응용통계연구 제 17권 3호, 2004년, pp. 475-488

다양한 오염 상황에서의 여러 로버스트 회귀추정량의 비교연구 *

김지연[1] 황진수[2] 김진경[3]

요 약

위치추정량에서 로버스트한 추정기법 중의 하나로 알려진 데이터 뎁스(depth)를 회귀추정에 적용한 회귀뎁스(regression depth)는 Rousseeuw and Hubert(1999)에 의하여 제안 되었다. 이 이외의 회귀뎁스 추정량으로는 심플리셜(simplicial) 뎁스와 사영(projection) 개념의 뎁스 회귀추정량들이 있다. 본 논문에서는 뎁스 기반 회귀추정량들의 성능에 대한 모의실험을 여러 오염 조건에서 행하여 비교하였으며 기존의 우수한 로버스트성을 지니는 추정량으로 최근에 제안된 HBR 추정량(Chang et al., 1999)들과의 비교연구도 하였다. 2차원 공간에서의 실험은 전반적으로 사영뎁스기반 회귀추정량이 좋은 결과를 보여주었다.

주요용어: 붕괴점, 영향력함수, 동등성, 회귀뎁스

1. 서론

회귀계수 추정으로 널리 알려진 최소제곱법은 오차항의 분포가 정규분포를 따르고 이상치가 없는 경우에 가장 최적의 방법이며 계산의 간편성 등으로 인하여 지금까지 널리 이용되고 있다. 그러나 오차항의 분포가 정규분포를 따르지 않거나 이상치가 존재하는 경우에는 최소제곱법이 최적의 결과를 주지 않으므로 여러 가지 로버스트한 회귀 추정방법을 사용하게된다. 대표적인 방법들로서는 전통적인 Huber(1973)의 M-추정법을 활용한 M-회귀추정량 계열을 생각할 수 있다. 또한 Hampel(1974)의 영향력함수(influence function)의 유한성을 고려한 회귀추정량들과 Jaeckel(1972)에 의해서 제안된 순위(rank)에 기반한 회귀추정량, 그리고 이를 발전시킨 HBR 추정량(Chang et al., 1999)등이 있다. 최근에 이르러서는 위치 추정량에서 로버스트한 추론방법으로 각광을 받고 있는 뎁스 기반 추정법을 회귀추정에 적용시킨 Rousseeuw and Hubert(1999)의 회귀뎁스추정과 심플리셜(simplicial) 뎁스 회귀추정, 그리고 사영(projection)뎁스의 개념을 이용한 회귀추정 등 여러 뎁스 기반 회귀추정방법이 제안되어 있다.

* 이 논문은 인하대학교 교내연구비로 지원 되었음 INHA-21962

1) (402-751)인천시 남구 용현동 253, 인하대 통계학과, 박사과정
E-mail: jeeyun@anova.inha.ac.kr
2) (402-751)인천시 남구 용현동 253, 인하대 통계학과, 교수
E-mail: jshwang@anova.inha.ac.kr
3) (402-751)인천시 남구 용현동 253, 인하대 통계학과, 교수
E-mail: jkkim@anova.inha.ac.kr

여러 추정량들의 로버스트성을 판정하는 척도로는 대표적으로 최대편의(maxmum bias)와 영향력함수(influence function), 그리고 붕괴점(breakdown point)등이 널리 쓰이고 있으며, 추정량이 만족해야하는 바람직한 성질로는 세가지의 동등성(equivariance)을 통상적으로 고려한다. 즉, 척도(scale)동등성, 회귀동등성, 그리고 유사(affine)동등성이 그것이다.

L_2(최소제곱)회귀추정량의 붕괴점 은 $1/n$로서 표본크기(n)가 커지면 그 값이 0 으로 수렴한다. 이러한 문제점은 L_1 회귀에서도 나타난다. 즉, 단 하나의 이상점이 추정량을 붕괴시킬 수 있게 된다. Huber(1973)의 M-추정량을 발전시킨 GM-추정량(Mallows, 1975)은 영향력함수 값은 유한하게 하였지만 독립변수 수가 증가함에 따라 붕괴점이 감소하는 문제는 해결되지 않았다. 최초로 1/2의 붕괴점을 가지는 추정량은 Siegel(1982)의 반복중위수 추정량(repeated median estimator)이지만, 이 추정량은 다차원에서 좌표축별로 계산한 것으로 동등성 성질을 만족하지 않으며 또한 변수의 차원(p)이 증가함에 따라 계산시간이 많이 증가하는($O(n^p)$) 단점을 가지고 있다. 붕괴점이 1/2인 추정량으로써 널리 쓰이기 시작한 것은 Rousseeuw(1984)등에 의하여 제안된 LMS(Least Median of Squares) 추정량이다. 이 추정량의 붕괴점은 $([n/2]-p+2)/n$이므로 n이 크면 최대값인 1/2에 수렴하게 된다. 이 추정량은 앞서 언급한 세가지 동등성은 만족하나 근사적 효율성(efficiency)이 좋지 않은 단점을 가지고 있으므로 이를 발전시켜 효율성을 향상시킨 추정량으로는 Rousseeuw(1984)의 LTS(Least Trimmed Square) 회귀추정량과 S-추정량(Rousseeuw and Yohai, 1984)등이 있다.

위에서 언급한 추정량과는 다른 계열의 추정량으로는 순위에 기반한 추정량이 있는데, Wilcoxon R-회귀추정량과 이를 발전시킨 GR-추정량 그리고 최근의 Chang et al.(1999)에 의해서 제안된 High Breakdown Rank(HBR) 회귀추정량등이 있다. 이 HBR 추정량은 붕괴점이 1/2이 되며 영향력함수도 유한하고 높은 효율성을 가지는 좋은 추정량이나, 추정량의 유도과정이 여러 단계를 거치는 단점이 있다.

Rousseeuw and Hubert(1999)가 제안한 회귀뎁스는 기존의 로버스트 회귀추정량과는 달리 모형 차원에서 비대칭 오차나 비등분산성(heteroscedacity)을 포함하는 새로운 차원의 로버스트한 추정량으로 알려져 있다. 뎁스의 개념은 위치추정량에서 자료의 상대적인 깊이를 나타내는 것으로서, 유클리디언 거리 뿐아니라 여러 형태의 뎁스 척도가 다차원에서 정의되어 있다(Zuo et al., 2000). 서로 다른 뎁스의 척도로는 현재 반공간(halfspace)뎁스, 심플리셜(simplicial)뎁스, 사영(projection)뎁스, 오자(Oja)뎁스 등이 있다. 위치 척도에서 여러 뎁스의 척도들 간의 부분적인 비교 연구는 Hwang et al.(2004)에 나와있다. 회귀뎁스는 각 회귀 직선들마다 뎁스값을 부여하여 순위를 정하는 것이다. 이중에서 뎁스값이 가장 큰 회귀 직선이 우리가 찾고자 하는 추정 회귀직선이다. 각 직선들마다 뎁스를 부여하는 방법에 따라서 여러 가지 뎁스 기반 회귀추정량이 있는데 우리가 고려하는 것은 Tukey가 제안한 반공간 뎁스를 이용한 회귀추정량과 Liu가 제안한 심플리셜 뎁스를 이용한 회귀추정량, 그리고 붕괴점을 1/2로 향상시킨 사영 뎁스 기반의 회귀추정량을 고려하고자 한다.

2 절에서는 뎁스 기반 회귀추정량과 HBR 추정량에 대한 정의와 성질에 대하여 소개하며 3 절에서는 여러 가지의 오염상황에서의 비교 모의실험 결과를 보여주고 끝으로 4절에서는 실험결과에 대한 분석과 토의를 하였다.

2. 로버스트 추정량의 종류와 성질

앞에서 소개한 바와 같이 수 많은 형태의 로버스트한 회귀추정량이 존재하나 본 연구에서 관심을 가지고 비교하려는 대상은 여러 측면에서 좋은 로버스트한 성질을 보이는 뎁스에 기반한 3가지 회귀추정량과 순위 기반한 HBR 추정량에 국한하였다.

2.1. 로버스트성 판정 척도

회귀 추정량들의 로버스트 정도를 판정하는 척도로는 다음의 대표적인 3 가지가 있다.

- Influence Curve or Function(영향력함수) :

$$IF(x, F, T) = \lim_{s \to 0} \frac{T((1-s)F + s\delta_x) - T(F)}{s}$$

여기서 F는 분포함수이고 $T(F)$는 회귀추정량을 나타내는 functional이다. 그리고 δ_x는 자료점 x에 전체의 매스가 집중 되어있는 함수이다. 따라서 이 영향력함수는 부분적으로 자료점 x가 추정량에 미치는 변화를 측정하는 것이다.

- Breakdown point(붕괴점) : 통계량의 붕괴점은 크게 3가지 방법으로 측정할 수 있는데 Donoho and Huber(1983) 에 의하여 제시된 유한표본에서 계산하는 방법과 "gross error neighborhood"를 이용한 모집단 분포에서의 계산방법과 적절한 metric을 사용하여 계산하는 방법 등이 있는데 본 논문에서 사용된 것은 유한표본의 붕괴점이다. 표본의 붕괴점은 표본의 일부를 얼마나 교체하여 통계량의 값이 무한대가 되는지를 파악하거나, 기존의 표본에 얼마의 새표본을 더하여 통계량이 무한대가 되는지를 판단하는 두가지가 있는데 본 논문에서의 붕괴점은 이중 전자인 교체붕괴점(finite sample replacement breakdown point)을 사용하였다.

  회귀추정량을 T라 하고 크기가 n인 전체 자료의 집합을 Z라 할때 n개의 자료 중 m개를 오염된 자료로 교체하였을 때 추정량의 편의가 무한대가 되는 최소 비율(m/n)로 붕괴점을 정의한다. 보통 이론적으로 얻을 수 있는 최대의 붕괴점은 50% 또는 1/2이 된다. 따라서 추정량 중에서 붕괴점이 1/2 에 가까운 것이 로버스트성이 좋은 통계량이 된다.

$$\epsilon_n^*(T, Z) = \min\left\{\frac{m}{n}; bias(m; T, Z) = \infty\right\}$$

- Maximum bias(최대편의) : 최대편의는 통계량이 붕괴점까지 이르지 않는 정도로 자료를 오염시켰을 때 발생할 수 있는 최대의 편의를 계산하는 것으로써 이러한 최대편의를 최소화하는 통계량이 좋은 통계량이라고 할 수 있다.

$$bias(m; T, Z) = \sup_{Z} \|T(Z') - T(Z)\|$$

여기서 $Z = \{(x_{11}, \ldots, x_{1p}, y_1), \ldots, (x_{n1}, \ldots, x_{np}, y_n)\}$로서 자료집합을 나타내고 Z'은 n개의 자료중 m개가 오염된(교체된) 자료집합을 나타낸다.

위의 세가지 판정기준을 보면 영향력함수는 국부적인 변화에 따른 추정량의 편의를 측정하는 것이고 붕괴점은 추정량의 편의가 무한히 커지는 정도를 측정하며 최대편의는 붕괴점까지 도달하기 전의 적당한 오염상황 에서의 최대편의를 측정한다. 각 판정기준들은 국부적 또는 전역적인 통계량의 성질을 파악하는 것이므로 세 가지의 기준에서 모두 제일 좋은 추정량 한가지를 찾는 것은 쉽지가 않다.

2.2. 로버스트 추정량의 성질

로버스트한 회귀추정량에서 보통 언급되는 동등성은 다음 3가지가 있다.

1. Scale equivariance :

$$T(\{(\mathbf{x}_i, cy_i); i = 1, \ldots, n\}) = cT(\{(\mathbf{x}_i, y_i); i = 1, \ldots, n\})$$

상수를 곱하여도 추정량이 유지가 되는 것으로서 측정 척도에 관계없이 회귀추정량이 일정하다는 것을 말한다.

2. Regression equivariance :

$$T(\{(\mathbf{x}_i, y_i + \mathbf{x_i v}); i = 1, \ldots, n\}) = T(\{(\mathbf{x}_i, y_i); i = 1, \ldots, n\}) + \mathbf{v}$$

회귀동등성은 위치(localtion) 동등성 또는 이동(shift)동등성으로도 불린다.

3. Affine equivariance :

$$T(\{(\mathbf{x_i A}, y_i); i = 1, \ldots, n\}) = \mathbf{A^{-1}} T(\{(\mathbf{x}_i, y_i); i = 1, \ldots, n\})$$

여기서 A는 임의의 $n \times n$ 비정칙행렬이다. 유사동등성은 회귀추정량이 설계행렬의 재배치에 대하여도 일관성을 유지하는 성질을 말한다.

2.3. 로버스트 회귀추정량

로버스트한 회귀추정량들 중에서 비교하고자 하는 주된 추정량은 뎁스를 기반으로한 추정량으로써 반공간 뎁스를 이용한 회귀추정량, 심플리셜 뎁스를 이용한 회귀추정량, 그리고 사영 뎁스의 개념을 이용한 회귀추정량이 있으며, 순위를 기반으로 한 것으로는 High Breakdown Rank(HBR) 회귀추정량이다.

2.3.1. 반공간 회귀뎁스

회귀모형에서 특정 회귀직선의 모수를 θ라 할 때 *nonfit*의 정의는 다음과 같다.

정의 2.1 자료의 집합을 Z_n이라 할 때 θ는 다음의 조건을 만족하면 *nonfit*이라 부른다.

$$r_i(\theta) < 0 \quad \forall x_i < \nu \quad \text{and} \quad r_i(\theta) > 0 \quad \forall x_i > \nu$$

또는

$$r_i(\theta) > 0 \quad \forall x_i < \nu \quad \text{and} \quad r_i(\theta) < 0 \quad \forall x_i > \nu$$

를 만족하는 어떤 자료값 x_i와도 일치하지 않는 실수 $\nu_\theta = \nu$가 존재한다.
여기서 $r_i(\theta)$는 i번째 관측자료의 잔차를 나타낸다고 하자.

정의 2.2 $rdepth(\theta, Z_n)$는 θ를 *nonfit*으로 만들기 위하여 Z_n 으로부터 제거해야 하는 최소의 관찰자료 수이다.

모든 θ에 대해서 $rdepth$의 최대값은 적어도 $n/3$보다 크다. 반공간 회귀뎁스에서 뎁스가 가장 큰 회귀추정량을 최대깊이 회귀(Deepest regression) 추정량이라고 하며 그 정의는 다음과 같다.

$$T_r^*(Z_n) = arg\max_\theta rdepth(\theta, Z_n).$$

최대깊이 회귀 추정량의 붕괴점은 일반적인 경우에 자료의 차원에 무관하게 최소한 1/3이 된다.

2.3.2. 심플리셜 회귀뎁스

심플리셜 뎁스를 이용한 회귀추정량에 대한 정의는 다음과 같다.

정의 2.3

$$rdepth^{(s)}(\theta, Z_n) = \binom{n}{p+1}^{-1} \sum_{i_1 < \cdots < i_{p+1}} I(\theta \in S(H_{i_1}, \ldots, H_{i_{p+1}}))$$

여기서 H_{i_1}은 관측치 Z_{i_1}에 해당하는 초평면(hyperplane)이고 S는 $p+1$개의 초평면으로 정의되는 심플렉스이다. 즉, 서로 다른 $p+1$개의 심플렉스 중에서 주어진 모수 θ를 포함하는 비율로써 심플리셜 회귀뎁스로 정의한다. 심플리셜 최대깊이 회귀추정량은 전과 마찬가지로 $T_r^{(s)}(Z_n) = arg\max_\theta rdepth^{(s)}(\theta, Z_n)$이며 붕괴점은 심플리셜 위치 뎁스의 붕괴점의 값인 $1/2^p$ 가 된다. 여기서 p는 독립변수의 수를 나타낸다. 따라서 다변수의 경우에는 좋은 추정량이 되지 않는다.

2.3.3. 사영회귀뎁스

반공간 회귀뎁스 추정량 $rdepth$는 잔차의 부호에 의하여 결정되는 통계량이다. 잔차의 부호 뿐 아니라 크기까지 고려한 추정량으로 사영뎁스 개념을 이용한 추정량이 Rousseeuw et al.(2001)에 의하여 제안되었다.

정의 2.4

$$rcent(\theta, Z_n) = \inf_{\substack{\|u\|\\ \nu \in R}} \frac{M_r}{\left(M_r + \left|med_i \frac{r_i(\theta)}{u^t x_i - \nu}\right|\right)}$$

단

$$M_r = \frac{med_i |y_i - med_j y_j|}{med_i |u^t x_i - \nu|}$$

사영회귀뎁스(rcent) 추정량은 중앙 위치추정량(lcent)과 같은 개념으로 형성되었으며 중앙위치 추정량은 Zuo et al.(2000)등에 의하여 사용되는 사영위치 추정량과 일치하며 그 붕괴점도 역시 1/2이 됨이 알려져 있다. 사영최대깊이 회귀추정량은 $T_r^c(Z_n) = arg\max rcent(\theta, Z_n)$이며 그 붕괴점은 1/2이며 모든 동등성을 만족한다.

2.3.4. HBR 회귀추정량

윌콕슨 순위에 기반을 두어 발전시킨 HBR 회귀추정량에 대한 정의는 다음과 같다.

정의 2.5

$$arg\min_{\beta} D(\beta), \quad \text{단} \quad D(\beta) = \sum\sum_{i<j} b_{ij}|e_i - e_j|$$

그리고

$$b_{ij} = \varphi\left[\left|\frac{cm_i cm_j}{(e_i(\hat{\beta}_0)/\hat{\sigma})(e_j(\hat{\beta}_0)/\hat{\sigma})}\right|\right], \qquad \hat{\beta}_0 = \text{초기 추정량}$$

초기추정량을 붕괴점이 좋은(1/2) LMS 추정량을 이용하여 얻어진 잔차를 $e_i(\hat{\beta}_0)$로 나타내며 나머지 상수와 함수들은 다음과 같다.

$$c = [med(a_i) + 3MAD(a_i)]^2, \quad a_i = e_i(\hat{\beta}_0)/(\hat{\sigma}m_i), \quad \hat{\sigma} = MAD(e_i(\hat{\beta}_0))$$

$$MAD = 1.438med_i\left|e_i(\hat{\beta}_0) - med_j\{e_j(\hat{\beta}_0)\}\right|$$

$$m_i = \varphi\left[b/(x_i - \hat{\mu})^t S^{-1}(x_i - \hat{\mu})\right], \quad \hat{\mu} \text{ 와 } \hat{S} \text{는 각각 중앙값과 MCD임.}$$

$$\varphi(t) = \begin{cases} 1 & t \geq 1 \\ t & -1 < t < 1 \\ -1 & t \leq -1 \end{cases}$$

이 HBR 추정량은 붕괴점이 1/2로 최적이며 영향력함수도 전체공간에서 유한하며 효율성도 좋은 추정량이며 계산시간도 비교적 빠르다. 그러나 붕괴점이나 효율성을 향상시키기 위하여 여러 단계를 거치게 되는 단점이 있다.

3. 모의 실험

모의실험에서는 앞절에서 언급한 뎁스 기반 추정량 세가지(R(H), R(S), Rcent) 와 순위기반 추정량 HBR, 그리고 전통적인 추정량인 L_1, L_2와의 성능을 여러 오염 상황에서 비교하여 보았다. 비교의 기준은 추정량의 안정성인 분산과 편의 그리고 MSE를 계산하여 보았다. 단순회귀인 2차원에서는 R(H), R(S), Rcent, HBR, L_1, L_2들간의 모든 비교를 하였으며, 3차원에서는 계산 알고리즘의 문제와 너무 다양한 오염 조건때문에 비교 대상을 축소하여 R(H), HBR, L_2의 성능만을 비교하였다.

모의실험은 크게 두가지 형태로 나누어 실시하였는데 그 하나는 오염원의 중심위치가 원자료의 중심위치와 다른 "비대칭오염"이고 다른 하나는 원자료와 오염원 자료의 중심

위치가 동일하고 분산이 다른 "대칭오염"을 고려하였다. 비대칭오염에서는 오염의 방향이 x축 방향인 경우와 y축 방향인 경우를 각각 고려하였으며 또한 원자료의 분산이 크고 작은 경우로 나누어서 보다 명확하게 원자료와 오염자료가 구분되는 경우와 그렇지 않은 경우를 나누어 실험하였다. 대칭오염에서는 오염자료의 분산과 오염 비율별로 실험을 하여 결과를 비교하여 보았다.

3.1. 비대칭오염

1. 2차원(단순회귀직선)

   원자료의 모형은 $Y_i = \beta_0 + \beta_1 x_i + e_i$, $e_i \sim N(0, \sigma^2)$ 이고 독립변수 $x_1, \ldots, x_n$ 는 $U(-1, 1)$ 에서 생성한다. (β_0, β_1 은 0과 1이다.) 오염자료는 평균이 μ_1 과 μ_2 이고 분산 공분산 행렬이 τI_2인 이변량정규분포에서 생성한다. ($\tau = 0.1$) 따라서 전체 생성자료는 $(1-\alpha)$ 만큼의 원자료와 α 만큼의 오염자료로 이루어졌다. 여기서 α는 오염비율을 나타낸다. 모의 실험 표본의 크기는 50이며 반복은 200번 행하였다. 각 표에서 L_1, L_2는 최소절대값회귀추정량과 최소제곱회귀추정량을 나타내고, R(H)는 반공간 뎁스를 이용한 회귀뎁스추정량, R(S)는 심플리셜 뎁스를 이용한 회귀뎁스추정량, 그리고 Rcent는 사영뎁스의 개념을 활용한 회귀뎁스추정량을 나타낸다.

   표 3.1의 경우는 원자료의 분산이 1 이고 오염자료는 원자료로부터 양의 x 축과 양의 y축(1사분면)으로 떨어져 있는 형태이다. 원자료에 대하여 x축 방향보다는 y 축방향으로 오염이 된 자료의 경우에 각 추정량들의 성능을 비교하여 보았다. 결과를 살펴보면 오염의 비율이 0.1, 0.2, 0.3 일때 Rcent 의 MSE와 편의^2 값이 다른 방법에 비해 작음을 보여준다.

   오염자료가 양의 x 축으로 떨어진 형태와 음의 x 축과 양의 y축(2사분면)으로 떨어진 형태에서는 오염의 비율이 증가하면서 각각 L_2와 R(S)의 MSE 값이 다른 방법에 비해 작아진다.

   표 3.2의 경우는 원자료의 분산이 0.01 이므로 상대적으로 오염자료와 원자료간의 구별이 좀더 명확한 경우라고 할 수 있다. 이 경우는 Rcent의 상대적인 우월성이 돋보이는 결과를 보여준다. 오염 비율이 0.1에서는 HBR이 좋은 결과를 보이지만 나머지 오염비율에서는 모두 Rcent 의 MSE와 편의^2 값이 다른 방법에 비해 훨씬 작음을 보여준다.

   표 3.3의 실험은 오염자료가 x축 방향에서 발생한 경우에 대한 모의실험이다. 이 경우도 역시 오염 비율이 0.1에서는 HBR이 제일 작지만 나머지 0.2, 0.3, 0.4 에서는 Rcent 의 MSE와 편의^2 값이 다른 방법에 비해 훨씬 작음을 보여준다. 또한 오염자료가 $y = x$ 축 주변으로 1사분면에 떨어진 형태에서도 같은 결과를 얻을 수 있었다. 표 3.2, 표 3.3 에서 HBR이 오염비율 0.1에서 제일 작지만 Rcent와의 차이가 아주 작음을 알 수 있다. 그러나 오염비율이 0.2, 0.3, 0.4 일 경우는 다른 방법들과의 차이가 많은것을 볼 수 있다. 또한 뎁스를 기반한 다른 추정량, 즉 R(H), R(S)도 이경우엔 다른 추정량들에 비해 MSE 값이 작음을 알 수 있다.

표 3.1: $\sigma^2 = 1$, $(\mu_1, \mu_2)^t = (1, 5)^t$

		Slope			Intercept		
		분산	편의^2	MSE	분산	편의^2	MSE
$\alpha = 0.1$	L_1	0.1137	0.1946	0.3083	0.0421	0.0173	0.0594
	L_2	0.0490	0.7768	0.8258	0.0298	0.0973	0.1271
	$R(H)$	0.0987	0.1134	0.2121	0.0447	0.0007	0.0454
	$R(S)$	0.1137	0.1126	0.2263	0.0495	0.0007	0.0502
	$Rcent$	0.1020	0.0985	0.2005	0.0438	0.0032	0.0470
	HBR	0.0801	0.3626	0.4427	0.0472	0.0204	0.0676
$\alpha = 0.2$	L_1	0.1749	1.4107	1.5856	0.0760	0.1357	0.2117
	L_2	0.0512	2.0598	2.1110	0.0381	0.2333	0.2714
	$R(H)$	0.1255	0.7370	0.8625	0.0603	0.0106	0.0709
	$R(S)$	0.1367	0.6956	0.8323	0.0678	0.0063	0.0741
	$Rcent$	0.1360	0.6050	0.7410	0.0637	0.0317	0.0954
	HBR	0.0772	1.8988	1.9760	0.0793	0.1476	0.2269
$\alpha = 0.3$	L_1	0.1120	3.6084	3.7204	0.0968	0.5416	0.6384
	L_2	0.0435	2.9759	3.0194	0.0473	0.4642	0.5115
	$R(H)$	0.2177	2.9568	3.1745	0.1170	0.0983	0.2153
	$R(S)$	0.2393	2.6064	2.8457	0.1385	0.0622	0.2007
	$Rcent$	0.3361	2.0346	2.3707	0.1421	0.1853	0.3274
	HBR	0.0498	3.2114	3.2612	0.0944	0.5633	0.6577
$\alpha = 0.4$	L_1	0.0841	4.5380	4.6221	0.0973	0.8877	0.9850
	L_2	0.0484	3.7659	3.8143	0.0504	0.6315	0.6819
	$R(H)$	0.1110	5.5539	5.6649	0.1401	0.4546	0.5947
	$R(S)$	0.1357	5.4539	5.5896	0.2027	0.4100	0.6127
	$Rcent$	0.1131	4.6715	4.7846	0.1655	0.6051	0.7706
	HBR	0.0517	4.0519	4.1036	0.0863	0.9657	1.0520

표 3.2: $\sigma^2 = 0.01$, $(\mu_1, \mu_2)^t = (1, 5)^t$

		Slope			Intercept		
		분산	편의^2	MSE	분산	편의^2	MSE
$\alpha = 0.1$	L_1	0.0011	0.0019	0.0030	0.0004	0.0002	0.0006
	L_2	0.0118	0.7908	0.8026	0.0066	0.0978	0.1044
	$R(H)$	0.0010	0.0011	0.0021	0.0004	0.0000	0.0004
	$R(S)$	0.0012	0.0011	0.0023	0.0005	0.0000	0.0005
	$Rcent$	0.0010	0.0010	0.0020	0.0004	0.0000	0.0004
	HBR	0.0008	0.0010	0.0018	0.0004	0.0001	0.0005
$\alpha = 0.2$	L_1	0.0028	0.0155	0.0183	0.0010	0.0016	0.0026
	L_2	0.0222	2.0270	2.0492	0.0164	0.2572	0.2736
	$R(H)$	0.0013	0.0074	0.0087	0.0006	0.0001	0.0007
	$R(S)$	0.0016	0.0071	0.0087	0.0006	0.0001	0.0007
	$Rcent$	0.0014	0.0060	0.0074	0.0006	0.0003	0.0009
	HBR	0.0502	0.0623	0.1125	0.0098	0.0043	0.0141
$\alpha = 0.3$	L_1	0.4605	1.8533	2.3138	0.1057	0.2954	0.4011
	L_2	0.0232	3.0699	3.0931	0.0217	0.4534	0.4751
	$R(H)$	0.0043	0.0342	0.0385	0.0013	0.0010	0.0023
	$R(S)$	0.0044	0.0296	0.0340	0.0013	0.0006	0.0019
	$Rcent$	0.0040	0.0141	0.0181	0.0010	0.0019	0.0029
	HBR	0.1526	2.6302	2.7828	0.0733	0.3792	0.4525
$\alpha = 0.4$	L_1	0.0623	4.0581	4.1204	0.0660	0.8627	0.9287
	L_2	0.0289	3.7254	3.7543	0.0264	0.6608	0.6872
	$R(H)$	0.1821	4.0228	4.2049	0.0876	0.2023	0.2899
	$R(S)$	0.8083	1.5435	2.3518	0.1209	0.0761	0.1970
	$Rcent$	0.0436	0.0245	0.0681	0.0200	0.0069	0.0269
	HBR	0.0340	3.9613	3.9953	0.0580	0.8980	0.9560

표 3.3: $\sigma^2 = 0.01$, $(\mu_1, \mu_2)^t = (4, 0)^t$

		Slope			Intercept		
		분산	편의^2	MSE	분산	편의^2	MSE
$\alpha = 0.1$	L_1	0.0709	0.2089	0.2798	0.0072	0.0033	0.0105
	L_2	0.0013	0.6894	0.6907	0.0055	0.0047	0.0102
	$R(H)$	0.0011	0.0014	0.0025	0.0004	0.0000	0.0004
	$R(S)$	0.0013	0.0016	0.0029	0.0004	0.0000	0.0004
	$Rcent$	0.0013	0.0012	0.0025	0.0004	0.0000	0.0004
	HBR	0.0008	0.0004	0.0012	0.0004	0.0002	0.0006
$\alpha = 0.2$	L_1	0.0023	0.8537	0.8560	0.0186	0.0120	0.0306
	L_2	0.0010	0.8260	0.8270	0.0070	0.0050	0.0120
	$R(H)$	0.0020	0.0084	0.0104	0.0006	0.0000	0.0006
	$R(S)$	0.0022	0.0072	0.0094	0.0006	0.0000	0.0006
	$Rcent$	0.0031	0.0040	0.0071	0.0005	0.0000	0.0005
	HBR	0.0013	0.8216	0.8229	0.0159	0.0179	0.0338
$\alpha = 0.3$	L_1	0.0017	0.8940	0.8957	0.0175	0.0129	0.0304
	L_2	0.0009	0.8829	0.8838	0.0080	0.0050	0.0130
	$R(H)$	0.0260	0.0757	0.1017	0.0026	0.0000	0.0026
	$R(S)$	0.0048	0.0356	0.0404	0.0013	0.0000	0.0013
	$Rcent$	0.0041	0.0053	0.0094	0.0006	0.0000	0.0006
	HBR	0.0012	0.8772	0.8784	0.0154	0.0173	0.0327
$\alpha = 0.4$	L_1	0.0020	0.9177	0.9197	0.0253	0.0124	0.0377
	L_2	0.0008	0.9104	0.9112	0.0098	0.0060	0.0158
	$R(H)$	0.0019	0.8669	0.8688	0.0243	0.0000	0.0243
	$R(S)$	0.0023	0.8867	0.8890	0.0304	0.0008	0.0312
	$Rcent$	0.1945	0.1962	0.3907	0.0339	0.0030	0.0369
	HBR	0.0011	0.9026	0.9037	0.0190	0.0164	0.0354

2. 3차원

원자료의 모형은 $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$, $e_i \sim N(0,1)$ 으로서 두 개의 독립변수의 퍼진 정도에 따라서 실험을 하였다.

표 3.4: $x_{1i} \sim U(-2,2)$, $x_{2i} \sim U(-0.5,0.5)$, (단 , $\beta_0 = 0, \beta_1 = 1, \beta_2 = 1$)

	Slope1		Slope2		Intercept	
	분산	편의^2	분산	편의^2	분산	편의^2
L_2	0.0168	0.0000	0.2821	0.0007	0.0235	0.0000
$R(H)$	0.0431	0.0014	0.0648	0.5587	0.0000	0.0000
HBR	0.0179	0.0000	0.2982	0.0005	0.0319	0.0000

표 3.4의 경우는 L_2 가 Slope1, Slope2 모두 분산 값이 작다. 그러나 x_{1i} 를 $U(-0.5,0.5)$에서 x_{2i}를 $U(-2,2)$에서 각각 생성했을 경우엔 Slope1에선 R(H)가 Slope2 에선 L_2의 분산 값이 작았다.

표 3.5: $x_{1i} \sim N(0,1)$, $x_{2i} \sim N(0,1)$, (단 , $\beta_0 = 0, \beta_1 = 0, \beta_2 = 0$)

	Slope1		Slope2		Intercept	
	분산	편의^2	분산	편의^2	분산	편의^2
L_2	0.0237	0.0000	0.0243	0.0001	0.0235	0.0000
$R(H)$	0.0163	0.0000	0.0181	0.0000	0.0000	0.0000
HBR	0.0247	0.0000	0.0256	0.0001	0.0320	0.0000

표 3.5의 경우는 R(H)가 Slope1, Slope2 모두 분산 값이 작다.

3.2. 대칭오염

원자료의 모형은 $Y_i = \beta_0 + \beta_1 x_i + e_i$, $e_i \sim f(t) = 2^{-1} \exp\left\{\frac{-|t|}{2}\right\}$ 이다. (단 , $\beta_0 = 0$, $\beta_1 = 0$) 독립변수 생성은 표 3.6 인 경우엔 Uniform에서, 표 3.7 에서는 오염정규분포에서 생성을 하였다. 표본의 크기는 30 이며 1000번 반복하여 실시하였다. 아래의 표에서 ARE는 근사적상대효율성(asymptotic relative efficiency)을 나타내는 값이다.

표 3.6을 보면 오염이 없는 일양분포에서는 Rcent의 분산 값이 제일 작음을 알 수 있다. 표 3.7을 보면 독립변수를 오염정규분포에서 생성한 경우에는 HBR의 분산 값이 제일 작음을 알 수 있다. 또한 오염비율과 오염부분의 분산이 각각 15%, 16 와 15%, 64 인 경우에도 HBR의 분산 값이 제일 작았다.

표 3.6: $x_i \sim U(0,1)$

	Slope			Intercept		
	분산	편의^2	ARE	분산	편의^2	ARE
L_1	0.1715	0.0000	1.1557	0.0572	0.0000	1.2115
L_2	0.1978	0.0004	1.0000	0.0693	0.0000	1.0000
$R(H)$	0.1708	0.0000	1.1604	0.0594	0.0000	1.1667
$R(S)$	0.1758	0.0000	1.1274	0.0646	0.0000	1.0728
$Rcent$	0.0462	0.0000	4.2900	0.0150	0.0000	4.6200
HBR	0.1504	0.0000	1.3178	0.0526	0.0000	1.3175

표 3.7: $x_i \sim CN(0.25, 100)$

	Slope			Intercept		
	분산	편의^2	ARE	분산	편의^2	ARE
L_1	0.0010	0.0000	1.0000	0.0112	0.0000	1.5536
L_2	0.0010	0.0000	1.0000	0.0174	0.0000	1.0000
$R(H)$	0.0021	0.0000	0.4762	0.0133	0.0000	1.3083
$R(S)$	0.0023	0.0000	0.4348	0.0164	0.0000	1.0610
$Rcent$	0.0020	0.0000	0.5000	0.0167	0.0000	1.0419
HBR	0.0009	0.0000	1.1111	0.0111	0.0000	1.5676

4. 결론 및 토의

현재까지 진행되어온 모의 실험결과를 분석해 보면 HBR 추정량이 반공간 뎁스에 기반한 회귀 뎁스 추정량보다 Chang et al.(1999)의 논문에서 주장한것 처럼 좋은 것은 아니라는 것을 우리의 모의실험에서 확인할 수 있었다. HBR 추정량이 상대적으로 우위를 보이는 경우는 그들의 논문의 실험조건처럼 대칭오염의 일부에서 라고 할 수 있다. 그러나 대칭오염중 일양분포에서는 Rcent가 HBR보다 좋은 결과를 보여주고 비대칭오염 상황에서는 전체적으로 Rcent가 제일 우수한 성능을 보임을 알 수 있었으며 HBR이 회귀뎁스 계열 추정량보다 좋지 않은 결과를 보여준다.

3차원의 경우에서는 각 추정량 별로 분산과 MSE등을 계산하여 보았고 또한 각 변수들의 변동성이나 MSE를 개별적으로 취급하지 않고 동시에 추론하는 방법을 사용하여 계산하여 보았다. 2차원의 경우에는 절편과 기울기를 동시에 할수도 있지만 각각의 모수가 가지는 의미가 다르기 때문에 따로 계산하였다.

회귀뎁스추정에서는 반공간 뎁스를 기반한 추정량이 Liu et al.(1999)의 심플리셜 뎁스

를 기반한 회귀추정량보다 좋은 결과를 보이는데 이 결과는 위치 추정량에서 모의실험결과(Hwang et al., 2004)와 동일한 양상을 보여주었다.

Rcent의 성능이 대부분의 경우에 좋게 나타난 것은 이론적인 방향과 일치 한다고 할 수 있다. 그러나 계산시간의 문제점과 이에 대한 이론적 성질의 규명등은 앞으로 해결해야 할 과제라고 생각한다.

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**, 445, Theory and Methods.

Donoho, D.L., and Huber, P.J. (1983). The notion of breakdown point, in: P.J. Bickel, K.A.Doksum and Hodges, Jr., eds, *A Fesrschrift foe Erich L. Lehmann*, Wadsworth, Belmont, CA, 157-184.

Hampel, F.R. (1974). The influence curve and its roles in robust estimation. *Journal of the American Statistical Association*, **69**, 383-393.

Huber, P.J.(1973). Robust regression: asymptotics, conjectures, and monte carlo. *Annals of Statistics*, **1**, 799-821.

Hubert, M., Rousseeuw, P.J., Van Aelst, S. (2001). Similarities between location depth and regression depth, *Trends in Mathematics*, 159-172.

Hwang, J.S., Jorn, H.S., and Kim, J.K. (2004). On the performance of bivariate robusst location estimators under contamination, *Computaional Statistics & Data Analysis*, **44**, 587-601.

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

Liu, R.Y., Parelius, J.M., and Singh, K. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference, *Annals of Statistics*, **18**, 405-414.

Mallows, C.L. (1975). On some topics in robustness. Technical Memorandum, Bell Telephone Laboratories, Murray Hill, New Jersey.

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

Rousseeuw,P.J., and Hubert, M. (1999). Regression depth, *Journal of the American Statistical Association*, **94**, 388-402.

Rousseeuw,P.J., and Yohai, V.J. (1984). Robust regression by means of s-estimators, *Robust and Nonlinear Time Series Analysis*, Lecture Notes in Statistics, **26**, 256-272, Springer, New York.

Siegel, A.F. (1982). Robust regression using repeated medians, *Biometrika*, **69**, 242-4

Zuo, Y., and Serfling, R. (2000). General notions of statistical depth function, *Annals of Statistics*, **28**(2), 461-482.

[2004년 1월 접수, 2004년 4월 채택]

A Comparison Study of Several Robust Regression Estimators under Various Contaminations *

Jeeyun Kim[1] Jinsoo Hwang [2] Jeankyung Kim [3]

ABSTRACT

Several robust regression estimators are compared under contamination. Symmetric and asymmetric contamination schemes are used to measure the variance and MSE of regression estimators. Under asymmetric contamination depth-based regression estimator, especially projection based regression estimator(rcent) outperforms the rest and under symmetric contamination HBR performs relatively well.

Keywords: Equivariance, Breakdown point, Influence function

* This work was supported by INHA University Research Grant(INHA-21962).

1) Graduate, Dept. of Statistics Inha University, 253 Yonghyun-Dong, Nam-Gu, 402-751, Incheon, Korea E-mail: jeeyun@anova.inha.ac.kr

2) Professor, Dept. of Statistics Inha University, 253 Yonghyun-Dong, Nam-Gu, 402-751, Incheon, Korea E-mail: jshwang@anova.inha.ac.kr

3) Professor, Dept. of Statistics Inha University, 253 Yonghyun-Dong, Nam-Gu, 402-751, Incheon, Korea E-mail: jkkim@anova.inha.ac.kr