추정량을 사용한 회귀계수 검정법들을 소개하고 오차항 중앙값의 추정량의 형태에 따른 검정통계량들의 점근적 성질을 모의실험을 통해 비교하였다.

2. 선형회귀계수의 LAD 추정량의 점근분포

본 연구에서 고려하는 선형모형의 수리적 모형은 다음과 같다.

$$y_i = \beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i; \quad i = 1, \cdots, n, \tag{2.1}$$

여기서, $\boldsymbol{x}_i$는 i번째 설명변수인 $k \times 1$ 벡터, $\boldsymbol{\beta}$는 회귀계수인 $k \times 1$ 벡터, β_0은 회귀직선의 절편, 그리고 i번째 오차항 ϵ_i는 독립적으로 공통 분포함수 F를 따르는 확률변수이다.

L_1-추정법은 다음 식을 최소화하는, $\hat{\beta}_0$, $\hat{\boldsymbol{\beta}}$을 유도한다.

$$\sum_{i=1}^{n} |y_i - \hat{\beta}_0 - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}| = \min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^{n} |y_i - \beta_0 - \boldsymbol{x}_i'\boldsymbol{\beta}|.$$

여기서 절편을 제외한 기울기 회귀계수의 LAD 추정량 $\hat{\boldsymbol{\beta}}$의 점근분포는 Bai et al.(1987)에 의하여 다음과 같이 유도되었다.

정리 2.1 모형 (2.1)에서 다음 조건이 만족한다고 가정하자.

i) $|u| \le \Delta$일 때 $f(u) = F'(u)$가 성립하는 Δ가 존재하고, f는 0에서 연속이고 양의 값을 가지며 $f(0) = 1/2$이다.

ii) $S_n = \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})'$가 (단, $\overline{\boldsymbol{x}} = \sum_i^n \boldsymbol{x}_i/n$) 어떤 n에 대해서도 정칙(nonsingular)행렬이고 다음을 만족한다고 하자.

$$\lim_{n\to\infty} \max_{1 \le i \le n} \boldsymbol{x}_i' S_n^{-1} \boldsymbol{x}_i = 0$$

이때 다음과 같은 분포근사가 성립한다.

$$2f(0) S_n^{\frac{1}{2}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{L} N(\mathbf{0}, I_k) \qquad \square$$

예를 들어, $k = 1$일 때 기울기 회귀계수 β_1의 LAD 추정량 $\hat{\beta}_1$의 점근분포는 다음과 같다.

$$2f(0)\left(\sum (x_i - \overline{x})^2\right)^{\frac{1}{2}} (\hat{\beta}_1 - \beta_1) \xrightarrow{L} N(0, 1).$$

이는 점근적으로 다음과 같은 분포를 따른다고 표현할 수 있다.

$$(\hat{\beta}_1 - \beta_1) \sim N\left(0, \frac{\tau^2}{\sum (x_i - \overline{x})^2}\right),$$

여기서, $\tau = \{2f(0)\}^{-1}$이다. 즉 τ는 LS 추정량의 분포와 비교하면 오차항의 표준편차 σ와 같은 의미인데, 이는 오차항의 분포에 의존하기 때문에 구간추정이나 검정을 위해서는 추정값을 사용할 수밖에 없다. 다음 장에서는 τ를 추정하는 방법들을 소개하고 이들의 특성을 비교하겠다.

3. LAD 추정량의 표준편차의 추정

Sposito와 Tveite(1986), Birkes와 Dodge(1993) 등이 τ를 추정하는 방법을 제안하였는데, 이들은 모두 잔차의 순서통계량의 함수이다. 여기서는 $k = 1$인 경우를 가정하고 두 방법을 간단히 소개하겠다. 우선 다음과 같이 회귀계수의 LAD 추정량에 의해 계산된 잔차를 크기 순으로 정렬한다.

$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i; \quad i = 1, 2, \cdots, n,$$

$$e_{(1)} \le e_{(2)} \le \cdots \le e_{(n)}.$$

Sposito와 Tveite(1986)는 특정 순서통계량의 차이의 분포가 평균이 $\{nf(0)\}^{-1}$인 지수분포를 따른다는 Cox와 Hinkley(1974)의 결과를 이용하여 τ의 추정값을 다음과 같이 제안하였다.

$$\hat{\tau} = \frac{(e_{(t)} - e_{(s)})}{2(t-s)/n}, \tag{3.1}$$

여기서 t, s는 각각 $[n/2] \pm v$에 가장 가까운 정수이고 v는 $v = O(n^d)$, $0 < d < 1$을 만족하는 임의의 양의 정수인데, v의 적당한 값을 찾기 위해 Sposito와 Tveite(1986)가 Monte Carlo 연구를 하였다.

Birkes와 Dodge(1993)는 중앙값도 표본의 크기가 증가하면 정규분포에 근사하므로 중앙값에 관한 95% 신뢰구간을 $\text{med}(\epsilon_i) \pm 2\hat{\tau}/\sqrt{n}$ 같이 근사할 수 있다는 사실로부터 다음과 같은 추정방법을 제안하였다.

$$\hat{\tau}^* = \frac{\sqrt{m}(e_{(k_2)} - e_{(k_1)})}{4}, \quad m = n - k - 1, \tag{3.2}$$

여기서, k_1, k_2는 각각 $(m+1)/2 \mp \sqrt{m}$에 가장 가까운 정수이다.

앞의 식 (3.1)과 (3.2)를 비교하면 어떤 순서 잔차를 이용하는가만 다르다고 할 수 있다. Cox와 Hinkley(1974)의 이론적인 결과에서 이들 t, s값에 관한 유일해가 존재하지 않기 때문에 이 값을 결정하는 것은 사용자의 몫인데, Birkes와 Dodge(1993)는 직관적인 유도를 통해 그 값을 제안한 반면, Sposito와 Tveite(1986)는 모의실험을 통해 제안하였다.

Sposito와 Tveite(1986)는 $v = 1/2[pn]$, $p \in (0,1)$이라 놓고 2~12까지의 $[pn] = t - s$에 대하여 각 모집단에 따른 τ의 추정값을 모의실험을 통하여 구하였다. 모의실험 결과 정규분포의 경우 어떤 $[pn]$에 대해서도 추정이 매우 정확하고, 두꺼운 꼬리를 갖는 분포에서도 표본의 크기가 100이상이거나 $[pn]$이 6 이상일 때 적합한 추정을 한다고 결론을 내렸다. 그러

나 직관적으로도 이와 같이 좁은 구간의 잔차를 사용하면 추정량의 분산이 클 거라고 추측할 수 있는데, 이 사실은 본 연구의 모의실험에서도 확인할 수 있었다. 모의실험에 의하면 Sposito와 Tveite가 제안한 바와 같이 작은 p에서 얻은 $\hat{\tau}$는 편의(bias)는 작으나, 가까운 잔차들을 사용하므로 추정량의 분산 $\widehat{Var}(\hat{\tau})$이 크게 나왔다.

Birkes와 Dodge(1993)는 $2\sqrt{m} = 2\sqrt{n-k-1}$ 정도 떨어진 잔차들을 이용하고 있으므로 표본의 크기에 따라 사용하는 잔차가 조정되지만, 95% 신뢰구간을 기준으로 유도한 추정량이기 때문에 두꺼운 꼬리를 갖는 오차항 분포에서는 편의가 존재할 것임을 직관적으로 알 수 있다. 본 연구의 모의실험 결과, Birkes와 Dodge의 $\hat{\tau}^*$는 추정량의 분산 $\widehat{Var}(\hat{\tau}^*)$은 작지만 편의가 큼을 표 3.1에서 확인할 수 있었다.

본 논문에서는 p값을 변화시키며 식 (3.1)의 $\hat{\tau}$ 값을 컴퓨터 모의실험에 의해 계산하여 실제 τ와 가장 근사한 경우의 p값을 찾으려하였다.

다음의 표 3.1은 $p = 0.1, 0.3, 0.5, 0.7, 0.9$일 때 식 (3.1)에 의한 LAD 추정량의 편차 추정값 $\hat{\tau}$의 평균과 그 평균제곱오차를 구한 모의실험 결과이다. 여기서는 모형을 단순회귀모형인 $y_i = \beta_0 + \beta_1 x_i + \epsilon_i;\ i = 1, \cdots, n\ (n = 20, 50, 100, 200)$으로 하였고, 실험점은 $x = (0, 1, \cdots, 9)$와 같이 등간격이다. 또한 고려된 오차 분포의 함수식은 다음과 같다.

$$\text{Uniform}: f(x) = \begin{cases} 1, & -1/2 \le x \le 1/2 \text{ 일때} \\ 0, & \text{기타의 경우} \end{cases}$$

$$\text{Normal}: f(x) = (\sqrt{2\pi})^{-1} \exp(-\tfrac{x^2}{2}) \tag{3.3}$$

$$\text{Laplace}: f(x) = 1/2 \exp(-|x|)$$

$$\text{Cauchy}: f(x) = \left\{\pi(1+x^2)\right\}^{-1}$$

모의실험은 FORTRAN IMSL 패키지에서 식 (3.3)의 각 분포를 따르는 난수를 얻어 각 경우에 10,000회 반복 실험하였다. 표 3.1에 추정값과 참값을 제시하였고, 각 추정값의 평균제곱오차는 괄호 안에 표기하였다. 또한 모의실험을 통해 식 (3.1)에서 n을 $m = n-k-1 = n-2$으로 대치하면 더욱 정확한 추정값을 얻을 수 있음을 확인하였기 때문에 여기서는 $n-2$를 사용하였다.

표 3.1은 p의 값에 따라 τ가 얼마나 정확하게 추정되는 지를 보여주고 있다. 또한 앞에서 소개한 두 방법의 결과도 이 표에서 유추할 수 있는데, Sposito와 Tveite(1986) 방법은 $p = 0.1$ 미만이고, Birkes와 Dodge(1993)는 표본의 크기에 따라 $p = 0.1 \sim 0.6$이다. 표 3.1의 결과에 따르면 꼬리가 두꺼운 모집단 분포에서는 p의 값이 0.3인 경우에 가장 정확한 추정이 이루어지고 있고, 정규분포에서는 $p = 0.3 \sim 0.7$, 일양분포에서는 $p = 0.3 \sim 0.9$일 때, 매우 정확하게 τ를 추정함을 알 수 있다.

회귀계수의 추정에서 LAD 추정량을 사용하는 이유는 로버스트한 추정값을 얻기 위함인데, 그러한 측면에서 두꺼운 분포에서 가장 우수한 행태를 보이는 p 값을 사용하는 것이 바람직 할 것이다. $p = 0.3$인 경우에 대략적으로 모든 오차항 분포에서 만족할 만한 추정값을 얻을 수 있기 때문에 본 논문에서는 표본의 크기에 관계없이 $v = 1/2[0.3n]$이고

표 3.1: τ의 추정값

분포	τ의 추정값 평균 (( )안은 추정량의 평균제곱오차)					
			p			
τ의 참값	0.1	0.3	0.5	0.7	0.9	n
Uniform	.220 (.124)	.357 (.039)	.389 (.022)	.411 (.013)	.451 (.006)	20
	.374 (.047)	.438 (.013)	.452 (.007)	.459 (.004)	.479 (.002)	50
	.436 (.023)	.468 (.006)	.475 (.003)	.478 (.002)	.488 (.001)	100
0.5	.467 (.011)	.484 (.003)	.488 (.001)	.489 (.001)	.493 (.000)	200
Normal	.589 (.749)	.997 (.218)	1.135 (.109)	1.289 (.078)	1.641 (.264)	20
	.977 (.298)	1.167 (.083)	1.262 (.044)	1.403 (.058)	1.739 (.284)	50
	1.117 (.141)	1.228 (.040)	1.307 (.026)	1.442 (.054)	1.786 (.308)	100
1.2533	1.188 (.071)	1.255 (.021)	1.329 (.018)	1.461 (.053)	1.805 (.317)	200
Laplace	.596 (.499)	1.025 (.201)	1.242 (.223)	1.559 (.513)	2.384 (2.444)	20
	.898 (.203)	1.135 (.108)	1.330 (.181)	1.649 (.507)	2.454 (2.318)	50
	.998 (.105)	1.166 (.074)	1.360 (.168)	1.687 (.519)	2.511 (2.389)	100
1	1.031 (.056)	1.174 (.053)	1.370 (.156)	1.702 (.516)	2.531 (2.398)	200
Cauchy	.875 (1.247)	1.544 (.548)	2.004 (.990)	3.141 (6.499)	32.486 (184009.8)	20
	1.316 (.472)	1.651 (.203)	1.998 (.412)	2.879 (2.388)	8.303 (71.569)	50
	1.449 (.241)	1.668 (.105)	1.997 (.287)	2.833 (1.877)	7.747 (46.694)	100
1.5708	1.516 (.117)	1.683 (.058)	1.996 (.232)	2.822 (1.702)	7.343 (36.396)	200

n을 $m = n - k - 1$으로 대치한 식 (3.1)을 사용하여 τ를 추정할 것을 제안한다. 예를 들어 $n = 10$인 경우에는

$$\hat{\tau} = \frac{e_{(7)} - e_{(4)}}{2(7-4)/8}$$

를 사용하여 τ를 추정한다.

4. LAD를 이용한 회귀계수의 검정

이 장에서는 검정통계량에 τ가 포함된 경우 그 추정량에 따른 검정통계량의 근사적 성질을 파악하고자한다. 모형 (2.1)의 일반적 다중회귀모형에서 다음 가설을 검정한다고 가정하자.

$$H_0 : \beta_{q+1} = \beta_{q+2} = \cdots = \beta_k = 0 \tag{4.1}$$

여기서 먼저 τ에 관계 없이 검정통계량을 구축할 가능성에 대하여 언급하고자 한다. Rao와 Toutenburg(1995)는 $\hat{\tau}$를 구하지 않고도 위의 가설을 검정하는 검정통계량을 만드는

방법을 다음과 같이 이론적으로 제안하였다.

정리 4.1 정리 2.1의 조건과 식 (4.1)의 H_0 아래서 다음이 성립한다.

$$\frac{2\tau}{\sigma^2}\sum_{i=1}^{n}\left[|y_i-\tilde{\beta}_0-\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}|-|y_i-\hat{\beta}_0-\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}|\right]\xrightarrow{L}\chi^2_{k-q}.$$

여기서, χ^2_{k-q}은 자유도가 $k-q$인 카이제곱분포이다. □

정리 4.1의 통계량에 포함된 장애모수 τ를 제거하기 위해 Rao와 Toutenburg(1995)는 다음과 같이 가상의 확장된 회귀모형을 고려하였다.

$$y_i=\beta_0+\boldsymbol{x}_i'\boldsymbol{\beta}+\boldsymbol{z}_i'\boldsymbol{\gamma}+\epsilon_i\,;\quad i=1,\cdots,n$$

여기서, $\boldsymbol{z}_i$는 $(s\times 1)$벡터이고 $Z=(\boldsymbol{z}_1,\cdots,\boldsymbol{z}_n)'$, $X=(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n)'$일 때 다음을 만족해야 한다.

$$Z'X=\boldsymbol{O},\ Z'Z=I_s,\ d_n=\max_{1\le i\le n}|\boldsymbol{z}_i|\to 0. \tag{4.2}$$

그리고 $(\beta_0^*,\ \boldsymbol{\beta}^*,\ \boldsymbol{\gamma}^*)$가 다음 식을 최소화하는 LAD추정량이라 하면,

$$\min_{\beta,\gamma}\sum_{i=1}^{n}|y_i-\beta_0-\boldsymbol{x}_i'\boldsymbol{\beta}-\boldsymbol{z}_i'\boldsymbol{\gamma}|,$$

정리 4.1에 의하여 다음 정리 4.2가 성립한다.

정리 4.2 (Rao와 Toutenburg(1995))
정리 4.1의 조건 아래서 다음이 성립한다

$$\frac{s\sum_{i=1}^{n}\left[|y_i-\tilde{\beta}_0-\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}|-|y_i-\hat{\beta}_0-\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}|\right]}{(k-q)\sum_{i=1}^{n}\left[|y_i-\hat{\beta}_0-\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}|-|y_i-\beta_0^*-\boldsymbol{x}_i'\boldsymbol{\beta}^*-\boldsymbol{z}_i'\boldsymbol{\gamma}^*|\right]}\xrightarrow{L}F(k-q,s).$$

□

완전모형과 축소모형에 따른 정리의 아이디어를 이용하여 어떤 제약조건을 만족하는 확장모형의 설명변수 행렬을 도입하면, 장애모수 τ가 소거되어 τ에 관계없이 검정을 행할 수 있다는 것이 Rao와 Toutenburg(1995)의 주장이다. 그러나 정리 4.2의 통계량을 사용하기에는 몇 가지 어려움이 있다. 우선 식 (4.2)을 만족하는 가상의 실험점을 유도하기 어려운 경우가 많고, 통계량의 분모가 0이 되는 경우가 자주 발생할 수 있다는 약점 때문이다. 본 연구의 과정 중에 간단한 설명변수 행렬의 경우 가상의 실험점을 더해 이 통계량의 특성을 모의실험하여 본 결과, 통계량의 분모가 0이 되는 경우가 매우 빈번히 발생하였고 점근분포 또한 적은 표본에서 매우 나쁜 근사를 보여주었기 때문에, Rao와 Toutenburg(1995)의 검정통계량의 사용을 권장하지 않는다. 그러므로 검정문제에서도 τ의 추정단계는 반드시 필요한 과정이다.

식 (4.1)을 검정하는 Birkes와 Dodge(1993)가 제안한 LAD 추정량에 기초한 검정통계량은 다음과 같다.

$$F = \frac{\sum_{i=1}^{n}\left[|y_i - \tilde{\beta}_0 - \boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}| - |y_i - \hat{\beta}_0 - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}|\right]}{(k-q)(\hat{\tau}/2)} \tag{4.3}$$

여기서, $\hat{\tau}$는 $m = n - k - 1$인 식 (3.2)와 같으며, $(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}})$는 가설 (4.1) 아래의 축소모형에서 구한 LAD 추정량 벡터이고, $(\hat{\beta}_0, \hat{\boldsymbol{\beta}})$는 (2.1)의 완전모형에서 구한 LAD 추정량 벡터이다. 이 검정통계량은 근사적으로 자유도 $k-q$, $n-k-1$을 갖는 F분포를 따른다. 그러나 Birkes와 Dodge(1993)는 실제적인 사용에서 n이 작은 경우 다음과 같은 값을 사용할 것을 제안하였다.

$$G = (k-q)(1-(k-q)/n)F \tag{4.4}$$

이는 자유도가 $k-q$인 카이제곱분포를 따른다.

본 논문에서는 주어진 가설 (4.1)에서 $k = 1$, $q = 0$인 경우에 검정통계량 (4.3), (4.4)의 모의실험에 의한 경험적 유의수준 결과를 제시한다. 다음의 표 4.1은 표본점이 $x = (-1, 1)$ 등과 같은 대칭형과 $x = (1, 2, 3, \cdots)$ 등의 증가형일 때, $p = 0.1, 0.3, 0.5, 0.7, 0.9$에 따른 두 검정통계량의 10% 경험적 유의수준 값(괄호 안은 5%)이다. 적은 표본에서의 F와 G의 비교에 관심이 있으므로 $n = 10, 20$만 고려하였고, 앞에서와 같이 FORTRAN IMSL 패키지에서 (3.3)의 각 분포를 따르는 난수를 얻어 각 경우에 10,000회 반복 실험하였다.

표 4.1의 실험결과에 의하면 $p = 0.3$에 의하여 추정한 $\hat{\tau}$를 사용한 경우에는 식 (4.3)의 F의 경험적 유의수준 값이 실제 유의수준 값과 매우 근사하였다. 이는 $p = 0.3$에 의해 τ를 추정하면 식 (4.4)와 같은 수정통계량이 필요 없음을 시사하고 있다. Birkes와 Dodge(1993)가 적은 표본에서 검정통계량 G를 사용할 것을 제안하였지만 구체적인 표본크기를 제시하지 않았기 때문에 실제적인 적용에 어려움이 있었다. 그러나 τ를 추정할 때 $p = 0.3$을 사용하면 수정 없이 원래의 검정통계량 F를 사용할 수 있다는 것이 본 논문의 모의실험에 의해 밝혀졌으므로 LAD를 이용한 검정문제에서 표본의 크기에 관계없이 검정통계량 F를 사용할 것을 제안한다. 또한 본 연구에서 $k = 2$인 경우 (4.1)을 검정하는 모의실험도 수행하였는데, 그 결과는 표 4.1과 유사하여 생략하였다.

5. 결 론

본 논문에서 소개한 회귀계수에 관한 LAD 추론은 컴퓨팅 계산능력의 확대와 분포이론의 발전으로 인하여 많은 분야에서 사용이 점차 확대되는 추세이다. 실제적으로 LAD 추정량을 사용하려면 오차항분포의 중앙값을 추정하는 절차가 반드시 필요한데, 우리는 이를 추정하는 기존의 두 가지 방법을 컴퓨터 모의실험을 통해 비교하고 바람직한 추정 방법을 제시하였다. 이는 기존에 제안된 점근적 이론에 의한 방법을 검증 없이 사용하여 LAD 회귀추정법으로 열악한 결과를 도출하는 것을 막고자하는 의도에서 비롯되었다.

표 4.1: 검정통계량의 경험적 유의수준

표본수와 실험점	분포	검정통계량의 경험적 유의수준($\times 10,000$)					
		p	0.1	0.3	0.5	0.7	0.9
$n = 10$ 실험점 = $(-1, \cdots, -1, 1, \cdots, 1)$	Uniform	F	2104(1411)	**1033(296)**	833(217)	828(268)	838(278)
		G	2383(1747)	1327(592)	1082(516)	1089(527)	1121(522)
	Normal	F	2186(1466)	**1014(285)**	800(204)	661(192)	437(127)
		G	2494(1802)	1328(619)	1046(472)	902(384)	612(249)
	Laplace	F	2331(1568)	**1140(391)**	772(197)	520(125)	235(36)
		G	2662(1938)	1455(729)	1013(428)	723(291)	360(113)
	Cauchy	F	2420(1647)	**1215(437)**	759(166)	411(87)	83(12)
		G	2734(2009)	1531(776)	1000(426)	561(206)	142(33)
$n = 10$ 실험점 = $(1, 2, \cdots, 10)$	Uniform	F	1844(1209)	**844(236)**	800(179)	837(196)	843(209)
		G	2124(1490)	1149(490)	1137(417)	1164(473)	1145(487)
	Normal	F	1890(1232)	**943(312)**	879(197)	712(166)	466(101)
		G	2169(1541)	1272(577)	1178(474)	1009(396)	686(231)
	Laplace	F	2078(1407)	**1053(400)**	819(190)	553(112)	233(33)
		G	2331(1727)	1399(661)	1124(440)	810(274)	388(96)
	Cauchy	F	2096(1446)	**1046(393)**	726(175)	393(85)	94(12)
		G	2335(1737)	1369(643)	1048(388)	562(203)	137(42)
$n = 20$ 실험점 = $(-1, \cdots, -1, 1, \cdots, 1)$	Uniform	F	1242(705)	**900(301)**	925(356)	944(380)	949(371)
		G	1330(789)	1000(402)	1018(460)	1037(471)	1036(474)
	Normal	F	1168(671)	**856(306)**	801(294)	656(226)	411(104)
		G	1256(765)	953(402)	898(391)	755(305)	476(147)
	Laplace	F	1214(723)	**795(298)**	643(243)	418(129)	137(26)
		G	1308(832)	873(380)	724(304)	483(186)	162(35)
	Cauchy	F	1248(715)	**796(273)**	553(193)	252(61)	20(7)
		G	1354(820)	893(366)	635(243)	301(88)	29(8)
$n = 20$ 실험점 = $(1, 2, \cdots, 20)$	Uniform	F	1341(731)	**859(300)**	872(271)	877(308)	883(318)
		G	1446(857)	988(388)	987(362)	994(407)	1002(407)
	Normal	F	1456(864)	**993(371)**	935(318)	763(266)	454(123)
		G	1571(976)	1098(477)	1025(433)	856(347)	525(170)
	Laplace	F	1406(820)	**934(359)**	718(253)	480(156)	183(32)
		G	1513(949)	1039(455)	819(319)	554(202)	209(63)
	Cauchy	F	1514(897)	**979(412)**	708(240)	334(107)	36(6)
		G	1631(1010)	1093(517)	811(322)	384(134)	45(8)

LAD를 사용하는 것은 로버스트한 결과를 얻기 위함인데, 기존의 추정법은 두꺼운 분포에서 τ를 정확히 추정하지 못하였다. 제안한 $p = 0.3$에 의한 추정법이 두꺼운 오차항의 분포에서 우수한 행태를 보였으므로 제안한 추정 방법은 기존의 추정법보다 로버스트한 관점에서 우수하다고 할 수 있다. 또한 중회귀모형의 회귀계수 검정에 있어서 제안한 추정법에 의하여 τ를 추정하면 적은 표본에서도 검정통계량을 수정 없이 사용할 수 있기 때문에 표본의 크기에 관계없이 동일한 통계량을 사용할 수 있다는 이점이 있다.

참고문헌

[1] Bai, Z.D., Rao, C.R. and Yin, Y.Q. (1987). Least absolute deviations analysis of variance, *Sankhya*, vol. **52**, 166-177.

[2] Basset, G. and Koenker, R. (1978). Asymptotic theory of least absolute error regression, *Journal of the American Statistical Associations*, vol. **73**, 618-622.

[3] Birkes, D. and Dodge, Y (1993). *Alternative Methods of Regression*, John Wiley & Sons, New York.

[4] Bloomfield, P. and Steige, W.L. (1983). *Least Absolute Deviations*, Birkhäuser, Boston.

[5] Boscovich, R.J. (1757). De Litteraria expeditione per pontificiam ditionem, et synopsis amplioris operis, *Binoniensi Scientiarum et Artum Instituto atque Academia Commentarii*, vol. **4**, 353-396.

[6] Charnes, A., Cooper, W.W. and Ferguson, R.O. (1955). Optimal estimation of executive compensation by linear programming, *Management Science*, vol. **1**, 138-151.

[7] Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*, Chapelman and Hall, London.

[8] Dupacovà, J. (1987). Asymptotic properties of restricted L_1-estimates of regression, *Statistical Data Analysis Based on the L1-norm and Related Methods*, North-Holland, Amsterdam, 263-274.

[9] Dupacovà, J. and Wets, R. (1988). Asymptotic behavior of statistical estimators and optima solutions of stochastic optimization problems, *The Annals of Statistics*, vol. **16**, 1517-1549.

[10] Laplace, P.S. (1793). Sur quelques points du syst me du monde, *Oeuvres*, vol **11**, 477-558.

[11] Rao, C.R. and Toutenburg, H. (1995). *Linear Models -Least Squares and Alternatives*, Springer, New York.

[12] Sposito, V.A. and Tveite, M.D. (1986). On the estimation of the variance of the median

used in L1 linear inference procedures, *Communications in Statistics*, vol. **15**, No. 4, 1367-1375

[2001년 4월 접수, 2001년 7월 채택]

A Study on the Estimation of Standard Deviation of Least Absolute Deviation Estimators of Regression Coefficients *

Ki Hoon Lee[1] Sung S. Chung[2]

ABSTRACT

The asymptotic distribution of the L_1-estimator in linear models depends on the median of the error distribution which can be estimated with a function of the ordered residuals. This paper investigates methods for estimating τ, the function of the median of the error distribution. A Monte Carlo study was conducted to compare the usefulness of the methods in estimating τ and to suggest the good estimator of τ. Also we examine behaviors of the test statistics for the regression coefficients by simulation studies.

Keywords: L_1-estimator; Regression coefficients; Median of the error distribution.

* This paper was supported by the Basic Science Research Institute Program of the Korea Research Foundation, 1988, Project No. 1998-015-D00048.

1) Associate Professor, School of Business, Jeonju University.
E-mail: khlee@jeonju.ac.kr

2) Associate Professor, Division of Mathematics and Statistical Informatics, Chonbuk National University.
E-mail: sschung@stat.chonbuk.ac.kr