The Korean communications in Statistics
Vol. 13 No. 1, 2006, pp. 29-38

Unmasking Multiple Outliers in Multivariate Data[1)]

Jong Young Yoo[2)]

Abstract

We proposed a procedure for detecting of multiple outliers in multivariate data. Rousseeuw and van Zomeren (1990) have suggested the robust distance RD_i by using the Resampling Algorithm. But RD_i are based on the assumption that X is in the general position.(X is said to be in the general position when every subsample of size $p+1$ has rank p) From the practical points of view, this is clearly unrealistic. In this paper, we proposed a computing method for approximating MVE, which is not subject to these problems. The procedure is easy to compute, and works well even if subsample is singular or nearly singular matrix.

Keywords : Mahalanobis distances; Minimum volume ellipsoid; Masking effects; Swamping effects; Resampling algorithm.

1. 서론

이 연구는 다변량자료에서 다중 이상점의 식별문제를 다루고자 한다. 일반적으로 이상점은 다수 자료의 형태에 따르지 않는 특정한 자료의 집단을 의미한다. 이런 이상점은 분석결과를 크게 왜곡시킬 수 있고 경우에 따라서는 가정한 모형이 잘못되었다는 중요한 정보를 갖고 있을 수도 있기 때문에 최종적인 분석에 앞서 이러한 점들을 구분해 낼 필요가 있다. 만약 자료에 하나의 이상점이 있다면 이 점을 식별하는 데는 이론적으로나 계산상에 별 문제가 없다. 그러나 두 개 이상의 이상점이 있을 경우 가장효과(masking effect)와 편승효과(swamping effect) 때문에 이러한 점들의 식별이 어려워진다. 가장효과란 두 개 이상의 이상점이 근접한 곳에 위치하여 서로 상대방을 이상점으로 식별되지 못하도록 방해하는 것을 말하며, 편승효과는 이상점이 아닌 점들을 이상점으로 식별될 때 나타나는 효과를 의미하고 있다. 이러한 편승효과와 가장효과를 줄이기 위하여 최근 들어 많은 로버스트한 방법들이 제안되고 있다.

이상점을 식별하는 고전적인 방법으로 마하라노비스의 거리(Mahalanobis Distances)가 있으며, 우리는 단순히 MD_i로 명칭한다.

---

1) This Research was supported by Yong In University Research Grants in 2004.

2) Associate Professor, School of Computer & Information, Yong In University, Yongin-si, 449-714, Korea. E-mail : jyyoo@yongin.ac.kr

$$MD_i = \sqrt{(x_i - \overline{X})^T S^{-1}(x_i - \overline{X})}, \qquad i = 1,2,\cdots,n. \tag{1.1}$$

X를 $n \times p$의 행렬로 n개의 관측값과 p개의 변수로 구성되는 자료라고 하면, (1.1)에서 $\overline{X}$는 표본평균, S는 공분산을 의미하며 유의수준 α에서 MD_i의 임계값은 $\sqrt{\chi^2_{p,1-\alpha/2}}$ 이다. 이 임계값을 벗어나는 MD_i에 해당하는 관측값은 이상점으로 판정한다. 그러나 MD_i를 이용하여 이상점을 식별하는 데는 실제로는 두 가지의 문제점이 발생한다. 첫 번째는 이상점들은 필연적으로 큰 MD_i를 갖지는 않는다는 것이다. 예를 들면 몇 개의 이상점들은 $\overline{X}$와 S를 같은 방향으로 팽창시켜 결과적으로는 MD_i가 크게 나오지 않을 수 있다는 것이다. 이런 현상은 앞에서 언급한 가장효과의 영향으로 나타난 결과이다. 두 번째는 MD_i가 크다고 해서 항상 그 관측값들이 이상점이라고 판정하는 데에는 문제가 있다는 것이다. 예를 들면 몇 개의 이상점들은 $\overline{X}$ 와 S를 같은 방향으로 팽창시키고 결과적으로 이상점이 아닌 관측값들의 MD_i를 크게 만들 수 있다는 것으로 이것은 앞에서 언급한 편승효과에 기인한 것으로 분석할 수 있다. 이러한 가장효과와 편승효과는 (1.1)에서 사용한 $\overline{X}$와 S가 로버스트한 통계량이 아니라는 데에서 기인한다.

일반적인 다중 선형 회귀모형에서 이상점을 식별하는 척도로서 행렬 $P = X(X^tX)^{-1}X^t$의 대각요소 p_{ii}를 사용하고 있다. Hoaglin and Welsch (1978)은 p_{ii}의 임계값을 $2(p+1)/n$으로 제시하였다. 이 방법도 마하라노비스의 거리와 마찬가지로 가장효과와 편승효과에 영향을 받고 있으며, p_{ii}와 MD_i는 다음과 같은 관계식이 있는 것을 분석할 수 있다.

$$p_{ii} = \frac{(MD_i)^2}{n-1} + \frac{1}{n}, \qquad i = 1,2,\cdots,n. \tag{1.2}$$

Rousseeuw (1985)는 블렉다운점이 50%까지 가능한 minimum volume ellipsoid (MVE)를 제안하였다. MVE는 (M, V)의 쌍으로 정의되어 지는데, 이때 M은 $p-$벡터이고 V는 $p \times p$ 양반정치행렬로 V의 행렬값이 (1.3)의 조건하에서 최소화되는 V를 의미한다.

$$\#\{i : (x_i - M)V^{-1}(x_i - M)^t \leq a^2\} > h. \tag{1.3}$$

단, 위에서 h는 $(n+p+1)/2$의 정수값이고, a^2는 상수로서 데이터가 정규분포에 따른다는 가정 하에서 $\chi^2_{p,0.50}$을 선택할 수 있다. MVE는 로버스트 표본평균과 로버스트 공분산을 사용하여 로버스트 통계량을 만드는 장점을 지니고 있다. 그러나 이론상으로는 매우 좋은 로버스트 통계량을 만든다고 하여도 이것의 계산이 너무 많아 실제의 사용에서는 상당한 문제점을 내포하고 있다. 이러한 계산상의 문제점을 해결하기 위하여 근사 MVE를 구하는 여러 가지 알고리듬이 제안되었는데 그중의 하나가 재표본 알고리듬이다.(Rousseeuw and Leroy, 1997) 재표본 알고리듬은 각 크기가 $p+1$인 여러 가지 부표본을 추출하여 다음의 식을 계산한다.

$$D_i(C_j, S_j) = \sqrt{(x_i - C_j)^T S_j^{-1}(x_i - C_j)}, \qquad i = 1,2,\cdots,n. \tag{1.4}$$

여기에서 C_j와 S_j는 j번째 부표본의 평균과 공분산을 의미한다. 앞의 (1.4)의 n개의 값 중에서 $100(h/n)$번째 백분위수에 해당하는 값을 m_j, S_j 행렬 값을 $\det(S_j)$라 정의하면,

h개의 관측값을 포함하고 C_j와 S_j에 근거한 타원의 볼륨은 $\{m_j^p \det(S_j)\}^{1/2}$에 비례하게 된다. 이때 j를 $m_j^p \det(S_j)$가 최소가 되는 부표본이라고 하면 부표본 j에 근거한 타원의 볼륨은 h개의 관측값을 포함하는 MVE의 근사값으로 사용할 수 있다. Rousseeuw and van Zomeren (1990)은 재표본 알고리듬을 이용하여 $m_j^p \det(S_j)$을 최소화시키는 j번째 부표본을 식별한 후 이상점을 구별하기 위하여 다음과 같은 로버스트 통계량을 제안하였다.

$$RD_i = D_i(C_j, c_j S_j) = \sqrt{(x_i - C_j)^T (c_j S_j)^{-1} (x_i - C_j)}, \qquad i = 1,2,\cdots,n. \tag{1.5}$$

위에서 $c_j = c_{np} m_j / \chi^2_{p,0.50}$는 조정요인이며, $c_{np} = 1 + 15/(n-p)^2$을 사용하고 있다. MD_i와 마찬가지로 유의수준 α에서 RD_i의 임계값은 $\sqrt{\chi^2_{p,1-\alpha/2}}$이다. (1.5)의 RD_i는 로버스트 통계량이고 가장효과와 편승효과의 문제점을 상당히 해결하고 있다. 그러나 RD_i는 다음과 같은 세 가지의 문제점을 내포하고 있음을 알 수 있다.(Hadi, 1992) 첫 번째는 Rousseeuw and Leroy는 부표본의 수를 $p+1$로 정하였는데 이 부표본의 수에 따라 이상점의 의사결정이 달라질 수 있다는 것이고, 두 번째는 매우 심각한 문제로 우리는 (1.4)와 (1.5)에서 S_j의 역행렬을 사용하였는데 이는 $\det(S_j)$가 0이 아니라는 강력한 가정에서 출발하고 있는 모순을 지니고 있고, 세 번째는 $\det(S_j)$가 0이 아니더라도 $\det(S_j)$의 값이 0에 가깝게 나타나면서 나타나는 여러 가지 문제점을 묵과하고 있다는 것이다. Rousseeuw and van Zomeren는 두 번째와 세 번째의 문제점을 $\det(S_j)$이 단순히 일정한 기준보다 적은 값에 대하여 삭제하는 방법을 택하고 있다. 이 연구에서는 $\det(S_j)$의 값이 0 또는 0에 가까울 때 나타나는 여러 가지 현상을 분석하고 실증하여 새로운 방법을 제안하고자 하는 데에 목적이 있다.

2. 제안된 방법

우리는 Singular Value Decomposition (SVD)의 개념을 이용하여 $m_j^p \det(S_j)$가 최소가 되는 부표본을 구하고자 한다. 먼저 S_j는 $p \times p$의 양반정치 행렬이고 대칭행렬이므로, SVD에 의하여 $S_j = U_j D_j U_j^t$로 분해할 수 있다. 이때 U_j는 $U_j^t U_j = I_p$를 만족하고 행렬 D_j는 행렬 S_j의 고유값으로 이루어진 $p \times p$ 대각행렬이다. 이때 S_j의 행렬의 값은 S_j의 고유값들의 곱으로 나타낼 수 있다. 즉 $\det(S_j) = \lambda_1 \lambda_2 \cdots \lambda_p = \lambda_{(1)} \lambda_{(2)} \cdots \lambda_{(p)}$로 쓸 수가 있는데 행렬의 값이 0 혹은 0에 가깝다고 하는 것은 $\lambda_{(p)}$의 값이 0 혹은 0에 매우 가깝다는 것을 의미하고 있다. 3장의 시뮬레이션 결과에서 후술한 바와 같이 $\lambda_{(p)}$가 0인 경우 $\det(S_j)$가 0이 되고 $m_j^p \det(S_j)$가 0이 되어 랭크가 p보다 작은 부표본 j가 $m_j^p \det(S_j)$를 최소로 만드는 부표본으로 선택되는 모순이 있고, $\lambda_{(p)}$가 0에 근접하면서 m_j의 값이 음수가 나오는 모순과 함께 $\lambda_{(p)}$가 음수로 나타나는 모순이 나타날 수 있다. Rousseeuw and van Zomeren이 $\det(S_j)$가 단순히 일정한 값보다 작았을 때 무시하였던

부표본에 대하여 우리는 $\lambda_{(p)}$를 조절하여 부표본의 $m_j^p \det(S_j)$를 다음과 같이 구하고자 한다. 우리는 1장의 (1.4)와 (1.5)를 SVD를 이용하여 다음의 식으로 변형시킬 수 있다.

$$D_i(C_j, S_j) = \sqrt{(x_i - C_j)^t U_j W_j U_j^t (x_i - C_j)} \tag{2.1}$$

$$RD_i = D_i(C_j, c_j S_j) = \sqrt{(x_i - C_j)^t c_j^{-1} U_j W_j U_j^t (x_i - C_j)} \tag{2.2}$$

단, (2.1)과 (2.2)에서 W_j는 $p \times p$인 대각행렬로 i번째 대각요소는 $w_i = 1/\lambda_i$이고 $\lambda_{(p)}$가 0이거나, $\lambda_{(p)}$가 0에 근접하면서 m_j의 값이 음수가 나오거나 $\lambda_{(p)}$가 음수로 나타나 $m_j^p \det(S_j)$가 음수로 계산되는 경우 $w_p = M$(임의의 큰 수)로 대처한다. 이는 부표본의 랭크가 p보다 작은 경우는 λ_p가 0이 되어 $1/\lambda_p$가 불능이 되는 것을 방지하고, 또한 3장의 시뮬레이션 결과에서 후술한 바와 같이 부표본의 랭크가 p이라고 하더라도 행렬 S_j의 값이 0에 근접한 경우 나타나는 여러 가지 모순을 해결하고자 함이다.

3. 예제와 제안된 방법의 근거

이 절에서는 여러 가지 예제를 시뮬레이션하여 우리의 제안된 방법의 근거와 함께 나타난 결과를 기술하고자 한다. 예제로서는 그동안 이상점 탐색에 많이 쓰이고 Rousseeuw and van Zomeren에 기술되었던 3개의 자료를 가지고 기술하고자 한다. 이 연구에서의 모든 계산은 S-PLUS 8.2를 이용하였다.

3.1. Artificial Data

인공자료는 75개의 관측값과 3개의 변수로 이루어져 있으며 가장효과를 나타내는 좋은 예제로 많이 사용되고 있으며, 관측값 {1,2,3,4,5,6,7,8,9,10,11,12,13,14}이 이상점이다. 이 자료는 ${}_{75}C_4 = 1,215,450$의 크기 4인 부표본이 존재하며, 부표본 중에서 191개의 부표본이 $m_j^p \det(S_j)$를 계산하는 것이 불가능한 것으로 분석이 되었다.

<표 1> Artificial Data에서 $\det(S_j)$의 계산이 불가능한 경우

$\det(S_j)$의 부호	m_j의 부호	갯수
$\det(S_j)<0$	$m_j>0$	32
$\det(S_j)<0$	$m_j<0$	51
$\det(S_j)=0$	$m_j<0$	13
$\det(S_j)=0$	$m_j>0$	63
$\det(S_j)>0$	$m_j<0$	32

<표 1>의 191개의 부표본에서 $dS_j = 0$인 경우는 부표본의 차수가 p보다 작은 경우에

나타나고 $dS_j < 0$인 경우와 $\{dSj > 0, m_j < 0\}$인 경우는 $m_j^p \det(S_j)$는 계산이 불가능한 상태로 Rousseeuw and van Zomeren는 이러한 부표본을 제외하고 결과를 도출하였다. 또한 $m_j^p \det(S_j)$를 계산할 수 있는 부표본에서 146개 부표본의 S_j 행렬값이 10^{-8}보다 작은 것으로 분석되었으며 <표 2>에서 분석한 것처럼 S_j의 행렬값이 작으면 $m_j^p \det(S_j)$의 값이 커지는 경향이 있는 것을 분석할 수 있다.

<표 2> Artificial Data의 행렬의 값과 $m_j^p \det(S_j)$의 값

$\det(S_j)$의 구간	갯수	$\det(S_j)$의 평균	$m_j^p \det(S_j)$의 평균
1E-09<$\det(S_j)$<1E-08	71	9.2E-09	4.9E09
1E-16<$\det(S_j)$<1E-09	21	1.7E-12	9.5E15
$\det(S_j)$<1E-16	54	2.7E-17	4.2E16

<표 2>에서 행렬의 값이 1E-09<$\det(S_j)$<1E-08인 71개의 부표본의 $m_j^p \det(S_j)$의 평균은 4.9E09인 반면에 $\det(S_j)$<1E-16인 54개의 부표본의 $m_j^p \det(S_j)$의 평균은 4.2E16로 분석되어 행렬의 값이 적어질수록 $m_j^p \det(S_j)$의 값이 커지는 경향이 있는 것으로 실증되었다. 우리는 여기에서 S_j의 행렬값이 적어지면 S_j의 값이 0으로 근접하고 $1/\lambda_{(p)}$의 값이 큰 수로 발산하고 m_j가 커지고 따라서 $m_j^p \det(S_j)$가 커지는 현상을 실증분석 할 수가 있다. 여기에서 우리가 내릴 수 있는 결론은 191개의 $m_j^p \det(S_j)$를 계산하지 못하는 부표본은 행렬의 값이 매우 작은 경우이거나 행렬의 값이 0인 경우이므로 임의로 $1/\lambda_{(p)}$에 대하여 큰 수 M을 배정하여 $m_j^p \det(S_j)$를 큰 수로 만들어 기본 부표본으로 선택되는 것을 방지하는 알고리듬을 사용할 수 있다는 것이다.
모든 부표본을 조사한 결과 {20,32,60,65} 부표본에서 m_j는 2.469, $\det(S_j)$는 2.322로 계산이 되어 $m_j^p \det(S_j)$는 34.957로 최소의 값을 나타내고 있으며, 이때의 RD_i는 MD_i, p_{ii}와 함께 <표 3>에 기술되어 있다. <표 3>에서는 MD_i는 관측값은 {12,14}, p_{ii}는 관측값 {12,13,14}을 이상점으로 식별하였으며, 부표본 {20,32,60,65}를 사용한 RD_i는 관측값 {1,2,3,4,5,6,7,8,9,10,11,12,13,14}이 이상점으로 식별되었다. 결과로는 우리가 제안한 방법은 기존의 Rousseeuw and Von Zomeren의 방법과 Hadi의 방법과 동일한 결과를 얻을 수 있었으며, MD_i와 p_{ii} 방법은 이상점을 제대로 식별하지 못하는 것으로 분석되었다.

3.2 Stackloss Data

스택로스 자료는 21개의 관측값으로 3개의 설명변수와 1개의 반응변수로 구성되어 있으며 선형회귀분석에서 이상점을 탐색하는 방법에 많이 사용되고 있는 자료이다. 이 자료는 관측값 {1,2,3,21}이 이상점으로 구성되어 있고 ${}_{21}C_4 = 5,719$의 부표본이 존재한다.

5,719개의 부표본 중에서 221개의 부표본을 계산하는 것이 불가능한 것으로 분석이 되었다.

<표 3> ART DATA의 MD_i, p_{ii}, RD_i의 값

case	MD_i	p_{ii}	RD_i	case	MD_i	p_{ii}	RD_i
1	1.92	0.06	28.65	39	1.27	0.03	2.53
2	1.86	0.06	29.88	40	1.11	0.03	0.77
3	2.31	0.09	30.79	41	1.70	0.05	1.43
4	2.23	0.08	32.43	42	1.77	0.06	0.89
5	2.10	0.07	31.62	43	1.87	0.06	1.58
6	2.15	0.08	29.68	44	1.42	0.04	1.23
7	2.01	0.07	30.06	45	1.08	0.03	2.70
8	1.92	0.06	29.04	46	1.34	0.04	1.24
9	2.22	0.08	31.41	47	1.97	0.07	3.13
10	2.33	0.09	30.30	48	1.42	0.04	1.23
11	2.45	0.09	36.24	49	1.57	0.05	1.07
12	3.11	0.14	36.63	50	0.42	0.02	0.99
13	2.66	0.11	36.80	51	1.30	0.04	1.94
14	6.38	0.56	43.29	52	2.08	0.07	2.68
15	1.82	0.06	1.20	53	2.21	0.08	1.24
16	2.15	0.08	2.02	54	1.41	0.04	1.18
17	1.38	0.04	1.53	55	1.23	0.03	0.93
18	0.85	0.02	0.89	56	1.33	0.04	0.95
19	1.15	0.03	0.98	57	0.83	0.02	2.10
20	1.59	0.05	1.22	58	1.40	0.04	1.00
21	1.09	0.03	0.81	59	0.59	0.02	0.68
22	1.55	0.05	1.54	60	1.89	0.06	1.22
23	1.09	0.03	1.06	61	1.67	0.05	2.22
24	0.97	0.03	1.07	62	0.76	0.02	2.52
25	0.80	0.02	2.57	63	1.29	0.04	1.20
26	1.17	0.03	1.60	64	0.97	0.03	1.15
27	1.45	0.04	1.16	65	1.15	0.03	1.22
28	0.87	0.02	1.27	66	1.30	0.04	1.09
29	0.58	0.02	1.65	67	0.63	0.02	0.74
30	1.57	0.05	1.01	68	1.55	0.05	1.97
31	1.84	0.06	2.08	69	1.07	0.03	2.53
32	1.31	0.04	1.22	70	1.00	0.03	0.91
33	0.98	0.03	0.83	71	0.64	0.02	0.51
34	1.18	0.03	2.24	72	1.05	0.03	0.62
35	1.24	0.03	1.08	73	1.47	0.04	0.76
36	0.85	0.02	0.72	74	1.65	0.05	1.06
37	1.83	0.06	1.87	75	1.90	0.06	1.16
38	0.75	0.02	1.95				

<표 4> Stackloss Data에서 $m_j^p \det(S_j)$의 계산이 불가능한 경우

$\det(S_j)$의 부호	m_j의 부호	갯수
$\det(S_j)<0$	$m_j>0$	27
$\det(S_j)<0$	$m_j<0$	50
$\det(S_j)=0$	$m_j<0$	39
$\det(S_j)=0$	$m_j>0$	75
$\det(S_j)>0$	$m_j<0$	30

또한 45개 부표본의 S_j의 행렬값이 10^{-8}보다 작은 것으로 분석되었으며 <표 5>에서 분석한 것처럼 S_j의 행렬값이 작으면 $m_j^p \det(S_j)$의 값이 커지는 것을 분석할 수 있다.

<표 5> Stackloss Data의 행렬의 값과 $m_j^p \det(S_j)$의 값

$\det(S_j)$의 구간	갯수	$\det(S_j)$의 평균	$m_j^p \det(S_j)$의 평균
1E-11<$\det(S_j)$<1E-08	12	2.0E-12	1.9E16
1E-12<$\det(S_j)$<1E-11	19	3.7E-13	4.9E17
$\det(S_j)$<1E-12	14	3.9E-14	3.8E18

우리는 <표 5>에서도 S_j의 행렬값이 적어지면 $m_j^p \det(S_j)$가 커지는 현상을 실증분석 할 수 가 있다.

<표 6> Stackloss Data의 MD_i, p_{ii}, RD_i의 값

case	MD_i	p_{ii}	RD_i	*case*	MD_i	p_{ii}	RD_i
1	2.25	0.30	5.23	12	1.84	0.22	0.79
2	2.32	0.32	5.27	13	1.48	0.16	0.55
3	1.59	0.17	4.01	14	1.78	0.21	0.64
4	1.27	0.13	0.84	15	1.69	0.19	2.23
5	0.30	0.05	0.80	16	1.29	0.13	2.11
6	0.77	0.08	0.78	17	2.70	0.41	2.07
7	1.85	0.22	0.64	18	1.50	0.16	2.09
8	1.85	0.22	0.64	19	1.59	0.17	2.29
9	1.36	0.14	0.83	20	0.81	0.08	0.64
10	1.75	0.20	0.64	21	2.18	0.28	3.30
11	1.47	0.16	0.58				

여기에서도 우리가 내릴 수 있는 결론은 221개의 $m_j^p \det(S_j)$를 계산하지 못하는 부표본에 대해서 임의로 $1/\lambda_{(p)}$에 대하여 큰 수 M을 배정하여 $m_j^p \det(S_j)$를 큰 수로 만들어 기본 부표본으로 선택되는 것을 방지하는 알고리듬을 사용할 수 있다는 것이다. 또한 모든 부표본을 조사한 결과 {7,10,14,20}, 혹은 {8,10,14,20}의 부표본에서의 m_j는 3.871, $det(S_j)$는 472.93으로 계산이 되어 $m_j^p \det(S_j)$는 27,434로 최소의 값을 나타내고 있으며 이때의 RD_i는 MD_i, p_{ii}와 함께 <표 6> 에 기술되어 있다. <표 6> 에서는 MD_i중 임계값 $\sqrt{\chi^2_{3,0.975}} = 3.06$을 넘는 관측값은 전혀 존재하지 않아 이상점을 전혀 식별하지 못하고 있으며, p_{ii}는 관측값 {17}을 이상점으로 식별하여, MD_i와 p_{ii} 방법은 이상점을 제대로 식별하지 못하는 것으로 분석되었다. 그러나 부표본 {7,10,14,20} 혹은 {8,10,14,20}을 사용한 RD_i는 관측값 {1,2,3,21}을 이상점으로 식별하였으며, 이는 Rousseeuw and Zomeren의 방법과 Hadi의 방법과 동일한 결과이다.

3.3 Brain Data

뇌의 자료는 28개 종족의 뇌와 몸무게에 로그함수를 사용한 두개의 변수를 가지고 있는 자료로 관측값 {6,14,16,17,25}이 이상점으로 구성되어 있다. <표 7> 에서 4개 자료의 부표본 행렬값이 0.0001보다 적은 것을 알 수 있으며 이때도 S_j의 행렬값이 적어지면 $m_j^p \det(S_j)$가 커지는 현상을 실증하고 있다.

<표 7> Brain Data의 행렬의 값과 $m_j^p \det(S_j)$의 값

부표본	$\det(S_j)$	$m_j^p \det(S_j)$
{ 2, 9,12}	5.87e-05	416.29
{ 4,12,22}	1.59e-07	3,859.91
{ 8,19,21}	1.69e-06	2,036.63
{11,17,21}	2.92e-07	22,412.8

우리는 여기에서도 S_j의 행렬값이 적어지면 $m_j^p \det(S_j)$가 커지는 현상을 실증분석 할 수가 있었다. 또한 모든 부표본을 조사한 결과 {1,2,22}의 부표본에서의 $m_j^p \det(S_j)$는 5.772로 최소의 값을 나타내고 있으며 이때의 RD_i는 MD_i, p_{ii} 와 함께 <표 8> 에 기술되어 있다. <표 8> 에서 MD_i는 관측값 {25}, p_{ii}는 관측값 {6,20,25}를 이상점으로 식별하여 MD_i와 p_{ii} 방법은 이상점을 제대로 식별하지 못하는 것으로 분석되었다. 그러나 부표본 {1,2,22}를 사용한 RD_i는 관측값 {6,14,16,17,25}을 이상점으로 식별하여 Rousseeuw and Zomeren의 방법과 Hadi의 방법과 함께 이상점을 정확하게 식별하는 것으로

분석되었다.

<표 8> Brain Data의 MD_i, p_{ii}, RD_i의 값

case	MD_i	p_{ii}	RD_i	*case*	MD_i	p_{ii}	RD_i
1	1.01	0.07	0.54	15	1.76	0.15	1.14
2	0.70	0.05	0.54	16	2.37	0.24	6.12
3	0.30	0.04	0.40	17	1.22	0.09	2.72
4	0.38	0.04	0.63	18	0.20	0.04	0.67
5	1.15	0.08	0.74	19	1.86	0.16	1.19
6	2.64	0.29	6.83	20	2.27	0.23	1.24
7	1.71	0.14	1.59	21	0.83	0.06	0.47
8	0.71	0.05	0.64	22	0.42	0.04	0.54
9	0.86	0.06	0.48	23	0.26	0.04	0.29
10	0.80	0.06	1.67	24	1.05	0.08	1.95
11	0.69	0.05	0.69	25	2.91	0.35	7.26
12	0.87	0.06	0.50	26	1.59	0.13	1.04
13	0.68	0.05	0.52	27	1.58	0.13	1.19
14	1.72	0.15	3.39	28	0.39	0.04	0.75

4. 맺음말

Rousseeuw and van Zomeren이 제안하였던 RD_i 통계량은 행렬의 값이 0 혹은 0에 가깝게 나타나는 부표본을 단순히 삭제하는 방법을 택하고 있다. 우리는 이러한 부표본에 대하여 다중 이상점과 지렛점의 식별에 많이 사용하는 Artificial Data, Stackloss Data, Brain Data에 대하여 시뮬레이션을 하여 문제점을 분석한 후 이러한 문제점을 SVD을 이용하여 해결하는 방법을 제시하였다. 결과로는 우리의 분석은 그 동안 다른 통계학자들에 의하여 분석된 결과와 동일한 결과를 얻어 우리가 제안한 방법은 이상점을 재확인하는 데에 유효한 방법으로 생각된다. 향후 과제로서는 심사위원께서 지적하였듯이 행렬의 값이 0 혹은 아주 작은 경우 epsilon을 대각행렬에 더하여서 부표본의 랭크가 p보다 적어지는 것을 방지하는 알고리듬을 연구할 필요가 있다고 생각된다.

참고문헌

[1] 유종영, 안기수 (2002). 다중 선형 모형에서 식별된 다중 이상점과 다중 지렛점의 재확인 방법에 대한 연구. 「응용통계연구」, 제15권 2호, 269-279.

[2] Atkinson, A.C. (1986). Masking unmasked. *Biometrika,* Vol. 73, 533-541.

[3] Fung, W.K. (1993). Unmasking Outliers and Leverage Points: A Confirmation. *Journal of the American Statistical Association,* Vol. 88, 515-519.

[4] Hadi, A.S. (1992). Identifying Multiple Outliers in Multivariate Data. *Journal of the Royal Statistical Society*, Ser.B, Vol. 54, 761-771.

[5] Hadi, A.S. and Simonoff, J.S. (1993). Procedures for the Identification of Multiple Outliers in Linear Models. *Journal of the American Statistical Association,* Vol 75, 1264-1272.

[6] Hawkins, D.M. , Bradu, D. and Kass, G.V. (1984). Location of several outliers in multiple regression data using elemental sets. *Technometrics,* Vol 26, 197-208.

[7] Rousseeuw, P.J. (1984). Least Median of Squares Regression. *Journal of the American Statistical Association,* Vol. 79, 871-880.

[8] Rousseeuw, P.J. (1985). Multivariate estimation with high breakdown points. In *Mathematical Statistics and applications* (eds W. Grossman, G. Pflug, I. Vincze and W. Wertz), Vol. B, pp.283-297. Dordrecht: Reidel.

[9] Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection,* John Wiley & Sons.

[10] Rousseeuw, P.J. and van Zomeren, B.C. (1990). Unmasking multivariate outliers and leverage points(with comments). *Journal of the American Statistical Association,* Vol. 75, 633-651.

[Received April 2005, Accepted December 2005]