The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.3.471
29(3), 471–485

A comparison study of various robust regression estimators using simulation

Soohee Jang[a] · Jungyeon Yoon[b,1] · Heuiju Chun[a]

[a]Department of Statistics & Information, Dongduk Women's University;
[b]Korea Banking Institute

(Received February 15, 2016; Revised March 29, 2016; Accepted April 3, 2016)

Abstract

Least squares (LS) regression is a classic method for regression that is optimal under assumptions of regression and usual observations. However, the presence of unusual data in the LS method leads to seriously distorted estimates. Therefore, various robust estimation methods are proposed to circumvent the limitations of traditional LS regression. Among these, there are M-estimators based on maximum likelihood estimation (MLE), L-estimators based on linear combinations of order statistics and R-estimators based on a linear combinations of the ordered residuals. In this paper, robust regression estimators with high breakdown point and/or with high efficiency are compared under several simulated situations. The paper analyses and compares distributions of estimates as well as relative efficiencies calculated from mean squared errors (MSE) in the simulation study. We conclude that MM-estimators or GR-estimators are a good choice for the real data application.

Keywords: robust regression, breakdown point, M-estimation, L-estimation, R-estimation

1. 서론

회귀모형(regression model)의 대표적인 방법인 최소제곱법(least squares method; LS)은 가정과 이상치(outlier)에 매우 민감하기 때문에, 자료가 회귀모형의 가정을 만족하지 않을 경우 또는 이상치를 포함하는 경우에 왜곡된 추정 결과를 준다. 따라서 이상치에 민감한 LS의 단점을 보완하기 위해 이상치의 가중치를 줄여 이상치에 민감하지 않은 로버스트 추정법이 사용된다. 대표적인 로버스트 추정방법에는 Huber (1973)에 의해 제안된 MLE를 기반으로 한 M 추정량 계열, Siegel (1982)에 의해 제안된 순서형 통계량을 기반으로 한 L 추정량 계열, Jaekel (1972)에 의해 제안된 잔차의 순위(rank)를 기반으로 한 R 추정량 계열이 있다. M 추정량 계열은 Huber (1973)의 M 추정법을 이용한 추정량, Tukey의 이중가중(biweight) M 추정량 (Mosteller와 Tukey, 1977) 등이 있고, Yohai (1987)가 제안한 둘 이상의 M 추정 프로시저를 이용한 MM 추정량이 있다. S 추정량 계열은 Rousseeuw와 Yohai (1984)가 제안한 S 추정량과 S 추정량의 낮은 점근적 상대효율(asymptotic relative efficiency; ARE)을 보완한 generalized S(GS) 추정량 (Croux 등, 1994)이 있다. L 추정량 계열은 오차제곱의 중앙값을 최소화하

[1]Corresponding author: Korea Banking Institute, 118 Samchungro, Jongro-gu, Seoul 03053, Korea.
E-mail: juneyoon@kbi.or.kr

는 least median of squares(LMS) 추정량 (Rousseeuw, 1984)과 잔차가 큰 관측치의 α%를 절단하여 잔차제곱합을 최소화하는 least trimmed squares(LTS) 추정량 (Rousseeuw, 1985)이 있다. R 추정량 계열은 가중치함수(weight function)에 따라 Wilcoxon 점수를 사용한 R 추정량 (Jaeckel, 1972), Mallows 가중치를 사용한 generalized rank(GR) 추정량 (Naranjo와 Hettmansperger, 1994), high-breakdown rank(HBR) 가중치를 사용한 HBR 추정량 (Chang 등, 1999)이 있다.

본 논문에서는 이상치가 존재할 때 최소제곱 회귀추정량의 대안으로 사용되는 로버스트 추정방법들을 다양한 오염이 존재하는 경우를 상정하여 각 추정량들의 성능을 모의실험을 통해 비교하고자 하였다. 로버스트 회귀추정량을 비교하는 대표적인 척도인 붕괴점과 효율성을 기준으로 각 계열의 로버스트 회귀추정량들을 네 그룹으로 나누고 각 그룹에 속하는 대표적인 추정량들을 선정하였다. 붕괴점은 추정량의 편의가 무한이 되지 않도록 하는 최소 오염 비율을 말하며, 이 중 표본의 일부를 얼마나 교체하여 통계량의 값이 무한대가 되는지를 판단하는 붕괴점의 개념을 사용하였다. 즉,

$$\epsilon_n^*(T, Z) = \min_{1 \le m \le n} \left\{ \frac{m}{n} : \sup_{Z'} \|T(Z') - T(Z)\| = \infty \right\},$$

여기서 n은 자료의 수, m은 오염된 자료의 수, $T(Z)$는 n개의 자료 Z의 회귀추정량이고, $T(Z')$은 n개의 자료 중에서 m개가 오염된 자료 Z'의 회귀추정량이다. 본 논문에서의 효율성은 최소제곱 회귀추정량 대비 상대효율을 의미하며 회귀추정량들의 평균제곱오차(mean squared error; MSE)의 비율이다. 이들 척도를 기준으로 M 회귀추정량, MM 회귀추정량, LMS 회귀추정량과 LTS 회귀추정량, R 회귀추정량, 가중 Wilcoxon 추정량을 기반으로 한 GR 회귀추정량과 HBR 회귀추정량을 비교 대상으로 택하였다. 기존 연구에서 밝혀진 회귀추정량들을 시뮬레이션을 통해 비교하였다는 점에 있어서 Yu 등 (2014)의 연구와 동일선상에 있지만 직관적으로 이해할 수 있도록 오염 상황을 설정하였다. 또한 Yu 등 (2014)의 연구에서는 단순하게 추정치의 평균제곱오차만을 보았지만 본 연구는 추정치의 평균제곱오차를 변이와 분산으로 나누어 각 회귀추정량의 MSE 변화에 대한 설명을 하고자 했고 추정치의 분포를 boxplot을 통해 살펴보면서 좀 더 심도 있게 회귀추정방법의 성능을 비교하였다.

2절에서 연구 대상이 되는 로버스트 회귀추정량들에 대해 살펴보고, 3절에서는 이상치들이 존재하는 다양한 오염상황에서 모의실험을 통해 로버스트 회귀추정량들을 비교하였으며, 4절에서는 본 논문의 연구결과에 대한 요약과 결론을 제시한다.

2. 로버스트 회귀추정량의 종류

일반적인 회귀모형은 $y_i = x_i^\top \beta + \varepsilon_i$, $i = 1, \ldots, n$로, y_i는 i번째 관측치의 종속변수, β는 절편을 포함한 p차 모수 벡터, $x_i^\top$는 i번째 관측치의 독립변수 벡터, ε_i는 i번째 관측치의 오차이다. 다음은 M 추정량 계열의 Huber의 M 추정량과 MM 추정량, L 추정량 계열의 LMS와 LTS 추정량, R 추정량 계열의 R, GR, HBR 추정량들에 대한 설명이다.

2.1. M 회귀추정량

MLE를 기반으로 한 M 회귀추정량의 정의는 다음과 같다.

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \rho\left(\frac{y_i - x_i^\top \beta}{\hat{\sigma}} \right),$$

여기서 $\hat{\sigma}$은 척도모수의 M 추정량으로 일반적으로 중위절대편차(median absolute deviation; MAD)를 사용한다. $\rho(\cdot)$은 목적함수(objective function)이다. M 회귀추정량은 목적함수를 최소화시키는 방법

으로 목적함수를 β에 대하여 미분한 함수를 영향력함수 $\psi(\cdot)$라 한다. 예를 들면 목적함수가 $\rho(x) = (1/2)x^2$인 경우, 영향력함수는 $\psi(x) = x$가 되고, 가중치함수는 $\psi(x)/x = 1$이 되어 보통최소제곱법(ordinary least squares; OLS)이 된다 (Yu 등, 2014).

Huber가 제안한 목적함수는

$$\rho_H(u) = \begin{cases} \dfrac{u^2}{2}, & \text{for } |u| \le k, \\ k|u| - \dfrac{k^2}{2}, & \text{for } |u| > k \end{cases}$$

이고, 가중치함수는

$$\omega_H(u) = \begin{cases} 1, & \text{for } |u| \le k, \\ \dfrac{k}{|u|}, & \text{for } |u| > k \end{cases}$$

로 잔차의 크기가 일정 값 이상이 되면 감소한다. 이 추정량은 붕괴점이 $1/n$이며 영향력함수가 유한하지 않다. 또한 가우스-마코프 가정 하에서 $k = 1.345$일 때 95%의 상대효율을 가지고, 낮은 지레점(leverage)를 갖는 이상치에 로버스트하다 (Bellio와 Ventura, 2005).

2.2. MM 회귀추정량

둘 이상의 M 추정량 프로시저를 기반으로 한 MM 회귀추정량은 다음의 세 단계를 거쳐 얻을 수 있다.

단계 1: 높은 붕괴점을 가지는 초기 추정량을 $\tilde{\beta}$라 하고, 초기 추정량의 잔차를 $r_i(\tilde{\beta}) = y_i - x_i^\top \tilde{\beta}$라 한다. 이 때 초기 추정량이 높은 효율성을 가질 필요는 없다.

단계 2: 단계 1에서 얻은 잔차 $r_i(\tilde{\beta})$를 이용한 50%의 붕괴점을 갖는 척도모수의 M 추정량을 계산하고, $s_n = s(r_1(\tilde{\beta}), \ldots, r_n(\tilde{\beta}))$로 정의한다. 여기서 척도모수의 M 추정량을 구하기 위해 사용된 목적함수를 ρ_0로 정의한다.

단계 3: 다음 식의 해로 MM 회귀추정량 $\hat{\beta}$을 얻을 수 있다.

$$\sum_{i=1}^{n} x_{i,j} \psi_1 \left(\frac{y_i - x_i^\top \beta}{s_n} \right) = 0, \quad j = 1, \ldots, p.$$

ψ_1은 목적함수 ρ_1의 영향력함수로 $\psi_1(u) = \partial \rho_1(u)/\partial u$이다. 여기서 목적함수 ρ_1과 ρ_0가 같을 필요는 없지만 다음을 만족해야 한다.

i) ρ는 대칭이며 연속미분가능이어야 하고, $\rho(0) = 0$이다.

ii) ρ가 $[0, a]$구간에서는 순증가하고, $[a, \infty)$ 구간에서는 일정한 값을 가지게 하는 양의 a가 존재한다.

iii) $\rho_1(u) \le \rho_0(u)$. 즉,

$$\sum_{i=1}^{n} \rho_1 \left(\frac{y_i - x_i^\top \hat{\beta}}{s_n} \right) \le \sum_{i=1}^{n} \rho_1 \left(\frac{y_i - x_i^\top \tilde{\beta}}{s_n} \right)$$

을 만족해야 한다.

MM 추정량은 흔히 단계 1에서 S 추정량이 사용되고, 목적함수 ρ_0와 ρ_1로는 Tukey의 이중가중함수가 사용된다. Tukey가 제안한 이중가중 목적함수는

$$\rho_{BW}(u) = \begin{cases} \frac{k^2}{6}\left[1-\left[1-\left(\frac{u}{k}\right)^2\right]^3\right], & \text{for } |u| \le k, \\ \frac{k^2}{6}, & \text{for } |u| > k \end{cases}$$

로 Tukey의 M 추정량은 조율상수 $k = 4.685$에서 95%의 상대효율을 가지고, Tukey의 이중가중 목적함수를 사용한 S 추정량은 $k = 1.548$에서 50%의 붕괴점을 가진다. MM 회귀추정량의 붕괴점은 처음 두 단계의 조율상수(tuning constant) 선택에 따라 결정되고, 상대효율은 세 번째 단계에서 조율상수 선택에 따라 결정된다. 그러므로 M 회귀추정량과 달리 MM 회귀추정량의 붕괴점과 상대효율은 서로 독립적이다. 따라서, 붕괴점을 50%로 유지하면서 상대효율을 높일 수 있다.

2.3. LMS 회귀추정량

순서형 통계량의 선형결합을 기반으로 한 LMS 회귀추정량의 정의는 다음과 같다.

$$\hat{\beta} = \arg\min_{\beta} \text{Med}\left[\left(y_i - x_i^\top \beta\right)^2\right],$$

여기서 Med는 중앙값이다. LMS 추정량은 붕괴점이 $([n/2]-p+2)/n$(여기서 $[r]$은 r 이하의 최대 정수를 의미함)로 높다 하더라도 (Rousseeuw, 1984), 가질 수 있는 최대 상대효율은 오차의 정규분포 가정을 만족했을 때 37%로 낮고 (Rousseeuw와 Croux, 1993), 영향력함수가 잘 정의되지 않는 한계를 가지고 있다. 이러한 이유로 일반적으로 높은 붕괴점과 높은 효율을 갖는 로버스트 방법의 초기값으로만 사용된다.

2.4. LTS 회귀추정량

순서형 통계량의 선형결합을 기반으로 한 LTS 회귀추정량의 정의는 다음과 같다.

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{q} r_{(i)}(\beta)^2,$$

여기서 $r_{(i)}(\beta) = y_{(i)} - x_{(i)}^\top \beta$이고, $r_{(1)}(\beta)^2 \le \cdots \le r_{(q)}(\beta)^2$로 정의된다. $q = [n(1-\alpha)+1]$이며, α는 절단비율이다. 만약 $q = n/2+1$이라면 붕괴점은 50%이다 (Yu 등, 2014). LTS 추정량은 $q = [n/2]+[(p+1)/2]$일 때 최대 붕괴점이 $([(n-p)/2]+1)/n$으로 높고 (Rousseeuw, 1984), 이상치에 매우 강하지만 상대효율이 8%로 매우 낮다 (Stromberg 등, 2000). 이러한 이유로 일반적으로 높은 붕괴점과 높은 효율을 갖는 로버스트 방법의 초기값으로만 사용된다.

2.5. R 회귀추정량

잔차의 순위를 기반으로 한 R 회귀추정량의 정의는 다음과 같다.

$$\hat{\beta} = \arg\min_{\beta} D(\beta), \quad \text{여기서 } D(\beta) = \sum_{i<j} b_{ij}\left|e_i(\beta) - e_j(\beta)\right|. \tag{2.1}$$

Wilcoxon 점수 $b_{ij} = 1$을 사용하면

$$D(\beta) = 2\sum_{i=1}^{n}\left[R(e_i) - \frac{n+1}{2}\right]e_i,$$

여기서 $R(e_i)$는 잔차 e_i의 순위이다. Wilcoxon 점수를 사용한 R 추정량은 오차의 정규가정 하에서 95.5%의 상대효율을 가지고, 긴 꼬리 분포를 하는 경우에는 상대효율이 더 높다 (Hettmansperger 등, 1997). 붕괴점은 $1/n$이고, 높은 지레점을 갖는 값에 민감한 특징이 있다 (Naranjo와 Hettmansperger, 1994).

2.6. GR 회귀추정량

가중 Wilcoxon 추정량을 기반으로 한 GR 회귀추정량은 식 (2.1)에서 Mallows 가중치함수 $b_{ij} = w_i w_j$를 사용한다.

$$w_i = \min\left[1, \left[\frac{b}{(x_i-\hat{\mu})^\top S^{-1}(x_i-\hat{\mu})}\right]^{\frac{k}{2}}\right],$$

여기서 $\hat{\mu}$과 S는 각각 독립변수의 위치모수와 공분산의 로버스트 추정치이다. b는 절사점(cutoff point)으로 $\chi^2(p)$의 0.95 분위수가 흔히 사용된다. GR 추정량은 붕괴점이 k값과 p값에 따라 다르지만 33% 이하로 낮고, 영향력함수가 X축과 Y축에서 유한하며, 효율성이 X의 분포와 k의 영향을 받지만 일반적으로 정규모델을 가정했을 때 단순회귀에서 90–95%를 유지한다 (Naranjo와 Hettmansperger, 1994).

2.7. HBR 회귀추정량

가중 Wilcoxon 추정량을 기반으로 한 HBR 회귀추정량은 식 (2.1)에서 아래에서 정의된 HBR 가중치 함수를 사용한다.

$$b_{ij} = \psi\left[\left|\frac{cm_i m_j}{(e_i(\hat{\beta}_0)/\hat{\sigma})(e_j(\hat{\beta}_0)/\hat{\sigma})}\right|\right],$$

여기서

$$\begin{aligned}
c &= [\mathrm{Med}(a_i) + 3\mathrm{MAD}(a_i)]^2,\\
a_i &= \frac{e_i\left(\hat{\beta}_0\right)}{\hat{\sigma}m_i},\\
\hat{\sigma} &= \mathrm{MAD}\left(e_i\left(\hat{\beta}_0\right)\right),\\
\mathrm{MAD} &= 1.438\mathrm{Med}_i\left|e_i\left(\hat{\beta}_0\right) - \mathrm{Med}_j\left(e_j\left(\hat{\beta}_0\right)\right)\right|,\\
m_i &= \psi\left[\frac{b}{(x_i-\hat{\mu})^\top S^{-1}(x_i-\hat{\mu})}\right],
\end{aligned}$$

여기서 c는 조율상수, $\hat{\beta}_0$는 초기추정량, $e_i(\hat{\beta}_0)$는 초기추정량의 i번째 잔차이며, $\hat{\mu}$과 S는 각각 독립변수의 위치모수와 공분산의 로버스트 추정치이다. HBR 추정량은 LMS와 LTS와 같은 50% 붕괴점을 가지는 추정량을 초기값으로 사용하면 붕괴점을 50%까지 가질 수 있고, 영향력함수가 모든 방향에서 유한하다. 또한 R 추정량보다는 낮은 효율성을 갖지만, GR 추정량에서 손실되는 효율을 회복하는 특징이 있다.

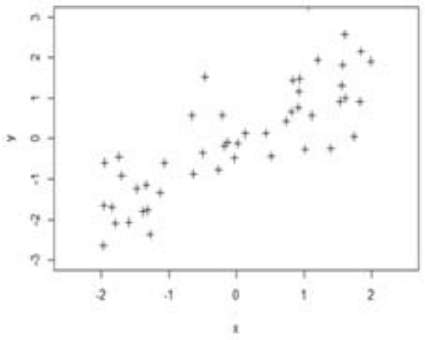

Figure 3.1. Non-contaminated data.

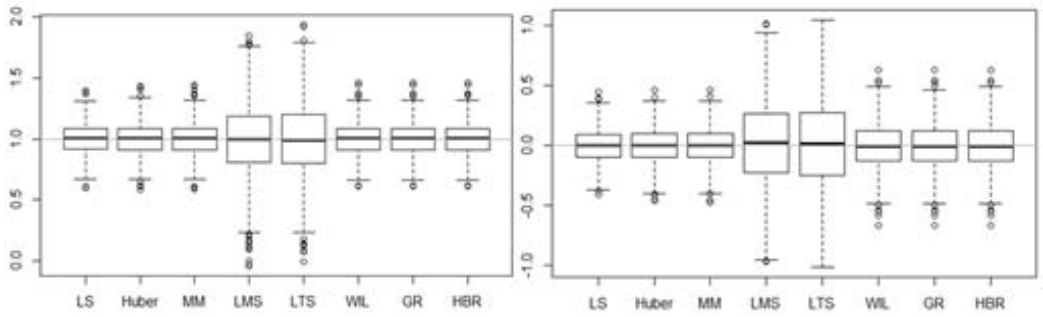

Figure 3.2. Box plot of regression coefficients(left) and intercepts(right) estimated with non-contaminated data.

3. 모의실험

본 절에서는 다양한 로버스트 회귀추정량(Huber-M, MM, LMS, LTS, R, GR, HBR)의 성능을 비로버스트 회귀추정량인 최소제곱추정량과 비교하고자 한다. 오염 상황에서 각 추정법의 성능을 두 가지 방법으로 비교하였는데, 첫 번째는 각 회귀추정량들의 MSE를 편의와 분산으로 나누어 살펴보고 LS의 MSE 대비 상대 MSE의 비율인 asymptotic relative efficiency(ARE)을 통해 효율성을 비교하는 것이고, 두 번째는 추정된 회귀계수의 분포를 비교하는 것이다. 모의실험에서 사용된 원자료의 모형은 $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $\varepsilon_i \sim N(0,1)$, $X \sim U(-2,2)$이고, $\beta_0 = 0, \beta_1 = 1$로 설정하고 크기가 50인 표본을 이용한 추정을 1,000회 반복하였다. 오염자료는 위치를 지정하기 위하여 이변량 정규분포(bivariate normal distribution; BVN)에서 생성했고, 오염자료의 X, Y의 분산은 1, X, Y 간 상관계수는 0으로 설정했다. 즉, $(X,Y) \sim \text{BVN}(\mu_1, \mu_2, \sigma_1^2 = 1, \sigma_2^2 = 1, \gamma = 0)$.

3.1. 비오염 자료

먼저 Figure 3.1은 이상치가 존재하지 않는 독립변수 X가 $(-2,2)$의 구간의 균일분포에서 생성된 비오염자료에 대한 그림이다. Figure 3.2는 비오염된 자료에서 추정된 회귀계수들과 절편들의 상자그림을 나타낸다. 수평선은 회귀계수의 참값이다. 여기서 Huber는 $k = 1.345$를 선택한 Huber의 M 추정량을 의미하고, MM은 초기 설정으로 S 추정량을 선택하고, 목적함수 ρ_0과 ρ_1로는 Tukey의 이중가중 함수($k_0 = 1.548, k_1 = 4.685$)를 선택한 MM 추정량을 의미한다. WIL은 Wilcoxon 점수를 사용한 R 추

Table 3.1. Variance, bias^2, MSE, ARE of coefficient and intercept estimates for non-contaminated data

	Coefficient				Intercept			
	Variance	bias^2	MSE	ARE	Variance	bias^2	MSE	ARE
LS	0.0153	0.0000	0.0153	1.0000	0.0197	0.0000	0.0197	1.0000
Huber	0.0162	0.0000	0.0162	0.9428	0.0207	0.0000	0.0207	0.9515
MM	0.0166	0.0000	0.0166	0.9221	0.0210	0.0000	0.0210	0.9368
LMS	0.0894	0.0001	0.0895	0.1708	0.1228	0.0001	0.1230	0.1600
LTS	0.0899	0.0001	0.0899	0.1700	0.1241	0.0001	0.1242	0.1584
WIL	0.0164	0.0000	0.0164	0.9343	0.0331	0.0000	0.0331	0.5939
GR	0.0163	0.0000	0.0163	0.9359	0.0332	0.0000	0.0332	0.5927
HBR	0.0164	0.0000	0.0164	0.9324	0.0331	0.0000	0.0331	0.5941

MSE = mean squared errors, ARE = asymptotic relative efficiency, LS = least squares, LMS = least median of squares, LMS = least median of squares, LTS = least trimmed squares, WIL = Wilcoxon rank, GR = generalized rank, HBR = high-breakdown rank.

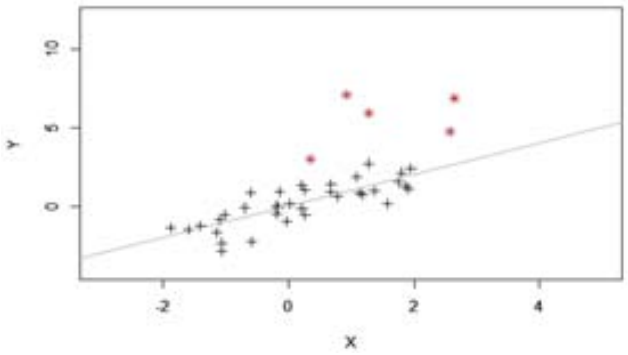

Figure 3.3. The contaminated data with $(X, Y) \sim \text{BVN}(1, 6, 1, 1, 0)$, $\alpha = 0.1$.

정량을 나타내며, GR은 $k = 2$를 선택한 GR 추정량을 의미하고, HBR은 초기 추정량으로 LMS를 사용한 HBR 추정량을 의미한다.

Table 3.1은 비오염 자료를 가지고 로버스트 회귀추정량들의 분산, 편의제곱, MSE, ARE를 최소제곱추정량(LS)과 비교한 표이다. Table 3.1에서 ARE가 모두 1보다 작아 효율성 측면에서 LS가 모든 로버스트 방법보다 우수한 것을 볼 수 있다. 또한 타 로버스트 추정방법들에 비해 LMS와 LTS의 분산이 크고 결과적으로 낮은 효율성을 보인다는 것을 알 수 있다.

3.2. X축 구간 내의 오염자료가 있는 경우

Figure 3.3은 이상치가 존재하도록 오염 자료를 $(X, Y) \sim \text{BVN}(1, 6, 1, 1, 0)$에서 생성하여 구성된 자료에 대한 그림이다. 검은색 점은 원자료, 빨간색 점은 오염자료이며 여기서 직선은 절편이 0이고 기울기가 1인 오염된 자료가 없을 때의 회귀식을 표현한 것이다. Figure 3.4는 오염된 자료의 비율을 10%, 20%, 30%로 주었을 때 추정된 회귀계수들과 절편들의 상자그림을 나타낸다. 오염비율이 늘어나더라도 LMS와 LTS의 편의는 타 방법들에 비해 0에 가깝게 유지하지만, 분산은 타 방법들에 비해 큼을 알 수 있다. Huber의 M 추정량은 낮은 지레점을 갖는 이상치에 로버스트한 특징이 있는 것으로 알려져 있으나 여기서는 타 추정량에 비해 높은 편의를 보이고 있다.

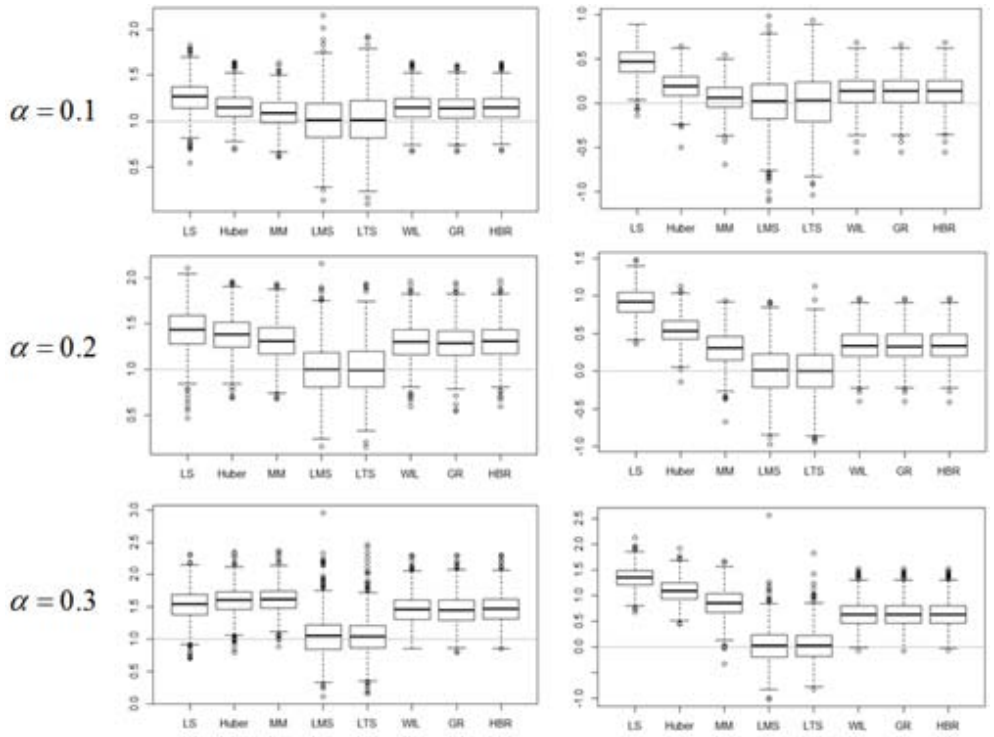

Figure 3.4. Box plot of regression coefficients(left) and intercepts(right) estimated with contaminated data with $(X, Y) \sim \text{BVN}(1, 6, 1, 1, 0)$.

Table 3.2는 Figure 3.3의 오염된 자료를 가지고 로버스트 회귀추정량들의 분산, 편의제곱, MSE, ARE를 최소제곱추정량과 비교한 표이다. 기울기의 추정량을 중심으로 오염비율이 30%로 높은 경우 Huber와 MM을 제외하면 모든 경우에서 로버스트 추정량의 ARE가 1보다 큰 것을 보아 LS에 비해 효율성이 높은 것을 볼 수 있다. 오염비율이 10%로 낮을 경우에는 MM 추정량은 ARE가 가장 높아 효율성은 가장 크지만, 오염비율이 늘어날수록 편의가 커짐으로 해서 그 효율성이 점차 떨어짐을 알 수 있다. 그러나 오염비율이 클수록 LMS와 LTS는 분산은 다른 추정량들에 비해 크게 나타나지만 편의제곱이 아주 작아 ARE가 가장 높게 나타나고 있다. 오염비율이 30%인 경우에 M 추정량과 MM 추정량의 ARE가 LS보다도 낮게 나타나는데, 이는 LS 방법의 절편 추정량은 크게 바뀌고 기울기 추정량에는 큰 변화를 주지 않아서 나타난 결과로 보인다. R 계열의 세 추정량은 비슷한 성능을 보이지만, 모든 오염 상황에서 GR이 가장 좋은 성능을 보인다.

3.3. X축 구간 외의 오염자료가 있는 경우

Figure 3.5는 이상치가 X축 구간 외에서 존재하도록 오염자료를 $(X, Y) \sim \text{BVN}(3, 8, 1, 1, 0)$에서 생성하여 구성된 자료에 대한 그림이다. Figure 3.6은 Figure 3.5의 오염된 자료의 비율을 10%, 20%, 30%로 주었을 때 추정된 회귀계수들과 절편들의 상자그림을 나타낸다. X축 구간 내의 오염인 Figure 3.4의 자료를 추정한 결과와 비교하였을 때 모든 추정방법의 편의가 증가하였고, LMS와 LTS의 경우는 분산이 더욱 증가하였다.

Table 3.2. Variance, bias^2, MSE, ARE of coefficient and intercept estimates for contaminated data with $(X, Y) \sim$ BVN$(1, 6, 1, 1, 0)$

		Coefficient				Intercept			
		Variance	bias^2	MSE	ARE	Variance	bias^2	MSE	ARE
$\alpha = 0.1$	LS	0.0318	0.0663	0.0981	1.0000	0.0273	0.2170	0.2443	1.0000
	Huber	0.0227	0.0233	0.0460	2.1323	0.0245	0.0372	0.0617	3.9621
	MM	0.0258	0.0084	0.0342	2.8659	0.0274	0.0042	0.0316	7.7323
	LMS	0.0803	0.0002	0.0804	1.2194	0.1019	0.0003	0.1022	2.3900
	LTS	0.0850	0.0003	0.0853	1.1501	0.1070	0.0002	0.1072	2.2794
	WIL	0.0231	0.0212	0.0443	2.2162	0.0337	0.0185	0.0522	4.6784
	GR	0.0228	0.0192	0.0419	2.3395	0.0336	0.0183	0.0519	4.7085
	HBR	0.0231	0.0215	0.0446	2.2001	0.0337	0.0185	0.0522	4.6805
$\alpha = 0.2$	LS	0.0520	0.1833	0.2352	1.0000	0.0349	0.8429	0.8778	1.0000
	Huber	0.0408	0.1455	0.1863	1.2624	0.0341	0.2951	0.3291	2.6672
	MM	0.0468	0.0980	0.1448	1.6250	0.0557	0.0986	0.1542	5.6913
	LMS	0.0850	0.0000	0.0850	2.7677	0.1024	0.0001	0.1025	8.5676
	LTS	0.0798	0.0001	0.0799	2.9442	0.1017	0.0000	0.1017	8.6306
	WIL	0.0395	0.0893	0.1289	1.8255	0.0442	0.1199	0.1641	5.3484
	GR	0.0398	0.0842	0.1240	1.8966	0.0442	0.1190	0.1632	5.3791
	HBR	0.0392	0.0947	0.1340	1.7560	0.0445	0.1212	0.1657	5.2969
$\alpha = 0.3$	LS	0.0618	0.2834	0.3452	1.0000	0.0456	1.8220	1.8676	1.0000
	Huber	0.0497	0.3543	0.4040	0.8545	0.0526	1.2066	1.2592	1.4832
	MM	0.0438	0.3808	0.4246	0.8130	0.0721	0.7221	0.7941	2.3517
	LMS	0.1016	0.0025	0.1041	3.3154	0.1170	0.0008	0.1179	15.8439
	LTS	0.0892	0.0027	0.0919	3.7563	0.1026	0.0010	0.1036	18.0269
	WIL	0.0563	0.2143	0.2706	1.2758	0.0715	0.4122	0.4837	3.8612
	GR	0.0590	0.2093	0.2683	1.2868	0.0716	0.4111	0.4827	3.8693
	HBR	0.0559	0.2208	0.2767	1.2475	0.0721	0.4160	0.0.4881	3.8259

MSE = mean squared errors, ARE = asymptotic relative efficiency, LS = least squares, LMS = least median of squares, LMS = least median of squares, LTS = least trimmed squares, WIL = Wilcoxon rank, GR = generalized rank, HBR = high-breakdown rank.

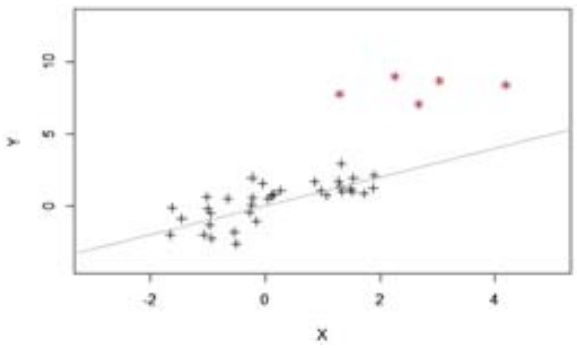

Figure 3.5. The contaminated data with $(X, Y) \sim$ BVN$(3, 8, 1, 1, 0)$, $\alpha = 0.1$.

Table 3.3은 Figure 3.5의 오염된 자료를 가지고 로버스트 회귀추정량들의 분산, 편의제곱, MSE, ARE를 최소제곱추정량과 비교한 표이다. 모든 경우에서 로버스트 추정량의 ARE가 1보다 큰 것을 보

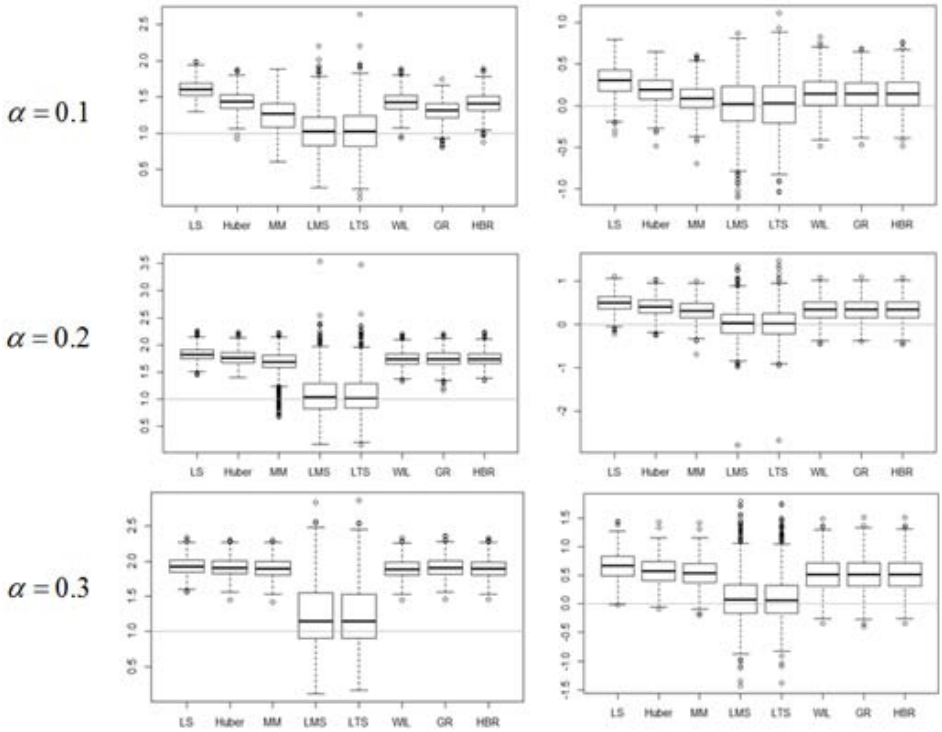

Figure 3.6. Box plot of regression coefficients(left) and intercepts(right) estimated with contaminated data with $(X, Y) \sim \text{BVN}(3, 8, 1, 1, 0)$.

아 LS에 비해 효율성이 높은 것을 볼 수 있다. 로버스트 추정방법 중에서는 오염비율이 높아지더라도 LMS와 LTS의 효율이 가장 높게 나타나는데 이는 추정량의 분산이 작은 타 방법들이 큰 편의를 보이기 때문이다. MM과 GR은 오염비율이 10%에서는 비교적 높은 ARE로 좋은 효율성을 가지지만 오염비율을 20%와 30%로 높였을 때는 편의가 커짐으로 효율성은 점점 떨어짐을 보이고 있다.

3.4. 원자료의 대각에 오염 자료가 있는 경우

다음은 오염된 자료가 원자료의 회귀직선과 대각선상에 존재하는 경우이다. Figure 3.7은 이상치가 회귀직선과 대각선상 위에 존재하도록 오염된 자료를 $(X, Y) \sim \text{BVN}(-2, 3, 1, 1, 0)$에서 생성하여 구성한 자료에 대한 그림이다. Figure 3.8은 Figure 3.7의 오염된 자료의 비율을 10%, 20%, 30%로 주었을 때 추정된 회귀계수들과 절편들의 상자그림을 나타낸다. Figure 3.8을 보면 LMS와 LTS를 제외한 추정방법들은 오염비율이 늘어날수록 편의가 급격히 커지는 것을 볼 수 있다.

Table 3.4는 Figure 3.7의 오염된 자료를 가지고 오염된 자료의 비율을 10%, 20%, 30%로 주었을 때 로버스트 회귀추정량들의 분산, 편의제곱, MSE, ARE를 최소제곱추정량(LS)과 비교한 표이다. Table 3.4를 보면 오염비율이 10%, 20%인 경우 로버스트 추정량의 ARE가 1보다 큰 것을 보아 LS에 비해 효율성이 높은 것을 볼 수 있다. 좀 더 자세히 보면 오염비율이 10%로 작은 경우에는 MM 방법이 ARE가 9.6674로 가장 효율적으로 나타나지만, 오염비율이 20% 이상인 경우에는 MM 추정량의 편의가 급격하게 증가하여 효율성은 떨어지는 것을 볼 수 있다. 반면에 LMS와 LTS는 오염비율이 늘어남에도 불구하

Table 3.3. Variance, bias^2, MSE, ARE of coefficient and intercept estimates for contaminated data with $(X, Y) \sim$ BVN$(3, 8, 1, 1, 0)$

		Coefficient				Intercept			
		Variance	bias^2	MSE	ARE	Variance	bias^2	MSE	ARE
$\alpha = 0.1$	LS	0.0147	0.3677	0.3824	1.0000	0.0343	0.0941	0.1285	1.0000
	Huber	0.0193	0.1924	0.2117	1.8062	0.0295	0.0371	0.0666	1.9274
	MM	0.0529	0.0608	0.1137	3.3630	0.0314	0.0078	0.0392	3.2797
	LMS	0.0891	0.0008	0.0898	4.2558	0.1041	0.0002	0.1043	1.2311
	LTS	0.0961	0.0013	0.0974	3.9242	0.1092	0.0003	0.1095	1.1732
	WIL	0.0194	0.1857	0.2051	1.8642	0.0420	0.0215	0.0635	2.0223
	GR	0.0211	0.0960	0.1170	3.2668	0.0374	0.0188	0.0562	2.2874
	HBR	0.0210	0.1691	0.1901	2.0114	0.0409	0.0209	0.0617	2.0811
$\alpha = 0.2$	LS	0.0161	0.6885	0.7046	1.0000	0.0457	0.2515	0.2973	1.0000
	Huber	0.0189	0.5840	0.6029	1.1687	0.0435	0.1669	0.2104	1.4128
	MM	0.0509	0.4480	0.4989	1.4125	0.0515	0.1034	0.1549	1.9192
	LMS	0.1440	0.0069	0.1509	4.6706	0.1255	0.0010	0.1266	2.3486
	LTS	0.1306	0.0062	0.1369	5.1485	0.1266	0.0006	0.1272	2.3373
	WIL	0.0208	0.5527	0.5735	1.2287	0.0669	0.1173	0.1841	1.6144
	GR	0.0226	0.5475	0.5701	1.2360	0.0673	0.1190	0.1863	1.5955
	HBR	0.0207	0.5615	0.5822	1.2104	0.0676	0.1167	0.1843	1.6134
$\alpha = 0.3$	LS	0.0169	0.8731	0.8900	1.0000	0.0584	0.4361	0.4945	1.0000
	Huber	0.0180	0.8282	0.8462	1.0518	0.0572	0.3247	0.3819	1.2948
	MM	0.0191	0.8064	0.8255	1.0781	0.0596	0.2830	0.3425	1.4436
	LMS	0.2071	0.0530	0.2601	3.4221	0.1817	0.0139	0.1955	2.5292
	LTS	0.1909	0.0507	0.2416	3.6840	0.1626	0.0117	0.1743	2.8376
	WIL	0.0196	0.7997	0.8192	1.0864	0.0853	0.2696	0.3548	1.3936
	GR	0.0190	0.8394	0.8583	1.0370	0.0889	0.2634	0.3523	1.4035
	HBR	0.0196	0.8066	0.8262	1.0773	0.0859	0.2680	0.3539	1.3971

MSE = mean squared errors, ARE = asymptotic relative efficiency, LS = least squares, LMS = least median of squares, LMS = least median of squares, LTS = least trimmed squares, WIL = Wilcoxon rank, GR = generalized rank, HBR = high-breakdown rank.

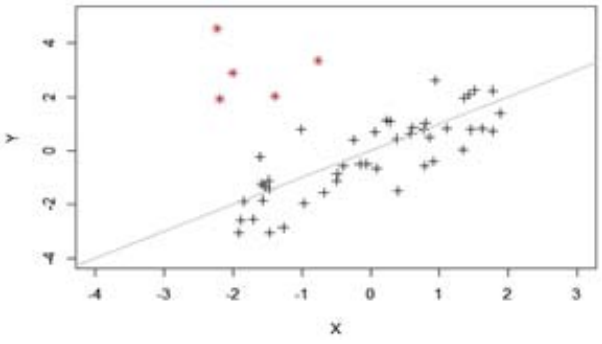

Figure 3.7. The contaminated data with $(X, Y) \sim$ BVN$(-2, 3, 1, 1, 0)$, $\alpha = 0.1$.

고 타 로버스트 방법에 비해 높은 효율성을 유지한다. 이는 LMS와 LTS는 원자료와 오염자료의 거리가 가깝고 30%의 데이터가 오염되어 실제 오염으로 구별하기 어려운 자료에도 기울기를 1에 가깝게 추정

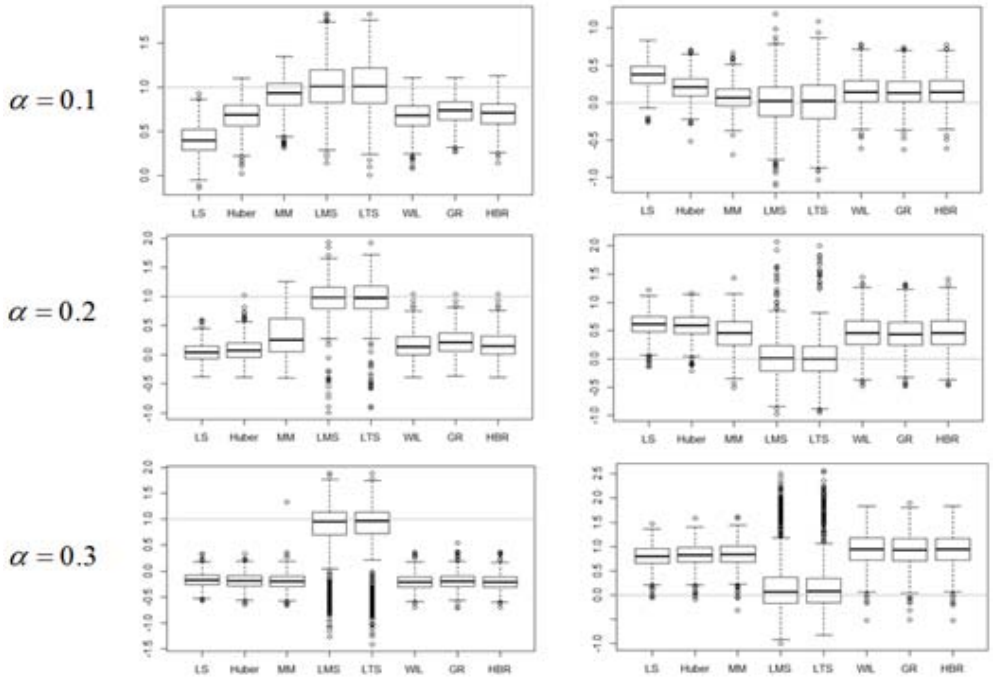

Figure 3.8. Box plot of regression coefficients(left) and intercepts(right) estimated with contaminated data with $(X, Y) \sim \text{BVN}(-2, 3, 1, 1, 0)$.

하고 있어 편의가 작기 때문이다.

4. 결론

본 연구의 모의실험에서는 실제 데이터에서 나타날 수 있는 다양한 형태의 오염을 고려하였다. 결과를 분석해보면 대부분의 경우 로버스트 회귀추정량이 LS 방법에 비해 우수한 성능을 가지는 것을 볼 수 있다. 낮은 효율성이 단점으로 부각된 LMS와 LTS는 이상치에 대해 저항력이 강하여 본 연구에서 시행된 모든 오염상황에서 일관적으로 가장 로버스트한 추정을 제공하며 편의제곱이 매우 작은 것을 볼 수 있었다. 결과적으로 효율성도 타 로버스트 방법에 비해 높은 것을 볼 수 있었다. 다만 추정량의 분산이 다른 추정량들에 비해 매우 크게 나타나므로 실제 데이터에 적용시에는 이를 주의할 필요가 있다. 높은 붕괴점과 효율성을 동시에 가지고 있다고 알려진 MM 추정량은 오염수준이 10%인 경우 아주 좋은 성능을 보였지만 오염비율이 높아질수록 편의제곱이 커지며 성능이 떨어지는 것을 볼 수 있었다. 한편, R계열 추정량을 살펴보면 HBR 추정량은 GR 추정량보다 높은 붕괴점을 가지며, GR 추정량에서 손실된 효율을 회복하는 추정량으로 알려져 있지만 (Chang 등, 1999), 본 연구의 모의실험에서는 R 계열의 세 추정량이 비슷한 성능을 보이며, GR 추정량이 대체로 R 추정량과 HBR 추정량보다 더 나은 성능을 보이는 것을 볼 수 있었다. 향후 연구는 위의 모의실험의 결과를 바탕으로 특이치가 존재하는 실 데이터에 본 연구에서 언급된 로버스트 회귀방법들을 적용하여 비교 연구하고자 한다. 단순하게 오염상황에 대한 추정량의 로버스트성을 높이는 것이 실 데이터 분석의 목적이 아니라면 로버스트 추정 방법 중 MM 추정방법이나 GR 추정방법이 적절할 것으로 보인다.

Table 3.4. Variance, bias^2, MSE, ARE of coefficient and intercept estimates for contaminated data with $(X, Y) \sim$ BVN$(-2, 3, 1, 1, 0)$

		Coefficient				Intercept			
		variance	bias^2	MSE	ARE	variance	bias^2	MSE	ARE
$\alpha = 0.1$	LS	0.0289	0.3592	0.3881	1.0000	0.0293	0.1398	0.1690	1.0000
	Huber	0.0290	0.1056	0.1346	2.8827	0.0275	0.0434	0.0710	2.3818
	MM	0.0325	0.0077	0.0401	9.6674	0.0302	0.0053	0.0355	4.7621
	LMS	0.0763	0.0000	0.0763	5.0877	0.1036	0.0002	0.1038	1.6292
	LTS	0.0803	0.0001	0.0803	4.8311	0.1071	0.0001	0.1072	1.5761
	WIL	0.0280	0.1051	0.1331	2.9170	0.0408	0.0237	0.0645	2.6217
	GR	0.0226	0.0726	0.0952	4.0784	0.0385	0.0215	0.0599	2.8214
	HBR	0.0258	0.0897	0.1155	3.3602	0.0396	0.0227	0.0624	2.7108
$\alpha = 0.2$	LS	0.0243	0.9163	0.9406	1.0000	0.0417	0.3752	0.4169	1.0000
	Huber	0.0350	0.8274	0.8625	1.0906	0.0452	0.3464	0.3915	1.0648
	MM	0.1365	0.4351	0.5716	1.6455	0.0796	0.2050	0.2846	1.4650
	LMS	0.1098	0.0013	0.1111	8.4646	0.1318	0.0005	0.1323	3.1507
	LTS	0.1124	0.0013	0.1137	8.2727	0.1401	0.0003	0.1405	2.9679
	WIL	0.0482	0.7169	0.7651	1.2294	0.0921	0.2231	0.3152	1.3226
	GR	0.0465	0.6057	0.6522	1.4422	0.0838	0.2041	0.2879	1.4480
	HBR	0.0493	0.6940	0.7433	1.2654	0.0909	0.2190	0.3099	1.3453
$\alpha = 0.3$	LS	0.0190	1.3659	1.3849	1.0000	0.0529	0.6367	0.6896	1.0000
	Huber	0.0207	1.4057	1.4264	0.9709	0.0587	0.6763	0.7351	0.9318
	MM	0.0253	1.4183	1.4436	0.9593	0.0633	0.7000	0.7632	0.9035
	LMS	0.3606	0.0526	0.4132	3.3513	0.3978	0.0509	0.4488	1.5367
	LTS	0.3502	0.0472	0.3974	3.4848	0.3946	0.0507	0.4453	1.5487
	WIL	0.0220	1.4547	1.4767	0.9378	0.1167	0.8658	0.9825	0.7019
	GR	0.0241	1.4197	1.4437	0.9593	0.1173	0.8617	0.9790	0.7044
	HBR	0.0223	1.4595	1.4818	0.9346	0.1168	0.8650	0.9818	0.7024

MSE = mean squared errors, ARE = asymptotic relative efficiency, LS = least squares, LMS = least median of squares, LMS = least median of squares, LTS = least trimmed squares, WIL = Wilcoxon rank, GR = generalized rank, HBR = high-breakdown rank.

References

Bellio, R. and Ventura, L. (2005). An introduction to robust estimation with R functions, In *Proceedings of the 1st International Workshop on Robust Statistics and R*, Treviso, Department of Statistics, Ca'Foscari University, Italy.

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.

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

Hettmansperger, T. P., McKean, J. W., and Sheather, S. J. (1997). Rank-based analyses of linear models, *Handbook of Statistics*, G.S. Maddala and C.R. Rao eds., Elsevier, 145–173.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo, *Annals of Statistics*, **1**, 799–821.

Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of residuals, *Annals of Mathematical Statistics*, **43**, 1449–1458.

Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression, A Second Course in Statistics*, Addison-Wesley, MA.

Naranjo, J. D. and Hettmansperger, T. P. (1994). Bounded-influence rank regression, *Journal of the Royal Statistical Society, Series B*, **56**, 209–220.

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

Rousseeuw, P. J. (1985). Multivariate Estimation with High Breakdown Point, 283-297 in Mathematical Statistics and Applications, Vol. B, edited by W. Grossman, G. Pflug, I. Vince, and W. Wetz. Dordrecht:Reidel.

Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the median absolute deviation, *Journal of the American Statistical Association*, **88**, 1273–1283.

Rousseeuw, P. J. and Yohai, V. (1984). Robust Regression by Means of S-Estimators, *Nonlinear Time Series Analysis*, Lecture Notes in Statistics, **26**, 256–272.

Siegel, A. F. (1982). Robust regression using repeated medians, *Biometrika*, **69**, 242–244.

Stromberg, A. J., Hossjer, O., and Hawkins, D. M. (2000). The least trimmed differences regression estimator and alternatives, *Journal of the American Statistical Association*, **95**, 853–864.

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

Yu, C., Yao, W., and Bai, X. (2014). Robust linear regression: a review and comparison, (Working Paper), Department of Statistics, Kansas State University, Manhattan, Kansas, USA 66506-0802.

시뮬레이션을 통한 다양한 로버스트 회귀추정량의 비교 연구

장수희[a] · 윤정연[b,1] · 전희주[a]

[a]동덕여자대학교 정보통계학과, [b]한국금융연수원

(2016년 2월 15일 접수, 2016년 3월 29일 수정, 2016년 4월 3일 채택)

요 약

회귀모형의 대표적인 추정법인 최소제곱법은 오차항의 분포가 정규분포를 따르고 이상치가 없는 상황에서는 최적이지만, 자료가 회귀모형의 가정을 만족하지 않을 경우 또는 이상치를 포함하는 경우와 같이 자료가 오염된 상황에서는 왜곡된 추정 결과를 준다. 따라서 이상치에 민감한 최소제곱법의 단점을 보완하기 위해 다양한 로버스트 추정방법이 제안되었다. 본 논문에서는 MLE를 기반으로 제안된 M 추정량, 순서형 통계량을 기반으로 제안된 L 추정량, 잔차의 순위를 기반으로 제안된 R 추정량 계열에서 높은 붕괴점 또는 높은 효율을 갖는 대표적인 추정량들을 다양한 모의실험을 통해 비교 연구하였다. 추정량의 성능을 비교하는데 효율성 뿐만 아니라 편의, 분산을 포함한 분포를 살펴보았다. 그 결과 실제 데이터 적용에는 MM 추정량과 GR 추정량이 좋은 성능을 가진 것으로 보였다.

주요용어: 로버스트, 로버스트 회귀, 붕괴점, M 추정, L 추정, R 추정

[1]교신저자: (03053) 서울시 종로구 삼청로 118, 한국금융연수원. E-mail: juneyoon@kbi.or.kr