The Korean Communications
in Statistics Vol. 11 No. 3, 2004
pp. 485-494

Identification of Regression Outliers Based on Clustering of LMS-residual Plots[1)]

Bu-Yong Kim[2)] and Mi-Hyun Oh[3)]

Abstract

An algorithm is proposed to identify multiple outliers in linear regression. It is based on the clustering of residuals from the least median of squares estimation. A cut-height criterion for the hierarchical cluster tree is suggested, which yields the optimal clustering of the regression outliers. Comparisons of the effectiveness of the procedures are performed on the basis of the classic data and artificial data sets, and it is shown that the proposed algorithm is superior to the one that is based on the least squares estimation. In particular, the algorithm deals very well with the masking and swamping effects while the other does not.

Keywords : regression outlier, robust residual, clustering, masking, swamping

1. 서 론

회귀분석 자료에 이상점들이 포함되어 있는지 확인하고, 이상점들이 포함되어 있다면 어느 관찰치가 이상점에 해당되는지 식별하는 작업은 회귀분석에서의 중요한 과정 중의 하나다. 이상점들은 회귀모형 적합에 악영향을 미치는데, 모수의 추정과 추정량의 정도 그리고 회귀모형의 전반적인 예측력을 왜곡시킨다. 그러므로 회귀분석에 앞서 이상점들을 식별하는 과정을 필히 거쳐야 한다. 특히, 자료수집 과정이 철저하게 관리되기 어려운 데이터마이닝 등의 분야에서는 회귀분석 자료에 이상점이 다수 포함되는 경우가 있기 때문에 이상점들을 정확히 식별하여 자료정제 과정에서 적절한 조치를 취하든지 로버스트 추정을 적용하여 회귀분석을 해야 한다.

회귀이상점을 식별하기 위한 다양한 방법들이 개발되었는데, Belsley *et al.* (1980), Cook and Weisberg(1980), Marasinghe(1985), Kianifard and Swallow(1990) 등이 대표적인 방법이. 그런데 이러한 식별방법들은 회귀자료에 단일이상점이 포함되어 있을 경우에는 효과적이지만, 복수이상점들이 존재하는 경우에는 효용성이 많이 떨어진다는 사실이 밝혀졌다. 그래서 복수이상점을 식별하기 위하여 관찰치의 모든 부분집합에 대해 진단척도들을 적용하려는 시도들이 있었지만 식별과정에 방대한 계산이 요구된다는 한계를 가지고 있다. 더욱이 이 방법들은 주로
최소제곱(LS)추정에 바탕을 두기 때문에 이상점의 악영향을 받은 결과를 바탕으로 이상점을 식별

1) This research was supported by Sookmyung Women's University Research Grants (2004).
2) Professor, Department of Statistics, Sookmyung Women's University, Seoul 140-742, Korea.
E-mail: buykim@sookmyung.ac.kr
3) Graduate student, Department of Statistics, Sookmyung Women's University, Seoul 140-742, Korea.

하는 모순에 빠진다. 한편, Rousseeuw and Leroy(1987)는 붕괴점이 높은 최소중위수추정(LMS)에서 얻은 잔차를 바탕으로 한 식별방법을 제시하였으나, 이 방법은 수직이상점의 식별에는 효과적이지만 나쁜 지렛점은 식별하지 못하는 한계를 가지고 있다.

일반적으로 복수이상점 식별과정에서 부딪치는 중요한 문제들이 있는데, 즉, 가림현상(masking effect)과 붙음현상(swamping effect)이다. 가림현상은 명백한 이상점들의 영향 때문에 이상점 일부를 정상점으로 잘못 판단하게 되는 상황이며, 붙음현상은 강력한 이상점의 영향 때문에 정상점을 이상점으로 잘못 식별하는 상황을 의미한다. 두 가지 문제 중에서 가림현상이 붙음현상보다 더 심각한 문제라는 주장이 있는데, 아무리 붙음현상이 덜 심각한 문제라고 하더라도 효과적인 식별방법은 붙음현상 역시 최소한으로 축소할 수 있는 것이라고 할 수 있다. 가림현상 문제를 극복하기 위하여 Rousseeuw and Zomeren(1990)은 로버스트 거리(robust distance)와 LMS/LTS-잔차를 동시에 고려하는 식별방법을 제안하였다. 이 식별방법은 관찰치들을 정상점, 수직이상점, 나쁜 지렛점, 좋은 지렛점으로 구분하여 식별할 수 있다는 특징을 가지고 있다. 그러나 이 방법에 의하면 붙음현상이 발생하여 정상점을 이상점으로 과도하게 식별하는 문제점이 있고, 모의실험에 의해 결정한 경계치를 적용해야 한다는 단점도 가지고 있다.

한편, Hadi and Simonoff(1993)는 전진탐색법에 바탕을 둔 식별방법을 제시하였는데, 이 방법은 가림현상을 상당히 극복할 수 있어서 이상점을 효과적으로 식별할 수 있으며, 객관적인 기준으로 경계치를 결정할 수 있다는 장점을 가지고 있다. 그러나 다수의 이상점들이 정상점들로부터 먼 위치에 있는 경우에는 식별과정에서 이상점들의 영향을 강력하게 받게 되어 이상점을 정확히 식별하기 어렵다는 사실을 Kim and Kim(2002)이 지적하고, 로버스트 추정에 바탕을 둔 식별방법을 제안하였다. 이 식별방법은 LMS-추정에서 얻은 잔차 중에서 $n-[n/2]+p-1$(n은 관찰치의 수, p는 설명변수의 수, [ ]은 최대정수함수임)개의 작은 크기의 절대표준화잔차에 대응하는 관찰치들로 초기 정상점 부분집합을 구성하고, 축차적으로 이상점 여부를 검정해 나가는 과정을 거친다. 제안된 방법의 효용성이 모의실험에 의해 평가되었는데, 강력한 영향력을 갖는 다수의 이상점이 존재하는 경우에는 Kim and Kim(2002) 방법이 Hadi and Simonoff(1993) 방법보다 우수하다는 사실이 확인되었다.

그러나 이러한 반복적인 식별방법들은 많은 계산을 필요로 하기 때문에 자료의 규모가 방대한 경우에는 실제로 적용하기가 어렵다. 따라서 계산효율성이 높은 식별방법에 관한 연구가 요구된다. Sebert *et al.*(1998)은 군집화 기법을 활용한 회귀이상점 식별방법을 제시하였는데, 붙음현상 문제를 해결하지 못하는 심각한 결점을 가지고 있음이 밝혀졌다. 그 원인은 이 방법이 LS-추정에 바탕을 두기 때문으로 파악된다. 따라서 본 논문에서는 가림현상과 붙음현상을 동시에 극복할 수 있도록 로버스트 추정인 LMS-추정으로부터 얻은 잔차의 군집화에 바탕을 둔 식별방법을 제시하고 그 효용성을 평가하고자 한다.

2. LMS-잔차산점도

본 논문에서의 이상점 식별은 일반적인 다중선형회귀모형: $y=X\beta+\epsilon$(y는 n차인 반응변수 관찰치 벡터, X는 절편을 포함한 $n\times p$차의 설명변수 행렬, β는 p차의 모수 벡터, ϵ는 n차의 오차 벡터임)을 전제로 한다. 이 회귀모형에 대하여 다양한 로버스트 추정법들이 사용되고 있는데, 그들 중에서 붕괴점이 가장 높다는 특성과 함께 계산효율성이 높은 추정알고리즘들이 잘 개발되었

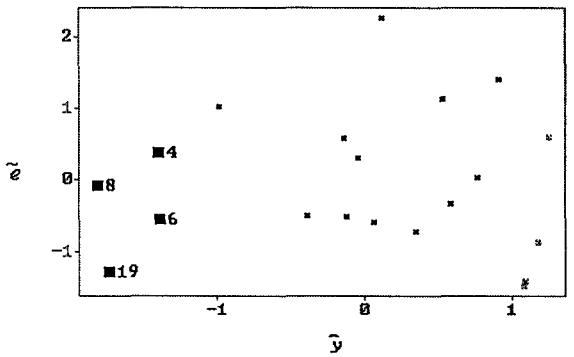
다는 이유로 회귀분석에 가장 많이 사용되고 있는 LMS-추정을 본 연구에서의 로버스트 추정법으로 채택하였다. LMS-추정량은 Rousseeuw(1984)에 의해 다음과 같이 정의되었는데,

$$\underset{\beta}{\text{minimize}} \ \underset{i}{median} \ (y_i - x_i^T\beta)^2,$$

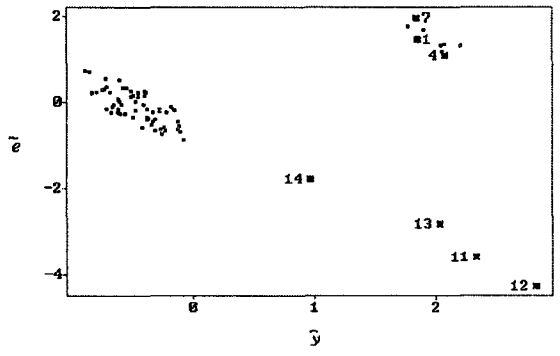
LMS-추정량의 중요한 특성인 붕괴점은 0.5로서 로버스트 추정량들의 붕괴점 중에서 가장 높은 값에 해당된다. 그리고 LMS-추정량의 통계적 특성과 추정을 위한 알고리즘은 Rousseeuw(1984), Basset(1991), Kim(1996) 등에 의해 제시되었다.

회귀모형의 타당성을 진단하거나 이상점을 식별하기 위하여 적합치와 잔차를 플롯한 잔차산점도가 많이 활용되는데, 대다수의 점들로부터 멀리 벗어난 위치에 한 개 혹은 몇 개의 점들이 군집을 형성하는 경우 이 점들을 이상점으로 식별한다. <그림 1>은 Rousseeuw and Leroy(1987)에 수록된 자료인 [wood gravity data]에 대한 LS-잔차산점도인데, 일정구역을 벗어난다고 할 수 있는 관찰치들을 명확히 구분할 수 없기 때문에 자료에 존재하는 실제의 이상점 {4, 6, 8, 19}를 전혀 식별할 수 없다. 한편, <그림 2>는 Hawkins, Bradu and Kass(1984)의 자료에 대한 LS-잔차산점도인데, 이 그림에서는 불음현상 때문에 정확한 이상점 {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}과 함께 좋은 지렛점인 {11, 12, 13, 14}도 이상점으로 분류하여 14개의 이상점이 존재하는 것으로 잘못 판단하게 된다. 이러한 오류는 이상점의 영향을 이미 받은 상태에서 얻어진 LS-잔차를 바탕으로 이상점을 식별하였기 때문에 발생한 것이다.

이와 같은 사례들에서 볼 수 있듯이, LS-잔차산점도에 의한 이상점 식별은 효과적인 방법이라고 할 수 없다. 따라서 다음과 같이 LMS-추정에 의한 잔차산점도를 바탕으로 이상점을 식별하는 방안을 고려할 수 있다. <그림 3>은 [wood gravity data]에 LMS-추정을 적용하여 구한 잔차산점도인데, <그림 1>에서와

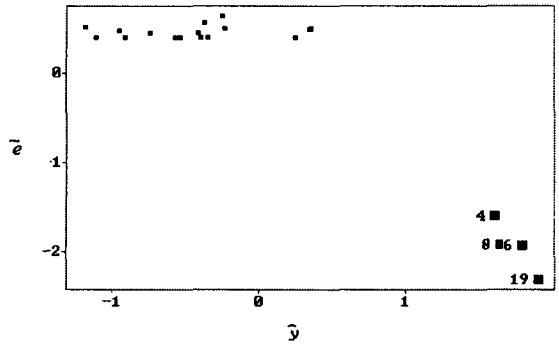


<그림 1> LS-잔차산점도 [wood gravity data]

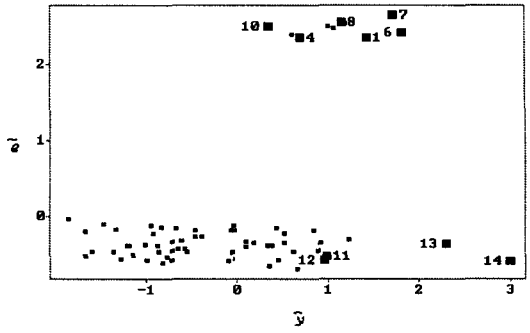


<그림 2> LS-잔차산점도 [Hawkins et al. data]

달리 <그림 3>에서는 이상점 {4, 6, 8, 19}를 정확하게 식별할 수 있다. 한편, <그림4>는 [Hawkins, Bradu, Kass data]에 대한 LMS-잔차산점도인데 <그림 2>에서와 달리 관찰치 {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}이 이상점이라는 사실을 정확히 파악할 수 있다. 그러나 LMS-잔차산점도를 바탕으로 이상점들을 식별하는 경우에도 최종 결정은 시각적 판단에 의존하게 되므로 정상점과 근접한 위치에 있는 이상점들을 찾아내기에는 한계가 있을 수 있다. 즉, 이상점들의 위치가 명백히 구분되는 자료에서는 시간적 판단만으로도 이상점들을 충분히 식별할 수 있지만, 시각적 판단만으로는 이상점 식별이 곤란한 상황이 있을 수 있다. 또



<그림 3> LMS-잔차산점도 [wood gravity data]



<그림 4> LMS-잔차산점도 [Hawkins et al. data]

한 이상점 식별결과를 후속 분석과정에서 활용할 수 있도록 전산프로그램을 작성하는 경우에는 시각적 판단은 무용지물이 된다. 따라서 LMS-잔차산점도를 바탕으로 이상점을 식별하고자 하는 경우에 군집화와 같은 접근방법을 통한 객관적인 판단이 요구된다.

3. LMS-잔차의 군집화에 의한 이상점 식별

잔차산점도를 바탕으로 이상점을 식별하는 경우 시각적 판단에 의존하게 되고 따라서 연구자들의 주관적인 결론일 수밖에 없다는 점을 제2장에서 지적하였다. 따라서 본 연구에서는 이상점에 대한 객관적 판단을 위하여 군집화 기법을 채택하고자 한다. 우선 LS-잔차산점도에 군집화를 적용하는 방법을 고려할 수 있는데, 이 식별방법을 [wood gravity data]에 적용해 보면 이상점을 전혀 식별할 수 없으며, [Hawkins et al. data]에 적용한 경우에도 붙음현상 때문에 정확한 이상점 식별이 불가능하다. 한편, 다양한 특성을 갖는 자료들에 이 방법을 적용해 본 결과, 많은 자료에서 붙음현상과 가림현상이 발생한다는 사실을 확인할 수 있었다. 따라서 본 연구에서는 LS-추정 대신에 LMS-추정을 적용하여 잔차와 적합치를 구하고 군집화를 통하여 이상점을 식별하는 방법을 제안하고자 한다.

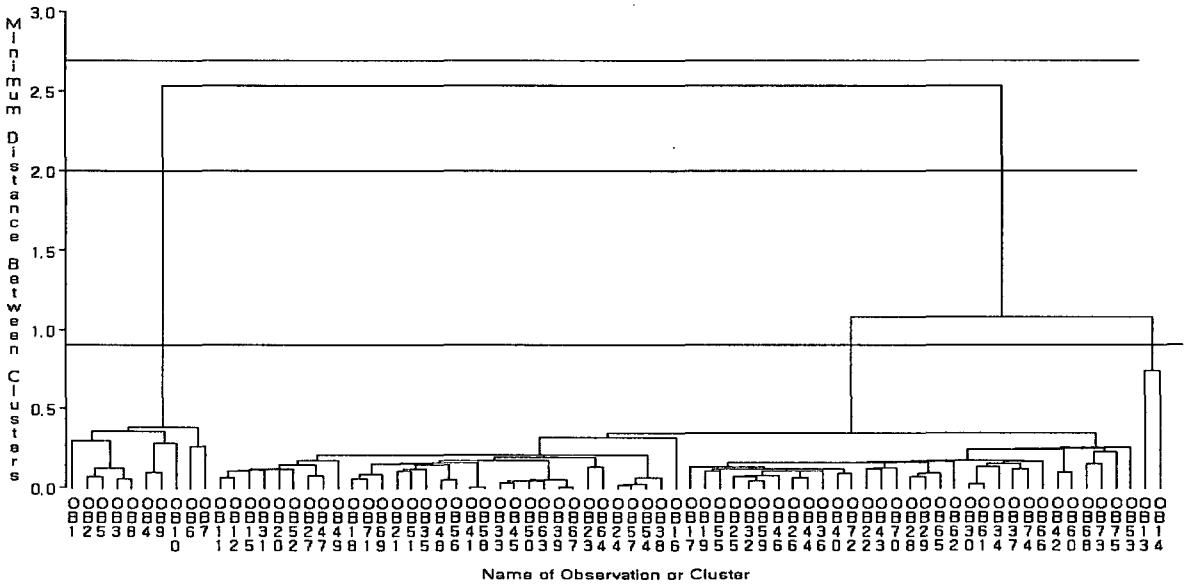
군집화 기법은 생물학분야에서 사용되기 시작하였으나(Hartigan, 1975), 최근에는 통계적인 방법과 연계시키는 노력들이 시도되고 있다. 군집화는 군집알고리즘을 사용하여 관찰치들을 몇 개의 집단으로 분류하는 기법인데, 회귀이상점 식별은 잔차산점도의 점들을 정상점과 이상점 군집으로 분류하는 군집화 문제에 해당된다고 할 수 있다. 이상점 식별과정에서 군집화의 대상은 잔차와 적합치 n개 값인데, 군집화를 위해 유사성 척도와 군집알고리즘을 선택해야 한다. Everitt(1993)는 유사성척도로서 유클리드거리가 가장 널리 사용된다고 하였는데, 본 연구에서는 LMS-잔차산점도의 두 점, $\boldsymbol{v}_i=(z_i,w_i)$와 $\boldsymbol{v}_j=(z_j,w_j)$(여기서 z는 표준화 LMS-적합치, w는 표준화 LMS-잔차 임)사이의 유클리드거리를 바탕으로 군집화하고자 한다. 군집알고리즘들은 자료의 형태와 군집화의 목적에 매우 의존적이기 때문에 모든 상황에서 가장 우수한 군집알고리즘은 존재하지 않는다고 할 수 있다. 본 연구에서는 군집화 기법 중에서 계층적 군집화(hierarchical clustering)를 채택하였는데, n개의 점들이 각기 군집을 구성하는 초기단계로부터 시작하여 한번에 한 개씩 군집을 줄여 나아감으로써 최종적으로 n개의 모든 점이 하나의 군집으로 묶일 때까지 군집의 수를 줄여 나가는 과정을 따른다. 한편, 군집을 합병하는 원칙이 어느 것이냐에 따라서 군집알고리즘이 달라진다. 본 연구에서는 단일결합군집화(single linkage clustering)을 적용하였는데, 이 원칙이 LMS-잔차산점도에서 흔히 볼 수 있는 체인형태의 군집을 효과적으로 식별해 낼 수 있기 때문이다.

군집알고리즘의 적용 결과는 군집나무로 표현되며 점들을 몇 개의 군집으로 분류할 것인지를 결정해야 하는데, 규정된 경계선의 높이에 따라 군집의 수가 결정된다. Mojena(1977)는 계층적 군집화 과정에서의 군집수 결정방법을 제시하였는데, 경계선 높이 h^*는 다음과 같이 결정하는 것이 적절하다고 하였다. 즉,

$$h^* = \bar{h} + \alpha s_h,$$

여기서 $\bar{h}$는 $n-1$개의 모든 군집에 대응하는 군집높이들의 평균이고, s_h는 군집높이들의 표준편차 불편추정량이며, α는 규정된 상수(Mojena는 α값으로 2.75~3.0이 가장 적절하다고 하였음)를 의미한다. h^*가 최대의 h보다 큰 경우는 군집의 수가 한 개인 것으로 판단되는데 이는 자료에 이상점이 전혀 포함되지 않았다는 것을 의미한다.

일반적으로 군집화를 하기 전에 관찰치를 표준화하는데 이는 변동성이 큰 변수의 관찰치가 유사성 척도의 크기를 좌우하기 때문이다. 특히, 유클리드 거리를 유사성 척도로 사용하는 경우에 표준화 작업이 필요한데 많은 회귀자료에서 적합치의 변동성이 잔차의 변동성보다 크기 때문이다. [Hawkins et al. data]의 표준화된 LMS-잔차와 적합치를 군집화한 결과는 군집나무 <그림 5>로 표현되었는데, 이를 바탕으로 회귀이상점을 식별하기 위해서는 군집나무의 높이를 규정된 수준에서 잘라야 한다. 즉, 정지규칙에서 α값을 결정하여 경계선 높이 h^*을 구하고 각 점을 이상점에 속하는 것들과 정상점에 해당하는 것들로 분류하게 된다. 그런데 <그림 5>에서 경계선 높이가 2.5371보다 높으면 자료에 이상점이 없다고 잘못 판정하게 되고, 1.0817보다 낮으면 정상점을 이상점으로 잘못 판정하는 붙음현상을 야기하게 된다. 그러나 경계선의 높이를 1.0817~2.5371사이로 결정하면 군집이 2개로 구성되며 이상점 {1~10}을 정확하게 식별하게 된다. 이와 같이 경계선 높이가 군집의 수를 결정하는데 절대적인 영향을 주며, 군집나무의 가지가 잘려지는 높이에 따라 가림현상과 붙음현상이 상반되게 발생할 수 있음을 알 수 있다. 따라서 Mojena(1977)가 제시한 α의 값



<그림 5> LMS-잔차의 군집나무와 경계선 높이

이 회귀이상점 식별에도 적절한 것인지 확인하고, 제시된 알고리즘의 정지규칙에 적용될 최적의 α값을 결정하기 위하여 <표 1>에 수록된 다양한 특성을 갖는 회귀자료를 대상으로 이상점을 식별해 보았다. <표 1>에는 회귀분석 관련 논문이나 문헌들에 자주 소개되는 자료들 중에서 이상점이 확인되어 이론의 여지가 없는 21가지의 자료가 선정되었으며, 연구대상 자료의 다양성을 위하여 여러 가지 형태의 이상점을 갖는 6가지의 자료가 인위적으로 생성되었다. LS-잔차의 군집화를 따르는 식별방법(LSOUT)과 LMS-잔차의 군집화에 의한 식별방법(CLUST)의 식별력을 측정하였는데, 가능영역에 속하는 다양한 α값에 대하여 붙음현상과 가림현상을 포함한 식별오류가 어느 정도 발생하는지 파악하였다.

<표 1> 연구대상 회귀자료 목록 및 특성

자료번호	자료 이름	설명변수 수	관찰치 수	이상점
1	Inflation in China 1940~1948	1	9	9
2	Monthly payments in 1979	1	12	12
3	Kootenay River data	1	13	4
4	Pension funds data	1	18	18
5	Cloud point data	1	19	1, 10, 16
6	Pilot-plant data	1	20	-
7	First word-Gesell adaptive score	1	21	18, 19
8	Number of telephone calls	1	24	15~24
9	Hadi-Simonoff data (1992)	1	25	19~25
10	Body and brain weight data	1	28	6, 16, 25
11	Hertzsprung-Russell diagram data	1	47	11, 20, 30, 34
12	Heart catheterization data	2	12	-
13	Phosphorus content data	2	18	17
14	Delivery time data	2	25	9
15	Hadi-Simonoff data (1993)	2	25	1~3
16	Stackloss data	3	21	1~4, 21
17	Education expenditure data	3	50	50
18	Hawkins-Bradu-Kass data	3	75	1~10
19	Aircraft data	4	23	22
20	Example of exact fit	4	25	25
21	Wood specific gravity data	5	20	4, 6, 8, 19
22	Generated data (1)	1	20	16~20
23	Generated data (2)	3	20	1~5
24	Generated data (3)	3	30	1~4
25	Generated data (4)	3	40	34~40
26	Generated data (5)	3	40	35~40
27	Generated data (6)	4	50	1~5

(-) 이상점이 포함되지 않은 자료임

다양한 크기의 α값을 적용하여 이상점을 식별한 결과는 <표 2>와 <표 3>에 요약되었는데 정상점을 이상점으로 판정한 경우는 붙음현상에, 이상점을 식별하지 못한 경우는 가림현상에 포함시킨 결과다. LSOUT의 경우 많은 자료에서 가림현상과 붙음현상이 발생하여 식별결과가 정확하지 않았으나 CLUST에 의해 이상점을 식별하는 경우에는 이상점을 비교적 정확하게 식별하는 것으로 나타났다. LSOUT에서 가림현상과 붙음현상의 발생빈도의 합계가 최소가 되게 하는 α값은

2.4와 2.5이므로 중간점인 2.45를 최적치로 선정하였고, CLUST에서 가림현상과 붙음현상의 발생빈도의 합계가 최소가 되게 하는 α값은 1.9, 2.0, 2.1이므로 중간점인 2.0을 최적치로 선정하였다.

<표 2> LSOUT에서 α값의 크기에 따른 식별력의 비교

α값	2.0	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0
붙음현상 자료 수	5	4	3	3	2	1	1	1	1	1	1
가림현상 자료 수	8	9	9	9	9	10	11	11	11	11	13
합 계	13	13	12	12	11	11	12	12	12	12	15

<표 3> CLUST에서 α값의 크기에 따른 식별력의 비교

α값	1.5	1.6	1.7	1.8	1.9	2.0	2.1	2.2	2.3	2.4	2.5
붙음현상 자료 수	5	5	4	4	2	2	2	2	2	2	2
가림현상 자료 수	1	1	1	1	1	1	1	2	4	5	6
합 계	6	6	5	5	3	3	3	4	6	7	8

따라서 α의 최적치 2.0을 적용한 LMS-잔차에 바탕을 둔 이상점 식별 알고리즘을 구체적으로 기술하면 다음과 같다.

알고리즘 CLUST:

[단계 1] LMS-적합치 u_i와 잔차 e_i를 구한 후 각각을 다음과 같이 표준화한다.

$$z_i = (u_i - \bar{u})/s_u, \quad w_i = (e_i - \bar{e})/s_e.$$

[단계 2] (z_i, w_i) 간의 유클리드 거리를 계산하고, 단일결합 군집알고리즘을 적용하여 계층적 군집나무를 구성한다.

[단계 3] 정지규칙에 따라 경계선 높이 $h^* = \bar{h} + \alpha s_h$ (단, α=2.0, $\bar{h}$는 군집나무 높이의 평균, s_h는 군집나무 높이의 표준편차 추정치임)에서 군집나무를 자르고 군집을 분류한다.

[단계 4] 대다수의 점들을 포함하는 군집에 대응하는 관찰치를 정상점으로 판별하고, 다른 점들에 대응하는 관찰치를 이상점으로 식별한다. 모든 점들이 한 군집으로 분류되는 경우는 이상점이 존재하지 않는 자료라고 선언한다.

4. 이상점 식별방법의 비교

이상점 식별방법 LSOUT과 CLUST의 효용성을 비교 분석하였다. <표 1>의 자료들을 대상으로 두 가지 식별방법을 적용하여 얼마나 효과적으로 이상점을 식별하는지 평가하였으며 식별결과의 일부가 <표 4>와 <표 5>에 수록되었다. LSOUT의 경우 제3장에서 최적값으로 확인한 α=2.45를 비롯하여 α=2.1, α=2.8을 적용한 이상점 식별결과가 <표 4>에 수록되었는데, 최적치 α=2.45을 적용한 경우 11과 18번 자료에서 붙음현상이 발생하였고 5, 8, 10, 16, 21, 23, 24, 26, 27번 자료에서 가림현상이 발생하였다. 이러한

<표 4> LSOUT의 식별효용성 측정결과

자료번호	이상점	α=2.1		α=2.45		α=2.8	
		식별된 이상점	붙음/가림 현상	식별된 이상점	붙음/가림 현상	식별된 이상점	붙음/가림 현상
1	9	9	0 / 0	9	0 / 0	-	0 / 1
2	12	12	0 / 0	12	0 / 0	12	0 / 0
3	4	4	0 / 0	4	0 / 0	4	0 / 0
4	18	18	0 / 0	18	0 / 0	18	0 / 0
5	1, 10, 16	-	0 / 3	-	0 / 3	-	0 / 3
6	-	-	0 / 0	-	0 / 0	-	0 / 0
7	18, 19	18, 19	0 / 0	18, 19	0 / 0	-	0 / 2
8	15~24	15~20	0 / 4	15~20	0 / 4	15~20	0 / 4
9	19~25	19~25	0 / 0	19~25	0 / 0	19~25	0 / 0
10	6, 16, 25	7, 15, 25	2 / 2	25	0 / 2	25	0 / 2
11	11, 20, 30, 34	7, 11, 20, 30, 34	1 / 0	7, 11, 20, 30, 34	1 / 0	11, 20, 30, 34	0 / 0
12	-	-	0 / 0	-	0 / 0	-	0 / 0
13	17	17	0 / 0	17	0 / 0	17	0 / 0
14	9	9	0 / 0	9	0 / 0	9	0 / 0
15	1~3	1~3	0 / 0	1~3	0 / 0	1~3	0 / 0
16	1~4, 21	2	0 / 4	-	0 / 5	-	0 / 5
17	50	50	0 / 0	50	0 / 0	50	0 / 0
18	1~10	1~10, 11~14	4 / 0	1~10, 11~14	4 / 0	1~10, 11~14	4 / 0
19	22	22	0 / 0	22	0 / 0	22	0 / 0
20	25	25	0 / 0	25	0 / 0	-	0 / 1
21	4, 6, 8, 19	-	0 / 4	-	0 / 4	-	0 / 4
22	16~20	16~20	0 / 0	16~20	0 / 0	16~20	0 / 0
23	1~5	-	0 / 5	-	0 / 5	-	0 / 5
24	1~4	2	0 / 3	2	0 / 3	2	0 / 3
25	34~40	34~40	0 / 0	34~40	0 / 0	34~40	0 / 0
26	35~40	35, 36, 37	0 / 3	35, 36, 37	0 / 3	35, 37	0 / 4
27	1~5	2, 3, 34, 50	2 / 3	-	0 / 5	-	0 / 5

(-) 이상점이 포함되지 않았거나, 이상점이 포함되지 않았다고 판정된 자료임

현상 때문에 정확한 식별비율이 0.59로서 만족스럽지 않은 수준인데, 특히 붙음현상보다 가림현상이 심한 것으로 나타났다.

한편, CLUST의 경우 최적값으로 확인한 α=2.0를 비롯하여 α=1.7, α=2.3을 적용한 식별결과가 <표 5>에 수록되었는데, 최적치 α=2.0을 적용한 경우 11과 18번 자료에서 붙음현상이 발생하였고 10번 자료에서만 가림현상이 발생하였다. 24개의 자료에서 완벽한 식별이 이루어졌으며 정확한 식별비율이 0.89로서 LSOUT보다 매우 높은 수준이다. CLUST에서는 가림현상보다 붙음현상이 약간 심한 것으로 나타났는데 이는 로버스트 잔차를 바탕으로 식별하였기 때문인 것으로 해석된다.

<표 5> CLUST의 식별효용성 측정결과

자료 번호	이상점	α=1.7		α=2.0		α=2.3	
		식별된 이상점	붙음/ 가림 현상	식별된 이상점	붙음/ 가림 현상	식별된 이상점	붙음/ 가림 현상
1	9	9	0 / 0	9	0 / 0	9	0 / 0
2	12	12	0 / 0	12	0 / 0	12	0 / 0
3	4	4	0 / 0	4	0 / 0	4	0 / 0
4	18	18	0 / 0	18	0 / 0	18	0 / 0
5	1, 10, 16	1, 10, 16	0 / 0	1, 10, 16	0 / 0	1, 10, 16	0 / 0
6	-	1, 4, 11, 13, 16~20	9 / 0	-	0 / 0	-	0 / 0
7	18, 19	18, 19	0 / 0	18, 19	0 / 0	18, 19	0 / 0
8	15~24	15~24	0 / 0	15~24	0 / 0	15~20	0 / 4
9	19~25	19~25	0 / 0	19~25	0 / 0	-	0 / 7
10	6, 16, 25	25	0 / 2	25	0 / 2	25	0 / 2
11	11, 20, 30, 34	14, 7, 11, 20, 30, 34	2 / 0	7, 11, 20, 30, 34	1 / 0	7, 11, 20, 30, 34	1 / 0
12	-	8	1 / 0	-	0 / 0	-	0 / 0
13	17	17	0 / 0	17	0 / 0	17	0 / 0
14	9	9	0 / 0	9	0 / 0	9	0 / 0
15	1~3	1~3	0 / 0	1~3	0 / 0	1~3	0 / 0
16	1~4, 21	1~4, 21	0 / 0	1~4, 21	0 / 0	21	0 / 4
17	50	50	0 / 0	50	0 / 0	50	0 / 0
18	1~10	1~10, 13, 14	2 / 0	1~10, 13, 14	2 / 0	1~10, 13, 14	2 / 0
19	22	22	0 / 0	22	0 / 0	22	0 / 0
20	25	25	0 / 0	25	0 / 0	25	0 / 0
21	4, 6, 8, 19	4, 6, 8, 19	0 / 0	4, 6, 8, 19	0 / 0	4, 6, 8, 19	0 / 0
22	16~20	16~20	0 / 0	16~20	0 / 0	16~20	0 / 0
23	1~5	1~5	0 / 0	1~5	0 / 0	1~5	0 / 0
24	1~4	1~4	0 / 0	1~4	0 / 0	1~4	0 / 0
25	34~40	34~40	0 / 0	34~40	0 / 0	34~40	0 / 0
26	35~40	35~40	0 / 0	35~40	0 / 0	35~40	0 / 0
27	1~5	1~5	0 / 0	1~5	0 / 0	1~5	0 / 0

(-) 이상점이 포함되지 않았거나, 이상점이 포함되지 않았다고 판정한 자료임

5. 결 론

대규모 회귀자료에서의 이상점 식별을 위하여 LMS-잔차산점도에 군집화 기법을 적용한 이상점 식별알고리즘을 제안하였다. 다양한 형태의 이상점들이 포함된 자료들을 대상으로 식별방법을 적용한 결과, LS-추정을 적용한 식별방법보다 로버스트 추정인 LMS-추정을 적용한 식별방법이 가림현상이나 붙음현상을 잘 극복하여 이상점을 매우 효과적으로 식별해낸다는 사실을 확인하였다. 한편, 자료탐색 결과를 통해 경계선 높이가 낮아질수록 정상점을 이상점으로 식별하는 붙음현상이 많이 나타나며 경계선 높이가 높아질수록 가림현상이 발생하는 자료가 많음을 확인할 수 있었다. 따라서 새로운 식별알고리즘의 군집분류 기준으로 적용되는 경계선 높이의 계산을 위해 사전에 결정해야 하는 α값의 최적화를 위한 연구를 실행하였는데 α

=2.0이 최적치임을 밝혀냈다.

참고 문헌

[1] Basset, Jr. G. W.(1991). Equivariant, monotonic, 50% breakdown estimators, *The American Statistician*, Vol. 45, 135-137.

[2] Belsely, D. A., Kuh, E. and Welsh, R. E.(1980). *Regression Diagnostics: Influential Data and Source of Collinearity*. Wiley, New York.

[3] Cook, R. D. and Weisberg, S.(1980). Characterizations of an empirical influence function for detecting influential cases in regression, *Technometrics*, Vol. 22, 495-508.

[4] Everitt, B. S.(1993). *Cluster Analysis*, Halsted Press, New York.

[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. 88, 1264-1272.

[6] Hartigan, J. A.(1975). *Clustering Algorithms*, Wiley, New York.

[7] 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.

[8] Kianifard, F. and Swallow, W. H.(1990). A Monte Carlo comparison of five procedures for identifying outliers in linear regression, *Commun. Statist.-Theory Meth.*, Vol. 19, 1913-1938.

[9] Kim, B. Y.(1996). L_∞-estimation based algorithm for the least median of squares estimator, *The Korean Communications in Statistics*, Vol. 3, 299-307.

[10] Kim, B. Y. and Kim, H. Y.(2002). A hybrid algorithm for identifying multiple outliers in linear regression, *The Korean Communication in Statistics*, Vol. 9, 291-304.

[11] Marasinghe, M. G.(1985). A multistage procedure for detecting several outliers in linear regression, *Technometrics*, Vol. 27, 395-399.

[12] Mojena, R.(1977). Hierarchical grouping methods and stopping rules: an evaluation, *Computer Journal*, Vol. 20, 359-363

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

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

[15] Rousseeuw, P. J. and Zomeren, B. C.(1990). Unmasking multivariate outliers and leverage points, *Journal of the American Statistical Association*, Vol. 85, 633-639.

[16] Sebert, D. M., Montgomery, D. C. and Rollier, D. A.(1998). A clustering algorithm for identifying multiple outliers in linear regression, *Computational Statistics & Data Analysis*, Vol. 27, 461-484.

[2004년 5월 접수, 2004년 8월 채택]