▨ 연구논문

다수 이상치 認識을 위한 외향성 검정 절차

염준근
동국대학교 통계학과
김종우
제주교육대학교 수학교육학과

Outward Testing Procedure for the Identification of Multiple Outliers

Joon-Keun Yum
Dept. of Statistics, Dongguk University
Jong-Woo Kim
Dept. of Mathematics Education, Cheju National University of Education

Abstract

This article is concerned with procedures for detecting multiple y outliers in linear regression. The outward-testing procedure, which is controled by the initial subset and the minimum residuals, is suggested by two phases. The performance of this procedure is compared with others by Monte Carlo techniques and found to be superior. The procedure, however, fails in detecting y outliers that are on high-leverage cases in Phase 1. Thus, we proposed ELMS algorithm for a set of suspect observations, in Phase 1. In Phase 2, the proposed testing is conducted using the studentized residuals to see which of the suspect cases are outliers. Several examples are analyzed.

1. 서론

선형회귀분석(linear regression analysis)은 여러 분야에서 자료를 분석하기 위하여 사용된다. 이러한 자료에는 이상치(outlier)와 영향력 관찰점(influential observation)을 흔히 포함하고 있으며, 이들을 인식하는 것은 자료를 분석하는데 매우 중요하다. 만일 자료에 하나의 이상치나 영향력 관찰점만이 존재할 경우는 자료의 특성을 파악하는데 상대적으로 단순하다. 그러나 대부분의 자료에서 문제가 되고 있는 다중 이상치와 다중 영향력 관찰점의 해석과 계산상의 문제점은 매우 복잡하다. 이것은 다중 이상치가 존재할 때 그것들이 군집(cluster)해 있는 방향으로 중심점을 끌어당기는 영향에 의하여 이상치를 숨기려는 은폐(masking)효과와 정상적인 점들이 중심점에서 멀리 떨어져 있는 점으로 인식되는 수렁(swamping)효과에 의해 이상치와 영향력 관찰점의 인식을 어렵게 하고 있기 때문이다. 따라서 자료를 이상치가 없는 부분집합과 잠재적인 이상치를 포함하고 있는 부분집합으로 나누어서 자료를 진단하는 다양하고 효과적인 방법들이 제시되어 왔다. 이러한 회귀 진단(regression diagnosis)분야는 접근 방법에 따라 크게 두 가지 분야로 나눌 수 있다: 즉, 직접접근방법과 로버스트(robust) 적합을 이용한 간접접근방법이 있다.

초기의 직접접근방법은 Prescott(1975)와 Tietjen, Moore와 Beckman(1973)이 제안한 방법으로 하나의 관찰점을 전체 자료에서 "단일 관찰점 진단방법(single case diagnostics)"을 사용하여 절대잔차가 가장 큰 관찰점부터 적은 관찰점순으로 이상치인지를 인식하고 검사하는 Inward testing 방법이다. 이를 활용한 방법으로 Marasinghe (1985)는 "단일 관찰점 진단방법"의 연속적 이용에 의한 잠재적인 K개 이상치 집단을 선정한 후에 이상치 여부를 확인하는 "다단계 접근법(multistage approach)"을 제안했다. Paul과 Fung(1991)은 잠재적인 K개 이상치 집단에 대한 진단으로 Cook의 거리(cook's distance)에 의한 이상치들과 GESR(generalized extreme studentized residuals)에 의한 이상치 집단을 잠재적인 이상치 집단을 대상으로 GESR을 사용할 것을 제안했다. Kianifard와 Swallow(1989)은 "단일 관찰점 진단방법"에 의해 자료를 정렬한 후에 최소 진단치를 갖는 K개 관찰점을 사용하여 이상치를 식별하는 순환잔차(recursive residual) 방법을 제안했다. Fung(1993)은 Rousseeuw(1984)의 LMS 방법이 과다하게 많은 관찰점을 이상치로 처리하는 것에 대한 조치로 수정된 Cook의 거리를 사용하여 이상치 여부를 확인하는 과정을 제안했다. 그러나 이들 진단통계량은 이상치 식별은 다중 이상치가 존재할 때 발생하는 은폐효과와 수렁효과에 의해 크게 영향을 받거나 또는 자료상의 다공선성에 민감한 것으로 알려져 있으며, 잠재적인 이상치 집단의 크기 K에 대한 사전 지식을 요구하는 어려운 점이 있다[[7], [14], [21]] Hadi와 Simonoff(1993)는 Hadi(1992)의 이상치 식별 알고리즘을 사용하여 이상치가 없으리라 예상되는 부분집합과 잠재적으로 이상치가 존재할지도 모르는 부분집합으로 나누어 이상치가 없으리라 예상되는 부분집합을 사용하여 관찰점들을 진단하고 이 부

분집합의 크기를 늘려 나가는 이상치 식별 방법을 제시했다. 이러한 외향성 검정(outward testing)은 붕괴점이 50%에 달하는 것으로 알려져 있다[Barnett와 Lewis, 1984; Davies와 Gather, 1993].

이상치 인식을 위한 간접방법으로 은폐효과와 수렁효과를 극복하기 위하여 최근에 널리 사용되고 있는 로버스트 추정량으로는 $(p+1)$개의 기저집합(elemental set)을 사용하여 잔차제곱의 중위수를 최소화하는 추정량을 선택하는 Rousseeuw(1984)의 LMS(least median of squares) 추정량과 Rousseeuw와 van Zomeren(1990)의 MD(mahalanobis 거리)의 $C(X)$와 $S(X)$에 로버스트 추정량을 사용하는 MVE(minimum volume ellipsoid)가 있다.

그러나 LMS와 MVE는 높은 붕괴점(breakdown point)을 갖는 추정량이지만 해석적으로 LMS는 이상치를 과도하게 지정하고 있으며[Atkinson, 1986; Fung, 1993], MVE는 선정된 기저집합이 최소 볼륨을 갖는 타원을 형성하는 지에 의심을 갖게 한다[Cook와 Hawkins, 1990]. 또한 이 추정량을 계산하기 위한 방법으로 제시되고 있는 무작위 재추출 방법(resampling method 혹은 random search algorithm)은 모수 추정량을 얻기 위하여 지나치게 많은 연산 횟수를 필요로 하고 있다[Woodruff와 Rocke, 1994]. Hadi(1992, 1994)는 MVE를 계산할 때 발생하는 표본 추출의 횟수를 줄이기 위한 방법으로 로버스트한 위치(location)와 척도(scale)를 갖는 추정량을 사용하여 기저집합에 원들을 추가해 가면서 표본을 선택함으로써 MVE를 구하는 알고리즘을 제시하고 있으며, Woodruff와 Rocke(1994)은 무작위 재추출 방법을 개선한 발견적 탐색(heuristic search)을 제안하였다. Atkinson(1994)는 LMS의 기저집합을 한정된 횟수의 무작위 부분집합으로 하고, 이를 사용하여 최소잔차를 갖는 원을 이 초기집합에 포함시켜 집합의 크기를 늘려 나가면서 잔차들의 변화를 그림으로 인식하는 알고리즘을 제시하였다.

본 연구에서는 선형회귀 구조를 갖는 모집단에서 이상치가 없으리라 예상되는 부분집합(clean subset)과 나머지들의 부분집합(remained subset)으로 자료를 분리하기 위하여 확장된 LMS(ELMS) 방법을 사용하고, clean subset을 사용하여 remained subset에서 이상치를 Bonferroni t-검정 방법을 이용하여 식별하고 정상적인 관찰값을 clean subset에 포함시켜 나아가는 외향성 검정을 제시하고자 한다.

선형회귀모형을 다음과 같이 설정하자.

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

여기서 $\boldsymbol{Y} = (y_1, y_2, \ldots, y_n)^T$는 반응변수인 $n \times 1$ 벡터이다.

$\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^T$는 $1 \times p$ 벡터인 $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$를 행으로 갖는 설명변수인 $n \times p$ 행렬이다 (단, $p < n$).

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

$\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n)$인 $n \times 1$ 벡터이다(즉, $E(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) = 0$ 이고 $\mathrm{var}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \sigma^2 \mathbf{I}_n$, $\boldsymbol{I}_n$은 계수 n인 단위 행렬).

이때, 최소제곱법(method of least squares: LS)에 의한 $\boldsymbol{\beta}$와 σ^2의 최소제곱 추정량은

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y},$$

$$\hat{\sigma}^2 = \boldsymbol{e}^T\boldsymbol{e}/(n-p),$$

여기서 $\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}$,

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$$

이다. 본 논문에서 사용할 잔차의 변형은

$$e_i / \hat{\sigma}\sqrt{1 - h_{ii}}$$

으로, 여기서 h_{ii}는 Hat 행렬 $\boldsymbol{H}$의 대각선상의 원소이다.

2. 제안된 방법

앞 절에서 언급한 바와 같이 이상치를 찾는 직접접근 방법은 자료를 이상치가 없으리라 예상되는 부분집합과 자료 전체에서 clean subset을 제외하고 남아 있는 나머지 부분집합으로 분리한 다음에 clean subset을 사용하여 이상치를 찾는 것이다. clean subset에 포함된 관찰점의 색인(index)으로 구성된 집합을 M이라면, 자료에서 이 색인들에 해당하는 설명변수와 반응변수만을 추출하여 각각 $\boldsymbol{Y}_M$, $\boldsymbol{X}_M$을 구성할 수 있고, 이에 대응하는 $\boldsymbol{\beta}_M$과 σ_M^2의 최소제곱 추정량은

$$\widehat{\boldsymbol{\beta}_M} = (\boldsymbol{X}_M^T\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M^T\boldsymbol{Y}_M.$$

$$\hat{\sigma}_M^2 = e_M^T e_M/(k-p), \quad \text{여기서} \quad e_M = Y - X\hat{\beta}_M$$

이다. 색인집합 M의 크기를 결정하는 한 가지 방법으로 Gentleman과 Wilk(1975)는 크기 k인 부분집합을 제거할 때 잔차의 합이 가장 크게 축소하는 값으로 할 것을 제안했다. 즉, remained subset의 색인집합 $M^c = \{i_1, \ldots, i_k\}$에 대응하는 관찰점들의 제거에 따르는 잔차제곱합은 다음과 같다.

$$Q_k = SSE - SSE_{(M^c)}$$

$$= e_{M^c}^T (I_{M^c} - H_{M^c})^{-1} e_{M^c},$$

여기서 $SSE_{(M^c)}$는 색인집합 M^c에 대응하는 관찰점이 제거된 잔차제곱합이고,

e_{M^c}는 색인집합 M^c에 대응하는 자료의 잔차집합이며,

H_{M^c}는 H에서 색인집합 M^c에 대응되는 관찰점들로 구성된 부분행렬이다.

이것은 곧 clean subset을 최소잔차합을 갖는 크기 $(n-k)$인 부분집합을 사용하여 구하는 것과 같다. 그러나 이러한 접근에서 k는 사전에 거의 알려져 있지 못하고, 설령 알고 있다 하더라도 크기 $(n-k)$인 부분집합을 구하기 위한 계산은 nCk가지의 경우에 달하게 된다. 따라서 본 논문에서는 최소잔차합을 갖는 부분집합을 구하기 위하여 다음과 같은 세 단계의 과정을 제안한다.

제안된 알고리즘

1 단계. 초기집합으로 적절한 방법(본 연구에서는 ELMS를 사용)을 사용하여 clean subset을 구성하고, 이에 따른 색인집합 M을 구한다. 이때, M의 크기는 $k = [(n+p-1)/2]$이다(단, $[\cdot]$는 자신보다 크지 않은 최대 정수).

2 단계. 잔차의 계산.

2.1. 초기집합의 크기가 k일 때 다음 식을 만족하는 크기 $(k+1)$인 부분집합을 구하고 이에 따른 색인집합 M을 구한다.

$$\min_{M_i} \sum (y_i - x_i^T \hat{\beta}_{M_i}) \quad \text{여기서} \quad M_i = M \bigcup \{i\}, \quad i = (k+1), \ldots, n.$$

2.2 각 관찰점마다 다음과 같은 변형된 잔차를 계산한다.

$$i \in \boldsymbol{M}\text{일 때}, \quad d_i = r_i / s_{(i)}\sqrt{1-h_{ii}}\ , \tag{1}$$

$$i \notin \boldsymbol{M}\text{일 때}, \quad d_i = r_i / s\sqrt{1+h_{ii}}\ , \tag{2}$$

여기서 $r_i = y_i - x_i^T \hat{\boldsymbol{\beta}}_{\boldsymbol{M}}$,

$$(k-p-1)s_{(i)}^2 = (k-p)s^2 - r_i^2/(1-h_{ii})\ ,$$

$$h_{ii} = \boldsymbol{x}_i^T(\boldsymbol{X}_{\boldsymbol{M}}^T \boldsymbol{X}_{\boldsymbol{M}})^{-1}\boldsymbol{x}_i\ ,$$

$$s^2 = \sum_{i=1}^{k} r_i^2/(k-p)\ .$$

(1)은 외적 스튜던타이즈드 잔차(externally studentized residual)이고, (2)는 부분집합 M을 사용한 예측 잔차(predicted residual, 혹은 deleted residuals)이다[Cook와 Weisberg 1982, p.20, p.33].

3 단계. 이상치 검정을 위하여 절대 잔차 $|d_{k+1}|$를 Bonferroni t-검정의 임계값 $t_{(\alpha/2(k+1),k-p-1)}$와 비교하여 임계값보다 적을 경우에 clean subset에 이 관찰점의 색인을 포함시켜 clean subset M의 크기를 증가시키고, 새로운 M을 사용하여 2 단계를 실행시킨다. 그렇지 않은 관찰점은 잠재적인 이상치가 포함된 remained subset으로 하는 과정을 반복 실행한다.

여기서 remained subset의 잔차들이 임계값에 의한 검정에서 더 이상 변화를 갖지 못할 때 즉, 색인집합 M이 더 이상 원을 추가시키지 못하면 remained subset을 이상치들의 집합으로 결정하고 실행을 종료한다.

σ^2의 추정량으로서 Hadi와 Simonoff(1993)는 로버스트한 $\hat{\sigma}_M^2$을 사용하여 전체 자료의 잔차를 계산하고 $(k+1)$번째 관찰점의 잔차의 크기를 임계값 $t_{(\alpha/2(k+1),k-p)}$와 비교하여 점진적으로 reminded subset에서 이상치를 검정하는 방법을 택하고 있다. 그러나 이러한 방법은 clean subset의 초기집합의 성향에 크게 의존하게 된다. 즉, 최

초에 결정된 $\hat{\beta}_M$은 이 방향의 원을 중심으로 clean subset에 원을 포함시키기 때문이다.

본 연구에서 제안한 방법은 초기 clean subset에 remained subset의 원을 하나씩 포함시킨 M_i중에 최소잔차합을 갖는 M_i를 사용하여 $\hat{\beta}_{M_i}$를 결정하므로 이상치들의 집단에 덜 민감하게 된다. 이러한 adding-back방법은 Atkinson(1986), Fung(1993)의 확인(confirm) 과정에서 효과적임을 보이고 있다. 또한 초기집합을 설정하기 위해 Hadi와 Simonoff(1993), Hadi(1992, 1994)은 MVE방법을 사용하여 자료 중심에 있는 관찰점을 취하는 방법을 제시하였다. 그러나 이 방법은 높은 레버레지 관찰점 즉, 1 방향의 이상치의 인식에 민감하지 못한 점을 나타내고 있다. 따라서 다음 절에 제시하고 있는 LMS방법을 응용한 ELMS 방법을 사용하여 극복시키고자 하였다.

2.1 초기 CLEAN SUBSET

Rousseeuw(1984)가 제안한 LMS는 $\hat{\beta}_M$을 결정하기 위해 자료에서 크기 $(p+1)$인 부분집합을 무작위로 추출하여 다음과 같은 최소중위수잔차를 갖는 부분집합을 기저집합 M으로 선택하는 것이다. 이 추정량은 50%에 달하는 붕괴점을 갖고 있으나 오차의 정규성 가정 아래서 매우 낮은 효율성을 갖고 있다.

$$\underset{\hat{\beta}_M}{\text{Minimize}} \ \underset{M}{\text{med}} \ r_i^2 ,$$

여기서 $\hat{\beta}_M = (X_M^T X_M)^{-1} X_M^T Y_M$,

$$r_i = y_i - x_i \hat{\beta}_M .$$

Atkinson(1994)가 제시하고 있는 것처럼 적절하게 선택된 기저집합 M의 크기를 확장시키는 것은 보다 빠르게 최소잔차집합에 도달할 수 있으며, Hadi와 Simonoff(1993)가 지적하고 있는 X변량 사이에 높은 상관성을 갖는 자료에서도 안정성을 갖게 한다. 따라서 본 연구에서는 LMS의 기법을 확장한 다음과 같은 확대된 LMS(ELMS) 방법을 제안한다. 이 방법은 다음과 같은 3단계로 구성되어 있다.

ELMS 알고리즘

1 단계. 초기 색인집합 M의 설정.

전통적인 최소제곱법을 사용하여 절대잔차 $|r_i|$를 구하고 이를 오름차순으로 정렬하

여 적은 순으로 크기 $k(=p+1)$인 초기 색인집합 M을 초기 기저집합으로 설정한다 (단, p는 X의 계수)

$|r_{i_1}|_{1:n} \le |r_{i_2}|_{2:n} \le , \ldots , \le |r_{i_n}|_{n:n}$일 때,

$M = \{ i_1, i_2, \ldots , i_k \}$ 이다.

2 단계. 최적 색인집합 M의 결정.

앞 단계에서 부분집합에 포함된 원들로 구성된 색인집합 M에서 $(k-1)$개의 원을 취하고, 나머지 부분집합에서 1개의 원을 취하여 새로운 색인집합 $M_{(i)j}$를 구성한다 이 색인집합 $M_{(i)j}$은 총 $k\times(n-k)$개이며, 이들 중에서 최적 색인집합 M의 결정은 (1)에서 제시한 방법으로 $\hat{\beta}_{M_{(i)j}}$를 구하고, 이들 중에서 LMS에서 제시한 최소잔차중위수 min $(\text{med } r_i^2)_{M_{(i)j}}$를 갖는 $M_{(i)j}$를 선택한다.

$M = \{ 1, 2, \ldots , k \}$ 와 $M^c = \{ k+1, k+2, \ldots , k+n \}$ 에서

$M_{(i)j} = M_{(i)} \bigcup M_j^c$.

여기서 $M_{(i)} = M - \{ i \}$, $i = 1, \ldots , k$,

$M_j^c = \{ j \}$, $j = k+1, k+2, \ldots , n$.

$$\underset{\hat{\beta}_{M_{(i)j}}}{\text{Minimize}} \ \underset{M_{(i)j}}{\text{med}} \ r_i^2 .$$

여기서 $\hat{\beta}_{M_{(i)j}} = (X_{M_{(i)j}}{}^T X_{M_{(i)j}})^{-1} X_{M_{(i)j}}{}^T Y_{M_{(i)j}}$,

$r_i = y_i - x_i \hat{\beta}_{M_{(i)j}}$.

3 단계. 색인집합 M의 크기 증가.

최적 색인집합 $M_{(i)j}$를 사용하여 구한 $\hat{\beta}_{M_{(i)j}}$에 의하여 결정된 색인집합의 원소 수가 $[(n+p-1)/2]$이 될 때까지 크기 k인 M에 $|r_i|_{(k+1):n}$에 해당하는 색인 i를 포함시켜 색인집합 $M_{(i)j}$의 크기를 하나 증가시키고, 2 단계를 반복 실행한다.

ELMS 알고리즘의 사용은 일반적으로 LS나 적절한 로버스트 추정을 사용한 초기

($p+1$)개 부분집합의 설정할 때 반드시 이상치가 포함되어 있지 않아야 한다는 가정을 배제시킬 수 있으므로 높은 붕괴점을 갖게 된다[염준근 외2, 1995].

3. 예

예1: Rousseeuw 자료.

Rousseeuw와 Leroy(1987, pp.67-68)에서 발췌한 Rousseeuw의 자료는 LMS 방법의 높은 붕괴점을 보여 주기 위하여 자료의 크기 $n=50$, $p=1$에서 정상적인 점 30개와 이상치 20개를 인위적으로 만든 것이다. 각 회귀 진단 방법에 의한 이상치 인식은 Huber의 M 추정량, Mallows의 GM 추정량, Sheweppe의 GM 추정량은 모두 높은 비율의 이상치에 의하여 정상적인 점들로 구성하는 $\hat{\beta}$를 구하지 못하고 있으며 [Rousseeuw와 Leroy, 1987], 또한 Cook(1986)의 국부영향법(local influence method)이 무의미함을 보이고 있다. 본 연구에서 제안한 방법은 Rousseeuw(1984)와 마찬가지로 잘 지적하고 있다<표 1 참조>.

예2: Hadi와 Simonoff 자료.

Hadi와 Simonoff(1993)의 자료는 LMS같은 추정량이 이상치 인식에서 과도하게 이상치를 인식함을 지적하기 위한 자료로서 크기 $n=25$, $p=2$에서 두 설명변수간에 상관성이 0.5인 관계를 갖고, 이상치를 3개와 22개의 정상적인 자료를 인위적으로 만든 것이다. 각 회귀 진단 방법에 의한 이상치 인식은 LMS의 경우에 관찰점 1, 2, 3, 5, 11, 13, 17, 24를 이상치로 인식하며, LMS에 의한 reminded subset을 사용하는 Paul과 Fung(1991)의 2 단계 과정은 이상치가 큰 잔차를 갖지 못하므로 이상치 인식에서 실패하고 있다. 그밖에 LTS, RWLS, MM은 관찰점 1, 2, 3을 다른 잔차에 비해서 큰 값을 갖게 하지만 의미가 있을 정도의 값을 나타내지 못하고 있다. 그러나 본 연구에서 제안한 방법은 관측치 1, 2 ,3을 Hadi와 Simonoff(1993)의 방법과 마찬가지로 잘 지적하고 있다<표 1 참조>.

예3: Body와 Brain 무게 자료

body와 brain 무게 자료는 Weisberg(1980)의 동물에 대한 body 무게(grams 단위)와 brain 무게(kilograms 단위)의 자료 중에 Rousseeuw와 Leroy(1987, p.57)가 28종의 동물을 발취한 $n=28$, $p=2$인 자료를 사용한다. LS에 의한 표준화잔차는 어떠한 과

측치도 이상치로서 인식하고 있지 않으며, 고전적인 Mahalanobis 거리는 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 LMS는 관측치 6, 14, 16, 17, 25를 이상치로 인식하고 있다. 본 연구에서 제안한 방법은 관측치 6, 16, 25를 이상치로 인식한다<표 1 참조>.

예4: Stack Loss 자료

stack loss 자료는 Rousseeuw와 Leroy(1987, p.76)에서 재인용한 것으로 3개의 설명변수와 1개의 종속변수로 구성되어 있는 선형회귀에서 이상치와 영향력 관측치를 제시하기 위하여 널리 사용되는 21개로 구성된 자료이다. 이상치 인식 정도를 살펴보면 표준화잔차와 Mahalanobis 거리는 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, 본 연구에서 제안한 방법은 관측치 1, 3, 4, 21을 이상치로 인식한다<표 1 참조>.

예5: Hawkins, Bradu와 Kass 자료

Hawkins, Bradu와 Kass 자료는 Hawkins 등(1984)이 $n = 75$, $p = 3$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 표준화잔차를 사용할 때 관측치 11, 12, 13을 이상치로 인식하고 Mahalanobis 거리는 관측치 12와 14를 이상치로 인식하고 이 두개의 이상치가 나머지 이상치들을 은폐시키고 있다. MVE와 Hadi(1992)의 진단 방법은 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. 본 연구에서 제안한 방법은 관측치 1부터 관찰점 10까지를 이상치로 인식한다<표 1 참조>.

< 표 1 > 예1 - 예5 자료의 제안한 방법에 의한 잔차표

관찰점	예1	예2	예3	예4	예5
1	0.1251666	3.7479845	-0.398568	3.8366437	16.068328
2	-0.884067	3.7551479	-1.040126	1.8102881	16.86458
3	0.976554	3.7623053	-0.095937	4.3214838	16.705118
4	-0.05449	1.5984253	0.1335065	6.0167932	15.608625
5	-0.261406	0.1145397	-0.672682	-0.543801	16.412158
6	0.1636334	-1.969949	6.648511	-1.06253	16.354672
7	1.407014	0.7406542	.5597553	-0.378033	17.79895
8	-0.563597	0.4705039	-0.06683	0.5169801	17.071868
9	1.6675992	0.0588329	-0.525858	-0.923089	15.801307
10	-1.351406	1.795753	.2272122	0.3167325	16.291226
11	-1.448047	-2.071913	0.2691636	0.8493969	-0.133961
12	1.2230168	-0.059149	0.488777	0.433914	-0.464598
13	-1.904231	-0.440356	0.216677	-2.724301	1.2926082
14	0.6949897	0.6344862	3.2317587	1.254347	-0.611395
15	-1.551859	0.8034159	-0.181147	1.2400989	-0.931051
16	0.4136632	0.8124197	-6.045432	0.1193925	0.8510999
17	-0.825262	-1.448047	2.4895887	-0.381786	-0.132813
18	0.8010274	1.0305945	-1.13185	0.081685	0.059053
19	0.1586514	-1.685902	-0.821776	0.5119953	0.3301286
20	0.1494755	-0.241806	-0.351108	1.7332089	0.5604391
21	1.9685125	0.2105931	-0.48351	-5.592254	1.6237539
22	0.1310467	0.2687907	-0.009637		0.7666665
23	0.0894939	0.3915379	-0.782647		-1.534599
24	-2.75536	-0.930108	1.3811205		1.2970819
25	0.465032	-0.398809	-6.804101		-0.32472
26	0.1662957		-0.799805		-1.274973
27	0.420497		0.7839223		-1.364236
28	-0.182564		-1.316952		0.7339725
29	-0.236969				0.7000952
30	0.5675821				-0.127027
31	-28.9119				-0.139657
32	-27.336				-0.620237
33	-29.18286				-1.070916
34	-22.86211				0.940587
35	-22.76364				0.5223366
36	(36-50 생략)				(36-75 생략)

(주) 이상치 예1: $|r_i|>$ 3.49, 예2: $|r_i|>$ 3.54, 예3: $|r_i|>$ 3.50, 예4: $|r_i|>$ 3.68, 예5: $|r_i|>$ 3.55

4. 모의실험

제안된 알고리즘의 안정성을 조사하기 위하여 단순 선형방정식일때

$$y_i = \beta_o + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, 25.$$

여기서, $\beta_o = 0$, $\beta_1 = 1$,

$$x_i \sim U(0, 15),$$

$$\varepsilon_i \sim N(0, 1)$$

에서 정상적인 관찰점을 25개 구하고, 이상치 발생을 위하여

$$y_i = x_i + 4, \quad x_i = k - .05(i-1),$$

$$y_i = x_i - 4, \quad x_i = k - .05(i-1),$$

여기서 $k = -5, 7.5, 20,$

$$i = 1, 2, 3, 4, 5$$

를 사용하여 이상치 개수가 5개일 때 대하여 1000회의 모의실험을 하여 이상치의 인식율을 파악한다.

< 표 2 > 제안된 방법(PW)과 Hadi와 Simonoff(1993) 방법(HS)의 비교

계획된 이상치 위치(x축,y축)		이상치를 모두 정확히 인식(p_1)	적어도 1개의 이상치인식(p_2)	정상점을 이상치로 잘못 인식(p_3)
Null	PW	.941*1	.000	.059
	HS	.940*1	.000	.060
(-5,-4)	PW	.444	.468	.051
	HS	.430	.479	.077
(-5,+4)	PW	.436	.455	.053
	HS	.426	.479	.077
(7.5,-4)	PW	.771	.800	.029
	HS	.749	.807	.058
(7.5,+4)	PW	.773(.791)	.804(.807)	.031(.038)
	HS	.752(.762)	.815(.807)	.063(.065)
(20,-4)	PW	.456	.472	.046
	HS	.431	.484	.078
(20,+4)	PW	.437(.464)	.467(.481)	.054(.041)
	HS	.423(.448)	.481(.495)	.075(.053)

(주) 1. *1 계획된 이상치가 없을 때 이상치가 없음으로 나타난 비율임.
2. ( )계획된 이상치가 1개일 때 비율임
3. 유의수준 $\alpha = 0.05$를 사용함.

P(masking) = $1 - p_2$, P(swamping) = p_3이다. 따라서 p_1과 p_2는 높고 p_3는 적을수록 좋은 방법이 될 것이다. 모의실험 결과는 일부를 제외하고는 대부분의 경우에 있어서 Hadi와 Simonoff(1993)의 방법인 HS보다 본 연구에서 제시하고 있는 PW 방

법이 더 높은 p_1과 p_2를 나타내고 있으며, p_3는 적게 유지하고 있다. 이러한 효과는 기저집합인 clean subset에서 원을 추가할 때, 최소잔차합을 갖는 원을 선택하는 것에 기인한다. 그 밖에 다른 이상치 식별 방법인 LMS, RWLS, LTS, MM95, MM70와의 비교 결과는 Hadi와 Simonoff(1993)에 제시되고 있다.

5. 결론

이상치와 영향력 관찰점을 파악하는 것은 주어진 자료를 분석하는데 결정적인 의미를 지니고 있다. 여기서 제안하고 있는 기저집합의 원을 증가시켜 나아가는 외항성 검정을 사용하여 이상치를 파악하는 방법은 앞 절의 예와 모의실험에서 보듯이 종전에 이상치 인식을 위해 제시하고 있는 다수의 방법들이 초기 부분집합과 β를 결정할 때 사용하는 부분집합의 안정성에 크게 의존하고 있으나, 본 연구에서 제안한 방법은 기저집합을 재구성하여 clean subset을 구성하고, 이 부분집합을 사용하여 외적 스튜던타이즈드 잔차와 예측 잔차를 계산하여 새롭게 기저집합을 구성하므로서 다중이상치에 의한 영향에서 벗어나도록 하였다.

그러나 나머지 집합에서 최소잔차합을 갖는 원을 취하여 clean subset에 포함시키는 방법은 초기 기저집합의 성향을 크게 개선하지 못하고 있으므로 향상된 방법이 필요하다고 여겨진다.

참고문헌

[1] 염준근, 박종구, 김종우(1995), "다변량 자료에서 다수 이상치 인식의 절차", 품질경영학회지, 제23권, 제4호, pp. 28-41.

[2] Atkinson, A. C.(1986), "Masking Unmasked," *Biometrika,* Vol. 73, No. 3, pp. 533-541.

[3] ——— (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] Barnett, V. and Lewis, T.(1984), *Outliers in Statistical Data* (2nd ed.), John Wiley & Sons, New York.

[5] Cook, R. D.(1986), "The Assessment of Local Influence(with discussion)," *Journal of the Royal Statistical Society,* Series-B, Vol. 48, pp. 133-169.

[6] Cook, R. D. and Hawkins, D. M.(1990), "Comment on Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association,*

Vol. 85, No. 411, pp. 640-644.

[7] Cook, R. D. and Weisberg, S.(1982), *Residuals and Influence in Regression,* Chapman and Hall, London.

[8] Davies L. and Gather U.(1993), "The Identification Multiple Outliers(with discussion)," *Journal of the American Statistical Association,* Vol. 88, No. 423, pp. 782-801.

[9] Fung, W-K.(1993), "Unmasking Outliers and Leverage Points: A Confirmation," *Journal of the American Statistical Association,* Vol. 88, No. 422, pp. 515-519.

[10] Gentleman, J. F and Wilk, M. B.(1975), "Detecting Outliers II: Supplementing the Direct Analysis of Residuals," *Biometrics,* Vol. 31, pp. 387-410.

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

[12] ——— (1994), "A Modification of a Method for the Detection of Outliers in Multivariate Samples," *Journal of the Royal Statistical Society,* Series-B, Vol. 56, No. 2, pp. 393-396.

[13] 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.

[14] 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.

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

[16] Kianifard F. and Swallow, W. H.(1989), "Using Recursive Residuals, Calculated on Adaptively Ordered Observations, to Identify Outliers in Linear Regression," *Biometrics,* Vol. 45, pp. 571-585.

[17] Marasinghe, M G.(1985), "A Multistage Procedure for Detecting Several Outliers in Linear Regression," *Techometric,* Vol. 27, No. 4, pp. 395-399.

[18] Paul, S.R. and Fung, K.Y.(1991), "A Generalized Extreme Studentized Residual Multiple-Outlier-Detection Procedure in Linear rgression," *Techometric,* Vol. 33, No. 3, pp. 339-348.

[19] Prescott, P.(1975), "An Approximation Test for Outliers in Linear Models," *Techometric,* Vol. 17, pp. 129-132.

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

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

[22] 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.

[23] Tietjen, G. L., Moore, R. H. and Beckman, R. J.(1973), "Testing for a Single Outlier in Simple Linear Regression," *Techometric,* Vol. 15, pp. 717-721.

[24] Woodruff, D. and Rocke, D. M.(1994), "Computable Robust Estimation of Multivariate Location and Shape in High Dimension Using Compound Estimator," *Journal of the American Statistical Association,* Vol. 89, pp. 888-896. No. 424, pp. 1264-1272.