응용통계연구 제 15권 2호, 2002년, pp. 269-279

다중 선형 모형에서 식별된 다중 이상점과 다중 지렛점의 재확인 방법에 대한 연구

유종영[1] 안기수[2]

요 약

다중 이상점과 다중 지렛점의 식별은 가장효과(masking effect)와 편승효과(swamping effect)에 영향을 받으므로 어려움이 존재한다. Rousseeuw와 van Zomeren(1990)은 LMS (Least Median of Squares) 회귀방법과 MVE(Minimum Volume Ellipsoid) 통계량을 이용하여 다중 이상점과 다중 지렛점을 식별하였다. 그러나 이들의 방법은 LMS와 MVE의 강한 로버스트성으로 인하여 이상점과 지렛점이 아닌 점들도 이상점과 지렛점으로 식별하는 경향이 있다. Fung (1993)은 식별된 이상점과 지렛점들에 대하여 재확인 방법을 제안하였는데 이 방법은 인근효과(adjacent effect)에 영향을 받아 이상점과 지렛점을 식별하는데 문제가 있는 것으로 분석되었다. 본 논문은 이러한 문제점을 지적하고 새로운 방법을 제안하여 식별된 이상점과 지렛점을 재확인하고자 한다.

주요용어: 이상점, 지렛점, 가장효과, 편승효과, 인근효과

1. 서론

우리는 선형모형에서 이상점과 지렛점을 식별하였을 경우 이 점들에 대한 재확인 방법에 대하여 논의하고자 한다. 일반적으로 이상점은 다수 자료의 형태에 따르지 않는 특정한 자료의 집단을 의미하며 지렛점은 설명변수의 자료 중에서 다수의 자료로부터 멀리 떨어져 있는 소수 집단의 자료를 의미한다. 이런 이상점과 지렛점은 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 이상점 혹은 지렛점이 있다면 이 점을 식별하는 데는 이론적으로나 계산상에 별 문제가 없다. 그러나 두 개 이상의 이상점이나 혹은 지렛점이 있을 경우 가장효과(masking effect)와 편승효과(swamping effect) 때문에 이러한 점들의 식별이 어려워진다. 가장효과란 두 개 이상의 이상점 혹은 지렛점이 근접한 곳에 위치하여 서로 상대방을 이상점 혹은 지렛점으로 식별되지 못하도록 방해하는 것을 말하며, 편승효과는 이상점 혹은 지렛점이 아닌 점들을 이상점 혹은 지렛점으로 식별될 때 나타나는 효과를 의미하고 있다. 이러한 편승효과와 가장효과를 줄이기 위하여 최근 들어 많은 로버스트한 방법들이 제안되고 있으며, Rousseeuw와

1) (449-714) 경기도 용인시 삼가동 470, 용인대학교 컴퓨터정보학부, 부교수
E-mail: jyyoo@yongin.ac.kr
2) (440-714) 경기도 수원시 장안구 695-1, 동남보건대학 컴퓨터응용과, 조교수
E-mail: ksahn@dongnam.ac.kr

van Zomeren(1990)은 LMS(Least Median of Squares) 회귀방법과 MVE(Minimum Volume Ellipsoid) 통계량을 이용하여 이상점과 지렛점을 식별하였다. 우리는 다음과 같은 일반적인 선형회귀모형에 관심을 갖고 있다.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1.1}$$

위에서 $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)^T$는 $n \times 1$ 의 반응변수의 벡터값, $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^T$는 $(p+1) \times 1$ 의 모수 벡터, $\boldsymbol{X} = (\boldsymbol{x_1^T}, \boldsymbol{x_2^T}, \cdots, \boldsymbol{x_n^T})^T$는 $n \times (p+1)$의 설명변수 행렬로 계수는 $(p+1)$이다(단, $p+1 < n$). 또 $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)^T$는 $n \times 1$인 오차항 벡터로 기대값은 $\mathbf{0}$, 분산은 $\sigma^2 \boldsymbol{I_n}$인 정규분포를 따른다고 가정하며, 여기서 $\boldsymbol{I_n}$은 차수 n인 단위행렬을 의미한다.

Rousseeuw(1984)는 회귀진단에 사용하는 모형을 기존의 최소제곱법과는 다른 개념으로 LMS를 제시하였다. LMS는 식 (1.2)를 만족하는 추정량 $\widetilde{\boldsymbol{\beta}}$로 ; $\widetilde{\varepsilon_i}(\widetilde{\boldsymbol{\beta}})$는 i번째 관측값의 잔차를 나타내고 있다.

$$\text{minimize}(\widetilde{\boldsymbol{\beta}}) \text{ median } \widetilde{\varepsilon_i}^2(\widetilde{\boldsymbol{\beta}}) \tag{1.2}$$

LMS 추정량은 붕괴점(breakdown points)이 거의 50% 에 달하는 강한 로버스트성의 장점을 갖는 반면, 계산상에 있어서는 효율성이 낮은 것을 분석할 수 있으며(Rousseeuw 1984), LMS의 분산은 $\widetilde{\sigma} = k\sqrt{median_i\, \widetilde{\varepsilon_i}^2(\widetilde{\boldsymbol{\beta}})}$ 로 추정할 수 있는데 이때 k는 양의 상수이다(Rousseeuw 1984, Rousseeuw와 Leroy 1987, Rousseeuw와 van Zomeren 1990). 이를 이용하여 Rousseeuw와 Leroy은 표준화된 LMS 잔차 $\widetilde{\varepsilon_i}/\widetilde{\sigma}$를 임계값 ± 2.5를 사용하여 이상점을 탐색하였다. 지렛점을 탐색하는 방법으로 Rousseeuw와 van Zomeren은 위에서 정의한 $\boldsymbol{x_i^T}$를 $(1, \boldsymbol{z_i^T})$로 분할하였을 때 임계값이 $\sqrt{\chi^2_{p,1-\alpha/2}}$ 인 RD_i(robust distances)를 제안하여 지렛점을 식별하였다.

$$RD_i = \sqrt{(\boldsymbol{z_i} - \boldsymbol{m(Z)})^T \boldsymbol{C(Z)}^{-1} (\boldsymbol{z_i} - \boldsymbol{m(Z)})} \tag{1.3}$$

위에서 $\boldsymbol{m(Z)}$와 $\boldsymbol{C(Z)}$은 Rousseeuw와 Leroy의 재표본 알고리즘과 Rousseeuw(1985)의 MVE 개념을 이용하여 얻은 로버스트 평균과 로버스트 공분산행렬을 나타내고 있다. RD_i의 계산하는 방법에 대해서는 Hadi(1992)에 자세히 기술되어 있다. Rousseeuw와 van Zomeren은 X축에는 그들이 제안한 RD_i, Y축에는 LMS의 표준화된 잔차를 플롯시켜 이상값과 지렛점을 동시에 식별하는 방법을 제안하였다. 즉, Rousseeuw와 van Zomeren은 2장에서 인공자료를 분석한 그림 2.1과 같이 X축의 임계값 $\sqrt{\chi^2_{p,1-\alpha/2}}$을 이용하여 지렛점을 식별하고, Y축의 임계값 ± 2.5를 사용하여 이상점을 식별하였다. 그들은 이 그래프를 이용하여 지렛점이면서 이상점인 나쁜 지렛점 집합과 지렛점이면서 이상점이 아닌 좋은 지렛점, 지렛점이 아니면서 이상점인 점들의 집합 및 지렛점도 이상점도 아닌 점들의 집합으로 구분하였다. 그러나 그들의 방법은 LMS 회귀방법과 MVE 통계량의 매우 높은 로버스트성으로 인하여 이상값과 지렛점이 아닌 점들도 이상점과 지렛점으로 식별하는 경향이 있는 것으로 분석되며, Atkinson(1986)은 이러한 이상점과 지렛점의 재확인 절차의 필요성을 나타내고 재확인 방법을 제안하였다. Fung(1993)은 Atkinson 방법의 한계성을 지적하고 다음과 같은 방법으로 이상점과 지렛점 재확인방법을 제안하였다.

집합 $\bar{\mathbf{I}}$를 Rousseeuw와 van Zomeren의 방법으로 이상점 혹은 지렛점으로 식별한 m개의 지표 집합이라고 가정하고, 집합 $\bar{\mathbf{I}}$를 생략된 집합이라고 명칭하였다. 나머지 $n_0 = n - m$개의 적합이 잘되는 집합의 지표를 $\mathbf{I}$라 정의하고 집합 $\mathbf{I}$를 적합된 집합이라고 명칭하였다. 또한 분석모형은 앞에서 정의한 식 (1.1)을 같은 모형으로 사용하였으며 생략된 관찰값 i, $i \in \bar{\mathbf{I}}$를 적합된 집합에 더하여 Fung은 지표 i 에 대하여 다음과 같은 재확인 모델(adding-back model)을 제안하였다.

$$\boldsymbol{y}_{+i} = \boldsymbol{X}_{+i}\boldsymbol{\beta}_{+i} + \boldsymbol{\varepsilon}_{+i}, \qquad i \in \bar{\mathbf{I}} \tag{1.4}$$

위의 식 (1.4)는 집합 $\mathbf{I}$의 개수(n_0) + 1개의 자료로 생성된 모형이다 . Fung은 식 (1.4)의 모델을 기존에 많은 연구가 되어온 단일 이상점이나 지렛점 식별방법을 사용하여 적합된 집합에 추가된 점에 대하여 이상점 혹은 지렛점으로 재확인하는 방법을 제안하였다. 예를 들면 재확인 모델에서의 햇요소인 $h_{+i} = \boldsymbol{x}_i^T(\boldsymbol{X}^T{}_{+i}\boldsymbol{X}_{+i})^{-1}\boldsymbol{x}_i$, $i \in \bar{\mathbf{I}}$ 는 지렛점을 식별하고, 표준화 잔차 $t_{+i} = e_i/\{s\sqrt{1+\boldsymbol{x}_i(\boldsymbol{X}^T{}_{+i}\boldsymbol{X}_{+i})^{-1}\boldsymbol{x}_i}\}$, $i \in \bar{\mathbf{I}}$ 와 수정된 Cook 통계량 $MC_{+i} = \{(n_0+1-p)h_{+i}/(p(1-h_{+i}))\}^{1/2}|t_{+i}|$, $i \in \bar{\mathbf{I}}$ 은 이상값을 식별하였다. Atkinson(1985) 방법이 이상값만을 재확인하였던 것과는 달리 Fung은 MC_{+i} 와 t_{+i}, h_{+i}로 이상점은 물론 지렛점을 재확인하는 방법을 제안하였다. 특히 Fung은 Barnett와 Lewis(1984)가 제안한 MD_i의 임계값과 $h_i = MD_i^2/(n_0-1) + 1/n_0$의 관계식을 이용하여 지렛점을 재확인하는 척도로 사용하였으며 재확인방법에 다단계방법을 적용하였다. 즉, 이상점이나 지렛점으로 식별되었던 점들이 적합한 점들로 판정되면 적합된 집합에 이들을 포함시키고 다시 식 (1.4)의 모델을 이용하여 나머지 이상점과 지렛점들에 대하여 이상점과 지렛점 재확인 작업을 지속하는 방법을 택하였다. 2장은 Fung의 분석을 이용하여 여러 가지의 예제들의 지렛점을 분석하고 동시에 Fung 방법의 문제점에 대하여 서술하였으며, 3장에서는 Fung의 문제점을 극복할 수 있는 새로운 방법을 제시하였으며, 4장에서는 우리가 제시한 새로운 방법으로 실증분석을 하여 제시된 방법의 유효성을 검증하였으며, 5장은 결론을 서술하였다.

2. 예 제

2.1. 인공자료(Hawkins 등 1984, Rousseeuw와Leroy 1987에서 재인용)

인공자료는 3개의 설명변수와 1개의 반응변수, 관측값이 75개인 자료로 관측값 지표 (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)은 나쁜 지렛점, 관측값 지표 (11, ,12, 13, 14)는 좋은 지렛점으로 알

표 2.1: Fung의 인공자료 지렛점 재확인 척도

	관측값 지표														
단계	1	2	3	4	5	6	7	8	9	10	11	12	13	14	임계값
1	0.93	0.94	0.94	0.95	0.95	0.94	0.94	0.94	0.94	0.94	0.96	0.96	0.96	0.96	0.96

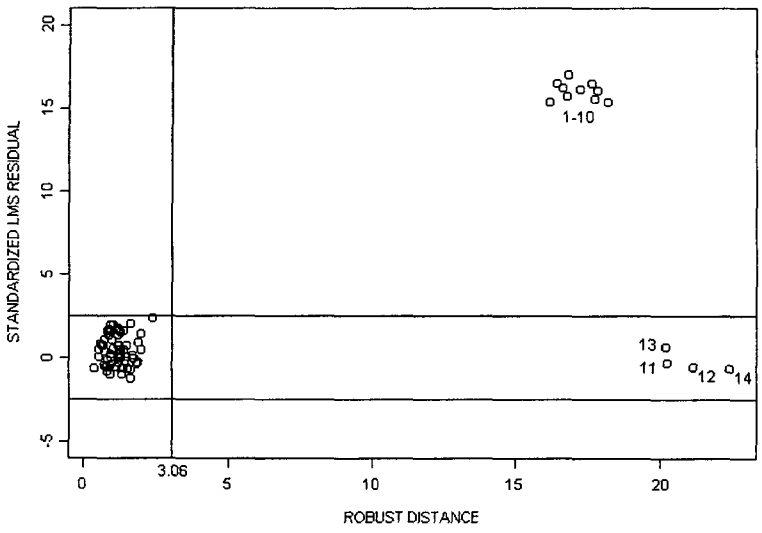

그림 2.1: 인공자료의 지렛점과 이상점 분석 그래프

려진 자료이다. 이 자료를 Rousseeuw와 van Zomeren의 방법으로 분석을 하여 그림 2.1을 플롯시키면 좋은 지렛점 지표 (11, 12, 13, 14)과 나쁜 지렛점 지표 (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)을 명확하게 식별할 수 있다. 따라서 삭제된 집합의 지표는 $\bar{\mathbf{I}}$ ={1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14 }임을 알 수 있다. 이를 Fung의 재확인 모델에서의 삭제된 집합에 해당하는 관측값들의 햇요소 h_{+i}를 나타낸 표 2.1를 보면 단계 1에서 관측값 지표 (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14) 모두 임계값 0.26을 훨씬 상회하여 지렛점이 14개 존재하는 것으로 분석할 수 있다. 이러한 분석은 많은 통계학자들이 그동안 주장하여온 결과와 동일한 결과이다.

2.2. 샐리니티 자료(Rousseeuw와 Leroy(1987))

이 자료는 3개의 설명변수, 1개의 반응변수 그리고 관측값이 28개인 자료이다. 우리는 여기에서 첫 번째와 세 번째 변수만을 사용하였다. 그림 2.2는 Rousseeuw와 van Zomeren의 이상점 및 지렛점 분석 그래프로 관측값 지표 (1, 8)은 이상값, 관측값 지표 (5, 10, 11, 23, 24)는 좋은 지렛점, 관측값 지표 (15, 16, 17)은 나쁜 지렛점으로 식별할 수가 있는데 이 점들은 관측값의 개수 28개중 약 36%에 해당한다. Fung은 이상값을 제외한 지렛점 지표 (5, 10, 11, 15, 16, 17, 23, 24)에 대한 재확인 모델을 사용하였는데 계산된 h_{+i}는 표 2.2에서 나타난 바와 같이 단계 1에서 관측값 지표 (10, 15, 17)이 임계값 0.51보다 낮아 지렛점에서 제외하고 이 점들을 적합된 집합에 삽입시킨후 단계 2에서 나머지 관측값 지표 (5, 11, 16, 23, 24)의 지렛점에 대한 식별을 지속하였다. 그 다음의 모형에서는 관측값 지표 (11, 24)가 지렛점에서 제외되는 것을 분석할 수 있었으며 이 작업을 지속한 후 관측값 지표 (16)만이 최

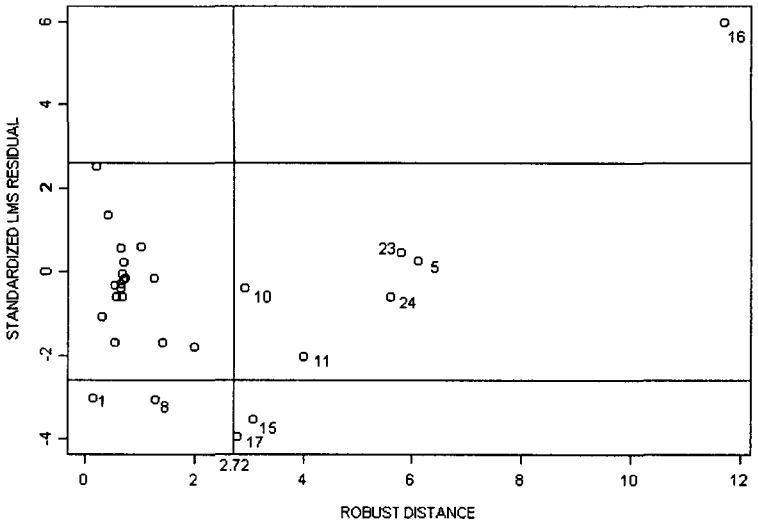

그림 2.2: 샐리니티자료의 지렛점과 이상점 분석 그래프

종적인 지렛점으로 판명되었다. 그러나 우리는 이 결과가 만족스럽지 못한 결과라고 분석하고 있다. Fung은 관찰값 지표 (16)만이 지렛점이라 분석하고 있는데 우리가 분석한 바로는 관찰값 지표 (11, 24)는 관찰값 지표 (10, 15, 17)이 적합된 모형에 포함됨으로서 생기는 효과(우리는 이를 인근효과(adjacent effect)라 명칭하고자 한다)로 인하여 지렛점인데도 지렛점이 아닌 것으로 판단되고 또 이 점들이 적합된 모형에 포함됨으로서 지속적으로 관측값 지표 (24), (5, 23)이 지렛점이 아닌 점으로 식별되는 경향이 있는 것으로 분석하였다.

2.3. 스택로스 자료(Brownlee 1965, Rousseeuw와 Leroy 1987에서 재인용)

스택로스 자료는 3개의 설명변수와 1개의 반응변수, 21개의 관측값으로 이루어진 자료로

표 2.2: Fung의 샐리니티자료 지렛점 재확인 척도

	관측값 지표								
단계	5	10	11	15	16	17	23	24	임계값
1	0.77	0.44	0.58	0.46	0.92	0.41	0.75	0.73	0.51
2	0.58		0.31		0.81		0.48	0.46	0.47
3	0.44				0.68		0.29		0.44
4	0.41				0.62				0.43
5					0.51				0.42

표 2.3: Fung의 스택로스자료 지렛점 재확인 척도

	관측값 지표				
단계	1	2	3	21	임계값
1	0.66	0.67	0.53	0.47	0.64
2	0.44	0.31			0.61

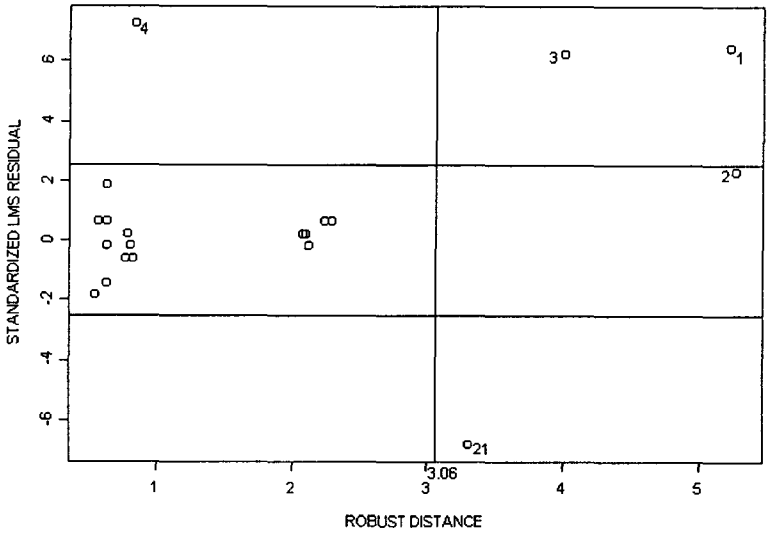

그림 2.3: 스택로스 자료의 지렛점과 이상점 분석 그래프

4개의 지렛점과 1개의 이상값이 존재하는 것으로 알려져 있다. Rousseeuw와 van Zomeren의 그림 2.3에서 관측값 지표 (4)는 이상값, 관측값 지표 (2)는 좋은 지렛점, 관측값 지표 (1, 3, 21)은 나쁜 지렛점으로 판단할 수가 있다. 이 자료에 대한 Fung의 방법으로 지렛점을 분석한 표 2.3을 살펴보면 관측값 지표 (3, 21)이 단계 1에서 임계값 0.64보다 낮아 지렛점에서 제외한 후 분석을 지속한 결과 스택로스 자료에는 지렛점이 존재하지 않는 것으로 나타났다. 이 결과에 대해서도 우리는 관측값 지표 (3, 21)이 지렛점이 아닌 점으로 분류된 후 인근효과에 의하여 관측값 지표 (1, 2)가 실제로 지렛점이면서도 지렛점이 아닌 것으로 분류되는 오류를 범하는 것으로 분석할 수 있다.

3. 새로운 재확인 방법의 제안

Atkinson의 방법을 수정하여 이상점과 지렛점의 재확인 방법을 제안한 Fung의 방법은 2장에서 분석한 샐리니티 자료와 스택로스 자료의 경우와 같이 우리가 명칭한 인근효과에

의하여 실제로는 지렛점으로 식별되어야 하는 자료들이 식별되지 못하는 결과를 나타냈다. 이 장에서는 우리는 이러한 인근효과에 로버스트한 방법을 이상값과 지렛점의 재확인 방법으로 제안하고자 하며 식 (1-1)의 모형을 기본모형으로 한다.

[단계 1] Rousseeuw와 van Zomeren 방법을 이용하여 이상점과 지렛점을 식별한 후 생략된 집합의 지표 $\bar{\mathbf{I}}$와 적합된 집합의 지표 $\mathbf{I}$를 정의한다.

[단계 2] 적합된 집합과 생략된 집합에 대하여 다음의 식을 이용하여 각각의 d_i를 계산한다.

$$d_i = \frac{y_i - \boldsymbol{x}_i^T \mathbf{b}_{\bar{\mathbf{I}}}}{\hat{\sigma}_{\mathbf{I}} \sqrt{1 - \boldsymbol{x}_i^T (\boldsymbol{X}_{\mathbf{I}}^T \boldsymbol{X}_{\mathbf{I}})^{-1} \boldsymbol{x}_i}} \qquad \text{if } i \in \mathbf{I}$$

$$d_i = \frac{y_i - \boldsymbol{x}_i^T \mathbf{b}_{\bar{\mathbf{I}}}}{\hat{\sigma}_{\mathbf{I}} \sqrt{1 + \boldsymbol{x}_i^T (\boldsymbol{X}_{\mathbf{I}}^T \boldsymbol{X}_{\mathbf{I}})^{-1} \boldsymbol{x}_i}} \qquad \text{if } i \in \bar{\mathbf{I}}$$

(위에서 $\mathbf{b}_{\mathbf{I}}$와 $\hat{\sigma}_{\mathbf{I}}$는 각각 적합된 집합의 최소제곱추정량인 $(p+1) \times 1$ 벡터와 표본 표준편차를 의미하고, $i \in \mathbf{I}$인 경우 d_i는 표준화된 잔차, $i \in \bar{\mathbf{I}}$ 인 경우 적합된 집합 $\mathbf{I}$에 기초한 예측오차임을 알 수 있다.)

[단계 3] d_i에 절대값을 취한 후 오름차순으로 나열하고, $d_{(s+1)}$을 $|d_i|$의 $(s+1)$번째의 순서통계량이라고 가정한다. 이때 s는 현재의 적합된 집합 $\mathbf{I}$의 크기이다.

(a) 만약 $d_{(s+1)} \geq t_{(\alpha/2(s+1),s-k)}$ 가 성립하면 $|d_i| \geq t_{(\alpha/2(s+1),s-k)}$ 를 만족하는 모든 관측점을 이상점이라 판정하고 종료한다.

(b) 위의 식이 성립하지 않는 경우 처음의 $(s+1)$개의 관측점을 선택하여 새로운 집합 $\mathbf{I}$를 구성한다. 만약 $n = s+1$이 성립하면 자료에는 이상점이 없는 것으로 판정하고, n>s+1인 경우 단계 2로 돌아간다.

[단계 4] Fung의 식 (1.4)를 이용하여 좋은 지렛점을 재확인한다. 단 이때 생략된 집합은 단계 2와 3에서 이상점 혹은 나쁜 지렛점으로 재확인된 관찰값과 단계 1에서 좋은 지렛점으로 식별된 관찰값만을 의미하며, 다단계방법을 이용하지않고 단일지렛점 검증방법을 이용한다.

앞에서 단계 1은 Rousseeuw와 van Zomeren의 방법을 이용하여 재확인 필요성이 있는 이상점과 지렛점을 식별하고, 단계 2와 단계 3에서는 이상점과 나쁜 지렛점을 재확인하는 단계로 이는 Hadi와 Simonoff(1993)에서 이론적 근거를 확인할 수 있으며, 단계 4는 좋은 지렛점을 재확인하는 절차이다.

표 4.1: 제안된 인공자료 지렛점 재확인 척도

관측값 지표				
11	12	13	14	임계값
0.96	0.96	0.96	0.97	0.26

4. 분석결과

4.1. 인공자료

이 자료는 Rousseeuw와 van Zomeren의 방법과 Fung의 방법이 동일하게 관측값 지표 (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14)를 지렛점으로 식별하였다. 따라서 따라서 삭제된 집합의 지표는 $\bar{\mathbf{I}}$={1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14} 임을 알 수 있다. 이를 우리가 제안한 방법에 적용시켰을 때 단계 2와 단계 3에서 관측값 지표 (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)은 나쁜 지렛점임을 밝혀낼 수 있었으며 [단계 4]에서 관측값 지표 (11, 12, 13, 14)의 햇요소의 값이 모두 임계값 0.26을 크게 상회하여 이들이 좋은 지렛점임을 식별할 수 있었다. 따라서 이 자료는 Fung의 방법과 동일하게 관측값 지표 (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14)의 지렛점이 존재하는 것을 분석할 수 있었다.

4.2. 샐리니티 자료

이 자료를 Rousseeuw와 van Zomeren의 방법을 이용하여 분석을 하면 앞의 그림 2.2에 나타난 바와 같이 이상점으로 관측값 지표 (1, 8), 나쁜 지렛점으로 관측값 지표 (15, 16, 17), 좋은 지렛점으로 관측값 지표 (5, 10, 11, 23, 24)를 식별할 수 있었다. 따라서 삭제된 집합의 지표는 $\bar{\mathbf{I}}$ ={1, 5, 8, 10, 11, 15, 16, 17, 23, 24 }임을 알 수 있다. 이를 우리가 제안한 단계 2에 적용시키면 관측값 지표 (16) 만이 이상점 혹은 나쁜 지렛점으로 식별할 수 있었으며 이상점으로 분류된 관측값 지표 (1, 8)과 나쁜 지렛점으로 식별된 관측값 지표 (15, 17)은 적합된 자료로 분류되었다. 따라서 관측값 지표 (1, 8, 15, 17)을 적합된 집합에 포함시키고 단계 4를 이용하면 자료 (5, 23, 24)가 좋은 지렛점으로 재식별되는 것을 분석할 수 있었으며, 이는 Fung의 결과와는 상이한 것으로 우리가 2장에서 지적한 Fung의 제안이 인근효과에 의하여 지렛점인 점들이 지렛점이 아닌 것으로 식별되는 오류를 방지할 수 있었음을 보여주고 있다.

표 4.2: 제안된 샐리니티자료 지렛점 재확인 척도

관측값 지표						
5	10	11	16	23	24	임계값
0.62	0.24	0.36	0.84	0.55	0.53	0.51

표 4.3: Fung의 방법과 제안된 방법의 결과 비교

		Fung의 방법	제안된 방법
인공자료	좋은 지렛점	11, 12, 13, 14	11, 12, 13, 14
	나쁜 지렛점	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
	이상점	-	-
샐리니티 자료	좋은 지렛점	-	5, 23, 24
	나쁜 지렛점	16	16
	이상점	-	-
스택로스 자료	좋은 지렛점	-	2
	나쁜 지렛점	-	1, 3, 21
	이상점	-	4

4.3. 스택로스자료

스택로스자료는 관측값 지표 (1, 3, 4, 21)은 이상점 혹은 나쁜 지렛점, 관측값 지표 (2)는 좋은 지렛점으로 잘 알려져 있는 자료이나 Fung의 방법으로는 지렛점이 존재하지 않는 것으로 분석된 자료이다. 그러나 이를 우리가 제안한 방법으로 분석하여 보면 삭제된 집합의 지표는 $\bar{\mathbf{I}}$={1, 2, 3, 4, 21}로 단계 2에서 관측값 지표 (1, 3, 4, 21)가 이상값 혹은 나쁜 지렛점으로 식별되고, 단계 4에서 관측값 지표 (2)의 h_{+i}의 값이 0.7로 나타나 그 임계값 0.64를 상회하므로 관측값 지표 (2)는 좋은 지렛점으로 식별되어 제시된 방법은 인근효과에 영향을 덜 받는 결과를 얻을 수 있었다.

5. 맺은말

Atkinson과 Fung이 지적하였듯이 Rousseeuw와 van Zomeren의 이상점 및 지렛점 식별 방법은 MVE 통계량과 LMS 회귀방법의 강한 로버스트성으로 인하여 이상점과 지렛점을 과중하게 식별하는 경향이 있는 것으로 분석된다. 그러나 그 대안으로 Fung은 재확인 모델을 제시하고 이상점을 탐색하는 방법으로 수정된 Cook 통계량을 사용하였으나 수정된 Cook 통계량의 불명확한 분포로 인하여 정확한 임계값을 제공하지 못하는 단점을 지니고 있으며, 앞의 재확인모델에서의 h_{+i}를 이용하여 지렛점을 다단계방법으로 식별하는 방법은 인근효과에 영향을 받아 지렛점을 지렛점이 아닌 점으로 식별하는 단점을 지니고 있다. 그러나 우리가 제안한 방법은 Hadi와 Simonoff (1993)이 제안한 t 통계량을 사용함으로써 통계량을 이용하여 이상점과 나쁜 지렛점을 식별할 수 있었으며 좋은 지렛점은 Fung의 방법을 수정하여 식별하였다. 결과로는 우리가 제안한 방법은 인근효과에 영향을 받지 않아 이상점과 지렛점을 재확인하는 데에 유효한 방법으로 생각된다.

참고문헌

[1] Atkinson, A.C. (1986). Masking unmasked, *Biometrika, 73, 3*, 533–541.

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

[3] Brownlee, K. A. (1965). *Statistical Theory and Methodology in Science and Engineering(2nd ed.)*, New York : John Wiley.

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

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

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

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

[8] Rousseeuw, P.J. (1984) Least Median of Squares Regression , *Journal of American Statistical Association*, **79**, 871–880.

[9] 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, 283–297. Dordrecht: Reidel.

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

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

[2001년 1월 접수, 2002년 5월 채택]

A Confirmation of Identified Multiple Outliers and Leverage Points in Linear Model

Jong Young Yoo[1] Kisoo Ahn[2]

ABSTRACT

We considered the problem for confirmation of multiple outliers and leverage points. Identification of multiple outliers and leverage points is difficult because of the masking effect and swamping effect. Rousseeuw and van Zomeren(1990) identified multiple outliers and leverage points by using the Least Median of Squares and Minimum Value of Ellipsoids which are high-breakdown robust estimators. But their methods tend to declare too many observations as extremes. Atkinson(1987) suggested a method for confirming of outliers and Fung(1993) pointed out Atkinson method's limitation and proposed another method by using the add-back model. But we analyzed that Fung's method is affected by adjacent effect. In this thesis, we proposed one procedure for confirmation of outliers and leverage points and compared three example with Fung's method.

Keywords: outliers, leveage points, masking effect, swamping effect, adjacent effect

1) Associate Professor, School of Computer & Information, Yongin University, Yongin
E-mail: jyyoo@yongin.ac.kr
2) Assistant Professor, Department of Computer Science Application, Dongnam Health College, Suwon
E-mail: ksahn@dongnam.ac.kr