▩ 연구논문

다변량 자료에서 다수 이상치 인식의 절차†

염준근
동국대학교 통계학과
박종구
원광대학교 컴퓨터공학과
김종우
제주교육대학교 수학교육과

A Procedure for Indentifying Outliers in Multivariate Data

Joon-Keun Yum
Dept. of Statistics, Dongguk University
Jong-Goo Park
Dept. of Computer Science, Wonkwang University
Jong-Woo Kim
Dept. of Mathematics Education, Cheju National University

Abstract

We consider the problem of identifying multiple outliers in linear model. The available regression diagnostic methods often do not succeed in detecting multiple outliers because of the masking and swamping effect. Recently, among the various robust estimator of reducing the effect of outliers, LMS(Least Meadian Square) estimator has been to be a suitable method proposed to expose outliers and leverage points. However, as you know it, the data analysis method with LMS estimator is to be taken the median of the squared residuals in the sample which is extracted the sample space. Then this model causes the trouble, for the number

† 이 논문은 1994년도 한국학술진흥재단의 공모과제 연구비에 의하여 연구되었음.

of the chosen sample is nCp, i.e. as the size of sample space n is increasing, the number is increasing fastly. And the covariance matrix may be the singular matrix, so that matrix is approching collinearity. Thus we propose a procedure ELMS for the resampling in LMS method and study the size of the effective elementary set in this algorithm.

1. 서론

선형회귀분석(linear regression analysis)에서 이상치(outlier)의 존재는 자료의 특성을 파악하는데 중요한 영향을 미치므로, 이들을 식별하기 위한 다양하고 효과적인 방법들이 많이 제시되어 왔다. 이러한 회귀 진단(regression diagnosis)분야는 접근 방법에 따라 크게 두가지 부류로 나눌 수 있다: 직접접근방법과 로버스트(robust) 적합을 이용한 간접접근방법이 있다.

직접접근방법은 Belsely, Kuh & Welch(1980), Rousseeuw and Leroy(1987) 등의 관련 서적을 통해 다수의 관측점들로 부터 이상치들을 찾는 방법으로 forward-stepping, backward-stepping, multistage, recursive residual 방법 등이 제시되고 있다. 그러나 이들 진단통계량의 이상치 식별에는 한계가 있다[Hadi & Simonoff, 1993]. 이러한 문제점의 주된 요인으로는 사용하는 이상치 식별을 위한 방법이 잔차 e_i와 hat matrix의 대각선상의 원 h_{ii}가 이상치에 크게 민감한 최소자승법(least square method : LS)을 기초로 하고 있기 때문에 다중 이상치의 존재시에는 이상치가 군집(cluster)해 있는 방향으로 중심점을 끌어 당기는 영향에 의하여 이상치를 숨기려는 은폐(masking)효과와 정상적인 점들이 중심점에서 멀리 떨어져 있는 점으로 인식되는 수렁(swamping)효과에 의해 이상치와 영향력 관찰점(influential observation)의 인식을 어렵게 하고 있다. 이상치 인식을 위한 간접방법으로 은폐효과와 수렁효과를 극복하기 위하여 로버스트 적합을 이용한 잔차분석을 하는 로버스트 추정량이 1970년대 중반부터 본격적으로 연구되기 시작했다.

보다 향상된 로버스트 추정량을 처음 도입한 사람은 Edgeworth로 알려져 있다 [Rousseeuw and Leroy 1987, p. 10].

$$\underset{\hat{\theta}}{\text{Minimize}} \sum_{i}^{n} | e_i |$$

그의 최소절대잔차 추정량은 이상치 저항성에 매우 강하지만 높은 지렛대점(leverage point)에는 매우 약한 것으로 지적되고 있으며, 이를 개선한 Huber(1973)의 M 추정량은 0에서 유일한 최소값을 갖는 미분가능한 함수 $\rho(\cdot)$에 대하여 잔차의 함수 $\rho(e_i)$를 최소화함으로써 얻는다.

$$\sum \varphi(e_i/\hat{\sigma})x_i = 0, \quad \text{여기서} \quad \varphi(\cdot) = \rho'(\cdot)$$

M 추정량은 효율적이고 Y상의 이상치에 매우 로버스트한 것으로 알려져 있으나, 하나의 모호한 지렛대점에 대하여도 완전히 붕괴(breakdown)될 수 있다. 따라서 이를 보완한 일반화된 M 추정량(generalized M-estimators 혹은 bounded influence estimators)이 제안되었다[Hampel et al., 1986, pp. 307-338].

$$\sum w(x_i)\varphi(e_i/\hat{\sigma})x_i = 0$$

이 추정량은 지렛대점에 덜 민감하지만 weighting 함수 $w(\cdot)$가 이상치에 높은 저항성을 가져야만 한다는 제약성을 갖고 있다. 그리고 특정 영역의 관찰치를 절사하는 방법을 사용하는 L 추정량과 repeated median(RM) 등이 있다[Siegel, 1982, Rosner, 1983]. 최근에 널리 사용되고 있는 로버스트 추정량으로는 Rousseeuw(1984)의 잔차제곱의 중위수를 최소화하는 least median of squares(LMS) 추정량, Rousseeuw(1985)의 least trimmed square(LTS)등이 있으며, 높은 붕괴점(breakdown point)을 갖는 추정량들로 S, MM, tau, GS 추정량등이 제시되고 있다[Yohai, 1987, Yohai and Zamar, 1988].

회귀 구조가 없는 다변량 자료에 대한 분석으로는 mahalanobis distance(MD)가 은폐와 수렴효과에 매우 민감하므로 이를 로버스트한 location과 shape을 사용하여 이상치를 파악하는 추정량으로 Rousseeuw & van Zomeren(1990)는 minimum volume ellipsoid (MVE)를 사용하여 다수의 이상치를 식별하고, 지렛대점을 찾아내는 방법을 제시하고 있다. 이러한 추정량으로서 최근에 Rousseeuw and van Zomeren(1990), Woodruff & Rocke(1993,1994) 등이 MD의 $C(X)$와 $S(X)$에 새로운 로버스트 추정량을 제시하였다. 이들은 거의 50%에 달하는 붕괴점을 가지므로써 이상치의 집단 식별을 매우 용이하게 해주고 있다.

그러나 이렇게 높은 붕괴점(breakdown point)을 갖는 추정량들은 극도로 많은 연산횟수를 필요로 하고 있다. 근래에 널리 사용되고 있는 추정량 LMS와 MVE에서 사용하고 있는 무작위 탐색 알고리즘(random search algorithm)은 로버스트한 모수 추정량을 얻기 위하여 지나치게 많은 연산 횟수를 필요로 하고 있다. 이러한 많은 표본 재추출의 문제점을 개선하고자, Hadi(1992)는 MVE의 계산시에 발생하는 표본 추출의 횟수를 줄이기 위한 방법으로 로버스트한 location과 scale을 갖는 추정량으로 MD를 구하여 중심에서 떨어져 있는 정도를 파악하고, 이를 기준으로 중심 근처에 위치한 원들과 나머지 원들중에서 오름차순(ascending ordering)에 의한 일정한 갯수의 원들을 추가해 가면서 표본을 선택함으로써 MVE에서 자료의 반복 추출 횟수를 줄이는 효과를 갖는 알고리즘을 제시하고 있으며, Woodruff and Rocke(1993)은 무작위 탐색를 개선한 heuristic 탐색를 제안하였고, Hawkins and Simonoff(1993), Hawkins(1993)는 elemental set 사용의 제안과 LMS, MVE, LTS, recursive residuals 등에 관한 개선된 알고리즘들을 제안하고 있다. Hadi and Siminoff(1993), Atkinson(1994) 등은 LMS, MVE에서 사용하고 있는 최적의 부분집합 대신에 이상치가 없는 초기 부분집합을 사용하고 이 부분집합의 자료의 갯수를 일정한 수준까지 늘려 나가면서 잠재적 이상치를 배제시킨 부분집합만을 이용하여 모수를 추정하는 알고리즘을 제시하고 있다.

본 연구에서는 선형회귀 구조를 갖는 모집단에서 다수 이상치 인식을 위해 Rousseeuw(1984)가 LMS에서 사용한 무작위 탐색 알고리즘과 Atkinson(1994)의 부분집합 크기 증가 방법을 바탕으로 초기 부분집합을 최소 잔차에 의하여 구성하고, 이 부분집합의 원을 늘려 나갈 때 부분집합에 포함된 이상치에도 강한 저항성을 갖도록 원을 재조정하면서 원의 갯수가 일정한 수에 달할 때까지 늘려나가는 알고리즘을 제시하며, 이에 의한 이상치 인식 방법을 제안하고자 한다.

2. Least Median Squares

선형회귀모형(linear regression model)을 다음과 같이 설정하자.

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

여기서 $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)'$는 종속변수인 $n\times 1$ 벡터이다.

$X = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)'$는 $p\times 1$ 벡터인 $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$을 행으로 갖는 독립 변수인 $n\times p$ 행렬이다 (단, $p<n$).

$\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)$인 모수 $p\times 1$ 벡터이다.

$\boldsymbol{\varepsilon} \sim N(0, \sigma^2)$인 $n\times 1$ 벡터이다.

(즉, $E(\boldsymbol{\varepsilon}|X) = 0$ and $\mathrm{var}(\boldsymbol{\varepsilon}|X) = \sigma^2 I_n$).

$\boldsymbol{\beta}$와 σ^2의 최소제곱 추정량은

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y}, \tag{2}$$
$$\hat{\sigma}^2 = \boldsymbol{e}\,\boldsymbol{e}/(n-p), \quad \text{여기서} \quad \boldsymbol{e} = \boldsymbol{y} - X\hat{\boldsymbol{\beta}}$$

으로 얻어진다. LS에서 이상치의 여부를 파악하기 위하여 잔차를 이용한 다양한 함수들이 개발되어 사용되고 있다[박성현 1991, pp. 521-543]. 그러나 이러한 최소제곱을 이용한 추정량은 이상치들이 군집되어 있을 때 0%에 가까운 붕괴점을 갖고 있다[Rousseeuw and Leroy, 1987, p. 69]. 주된 원인은 다수의 이상치가 존재하는 선형 회귀 모형에서 고전적인 LS 방법은 모수의 결정시에 군집되 있는 이상치들에 의하여 은폐효과와 수렁효과가 발생하기 때문이다. 이러한 효과를 제거하기 위한 방법으로 최근에 널리 사용되고 있는 LMS, MVE 방법은 붕괴점이 거의 50%에 달하는 높은 안정성을 보이고 있다[Rousseeuw, 1984; Rousseeuw and van Zomeren, 1990]. 그러나 이 방법의 적용을 위한 LMS에서 사용된 무작위 탐색 알고리즘은 크기 p인 부분집합을 구하기 위하여는 $n\times p$ 행렬에서 nCp회의 표본추출 즉, $n!/p!(n-p)!$회에 달하는 계산을 필요로 한다. 즉, 모수벡터 $\hat{\boldsymbol{\beta}}$를 결정하기 위해 행렬 X 대신에 크기 $p+1$인 부분집합을 무작위 추출하여 중위수

를 구하고, 각각의 부분집합에서 얻어지는 중위수들 중에 최소인 부분집합을 택하여 전체 관측치의 잔차를 구한 다음에 scale factor에 의하여 잔차를 표준화시키고, 기각치 (cuffoff value) 2.5와 비교하여 잔차에 의한 이상치 여부를 결정한다[Rousseeuw, 1984]. 그러면 LMS에 의한 이상치 결정 방법을 살펴 보기로 한다.

LMS의 주된 효과는 크기 $p+1$인 부분집합 J를 구하는데 있다.

$$\underset{\hat{\boldsymbol{\beta}}}{\text{Minimize}} \ \underset{J}{\text{med}} \ e_i^2 \tag{3}$$

모수 $\hat{\boldsymbol{\beta}}$을 얻기 위하여 부분집합 J로 구성된 X_J와 $\boldsymbol{y}_J$를 이용하여

$$\hat{\boldsymbol{\beta}} = (X_J^t X_J)^{-1} X_J^t \boldsymbol{y}_J \tag{4}$$

를 계산한다. 식(4)의 $\hat{\boldsymbol{\beta}}$를 사용하여 잔차

$$e_i = \boldsymbol{y}_i - X_i \hat{\boldsymbol{\beta}} \tag{5}$$

를 구한다. 관측치의 이상치 여부를 판정하기 위하여 scale factor σ^*를 구하고 n과 p에 의해 결정되는 유한표본수정계수를 사용하여 초기 수정계수 s^o와 가중치 w_i를 결정한다.

$$\sigma^* = \sqrt{\frac{\sum_{i=1}^{n} w_i e_i^2}{\sum_{i=1}^{n} w_i - p}} \tag{6}$$

여기서 $w_i = \begin{cases} 1 & |e_i/s^o| \le 2.5 \\ 0 & \text{otherwise.} \end{cases}$

$$s^o = 1.4826\left(1 + \frac{5}{n-p}\right)\sqrt{\min \text{med}_J e_i^2}$$

식(5)와 식(6)을 사용하여

$$\left|\frac{e_i}{\sigma^*}\right| > 2.5 \tag{7}$$

이면 i번째 관찰치를 이상치로 간주한다.

3. Extended Least Median Squares(ELMS)

ELMS 방법은 4단계로 구성되어 있다. 1단계에서는 초기 부분집합을 결정한다. LS를 사용하여 최소 잔차를 구하고 오름차순으로 X의 계수(rank)가 p일때, $p+1$개의 원을 초기 부분집합의 원으로 한다. 2단계에서 부분집합의 크기를 1개씩 증가 시켜 나가는 방법이다. 이 방법은 Atkinson(1994)가 제시하고 있는 forward search와 Rousseeuw(1984)의 무작위 탐색 방법을 결합한 구조와 유사하며, 부분집합 J의 크기가 k일때, 이 부분집합에서 $\binom{k}{k-1}$개 즉, k개의 부분집합와 나머지 $n-k$로 구성된 부분 집합에서 1개의 원을 취해 결합시킨 부분집합 J를 대상으로 LMS에서 제시하는 최적의 부분집합 선택 방법인 $\min(\text{med } r_i^2)$를 사용하여 새로운 부분집합 J를선택 한다.

3단계에서는 2단계에서 구한 부분집합를 사용하여 잔차와 scale factor를 구하여 관측치의 이상치 여부를 분석하고, 4단계에서는 3단계에서 이상치로 판별된 관측치들을 부분집합의 크기 증가에 따라 분석하여 관측치의 상태를 분석한다.

ELMS Algorithm

1단계. 초기 부분집합 J의 설정.

식(2)에서 제시된 LS의 잔차를 사용하여 잔차 e_i를 오름차순으로 정렬하여 크기 순으로 $k=p+1$인 초기 부분집합 J를 설정한다(단, p는 X의 계수).

$$e_1 \le e_2 \le, \cdots, \le e_n \text{ 일때,}$$
$$J = \{ y_1, y_2, \cdots, y_k \}, \quad \text{여기서} \quad k=p+1$$

2단계. 최적 부분집합 J의 결정.

앞 단계에서 부분집합에 포함된 원들로 구성된 부분집합 J에서 $k-1$개의 원을 취하고 부분집합 J에 포함되지 않은 원으로 구성된 부분집합에서 1개의 원을 취하여 새로운 부분집

합 J를 구성한다. 이 부분집합 J는 총 $\binom{k}{k-1}$개의 부분집합로 구성되며, k개 중에서 $\min(\text{med } r_i^2)$을 사용하여 부분집합 J를 결정한다.

$J = \{ y_1, y_2, \cdots, y_k \}$과 $\{ y_{k+1}, y_{k+2}, \cdots, y_n \}$에서

$$J_{1(k+1)} = \{ y_2, y_3, \cdots, y_k \} \cup \{ y_{k+1} \},$$
$$J_{1(k+2)} = \{ y_2, y_3, \cdots, y_k \} \cup \{ y_{k+2} \},$$
$$\cdots$$
$$J_{1(n)} = \{ y_2, y_3, \cdots, y_k \} \cup \{ y_n \},$$

$$J_{2(k+1)} = \{ y_1, y_3, \cdots, y_k \} \cup \{ y_{k+1} \},$$
$$J_{2(k+2)} = \{ y_1, y_3, \cdots, y_k \} \cup \{ y_{k+2} \},$$

$$\cdots$$

$$J_{2(i)} = \{ y_1, y_3, \cdots, y_k \} \cup \{ y_n \},$$

$$\cdots$$

$$\cdots$$

$$\cdots$$

$$J_{k(k+1)} = \{ y_1, y_2, \cdots, y_{k-1} \} \cup \{ y_{k+1} \},$$

$$J_{k(k+2)} = \{ y_1, y_2, \cdots, y_{k-1} \} \cup \{ y_{k+2} \},$$

$$\cdots$$

$$J_{k(n)} = \{ y_1, y_2, \cdots, y_{k-1} \} \cup \{ y_n \},$$

$$\min_{\beta} (\text{med } r_i^2)_{J_{ij}}$$

여기서 i : 부분집합 J의 index

j : 나머지 부분집합의 index

3단계. 부분집합 J에서 관측치 분석.

2단계에서 구한 부분집합 J를 사용하여

$$\hat{\boldsymbol{\beta}} = (X_J^T X_J)^{-1} X_J^T \boldsymbol{y}_J$$

$$\sigma^* = \sqrt{\frac{\sum_{i=1}^{n} w_i e_i^2}{\sum_{i=1}^{n} w_i - (k-1)}}$$

여기서 $w_i = \begin{cases} 1 & , |e_i/s^o| \le 2.5 \\ 0 & , \text{otherwise.} \end{cases}$

$$s^o = 1.4826 \left(1 + \frac{5}{n-p}\right) \sqrt{\min \text{med}_j e_i^2}$$

$$\boldsymbol{e} = \boldsymbol{y} - X \hat{\boldsymbol{\beta}}$$

를 구하고 전체자료를 대상으로 표준화된 잔차

$$e_i^* = \left| \frac{e_i}{\sigma^*} \right| > 2.5$$

를 계산하여 e_i^*의 크기가 2.5보다 크면 잔차로 간주한다.

4단계. 부분집합 J의 크기 증가.

e_i^*순으로 전체 자료를 정렬한 후에 다시 최소 잔차 e_i부터 $k+1$개의 원을 선정하고 이들로 부분집합 J를 설정한 다음에 부분집합의 원소 수를 증가시키기 위하여 2, 3, 4단계를

반복 실행한다.

ELMS 알고리즘의 사용시에 발생하는 장점은 다음과 같다. 첫째, 일반적으로 LS나 적절한 로버스트 추정을 사용한 초기 부분집합의 설정시에 반듯이 이상치가 포함되어 있지 않아야 한다는 가정을 배제시킬 수 있다. 더불어 부분집합 J에서도 LS를 사용한 잔차가 다수의 이상치에 의한 은폐에서 포함되더라도 최소 중위수를 갖는 부분집합을 재선정하므로서 이 원을 배제 시킬 수 있다. 둘째, LMS에서와 같은 높은 붕괴점과 아핀등변성(affine equivarance)을 갖는다. 세째, 무작위 탐색 알고리즘에 비하여 매우 낮은 연산횟수 최대 $(n-k)\times(n-p)$회의 계산만을 필요로 한다.

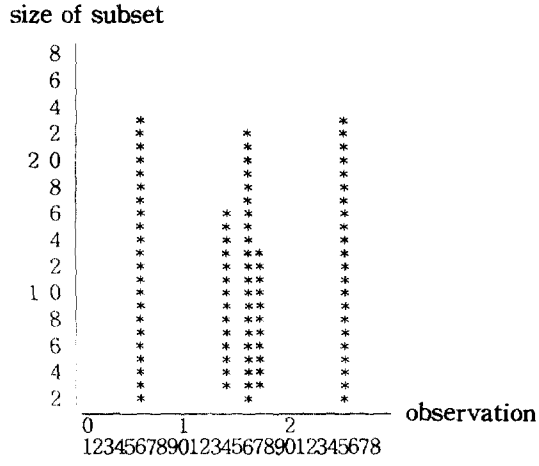
4. 예

회귀진단에서 널리 사용되고 이상치 진단 방법들로서 LS에 의한 표준화잔차, Mahalanobis distance, Rousseeuw and van Zomeren(1990)의 MVE, Hadi(1992)가 제시한 MVE의 개선된 로버스트 추정에 의한 진단, Rousseeuw(1984)의 LMS와 본 논문에서 제시한 ELMS에 의한 이상치 진단을 사용하여 이상치 인식정도를 비교한다(여기서 ELMS에서 사용한 부분집합의 크기는 $[\mathrm{Log}(n^3/p)]+p$개이며, 3가지 자료에 대한 각각의 진단방법에 따른 관측치 분석은 참고문헌 5, 6, 9, 14, 15에서 제시한 자료를 사용한다). ELMS 사용시에 따르는 부분집합 크기에 따른 이상치의 인식 정도를 Atkinson & Mulira(1993)이 제안한 종유석 그림을 이용하여 보인다.

4.1 예1 : Body and Brain Weight Data

body and brain weight 자료는 Weisberg(1980)의 동물에 대한 body weight(grams 단위)와 brain weight(kilograms 단위)의 자료중에 Rousseeuw and Leroy(1987, p. 57)가 28종의 동물을 발취한 $n=28$, $p=2$인 자료를 사용한다. LS에 의한 표준화잔차는 어떠한 관측치도 이상치로서 인식하고 있지 않으며, 고전적인 Mahalanobis distance는 관측치 25를 유일한 이상치로 인식하고 있고, MVE와 Hadi(1992)는 관측치 6, 14, 16, 25만을 이상치로 인식하고 관측치 17은 기각치의 경계에 위치한 것으로 판단하고 있다. 그리고 LMS와 ELMS(부분집합의 크기가 10)일때 잔차는 모두 동일하게 관측치 6, 14, 16, 17, 25를 이상치로 인식하고 있다. 따라서 LMS와 ELMS는 이상치를 정확하게 인식하고 있다. ELMS를 사용시에 각 부분집합의 크기에 따른 이상치의 인식을 〈그림 1〉에 제시한다.

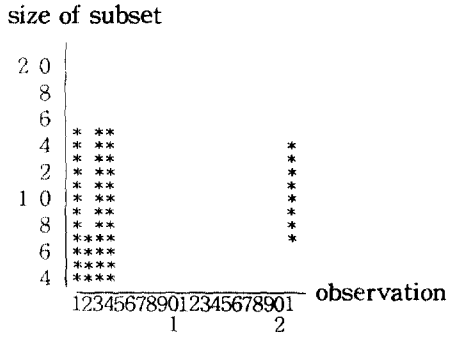
〈그림 1〉에 따른 각 부분집합별 이상치 인식정도는 부분집합의 크기가 3에서 부터 13에 이를 때까지 이상치의 인식을 정확하게 하고 있음을 보이고 있으며, 부분집합의 크기가 13보다 커짐에 따라 은폐효과에 의하여 이상치가 점차 인식되지 않고 있음을 알 수 있다. 크기가 24보다 커질 때는 전체 관측치를 부분집합으로 갖을 때인 LS의 잔차분석과 같이 어떠한 이상치도 인식하지 못하고 있음을 알 수 있다.



〈 그림 1 〉 Body and Brain Weight 자료의 종유석 그림

4.2 예2 : Stack Loss Data

stack loss 자료는 Rousseeuw and Leroy(1987, p. 76)에서 재인용한 것으로 3개의 설명 변수와 1개의 종속변수로 구성되어 있는 선형회귀에서 이상치와 영향력 관측치를 제시하기 위하여 널리 사용되는 자료이다. 이상치 인식정도를 살펴보면 표준화잔차와 Mahalanobis distance는 어떠한 관측치도 이상치로 인식하지 못하고 있으며, MVE와 Hadi(1992)는 관측치 1, 2, 3, 21을 이상치로 나타내고, LMS는 관측치 1, 2, 3, 4, 21을 이상치로 나타내고 있으며, ELMS는 부분집합의 크기가 11일때 관측치 1, 3, 4, 21을 이상치로 인식하고 관측치 2는 인식하지 못하고 있는데, 이는 관측치 2가 경계 영역 바로 아래에 위치한 것이다. 〈그림 2〉는 ELMS에서 각 부분집합의 크기에 따른 이상치의 인식을 보인 것이다.

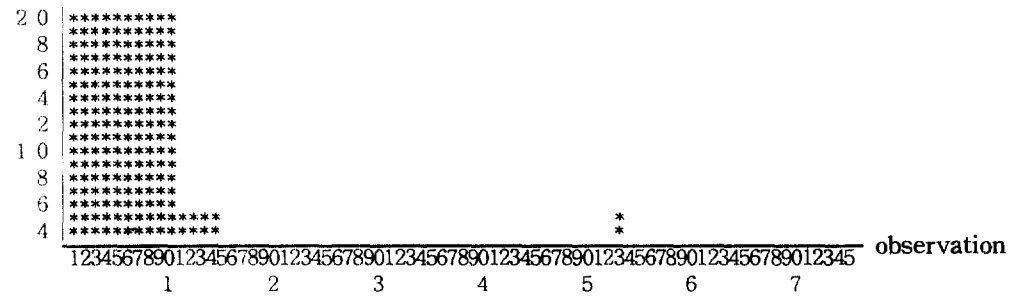


〈 그림 2 〉 Stack Loss 자료의 종유석 그림

〈그림 2〉에 따른 각 부분집합별 이상치 인식정도는 부분집합의 크기가 4에서 부터 6에 이를 때까지 초기 부분집합에 구성된 관측치의 영향에 의하여 관측치 21이 은폐되다가 크기가 7에서 부터 14에 이를 때까지 정확하게 이상치를 인식 하고 있음을 보이고 있으며, 부분집합의 크기가 15보다 커짐에 따라 은폐효과에 의하여 이상치가 점차 인식되지 않고 있음을 보이고 있다.

4.3 예3 : Hawkins, Bradu and Kass Data

Hawkins, Bradu and Kass 자료는 Hawkins et al.(1984)가 $n=75$, $p=3$인 자료를 인위적으로 만든 것으로 10개의 나쁜 지렛대점, 4개의 좋은 지렛대점과 61개의 내부점(inliers)을 포함하고 있는 은폐효과를 보여주는 좋은 예이다. 각 회귀 진단 방법에 의한 이상치 인식은 표준화잔차의 사용시에 관측치 11, 12, 13을 이상치로 인식하고, Mahalanobis distance는 관측치 12와 14를 이상치로 인식하고 이 두개의 이상치가 나머지 이상치들을 은폐시키고 있다. MVE와 Hadi(1992)의 진단방법은 모두 14개의 이상치(관측치 1에서 14까지)를 인식하고 있으나 좋은 지렛대점(11 - 14) 조차도 이상치로 인식하고 있다. 이러한 점은 관측치 11, 12, 13, 14가 큰 잔차를 갖고 있기 때문이다. LMS와 ELMS는 관측치 1부터 10까지를 정확하게 이상치로 인식하고 있다. ELMS를 사용시에 각 부분집합의 크기에 따른 이상치의 인식을 〈그림 3〉에 제시한다.



〈 그림 3 〉 Hawkins, Bradu and Kass자료의 종유석 그림

Hawkins et al. 자료의 종유석 그림은 부분집합의 크기가 20에 달할 때까지로 제한 하였다. 각 부분집합의 크기별 이상치 인식정도는 부분집합의 크기가 4, 5일때 LS 방법에 따른 초기 부분집합 설정으로 인하여 관측치 11, 12, 13, 14, 53을 이상치로 인식하다가 점차 부분집합의 크기가 커지면서 크기가 6에서 부터는 관측치 1에서 부터 관측치 10까지를 이상치로 잘 인식하고 있음을 볼 수 있다.

5. 최적의 SUBSET의 결정

4절에서 제시한 각 자료에 대한 ELMS의 결과는 알고리즘을 실행하면서 중간에 얻어진 결과의 일부이다. 즉, 부분집합 J가 반복되면서 그것의 크기는 원소 수를 점차로 증가시키고 있다. 따라서 어느 크기에서 최적의 이상치를 인식할 수 있는 지를 알아보기 위하여 Atkinson & Mulira(1993)이 제안한 종유석 그림을 사용하여 이상치 인식 상태를 분석하였다. ELMS에서 특정 부분집합의 이상치 인식은 scale factor σ^*에 의하여 결정되어진다. 따라서 부분집합의 크기에 따른 scale factor의 변화를 4절의 예를 사용하여 조사하면 〈표 1〉과 같다. 〈표〉에는 일부의 자료만을 제시하였다. 이 자료와 4절의 종유석 그림을 비교하면 scale factor σ^*를 가장 작게 갖는 크기의 부분집합이 최적의 이상치 인식을 나타내고 있음을 제시하고 있다. 또한 〈그림 1, 2, 3〉과의 비교에서 ELMS 알고리즘에서 사용하는 부분집합의 크기가 $p+2$에서 부터 $[\mathrm{Log}(n^3/p)]+p$ 근처에서 매우 안정적으로 이상치를 인식함을 보이고 있다.

〈 표 1 〉 부분집합의 크기에 따른 scale factor

부분집합의 크기	Body and Brain Weight Data	Stock Loss Data	Hawkins et al. Data
2	.3656611		
3	.2229023		
4	.2267361	2.0000000	.7117496
5	.2329567	2.0754981	.7144174
6	.2393541	2.1602469	.7253957
7	.2460964	1.3605854	.7324927
8	.2534760	1.6892913	.7383335
9	.2633011	1.7871952	.7139819
10	.2728871	1.6248434	.7476532
11	.2816110	2.0829596	.7525358
12	.2948925	2.1808317	.7614278
13	.3055763	2.4573701	.7649756
14	.3255075	2.5542503	.7724974
15	.3379270	3.3241610	.7778235
16	.3620407	4.3680200	.7813042
17	.3914888	5.5202496	.7913709
18	.4204998	7.5002039	.7970885
19	.5261540	7.8126888	.8191242
20	.5761175	9.6000745	.8134183

6. Simulation

ELMS 알고리즘의 안정성을 조사하기 위하여 자료의 수 $n=25$, 독립변수 x의 계수가 $b=1, 2, 3$에서 이상치의 갯수가 0, 3, 6개 일 경우에 인식 정도를 Monte Carlo simulation 방법으로 조사하여 본다[Kalos and Whitlock, 1989]. 여기서 사용하는 최종 판정을 위한 부분집합의 크기는 12로 지정한다.

simulation을 위한 선형회귀방정식은 단순 선형방정식일때 $y_i=\beta_0+\beta_1 x_i+\varepsilon_i$, $i=1, 2, \cdots, 25$, 계수는 $\beta_0=0$, $\beta_1=1$이고, 다중선형방정식은 $y_i=\beta_0+\beta_1 x_i+\beta_2 x_{2i}+\varepsilon_i$, $i=1, 2, \cdots, 25$, 계수는 $\beta_0=0$, $\beta_1=1$, $\beta_2=1$, 각 x_i, $i=1, 2$는 UNIFORM(0, 15)에서 발생된 값으로 각각 (25-이상치 갯수) 만큼을 구한다. 오차항 ε_i는 $N(0, 1)$에서 (25-이상치 갯수)회를 발생시켜 y_i값을 결정한다. 이상치의 발생을 위한 방정식은 $y_i=x_i+8$, $x_i=20-.05(i-1)$, i=이상치 갯수가 0개 일때, 3개 일때, 6개 일때로 하고, 다중선형방정식일때는 $y_i=x_{1i}+x_{2i}+8$, $x_{1i}=20-.05(i-1)$, $x_{2i}=20-.05(i-1)$, i=(이상치 갯수)로 한다. 각각의 결과는 $n=25$에 대하여 100회의 simulation을 하여 이상치의 인식율을 파악한다.

〈 표 2 〉 ELMS의 이상치 인식율

이상치 포함율	단순선형방정식	다중선형방정식
0%	.82	.70
12%	.96	.79
24%	.71	.82

ELMS에 의한 이상치 인식율은 Hadi and Simonoff(1993)의 〈표 2〉에 제시된 여러가지 로버스트 추정량들과 비교할 때, 높은 지렛대점에서 우수한 인식율을 보이고 있으며, LMS에서와 마찬가지로 다중선형방정식 구조에서 비교적 낮은 인식율을 보이고 있다.

7. 결론

이상치와 영향력 관측치를 파악하는 것은 주어진 자료를 분석하는데 결정적인 의미를 지니고 있다. 여기서 제시하고 있는 ELMS 알고리즘을 사용하여 regression outlier를 파악하는 방법은 종전에 이상치 인식을 위해 제시하고 있는 다수의 방법들이 초기 부분집합과 $\hat{\boldsymbol{\beta}}$의 결정시에 부분집합의 안정성에 크게 의존하고 있으나, 본 연구에서 제시하고 있는 ELMS 방법은 부분집합의 재구성과 최소 중위수법(LMS방법)을 사용하여 이상치의 잔차가 LS에 의해 작게 나타날 경우도 부분집합의 구성 관측치에서 제외시킬 수 있는 방법을 제시하고 있다.

LMS 방법에서의 주된 문제점으로 나타나고 있는 비효율성은 연산횟수에 비하여 상대

적으로 매우 작은 연산 처리 시간을 필요로 하고 있다. 이러한 점은 무작위 추출에 의한 반복선택을 취하기 때문이다. 따라서 ELMS에서는 초기 부분집합을 구성하고 이부분집합의 크기를 증가시키는 방법을 취하고 있기 때문에 앞서 제시한 바와 같이 연산횟수를 크게 줄일 수 있다.

ELMS에서의 주된 문제점으로는 부분집합의 크기를 어느 정도로 하는 것이 가장 정확한 이상치 식별 능력을 보일 것인가에 관한 점이다. Atkinson & Mulira(1993)에서 제시하고 있는 종유석 그림을 사용한 부분집합의 크기 평가에서 X의 계수가 p일때, 크기가 $p+2$에서 부터 $[\log(n'/p)]+p$ 근처의 부분집합에서 안정적으로 이상치를 잘 인식하고 있음을 회귀진단에서 널리 사용되고 있는 자료의 사용하여 나타나고 있다. 최적의 부분집합의 크기는 각 부분집합별 scale factor들 중에 최소의 scale factor를 갖는 크기의 부분집합이 이상치 식별에 적합하다.

로버스트 추정량으로서 이상치 인식에 매우 안정적으로 알려진 LMS와의 비교는 6절에서 제시한 simulation 결과와 4절의 예를 통하여 본 바와 같이 LMS와 매우 유사한 결과를 얻을 수 있으며, 잔차의 비교에서 ELMS를 사용할 때의 각 관측치들의 잔차 크기는 LMS에 비하여 낮게 나타난다. Hadi & Simonoff(1993)이 LMS의 문제점으로 지적하고 있는 X공간 상의 상관성이 높을 경우(약 .5이상)에 ELMS에서도 인식의 정도가 떨어짐을 조사를 통하여 알아 볼 수 있었다.

7. 참고문헌

[1] 박성현(1991), 「회귀분석」, 민영사.

[2] Atkinson, A. C., and Mulira, H.-M. (1993), "The Stalactite Plot for the Detection of Multivariate Outliers," *Statistics and Computing*, Vol. 3, pp. 27-35.

[3] Atkinson, A. C. (1994), "Fast Very Robust Methods for the Detection of Multiple Outliers," *Journal of the American Statistical Association*, Vol. 89, No. 428, pp. 1329-1339.

[4] Belsey, D. A., Kuh, E., and Welsch, R. E. (1980), "Regression Diagnostics," *Wiley-Interscience*.

[5] Hadi, A. (1992), "Identifying Multiple Outliers in Multivariate Data," *Journal of the Royal Statistical Society Series-B*, Vol. 54, No. 3, pp. 761-771.

[6] Hadi, A., and Simonoff, J. S. (1993), "Procedures for the Identifying of Multiple Outliers in Linear Models," *Journal of the American Statistical Association*, Vol. 88, No. 424, pp. 1264-1272.

[7] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, E. (1986), *Robust Statistics: The Approach Based on Influence Functions*, John Wiley & Sons, New York.

[8] Hawkins, D. M., Bradu, D., and Kass, G. V. (1984), "Location of Several Outiers in Mutiple Regression Data Using Elemental Sets," *Tecnometrics*, Vol. 26, pp. 197-208.

[9] Hawkins, D. M. and Simonoff, J. S. (1993), "High Breakdown Regression and Multivariate Estimation," *Applied Statistics*, Vol. 42, pp. 423-432.

[10] Hawkins, D. M. (1993), "The Accuracy of Elemental Set Approximations for Regression," *Journal of the American Statistical Association*, Vol. 88, No. 422, pp. 580-589.

[11] Huber, P. J. (1973), "Robust Regression: Asymptotics, conjectures and Monte Carlo," *The Annals of Statistics*, Vol. 1, pp. 799-821.

[12] Kalos, M. H., and Whitlock, P. A. (1989), *Monte Carlo Methods*, Vol. I: Basics, New York: John Wiley & Sons.

[13] Rousseeuw, P. J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, Vol. 79, No. 388, pp. 871-880.

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

[15] Rousseeuw, P. J., and van Zomeren, B. C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association*, Vol. 85, No. 411, pp. 633-639.

[16] Siegel, A. F. (1982), "Robust Regression Using Repeated Medians," *Biometrika*, Vol. 69, pp. 242-244.

[17] Woodruff, D. L., and Rocke, D. M. (1993), "Heuristic Search Algorithms for the Minimum Volume Ellipsoid," *Journal of Computational and Graphical Statistics*, Vol. 2, pp. 69-95.

[18] Woodruff, D. L., and Rocke, D. M. (1994), "Computable Robust Estimation of Multivariate Location and Shape in High Dimension Using Compound Estimation," *Journal of the American Statistical Association*, Vol. 89, No. 427, pp. 888-896.

[19] Yohai, V. J. (1987), "High Breakdown Points and High Efficient Robust Estimates for Regression," *The Annals of Statistics*, Vol. 15, pp. 642-656.

[20] Yohai, V. J., and Zamar, R. H. (1988), "High breakdown-Points Estimates of Regression by Means of Minimization of an Efficient Scale," *Journal of the American Statistical Association*, Vol. 83, pp. 406-413.