응용통계연구 제 19권 1호, 2006년, pp. 97–110

로버스트 회귀추정에 의한 신뢰구간 구축

이동희[1] 박유성[2] 김기환[3]

요 약

대부분의 자료는 여러가지 원인으로 인한 특이치로 오염되어 있으며, 이러한 상황에서 신뢰성 있는 추정량을 얻어내고 이에 대한 통계적 추론을 시행하는 것은 중요한 문제이다. 그러나 이제까지 제안된 로버스트 회귀추정량들은 계산상의 어려움과 정규오차모형에서 최소제곱추정량에 비하여 떨어지는 효율성때문에 통계적 추론의 정확성을 확신할 수 없었다. 최근 제안된 Lee(2004)의 가중자기조율회귀추정량(weighted self-tuning estimator, WSTE)은 다른 로버스트 회귀추정량에 비하여 정확한 계산과정과 그에 따른 추정량의 점근적 정규성 및 고붕괴점을 갖는다. 그러나 통계적 추론을 위하여 이제까지 널리 사용해왔던 로버스트 추정량에 기반한 가중최소제곱추정방법(weighted least squares estimator)은 WSTE에서조차 정규오차모형하에서 최소제곱추정량과 동일한 수준의 효율성을 제공해주지는 못한다. 본 논문에서는 WSTE에 기반한 또다른 통계적 추론 방법을 제안하고, 이 방법을 사용함으로써 정규오차모형 및 대표본에서 보다 정확한 결과를 얻을 수 있음을 몬테칼로 모의실험을 통해 제시하였다.

주요용어: 고붕괴점, 공동신뢰영역, 몬테칼로 모의실험, 로버스트 회귀추정, 신뢰구간에 대한 포함확률, 특이치

1. 서론

회귀모형(regression model)은 여러 분야에서 가장 널리 사용되는 통계분석 방법의 하나이며, 모수의 추론절차는 대부분 최소제곱추정량(least squares estimator, LSE)에 기반하고 있다. 반면 LSE는 특이치에 대하여 매우 민감하기 때문에 이를 보완하기 위한 다양한 추정방법들이 지난 30년 동안 제안되었으며, 특히 이들 가운데 로버스트 회귀추정 방법들은 고붕괴점 (high breakdown point)의 획득을 주요 목적으로 하고 있다. 붕괴점(breakdown point)이란 간단히 말하여 주어진 자료로부터 오염된 부분이 추정량을 왜곡시키기 시작하는 비율을 의미한다(Donoho and Huber, 1983). 이러한 의미에서 LSE는, 하나의 관찰값이 오염되어 있을 경우에도 영향을 받기 때문에 LSE의 붕괴점은 $1/n$, 즉 근사적으로 0이 된다. 특히 추정량의 붕괴점이 50%가 되는 경우, 즉 자료의 절반 가까이 오염되었을 경우에

1) (136-701) 서울시 성북구 안암동 5가 1, 고려대학교 통계학과, 연구조교수
E-mail: ld0351@korea.ac.kr
2) (136-701) 서울시 성북구 안암동 5가 1, 고려대학교 통계학과, 교수
E-mail: yspark@korea.ac.kr
3) (339-700) 충청남도 연기군 조치원읍 서창리 208, 고려대학교 자연과학대학 정보통계학과, 조교수
E-mail: korpen@korea.ac.kr

도 영향을 받지 않는 추정량들을 일컬어 고붕괴점 추정량(high breakdown estimator)이라 한다.

회귀모형에서 고붕괴점을 갖는 최초의 로버스트 추정방법은 Hampel(1975)에 의해 처음 제안되어 Rousseeuw(1984)에 의해 개발된 최소중위제곱추정량(least median of squares estimator, LMS)으로, 다음과 같이 정의된다.

$$\hat{\boldsymbol{\beta}}_{LMS} = \arg\min_{\boldsymbol{\beta}} \text{median } r_i^2(\boldsymbol{\beta}), \tag{1.1}$$

여기서 r_i는 회귀모형에 대한 잔차항이다. 그러나 LMS는 수렴률(rate of convergence)이 $n^{-1/3}$이라는 치명적인 약점을 가지고 있기 때문에, 정규오차모형하에서 LSE에 대한 상대효율(relative efficiency)이 0이 된다. LMS보다 높은 효율성을 얻기위하여 제안된 또다른 로버스트 회귀추정량은 최소절사제곱추정량(least trimmed squares estimator, LTS)이다. LTS는 다음과 같이 정의된다.

$$\hat{\boldsymbol{\beta}}_{LTS} = \arg\min_{\boldsymbol{\beta}} \sum_{l=1}^{h} \{r_i^2(\boldsymbol{\beta}) : i = 1, \cdots, n\}_{(l)}, \tag{1.2}$$

여기서 (l)은 l번째 순위통계량을 나타내는 지표이다.

Rousseeuw와 Yohai(1984)에 의해 제안된 S-추정량은 정칙조건(regularity condition)하에서 $\sqrt{n}$-일치성($\sqrt{n}$-consistency)에 도달한 최초의 고붕괴점을 갖는 로버스트 회귀추정량이다(Davis 1990). 이는 (1.1)과 (1.2)를 일반화한 것으로서 다음과 같이 정의된다.

$$\hat{\boldsymbol{\beta}}_S = \arg\min_{\boldsymbol{\beta}} S_n(\boldsymbol{\beta}), \tag{1.3}$$

여기서 $S_n(\boldsymbol{\beta})$은 척도모수(scale parameter)에 대한 추정량으로서 잔차 $r_i(\boldsymbol{\beta})$로 부터 얻어지며, 적절한 함수 ρ와 상수 K에 대하여 다음과 같이 정의된다.

$$S_n(\boldsymbol{\beta}) = \inf\{s > 0 : n^{-1} \sum_{i=1}^{n} \rho\left(\frac{r_i}{s}\right) = K\}.$$

S-추정량은 정규오차회귀모형에서 LMS와 LTS에 비하여 보다 높은 효율성을 갖는 것으로 알려져 있다. 그러나 S-추정량 역시 고붕괴점과 고효율성을 동시에 달성하지 못한다(Hössjer, 1992). 이와 같이 로버스트 회귀추정량들은 다양한 형태와 유형의 특이치로 오염된 자료들에 대하여 고붕괴점을 유지할 수는 있으나 그들의 낮은 효율성으로 인하여 신뢰구간의 구축이나 검정과정에서 정해진 명목수준이나 유의수준하에서 이에 미치지 못하는 결과를 제공한다. 이를 보완하기 위한 최근까지의 많은 시도들이 로버스트 추정분야의 중요한 부분을 이루어 왔으며, 이와 관련된 다양한 방법들이 제안되어 왔다. 그러나 이러한 연구들은 대부분 고붕괴점과 효율성간의 교섭(trade-off)으로 인하여 두 가지 모두에 대해서 만족할 만한 결과를 제공하는 것은 없는 실정이다(He and Portnoy, 2000). 더구나 로버스트 회귀추정량에 있어 제기되는 계산 상의 어려움이 이러한 효율성과 고붕괴점 간의 문

제에 있어서 어떻게 결부되는가의 문제는 아직 제대로 규명되지 못한 상태이다(Olive and Hawkins, 2002).

반면에 Lee(2004)가 제안한 가중자기조율회귀추정량 (weighted self-tuning estimator, WSTE)은 이제까지 제안된 로버스트 회귀추정량들과는 다른 방식으로 고붕괴점에 도달한 로버스트 회귀추정량이다. 특히 WSTE는 기존 로버스트 회귀추정 방법에서 드러난 계산상의 어려움을 극복함으로써 로버스트 추정량의 문제점으로 지적된 낮은 효율성의 문제를 동시에 보완하였다.

본 연구에서는 고붕괴점과 고효율성을 달성하기 위하여 가장 널리 사용되고 있는 방법 가운데 하나인 고붕괴점 회귀추정량에 기반한 가중최소제곱추정방법(weighted least squares, WLS)을 살펴보고, 몬테칼로 모의실험을 통해 이들을 평가해 보고자 한다. 이를 위하여 로버스트 회귀추정량의 회귀계수에 대한 신뢰구간(confidence intervals)과 공동신뢰영역(joint confidence region)의 정확성을 기준으로 살펴볼 것이다. 특히 WSTE 역시 다른 로버스트 회귀추정량과 같이 점근적 정규성을 가지고 있다는 것이 알려져 있으나 이에 대한 점근적 분산추정량(asymptotic covariance estimator)을 정확하게 계산하는 과정은 복잡함을 수반하며 말 그대로 점근적인 결과이기 때문에, 현실에서 사용하기 위해서는 다른 로버스트 추정량과 마찬가지로 이에 대한 적절한 수정 및 보완이 필요하다. 이를 위한 방법으로서 본 연구에서는 WLS방법을 보완한 재가중자기조율추정(reweighted self-tuning estimation, RWSTE)방법을 제안하였다. 이를 위하여 우선 2절에서는 Lee(2004)가 제안한 WSTE를 소개하고 이와 더불어 로버스트 회귀추정을 이용한 통계적 추론방법 및 RWSTE 방법을 살펴보도록 하며, 3절은 몬테칼로 모의실험 결과를, 그리고 마지막 4절은 요약과 결론이다.

2. 가중자기조율회귀추정을 이용한 통계적 추론

2.1. 가중자기조율회귀추정량

대부분의 고붕괴점을 갖는 회귀추정량은 잔차를 이용한 적절한 목적함수의 최적화를 통해 얻어지며, 이는 자료의 조합(combination)을 통해서만 얻어질 수 있기때문에 단순회귀모형과 같은 특수한 경우가 아닌 이상은 정확한 추정량을 얻는 것은 불가능하다. 반면 WSTE는 다음과 같은 자료분할을 기초로 하여 추정량이 계산된다. 우선 반응변수 y에 대한 관찰값을 통해 얻어진 1사분위수, 중위수, 3사분위수를 각각 $\tilde{y}_{q_1}, \tilde{y}_{q_2}, \tilde{y}_{q_3}$라 하자. 이를 이용하여 회귀모형을 위한 자료쌍 $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$을 다음과 같이 분할하도록 한다.

$$O_{01} = \{(\mathbf{x}_i, y_i) | \tilde{y}_{q_1} \le y_i < \tilde{y}_{q_2}\},\ O_{02} = \{(\mathbf{x}_i, y_i) | \tilde{y}_{q_2} < y_i \le \tilde{y}_{q_3}\}.$$

이와 더불어 j번째 독립변수에 대하여 다음과 같이 4개의 부분으로 분할하도록 한다.

$$O_{j1} = \{(\mathbf{x}_i, y_i) | x_{ji} \ge \bar{x}_j,\ y_j \ge \bar{y}_{U_i}\},\ O_{j2} = \{(\mathbf{x}_i, y_i) | x_{ji} \ge \bar{x}_j,\ y_j < \bar{y}_{U_i}\},$$
$$O_{j3} = \{(\mathbf{x}_i, y_i) | x_{ji} < \bar{x}_j,\ y_j \ge \bar{y}_{L_i}\},\ O_{j4} = \{(\mathbf{x}_i, y_i) | x_{ji} < \bar{x}_j,\ y_j < \bar{y}_{L_i}\}.$$

여기서

$$\bar{y}_{U_i} = \frac{\sum_{i=1}^{n} y_i I[x_{ji} \geq \bar{x}_i]}{\sum_{i=1}^{n} I[x_{ji} \geq \bar{x}_i]}, \quad \bar{y}_{L_i} = \frac{\sum_{i=1}^{n} y_i I[x_{ji} < \bar{x}_i]}{\sum_{i=1}^{n} I[x_{ji} < \bar{x}_i]}, \quad \bar{x}_j = n^{-1} \sum_{i=1}^{n} x_{ji}$$

이며, $I[\cdot]$는 지시함수(indicator function)이다. 이러한 기준으로 분할된 관찰값들로 이루어진 집합들에 대한 조합을 고려할 경우 p를 독립변수의 수라 하면 최대 $14p+4$개의 관찰값들의 집합을 구성할 수 있다. 이렇게 구성된 새로운 집합을 E_k $(k = 1, \cdots, K)$라고 정의하도록 한다. 이들 E_k를 이용하여 다음과 같은 계산과정을 통해 WSTE를 얻을 수 있다 (Lee, 2004).

단계 1 각 E_k를 구성하고 있는 관찰값만을 이용한 최소제곱추정량 $\mathbf{b}_k$를 구하고, 이를 이용하여 모든 관찰값들에 대한 표준화 잔차 $\tilde{r}_i(\mathbf{b}_k) = r_i(\mathbf{b}_k)/s(\mathbf{b}_k)$를 계산한다. 여기서 표준화를 위한 척도모수 추정량으로서 $s(\mathbf{b}_k) = (.6745)^{-1}\text{median}(|r_i(\mathbf{b}_k)|)$를 사용한다. 이들 $\mathbf{b}_k$를 이용하여 다음과 같은 추정량 $\tilde{\boldsymbol{\beta}}_n$을 얻는다.

$$\tilde{\boldsymbol{\beta}}_n = \arg\min_{\mathbf{b}_k} \sum_{|\tilde{r}_i(\mathbf{b}_k)|<c_1, |\tilde{r}_i(\mathbf{b}_k)|<c_1} \left[r_i^2(\mathbf{b}_k) - r_j^2(\mathbf{b}_k)\right]_+,$$

여기서 $[x]_+ = \max(0, x)$이며, c_1은 $\tilde{\boldsymbol{\beta}}_n$의 계산과정에서 특이치의 효과를 제한하기 위한 절사값으로 2.5와 3 사이의 값을 이용한다. 이와 더불어 $O(\tilde{\boldsymbol{\beta}}_n)$를 모든 관찰치 가운데 $|r_i(\tilde{\boldsymbol{\beta}}_n)/s(\tilde{\boldsymbol{\beta}}_n)| < c_1$을 만족하는 관찰치로 정의하도록 한다.

단계 2 앞에서 정의된 $\tilde{\boldsymbol{\beta}}_n$을 초기값으로 하여 다음을 만족하는 추정량을 계산한다.

$$\tilde{\boldsymbol{\beta}}_{n_1} = \arg\min_{\boldsymbol{\beta}} \sum_{(\mathbf{x}_i, y_i) \in O(\tilde{\boldsymbol{\beta}}_n)} \lambda(r_i) r_i$$

여기서 $\lambda(r_i)$는 다음과 같이 정의된다.

$$\lambda(r_i) = \left(1 + \frac{{n_1}^{1-R^2} \sum_{j=1}^{n_1} [r_i^2 - r_j^2]_+}{\sum_{j=1}^{n_1} \sum_{k=1}^{n_1} [r_j^2 - r_k^2]_+}\right)^{-1}, \; i = 1, \cdots, n_1, \tag{2.1}$$

단 n_1은 $O(\tilde{\boldsymbol{\beta}}_n)$에 포함된 관찰치의 수이며, R^2는 결정계수(the coefficient of determination)이다.

단계 3 $\tilde{\boldsymbol{\beta}}_{n_1}$를 이용하여 모든 관찰값에 대한 표준화잔차를 계산한 다음, 그 절대값이 c_2보다 크거나 같은 관찰값을 특이치로 간주하여 제거하도록 한다. 이 때 절사값 c_2는 일반적으로 c_1과 같은 값으로 결정하도록 한다.

단계 4 **단계 3**에서 제거된 관찰치만을 대상으로 하여 앞서 진행한 자료분할 과정 및 **단계 1**과 **단계 2**를 수행하여 최종 추정량 WSTE를 얻는다.

다른 로버스트 추정량들이 주어진 모든 관찰값들 가운데 일부분만으로 구성된 최적 관찰값들에 대한 탐색을 통해 얻어지는 것과는 달리, WSTE는 $\tilde{y}$, $\tilde{x}$와 같이 특이치에 민감한 표본평균값들을 기준으로 하여 전체 자료에 대한 분할을 사용함으로써 계산상의 어려움이 발생하지 않는다. 이와 더불어 WSTE는 정칙조건 하에서 점근적 정규성을 가지고 있다 (*Lee, 2004*). 그러나 이러한 장점에도 불구하고 WSTE는 표본평균값 및 결정계수 등을 이용함으로써 다른 로버스트 회귀추정량들이 가지고 있는 회귀 불변성(regression equivariance) 및 유사변환에 대한 불변성(affine equivariance) 등을 만족하지 못하기 때문에 다양한 형태의 특이치의 존재와 형태에 대한 로버스트성을 이론적으로 규명하기 어렵다. 그러나 Lee(2004)가 진행한 몬테칼로 모의실험 및 실제 자료에 대한 결과를 살펴봤을 때 WSTE는 현실적으로 다양한 특이치의 형태 및 유형, 그리고 오차 분포에 대하여 충분히 로버스트함을 알 수 있다.

2.2. 로버스트 회귀추정을 이용한 통계적 추론

앞에서 살펴봤듯이 고붕괴점을 갖는 로버스트 회귀추정량은 낮은 효율성과 계산상의 문제점으로 인하여 추정량을 직접 이용하여 통계적 추론에 사용하는 것이 거의 불가능하다. 따라서 로버스트 회귀추정량을 이용한 통계적 추론은 이들 고붕괴점을 갖는 추정량을 초기추정량으로 사용함으로써 이루어진다.

가장 널리 사용되는 일반적인 방법은 가중최소제곱추정방법(weighted least squares, WLS)이다. 이는 고정된 절사값 (cut-off value)을 기준으로 하여, 고붕괴점을 갖는 초기추정량을 이용한 표준화 잔차(standardized resudulas)의 절대값이 이보다 큰 관찰치만을 대상으로 하여 LSE를 재계산하는 과정으로써, Rousseeuw와 Leroy(1987)에 의하여 제안되었다. 그러나 이 방법은 특이치에 의한 변동성을 감소시킬 수는 있으나 LMS처럼 낮은 수렴률을 갖는 경우 WLS에 의한 추정량 역시 초기추정량의 수렴률과 동일하기때문에 효율성은 여전히 낮은 상태를 유지한다 (He and Portnoy, 1992). 그러나 S-추정량과 같이 $\sqrt{n}$-일치성을 갖는 추정량에 대해서는 Gervini와 Yohai(2004)에서 보여지듯이 효율성의 증가를 기대할 수 있다.

이외에 최적의 효율성과 고붕괴점을 M-함수 등을 통해 조율함으로써 이루어지는 MM-추정량(Yohai, 1987) 및 τ-추정량 (Yohai and Zamar, 1988) 등이 있다. 그러나 이들 추정량들은 효율성의 증가와 함께 편의 (bias)의 증가 역시 동반하기 때문에 결코 고붕괴점과 고효율성을 함께 이룰 수 없다. 이러한 사실은 최근 Salibian-Barrera와 Zamar(2004)에 의해 위치모형 (location model)에서의 현상이 규명되었고, 회귀모형에서는 Lee(2004)의 몬테칼로 실험 결과를 통해 나타나고 있다.

이와 더불어 일단계 GM-추정량(one-step Generalized M-estimators) 역시 고붕괴점을 갖는 추정량을 기초로 하여 효율성의 제고 및 지렛대 효과(leverage effect)에 대한 제약을 줌으로써 추정량의 고효율성과 로버스트성을 유지하며 동시에 지렛대 효과에 의한 특이치의 영향력을 제한하기 위하여 제안되었으나(Simpson *et al.*, 1992; Coakley and Hettamansperger, 1993), 이 역시 효율성과 고붕괴점 사이의 교섭(trade-off)이 발생한다. 이를 보완하기 위하여 일단계 추정량을 사용하나 독립변수의 수 및 오염수준에 따라 추정량의 편의 및 변동성이

점진적으로 증가하는 현상이 발생한다. Lee(2004)의 모의실험과 Simpson과 Yohai(1998)에 의하여 이러한 현상이 밝혀져 있다. 마찬가지로 고붕괴점 추정량에 기반하여 순위추정량(rank estimator)에 기반한 평활함수(smooth function)를 사용하는 고붕괴점순위회귀(high breakdown rank regression) 추정방법이 있으나 Chang et al.(1999)의 모의실험에서 나타나듯이 GM-추정량과 크게 다르지 않은 결과를 제공한다.

따라서 로버스트 회귀추정과 관련된 통계적 추론은 현재까지 WLS를 사용하는 것이 가장 일반적이다. 즉 회귀계수에 대한 로버스트 추정량과 척도모수에 대한 추정량을 각각 $\hat{\boldsymbol{\beta}}$와 $\hat{\sigma}$라 하면 표준화 잔차는 다음과 같이 정의된다. 즉

$$r_i^s = \frac{y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}}{\hat{\sigma}}$$

로서 만약 $|r_i^s|$가 크다면 특이치로 고려할 수 있다. 정규오차 모형하에서 $|r_i^s| \geq 2.5$인 관찰값을 특이치로 간주하는 것은 자연스런 일이다. 이러한 연장선상에서 Rousseeuw와 Leroy(1987)는 다음과 같은 방법을 제안하였다. 우선

$$w_i = \begin{cases} 1, \text{ if } |r_i^s| < c, \\ 0, \text{ otherwise} \end{cases} \tag{2.2}$$

를 계산하고, 다음과 같은 WLS를 추정한다. 일반적으로 절사값 c는 2.5로 고정된다.

$$\hat{\boldsymbol{\beta}}_{WLS} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y},$$

여기서 $\mathbf{W} = \text{diag}(w_1, \cdots, w_n)$이며, $\mathbf{X}$는 절편항을 포함한 $n \times (p+1)$ 계획행렬 (design matrix), $\mathbf{y} = (y_1, \cdots, y_n)'$ 이다. 이러한 단계를 이용함으로써 정규오차하에서의 효율성이 증진되며 동시에 초기추정량의 고붕괴점 역시 동시에 유지될 수 있는 것으로 알려져 있다. 이와 더불어 회귀계수에 대한 t-검정 등과 같이 LSE에 기반한 통계적 추론절차를 그대로 사용할 수 있다는 또다른 장점을 가지고 있다. 이러한 이유로 인하여 고붕괴점을 유지할 수 있는 로버스트 추론과정에서 가장 일반적으로 사용되는 방법이 바로 WLS이다.

2.3. 재가중자기조율회귀추정을 이용한 통계적 추론

WSTE 역시 점근적 정규성과 S-추정량과 같은 $\sqrt{n}$-일치성을 갖는 것으로 알려져 있다. 그러나 Gervini와 Yohai(2004)가 지적했듯이 이들 추정량에 기반한 WLS가 비록 다른 방법들과는 달리 초기 추정량의 고붕괴점을 유지하면서 동시에 효율성의 상승을 유도할 수는 있지만 만약 주어진 자료가 가정된 모형, 즉 정확히 정규오차를 갖는 회귀모형을 따른다면 절사값 밖에 표준화 잔차가 존재할 가능성은 조금이라도 존재하게 된다. 따라서 이러한 고정된 절사값에 의한 추정결과는 LSE와 동일한 수준의 효율성을 기대할 수 없게 된다. 반면 만약 이러한 가능성을 줄이기 위해 절사값 c를 증가시킨다면 다시 정규오차를 벗어나는 경우, 즉 특이치가 존재하게 되면 추정량의 편의를 증가시키게 된다. 뒤의 모의실험 결과에서 나타나듯이 WSTE 역시 다른 추정량들에 비하여 우수한 결과를 보여주지만 이러한

경향에서 벗어나는 것은 아니다. 이를 보완하기 위하여 본 논문에서는 WLS에서 사용하는 (2.2)를 다음과 같이 수정하여 사용하고자 한다.

$$w_i^* = \begin{cases} n_1 \dfrac{\lambda_i}{\sum_{i=1}^{n_1} \lambda_i}, \text{ if } |r_i^s| < c, \\ 0, \text{ otherwise,} \end{cases}$$

여기서 λ_i는 WSTE 계산과정에서 사용되는 (2.1)이며, n_1은 $|r_i^s| < c$를 만족하는 관찰치의 수를 의미한다. WSTE는 정칙조건하에서 점근적 정규성을 따르지만 회귀계수에 대한 점근적 공분산을 계산하는 것이 쉽지않기 때문에 추정량의 계산과정에서 이용된 (2.1)를 가중치로서 사용하는 것이다. 따라서 WLS 방법에서와는 달리 관찰치를 제거하고 난 나머지를 이용한 LSE 추정과정을 거치지 않고 WSTE 추정량을 직접 사용한다. 일반적으로 로버스트 추정량을 초기값으로 하여 M-추정량과 같은 평활함수를 이용한 가중치를 사용하는 것은 MM-추정량에서와 같이 효율성은 증가하지만 오염도의 증가에 따라 편의 역시 증가하기 때문에 WLS에서와 같이 절사값을 기준으로 하여 이보다 큰 값에 대해서는 0을 부여함으로써 이러한 현상을 방지할 수 있다.

이와 더불어 통계적 추론을 위한 오차항의 분산추정량으로서 Rousseeuw와 Leroy (1987)에서와 같이 다음과 같이 수정된 값을 사용한다. 즉

$$s_{RWSTE}^2 = \frac{1}{n_1 - p - 1} \sum_{i=1}^{n_1} w_i^* r_i^2$$

를 이용하여 신뢰구간 구축이나 검정과 같은 통계적 추론 과정에서 사용한다. 우리는 이러한 과정을 RWSTE(reweighted WSTE)라 부를 것이다.

3. 모의실험

본 절에서는 앞서 언급한 로버스트 회귀추정량들에 대한 통계적 추론결과의 정확성을 검토하기 위하여 신뢰구간 및 공동신뢰영역에 대한 모의실험을 수행하고 그 결과를 살펴보도록 한다. 앞에서 설명한 LMS, LTS, S, WSTE 등을 초기추정량으로 사용한 WLS를 이용한 방법과 본 논문에서 제안한 RWSTE를 비교해 보도록 한다. LMS, LTS, S 추정량의 계산은 모두 SAS/IML 및 SAS의 ROBUSTREG 프로시져를 사용하였으며, WSTE는 Lee(2004)에 의해 C++로 작성된 프로그램을 이용하였다. 그리고 모든 추정량의 계산에서 필요한 절사값 c는 Rousseeuw와 Leroy(1987)에서 제안한 2.5를 사용하였다.

모의실험에서는 다음과 같은 다중선형회귀모형을 고려하였다.

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + e_i, \tag{3.1}$$

여기서 $i = 1, \cdots, n$이고, 오차항 e_i는 $N(0,1)$을, 그리고 각 x_{ij}는 $N(7, 4^2)$을 따르도록 하였으며, 절편 β_0는 0으로 모든 회귀계수 $\beta_1, \cdots, \beta_p$는 1로 고정하여 사용하였다. 이와 더불어 각각 95%와 99% 명목수준에서 얻어지는 각 회귀모수에 대한 평균 신뢰구간과 공동신뢰영

역에 대하여 1000번 반복실험을 이용한 포함확률(coverage probability)을 계산하였다. 이를 위한 계산과정은 특이치를 제거한 후 나머지 자료를 이용한 LSE와 동일하다. 즉 (3.1)하에서

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{s^2((\mathbf{X}'\mathbf{X})^{-1})_{ii}}},\ i = 0, \cdots, p$$

는 자유도가 $n-p-1$인 t-분포를 따르며, 여기서 s^2는 오차항의 분산에 대한 LSE추정량이다. 이를 이용하여 명목수준 $(1-\alpha/2)\times 100\%$하에서의 β_i에 대한 신뢰구간을 다음과 같이 얻을 수 있다.

$$\left[\hat{\beta}_i - t_{n-p-1,\alpha/2}\sqrt{s^2((\mathbf{X}'\mathbf{X})^{-1})_{ii}}, \hat{\beta}_i + t_{n-p-1,\alpha/2}\sqrt{s^2((\mathbf{X}'\mathbf{X})^{-1})_{ii}}\right],\ i = 0, \cdots, p.$$

그러나 각 회귀계수에 대한 신뢰구간은 동시신뢰영역에 포함되지 않는 부분들을 포함하고 있기때문에 주어진 명목수준하에서 추정량의 정확성을 살펴보기위해서는 동시신뢰영역에 대한 포함확률을 고려해야 한다. 마찬가지로 $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)'$에 대한 공동신뢰영역(joint confidence region)을 얻기 위해서

$$\frac{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}{s^2(p+1)}$$

가 (3.1)하에서 $F_{p+1,n-p-1}$을 따른다는 사실을 이용함으로써 결국 명목수준 $(1-\alpha/2)\times 100\%$하에서의 공동신뢰영역은 다음과 같이 얻을 수 있다.

$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \le (p+1)s^2 F_\alpha(p+1, n-p-1).$$

3.1. 특이치가 없는 정규오차항 회귀모형

다음의 표 3.1과 표 3.2는 각각 95%, 99% 명목수준에 대하여 회귀계수에 대한 개별 신뢰구간의 포함확률의 평균 및 공동신뢰영역에 대한 포함확률을 나타낸 것이다. WLS의 초기 추정량으로 사용된 추정량들 가운데 WSTE를 이용한 경우가 가장 명목수준에 근접한 결과를 제공하고 있으며, RWSTE를 이용함으로써 이보다 10%정도 정확성이 향상된 결과를 보여주고 있다. 특히 표본의 수가 적고 독립변수의 수가 클수록 이러한 차이가 커지고 있으며, 표본의 수가 증가함에 따라 두 방법 모두 거의 비슷한 결과가 나타나고 있다. 특히 LTS에 비하여 수렴률이 떨어지는 것으로 알려진 LMS가 LTS보다 나은 결과를 보여주고 있으며, 공동신뢰영역에 대한 포함확률에 있어서는 LMS, LTS 모두 다른 추정량들에 비하여 더욱 부정확한 결과를 제공하고 있음을 알 수 있다.

3.2. 오염수준에 따른 정규오차항 회귀모형

오차항의 오염수준을 변화시켜가며 각 포함확률의 변화를 살펴보았다. 이를 위하여 식 (3.1)로부터 자료를 생성한 다음, 정해진 비율의 특이치를 생성하기 위하여 다음과 같은 방법을 사용하였다. 즉 각 i번째 독립변수에 대한 j번째 관찰값 x_{ij}는 i번째 변수에서의 관찰값의 평균 $\bar{x}_i$로부터 $5\sigma_x + N(0, \sigma_x^2)$만큼 거리를 두었으며, 종속변수에 대한 j번째 관찰값

표 3.1: 정규오차하에서의 회귀계수에 대한 신뢰구간의 평균 포함확률

명목수준	n	p	WLS의 초기추정량				RWSTE
			LMS	LTS	S	WSTE	
95%	60	5	.649	.496	.880	.905	.911
		10	.629	.229	.854	.884	.892
	100	5	.780	.682	.906	.922	.930
		10	.806	.448	.886	.914	.922
99%	60	5	.756	.600	.955	.966	.974
		10	.744	.298	.934	.951	.957
	100	5	.876	.795	.969	.972	.978
		10	.897	.552	.960	.972	.978

표 3.2: 정규오차하에서의 회귀계수에 대한 공동신뢰영역의 포함확률

명목수준	n	p	WLS의 초기추정량				RWSTE
			LMS	LTS	S	WSTE	
95%	60	5	.348	.173	.793	.852	.873
		10	.274	.001	.677	.763	.769
	100	5	.546	.387	.847	.896	.916
		10	.539	.067	.748	.840	.860
99%	60	5	.464	.259	.920	.943	.956
		10	.395	.006	.818	.863	.870
	100	5	.688	.508	.936	.960	.973
		10	.684	.097	.884	.930	.943

y_j는 $\bar{y}+N(0,1)$로부터 생성하여 사용하였다. 이는 큰 지렛값을 갖는 특이치를 생성하기 위한 것이다. Lee(2004)에 의해 이루어진 모의실험에 따르면 이 외의 여러 유형의 특이치 형태 및 분포에 대하여 각 추정량의 평균제곱오차(mean squared error)는 거의 동일한 결과를 제공하고 있기 때문에, 더 이상의 다양한 특이치 유형을 고려하여도 역시 이와 비슷한 결과를 얻게 될 것이다. 본 연구에서는 이러한 특이치에 의한 오염수준을 5%부터 45%에 이르기까지 5%씩 증가시켜가면 그 결과를 살펴보았다.

모의실험은 앞에서 설정된 표본크기와 독립변수의 수, 그리고 95%, 99% 명목수준에 따라 모두 진행하였으나 결과의 큰 차이가 없으므로 독립변수의 수가 10인 경우에서의 99% 명목수준에 대한 개별 회귀계수의 신뢰구간의 평균 포함확률과 독립변수의 수가 5인 경우의 공동신뢰영역의 95% 명목수준에 대한 포함확률만을 각각 표 3.3과 표 3.4에 나타내었다. 다른 추정량들에 비하여 WSTE를 이용한 두가지 방법과 S-추정량들이 LMS나 LTS에 비하여 주어진 명목수준에 가까운 포함확률을 보여주고 있다. 그러나 S-추정량은 오염수준 25%를 기점으로 하여 그 이후에서는 급격하게 명목수준으로부터 멀어지고 있음을 볼 수 있다. LMS나 LTS 역시 오염수준이 25%에 이르기까지 점진적으로 포함확률이 명목수준에 근접하고 있으나 그 이후부터는 정도의 차이는 있으나 떨어지고 있음을 알 수 있다. 이러한 경향은 독립변수와 표본크기의 비에 의존하는 것으로 보인다. 즉 독립변수의 수가 10인

경우는 오염수준 25% 근방에서 반면 독립변수의 수가 그보다 적은 5인 경우에서는 오염수준 30%까지 점진적으로 정확성이 증가하고 있다. 이와 달리 WSTE에 기반한 두 가지 방법은 모두 오염수준 45%까지 주어진 명목수준과 거의 일치된 수준을 유지하고 있다. 단지 RWSTE를 사용했을때 WLS 방법을 사용했을때보다 오염수준이 높은 경우는 약간 떨어지는 결과를 보여주고 있다. 특히 공동신뢰영역의 포함확률에서 표본크기가 작은 경우 이들 차이는 보다 크게 나타나고 있다.

표 3.3: 오염수준에 따른 99% 명목수준에서의 개별 회귀계수의 신뢰구간에 대한 평균 포함확률

p	n	오염수준	WLS의 초기추정량				RWSTE
			LMS	LTS	S	WSTE	
10	60	.05	.833	.366	.965	.959	.959
		.10	.895	.477	.978	.957	.956
		.15	.939	.610	.986	.951	.949
		.20	.969	.731	.991	.952	.950
		.25	.983	.822	.989	.954	.938
		.30	.989	.929	.534	.966	.944
		.35	.591	.247	.180	.970	.942
		.40	.410	.094	.121	.983	.949
		.45	.250	.007	.063	.991	.952
	100	.05	.931	.645	.970	.977	.982
		.10	.961	.741	.984	.982	.981
		.15	.972	.819	.986	.980	.978
		.20	.982	.897	.990	.981	.981
		.25	.989	.837	.991	.986	.983
		.30	.990	.961	.972	.988	.983
		.35	.941	.934	.031	.989	.981
		.40	.520	.017	.013	.990	.980
		.45	.053	.000	.003	.990	.974

3.3. 대표본에서의 모의실험

앞절에서의 모의실험 가운데 모든 추정량들이 비교적 다른 경우에 비하여 보다 정확한 결과가 나타난 표본크기가 100이고 독립변수의 수가 5인 경우를 표본크기 및 독립변수의 수의 비율을 유지한 채 각각 표본크기를 1000으로 독립변수의 수를 50으로 확장하여 각각 개별 회귀계수의 신뢰구간에 대한 평균 포함확률 및 공동신뢰영역에 대한 포함확률을 계산하였다. 표 3.5와 표 3.6에서 추정량으로 나타낸 LMS, LTS, S, WSTE는 모두 이들 추정방법에 의한 값을 초기 추정량으로 하여 WLS를 이용한 결과이며, 표 3.5는 평균 포함확률을 표 3.6은 공동신뢰영역에 대한 포함확률을 나타내고 있다. WSTE와 RWSTE를 제외한 다른 모든 추정량들의 포함확률이 15%를 전후로 하여 명목수준으로부터 멀어지고 있음을 확인할 수 있다. 그러나 이들 추정량들 간의 관계를 살펴봤을 때 일반적으로 수렴률이 가장

표 3.4: 오염수준에 따른 95% 명목수준에서의 공동신뢰영역에 대한 평균 포함확률

p	n	오염수준	WLS의 초기추정량				RWSTE
			LMS	LTS	S	WSTE	
5	60	.05	.493	.277	.879	.881	.879
		.10	.619	.418	.923	.891	.877
		.15	.691	.547	.943	.895	.871
		.20	.779	.639	.922	.889	.854
		.25	.833	.734	.949	.919	.872
		.30	.884	.832	.949	.930	.869
		.35	.901	.843	.000	.938	.867
		.40	.759	.720	.000	.943	.854
		.45	.074	.166	.000	.940	.807
	100	.05	.665	.514	.892	.909	.898
		.10	.750	.629	.928	.920	.911
		.15	.794	.715	.933	.910	.901
		.20	.855	.790	.946	.927	.910
		.25	.904	.841	.963	.950	.923
		.30	.906	.865	.954	.949	.916
		.35	.916	.878	.002	.948	.916
		.40	.847	.818	.000	.938	.896
		.45	.124	.266	.000	.944	.902

낮은 것으로 알려진 LMS가 가장 명목수준과 가까운 결과를 보여주고 있다. 이는 현실적으로는 WLS를 이용한 효율성의 제고가 LMS에서 가능함을 보여주고 있는 것이다. 본 실험에서 사용한 LTS는 Rousseeuw와 Van Driessen(2002)이 제안한 Fast-LTS 알고리즘을 사용한 것으로, LMS에서 사용된 재표본추출(resampling)방법보다 대표본에 있어서 추정량의 계산과정의 정확성 및 속도를 개선할 수 있는 것으로 알려져 있다. 그러나 실제로 LTS보다 LMS가 모든 경우에서 명목수준과 보다 가까운 결과를 제공하고 있으며, 부분적으로는 WSTE와 RWSTE에 거의 근접한 결과를 보여주고 있다. 이러한 사실은 이론적인 추정량의 성격과는 반대의 결과로 정확한 계산을 위해서는 이를 위한 알고리즘의 개발이 필요하다는 것을 반증하는 것이다. 반면에 WSTE 추정량을 이용한 두 가지 방법에서는 모두 개별 신뢰구간에서는 물론이고 공동신뢰영역에서조차 주어진 명목수준에 거의 근접하거나 동일한 결과를 제공하고 있는 것을 볼 수 있다.

4. 결론

고붕괴점을 갖는 로버스트 회귀추정량들은 목적함수에 대한 계산의 난해함으로 인하여 현실적인 상황에서 근사적인 해로서 대체된다. 특히 이들 로버스트 회귀추정량들은 이론적인 차원에서 점근적 정규성과 일치성 등이 규명되어 있으나 현실적인 근사적인 해를 찾기위한 계산과정에서 나타날 수 있는 현상들에 대해서는 많은 부분이 알려져 있지 않다. 본

표 3.5: $n = 1000$, $p = 50$인 경우의 로버스트 추정량의 개별 회귀계수의 신뢰구간에 대한 평균포함확률

	95% 명목수준					99% 명목수준				
오염수준	LMS	LTS	S	WSTE	RWSTE	LMS	LTS	S	WSTE	RWSTE
.00	.95	.78	.89	.92	.94	.99	.89	.95	.98	.99
.05	.95	.82	.93	.94	.94	.99	.92	.98	.98	.98
.10	.95	.85	.93	.93	.95	.99	.93	.98	.98	.99
.15	.64	.89	.88	.94	.94	.68	.95	.92	.99	.99
.20	.69	.01	.00	.95	.94	.73	.02	.01	.99	.99
.25	.67	.00	.00	.94	.94	.70	.01	.00	.99	.98
.30	.61	.00	.00	.95	.94	.64	.01	.00	.99	.99
.35	.49	.00	.00	.95	.95	.51	.00	.00	.99	.99
.40	.46	.00	.00	.95	.94	.48	.00	.00	.99	.99

표 3.6: $n = 1000$, $p = 50$인 경우의 로버스트 추정량의 공동신뢰영역에 대한 포함확률

	95% 명목수준					99% 명목수준				
오염수준	LMS	LTS	S	WSTE	RWSTE	LMS	LTS	S	WSTE	RWSTE
.00	.96	.04	.58	.74	.89	1.0	.09	.77	.89	.97
.05	.93	.14	.82	.84	.86	.99	.24	.94	.93	.95
.10	.97	.28	.89	.84	.93	1.0	.47	.94	.95	.98
.15	.62	.59	.87	.92	.92	.66	.76	.91	.98	.98
.20	.70	.01	.00	.95	.95	.72	.01	.00	.98	.99
.25	.65	.00	.00	.92	.90	.71	.00	.00	.98	.98
.30	.64	.00	.00	.92	.89	.64	.00	.00	.98	.98
.35	.50	.00	.00	.98	.97	.50	.00	.00	.99	.99
.40	.45	.00	.00	.96	.93	.47	.00	.00	.99	.99

논문에서는 이러한 문제를 보완하면서 이전의 로버스트 회귀추정량에 비하여 높은 효율성을 갖는 추정량으로 알려진 WSTE와 통계적 추론 과정에서 이를 보완하기 위하여 새롭게 제안한 RWSTE를 포함하여 이전에 제안된 로버스트 회귀추정량들에 대한 통계적 추론과 관련된 문제를 몬테칼로 모의실험에 의한 신뢰구간 및 공동신뢰영역을 통해 살펴 보았다. WSTE가 다른 로버스트 회귀추정량들에 비하여 우수하다는 사실은 Lee(2004)에 의해서 어느정도 밝혀진 바 있다. 본 연구를 통하여 통계적 추론 과정에서도 이러한 결과가 유효함을 확인할 수 있었지만, 오차정규모형과 대표본에서 WSTE에 기반한 WLS 방법은 주어진 명목수준과 일치된 결과를 제공하고 있지는 못하였다. 반면 본 연구에서 제안한 RWSTE는 기존에 널리 사용되어 왔던 WLS 방법에 비하여 명목수준과 일치된 결과를 정규오차모형과 대표본 자료에 대한 실험에서 보여주고 있음을 확인할 수 있었다.

참고문헌

Chang, W. H., Mckean, J. W., Naranjo, J. D. and Sheather, S. J. (1999). High-breakdown rank regression. *Journal of the American Statistical Association*, **94**, 205-219.

Coakley, C. W. and Hettamansperger, T. P. (1993). A bounded influence, high breakdown, efficiency regression estimator. *Journal of the American Statistical Association*, **88**, 872-880.

Croux, C., Rousseeuw, P. J., and Hössjer, O. (1994). Generalized S-estimators. *Journal of the American Statistical Association*, **89**, 1271-1281.

Davies, P. L. (1990). The asymptotics of S-estimators in the linear regression model. *The Annals of Statistics*, **18**, 1651-1675.

Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann*(Bickel, P. J., Doksum, K. A. and Hodges, J. L., eds.) 157-184. Wadsworth, Belmont, California.

Gervini, D. and Yohai, V. J. (2002). A class of robust and fully efficient regression estimators. *Annals of Statistics*, **30**, 583-616.

Hawkins, D. M. and Olive, D. J. (2002). Inconsistency of resampling algorithms for high breakdown regression estimators and a new algorithm. *Journal of the American Statistical Association*, **97**, 136-148.

He, X. and Portnoy, S. (1992). High breakdown point and high efficiency robust estimates for regression. *The Annals of Statistics*, **20**, 2161-2167.

Hössjer, O. (1994). Rank-based estimates in the linear model with high breakdown point.

Lee, Dong-Hee (2004). Self-tuning Robust Regression Estimator. Ph.D. Thesis, Korea University, Seoul.

Rousseeuw, P. J. (1984). Least median of squares. *Journal of the American Statistical Association*, **79**, 871-880.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection.* Wiley, New York.

Rousseeuw, P. J. and Van Driessen, K. (2002). Computing LTS regression for large data sets. *Estadistica*, **54**, 163-190.

Rousseeuw, P. J. and Yohai, V. J. (1984). "Robust regression by means of S-estimators" in *Robust and Nonlinear Time Series Analysis*, eds. J. Frank, W. Härdle, and R.D. Martin, Springer-Verlag, New York.

Salibian-Berrera, M. and Zamar, R. H. (2004). Uniform asymptotics for robust location estimator when the scale is unknown. *The Annals of Statistics*, **32**, 1434-1447.

Simpson, D. G., Ruppert, D. and Carroll, R. J. (1992). On one-step GM estimates and stability of inferences in linear regression. *Journal of the American Statistical Association*, **87**, 439-450.

Yohai, V. J. (1987). High breakdown point and high efficient robust estimates for regression. *The Annals of Statistics*, **15**, 642-656.

Yohai, V. J. and Zamar, R. H. (1988). High breakdown point estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association*, **83**, 406-414.

[2005년 4월 접수, 2005년 9월 채택]

On Confidence Intervals of Robust Regression Estimators

Dong-Hee Lee[1] YouSung Park[2] Kee Whan Kim[3]

ABSTRACT

Since it is well-established that even high quality data tend to contain outliers, one would expect far greater reliance on robust regression techniques than is actually observed. But most of all robust regression estimators suffers from the computational difficulties and the lower efficiency than the least squares under the normal error model. The weighted self-tuning estimator (WSTE) recently suggested by Lee (2004) has no more computational difficulty and it has the asymptotic normality and the high breakdown point simultaneously. Although it has better properties than the other robust estimators, WSTE does not have full efficiency under the normal error model through the weighted least squares which is widely used. This paper introduces a new approach as called the reweighted WSTE (RWSTE), whose scale estimator is adaptively estimated by the self-tuning constant. A Monte Carlo study shows that new approach has better behavior than the general weighted least squares method under the normal model and the large data.

Keywords: confidence interval, coverage probability, high breakdown point, joint confidence region, outliers, robust regression estimation

1) Research-Assistant Professor, Institute of Statistics, Korea University. 5-1 Anam-Dong, Sungbuk-gu Seoul 136-701, Korea
E-mail: ld0351@korea.ac.kr
2) Professor, Dept. of Statistics, Korea University. 5-1 Anam-Dong, Sungbuk-gu, Seoul 136-701, Korea
E-mail: yspark@korea.ac.kr
3) Associate Professor, Dept. of Informational Statistics, Korea University. 208 Seochang-Ri, Jochiwon-Eup, Yeonki-Gun, Chung-Nam, Korea
E-mail: korpen@korea.ac.kr