The Korean Journal of Applied Statistics (2012)
25*(3), 503–512*

DOI: http://dx.doi.org/10.5351/KJAS.2012.25.3.503

Clustering Observations for Detecting Multiple Outliers in Regression Models

Han Son Seo[1] · Min Yoon[2]

[1]Department of Applied Statistics, Konkuk University
[2]Department of Statistics, Pukyong National University

(Received March 2, 2012; Revised April 6, 2012; Accepted April 17, 2012)

Abstract

Detecting outliers in a linear regression model eventually fails when similar observations are classified differently in a sequential process. In such circumstances, identifying clusters and applying certain methods to the clustered data can prevent a failure to detect outliers and is computationally efficient due to the reduction of data. In this paper, we suggest to implement a clustering procedure for this purpose and provide examples that illustrate the suggested procedure applied to the Hadi-Simonoff (1993) method, reverse Hadi-Simonoff method, and Gentleman-Wilk (1975) method.

Keywords: Clustering, linear regression model, outliers, regression diagnostics.

1. 서론

모형추정과정에서 자료에 이상치가 포함되는 경우 여러 문제가 발생한다. 특히 모수의 추정값이 이상치에 민감하게 영향을 받는 선형회귀모형 분석에서 이상치 탐지를 위한 다양한 방법들이 제안, 연구되고 있다. Gentleman과 Wilk (1975)는 이상치군의 크기를 미리 추정한 후 해당되는 크기의 관찰치군들 중에서 관찰치군을 제외하고 추정된 모형의 잔차 합을 기준으로 이상치군을 선정하였다. Marasinghe (1985)은 이상치군 탐지에 요구되는 계산량을 축소하기 위해 다단계 방식으로 이상치 후보군을 구성한 후 관찰치 각각에 대하여 이상치 여부를 검정하였다. Paul과 Fung (1991)은 이상치 후보군의 구성을 위해 일반적 극단스튜던트화 잔차(generalized extreme studentized residual; GESR)와 영향치 정도에 관련된 측도를 함께 고려하였다. Kianifard와 Swallow (1989, 1990)는 반복잔차(recursive residual)를 이용하여 이상치 후보군을 선정하고 후보군에 속하는 관찰치에 대하여 순차적 검정을 수행하였다. Hadi와 Simonoff (1993)은 일정 크기의 양호 관찰치군으로부터 모형을 추정한 후 검정을 통해 이상치군의 크기를 한 개씩 줄여가는 방법을 제안하였다. 이상치군 탐지방법에서 계산량을 축소하기 위하여 Atkinson (1994)은 전진탐색법을 제안하였으며 Pena와 Yohai (1999)는 설명변수들이 많을 때 적용될 수 있는 이상치군 탐지법을 제시하였다. Jajo (2005)는 제안된 여러 가지 이상치군 탐지방법에 대한 장단점을 비교하였다. 이상치군 탐지에서 이상치군의 크기를 결정하는 과정은 이상치군의 크기를 사전

This work was supported by Konkuk University (2011).
[2]Corresponding author: Assistant Professor, Department of Statistics, Pukyong National University, 599-1 Daeyeon 3-Dong, Nam-Gu, Busan 608-737, Korea. E-mail: myoon@pknu.ac.kr

에 정해두거나 임의적인 판단에 의해 정해지는 것과 (Gentleman과 Wilk, 1975; Marasinghe, 1985) 이상치군의 크기를 순차적 검정을 통해 결정하는 것 (Hadi와 Simonoff, 1993)으로 구분될 수 있으며 또한 이상치군 크기에 따른 각 단계에서 그 이전 단계에서 탐지한 이상치군을 이용하는 방법과 각 단계별로 독립적으로 이상치군을 탐지하는 방법으로 분류할 수 있다. 두 종류의 이상치 탐지방법에서 공통적으로 중요하게 고려되는 점은 이상치군의 크기별로 전체 관찰치를 이상치군과 양호치군으로 나누게 되는 과정에서 요구되는 계산량과 이상치 탐지의 정확도이다. 두 종류의 방법은 모두 각 단계별로 이상치군을 왜곡되게 탐지할 수 있으며 특히 단계별 순차적 방법은 이전 단계에서 이상치군을 잘못 선정하였을 때 그 영향이 지속되어 결과적으로 오류를 범할 가능성이 커진다. 이와 같은 문제를 해결하기 위하여 상대적으로 계산량에서 유리한 단계별 순차적 방법을 수행하면서 일관성이 결여되는 양호치군을 점검하는 방법이 제안되었다 (Ahn과 Seo, 2011). 본 논문에서는 인접한 관찰치가 근소한 차이로 인하여 이상치군과 양호치군으로 상반되게 평가될 때 전체적으로 이상치군 탐지에서 가면화 효과(masking effect)나 수렁효과(swamping effect)와 같은 오류가 발생하는 경우가 많은 점을 고려하여 관찰치의 군집화를 먼저 수행한 후 이상치군 탐지방법을 적용하는 과정을 제안한다. 가면화 효과등에 강건한 이상치 탐지법으로 LMS를 이용한 방법 (Rousseeuw, 1984) 등이 제안되어 있지만 LMS를 적용함에 있어서 가장 큰 문제점은 과다한 계산량에 있다. 군집화를 순차적 이상치 탐지방법에 적용한다면 상대적으로 계산량이 작으면서 수렁화나 가면화로부터 자유로운 절차를 구현할 수 있다. 군집화에 의한 진단법은 영향치 분석 등에서 제안된 바 있으며 (Gray와 Ling, 1984) 군집화에 의하여 이상치 탐지의 정확성을 높일 수 있을 뿐만 아니라 자료의 크기가 줄어짐에 따라 계산량의 감소도 기대할 수 있다. 본 논문에서는 각 단계별로 독립적으로 이상치군을 탐지하여 많은 계산량이 필요한 Gentleman-Wilk (1975)의 방법과 계산량에서 유리한 순차적 방법인 Hadi-Simonoff (1993)의 방법 그리고 Hadi-Simonoff 보다 적은 계산량으로 수행할 수 있는 역 Hadi-Simonoff 방법을 제안하여 세 가지 방법에 의해 군집화 과정의 효율성을 검증한다. 2장에서는 본 논문에서 고려하는 세 가지 이상치 탐지법 과정을 설명하며 이 방법들을 적용하여 이상치군 탐지에 실패하게 되는 예를 제시한다. 3장에서는 군집화 과정을 설명하고 군집화 과정을 통해 개선된 효과를 검증한다. 4장은 본 연구의 결과를 요약 정리하기로 한다.

2. 이상치 탐지법

통계분석의 대표적인 기법인 선형회귀모형 분석은 설명변수와 반응변수 간에 다음과 같은 관계식을 가정한다.

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

여기서 Y는 $n \times 1$ 반응변수 벡터, $\boldsymbol{\beta}$는 $p \times 1$ 회귀계수벡터, $\boldsymbol{X}$는 p개의 설명변수를 나타내는 $n \times p$ 행렬이며 $\boldsymbol{\varepsilon}$는 평균이 0이고 분산행렬이 $\sigma^2 I_n$인 $n \times 1$ 오차벡터이다.

선형회귀분석에서 다수의 이상치를 검색하기 위해 여러 가지 방법 중 본 논문에서 다루게 될 기존의 방법은 Gentleman-Wilk (1975)의 방법과 Hadi-Simonoff (1993) 방법이며 Hadi-Simonoff (1993) 방법에서 탐지방향을 역순으로 진행하는 역 Hadi-Simonoff 방법을 제안하여 적용한다.

세 가지 방법 중 Gentleman-Wilk의 방법은 이상치군의 크기별로 전수조사를 하므로 계산량이 가장 많다. Gentleman-Wilk의 방법은 간략히 다음과 같이 설명될 수 있다.

1. 잠재적인 이상치의 숫자 k를 정한다.
2. 전체 데이터 n개에서 k개로 구성되는 가능한 모든 조합 ${}_nC_k$에 대하여 k개의 관찰치를 제거했을 때

발생하는 잔차 제곱합 축소치인 Q_k를 계산한다.

$$Q_k = \sum_{i \in I_k} t_i^2,$$

여기서 $t_i = e_i/\sqrt{(1-h_{ii})},\ i = 1, \ldots, n,\ e_i$는 잔차, h_{ii}는 해트행렬의 대각원소이다.

3. Q_k 중 가장 큰 값 Q_k^*를 찾는다.
4. 만약 Q_k^*가 통계적으로 충분히 크다면 Q_k^*에 해당하는 k개 관찰치를 이상치로 판단하며 Q_k^*가 충분히 크지 않다면 $(k-1)$개의 관찰치를 가지고 위의 과정을 반복한다.

일반적으로 단계 3에서 Q_k^*를 찾는 알고리즘은 알려져 있지 않다. 또한 단계 4에서 이상치의 크기를 결정하는 과정에서 주로 Q-Q 그림이나 P-P 그림이 사용되며 주관적인 판단을 따르게 된다.

Hadi-Simonoff 방법은 기초양호군에서 시작하여 각 단계별로 이상치군의 크기를 줄여가면서 이상치군에 대한 최종적인 이상치 여부를 순서통계량에 대한 t 검정의 결과를 적용하여 결정한다. Hadi-Simonoff 방법의 대략적인 절차는 다음과 같다.

1. 크기 s개의 양호치군을 생성한다. 초기 양호치군의 크기 s는 $\mathrm{int}[(n+p-1)/2]$이다.
2. 양호치군만으로 모형을 추정한 후 양호치군(M)과 이상치군에 속한 관찰치에 내적 스튜던트화 잔차(internally studentized residual) d_i를 계산한다.

$$d_i = \begin{cases} \dfrac{\left(y_i - x_i^T \hat{\beta_M}\right)}{\hat{\sigma}_M \sqrt{1 - x_i^T \left(X_M^T X_M\right)^{-1} x_i}}, & \text{if } i \in M, \\ \dfrac{\left(y_i - x_i^T \hat{\beta_M}\right)}{\hat{\sigma}_M \sqrt{1 + x_i^T \left(X_M^T X_M\right)^{-1} x_i}}, & \text{if } i \notin M. \end{cases}$$

3. $|d|_{(j)}$를 $|d_i|$의 오름차순으로 정리한 j번째 순서 통계량이라고 할 때 만약 $|d|_{(s+1)} \geq t_{(\alpha/\{2(s+1)\},s-k)}$라고 하면 $(n-s)$개의 후 순위 순서통계량에 속하는 관찰치를 최종 이상치군으로 판단하며 그렇지 않을 경우 양호치군의 숫자를 한 개 늘려서 앞의 절차를 반복한다.

Hadi-Simonoff 방법과 같은 순차적인 방법은 초기 양호치의 선정이 전체 절차의 정확성에 매우 중요한 역할을 하며 이와 관련하여 Hadi와 Simonoff도 두 가지의 기초군 선정 방법을 제시하고 있다. Hadi-Simonoff 방법은 대략 전체 데이터의 반 정도에 해당하는 부분 데이터로부터 시작하여 순차적인 검정을 진행하므로 데이터의 크기가 크면 초기 검정 단계에서 계산량이 많다는 단점이 있다. 이를 극복하기 위한 방안으로 Hadi-Simonoff 방법을 역방향으로 적용하여 이상치를 늘여가는 방법을 생각해 볼 수 있다. 역방향 Hadi-Simonoff 방법은 다음과 같이 정리될 수 있다.

1. 전체 자료에 의해 모형추정을 한 후 $d_{(n)}$에 해당하는 관찰치를 제외한 나머지 $(n-1)$개의 관찰치를 양호군으로 간주한 후 앞서 설명한 Hadi-Simonoff의 이상치 검정을 수행한다.
2. 이상치의 개수를 한 개씩 늘려서 Hadi-Simonoff의 이상치 검정을 반복한다. 어느 단계에서 귀무가설이 채택되고 그 이후 일정 정도의 크기까지 검정결과가 계속 기각되면 귀무가설이 채택된 단계의 결과를 이상치군으로 판단한다.

Hadi-Simonoff와 역 Hadi-Simonoff 방법은 특정 단계에서 이상치 탐지에 오류가 발생할 경우 그 이후 단계에서도 영향을 받게 된다. 반면에 Gentleman-Wilk 방법은 이전 단계의 결과와 상관없이 이상치군

Table 2.1. Artificial data with 7 outliers

번호	X	Y	번호	X	Y
1	−4	0	14	9.87	10.11
2	20	24	15	2.55	3.03
3	19.9	23.9	16	7.51	6.86
4	19.8	23.8	17	2.67	2.1
5	−5	−9	18	4.4	3.74
6	−4.9	−8.9	19	7.65	7.57
7	−4.8	−8.8	20	7.01	6.4
8	11.36	11.1	21	1.28	1.05
9	11.66	11.92	22	4.48	4.72
10	0.2	−0.27	23	8.73	9.39
11	5.27	4.95	24	4.36	4.63
12	10.52	11.38	25	5.47	6.04
13	6.16	6.34			

Table 2.2. Results of applying Gentleman-Wilk method, Hadi-Simonoff method, reverse Hadi-Simonoff method for Table 2.1 data

크기	G-W 이상치	Hadi-Simonoff 이상치	검정	역 Hadi-Simonoff 이상치	검정
8	1 8 9 10 15 21 22 24	1 8 10 15 21 22 24 25	N	1 8 9 10 15 21 22 24	N
7	1 8 9 10 15 21 25	1 8 9 10 15 21 22	N	1 8 9 10 15 21 25	N
6	1 8 9 10 15 21	1 8 9 10 15 21	N	1 8 9 10 15 21	N
5	1 8 10 15 21	1 8 10 15 21	N	1 8 10 15 21	N
4	1 8 15 21	1 8 15 21	N	1 8 15 21	N
3	1 8 15	1 8 15	N	1 8 15	N
2	1 8	1 8	N	1 8	N
1	1	1	Y	1	Y

Table 2.3. Modified Paul-Fung data

ID	1	2	3	4	5	6	7	8	9	10
X	1	2	3	4	5	6	−5	−4.9	10	9.9
Y	2.01	3.01	4.03	5.03	6.02	7.01	−4.01	−4	5.02	5

을 탐지함으로써 이전 단계에서의 오류에 영향을 받지 않지만 이 방법도 가면화 현상 등으로 부터 자유롭지 못하다. 다음의 예제들은 세 가지 방법들이 이상치군을 제대로 탐지하지 못하는 경우를 보여주고 있다.

예제 2.1: Table 2.1의 모의자료에서 X_i는 균등분포 $U(0, 15)$에서 생성되었다. 18개의 정상적인 Y_i는 X_i와 $\varepsilon_i \sim N(0, 0.5^2)$에 의해 모형 $Y_i = X_i + \varepsilon_i$, $i = 8, \ldots, 25$로 부터 생성되었고 나머지 7개 자료 $i = 1, \ldots, 7$는 모형에서 4 또는 6만큼 벗어나 이상치군을 형성하도록 생성되었다. Table 2.1의 자료에 대하여 이상치 탐지방법을 실행한 결과 Table 2.2에서 보듯이 Gentleman-Wilk 방법과 Hadi-Simonoff 방법, 역 Hadi-Simonoff 방법 모두 이상치인 관찰치 1−7을 탐지하는데 실패하는 것을 알 수 있다.

예제 2.2: 두 번째 예제에서 사용하는 자료는 Paul과 Fung (1991)의 자료에서 Figure 2.1과 같이 관찰치 9와 10이 이상치군을 형성하도록 수정된 자료이다 (Table 2.3). 이 자료에 대하여 세 가지 방법을 적용하여 이상치를 탐지한 결과 Table 2.4에서 알 수 있듯이 Gentleman-Wilk 방법은 이상치의 각 단계별

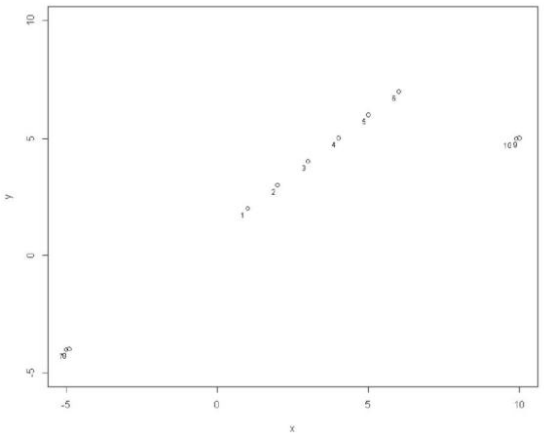

Figure 2.1. Scatter plot of modified Paul-Fung data

Table 2.4. Results of applying Gentleman-Wilk method, Hadi-Simonoff method, reverse Hadi-Simonoff method for Table 2.3 data

크기	G-W 이상치	Hadi-Simonoff 이상치	검정	역 Hadi-Simonoff 이상치	검정
5	3 4 8 9 10	2 6 8 9 10	N	3 6 8 9 10	N
4	6 8 9 10	6 8 9 10	N	6 8 9 10	N
3	8 9 10	8 9 10	Y	8 9 10	Y
2	9 10	9 10	Y	9 10	Y
1	9	10	N	10	N

Table 2.5. Artificial data with 2 outliers

ID	1	2	3	4	5	6	7	8	9	10
X_1	5.3251	6.3405	9.2053	8.9531	8.557	5.7429	5.5882	8.8267	9.0339	7.0505
X_2	1	2	3	4	5	6	-5	-4.9	10	9.9
Y	8.1066	9.9107	13.911	14.024	17.391	13.221	1.4282	4.6567	14.164	12.231

로 역 Hadi-Simonoff 방법에서와 동일한 관찰치를 이상치로 탐지하고 있고 세 방법은 모두 관찰치 8이 수렁효과에 영향을 받는 것을 알 수 있다.

예제 2.3: 이상치 탐지과정에서 사용되는 잔차는 일반적으로 전체 모형으로부터 계산되지만 때로는 부분잔차(partial residual)가 더 유용한 경우가 있다. 부분잔차는 부분선형모형(partial linear model)에서 비선형 함수의 모양을 식별하는데 사용되며 이 과정에서 이상치를 탐지할 수 있다. 이와 관련된 이상치 탐지를 위하여 다음과 같이 크기 $n = 10$의 모의자료를 생성한다 (Table 2.5).

X_1은 $X_1 \sim 5 + 5 \times U(0, 1)$에서 생성되었으며 X_2는 Table 2.3의 X와 동일하다. Y는 모형 $Y = X_1 + Y_0 + \varepsilon$으로부터 생성되었으며 이때 Y_0는 Table 2.3의 Y와 동일하며 ε은 표준정규분포를 따른다. 따라서 모의자료에 대하여 선형모형이나 부분선형모형 $Y = \beta_1 X_1 + f(X_2) + \varepsilon$을 구현하면 $f(X_2)$은 선형함수로 추정되며 이 과정에서 관찰치 9, 10은 이상치로 탐지되어야 한다.

Table 2.6은 부분선형모형 $Y = \beta_1 X_1 + f(X_2) + \varepsilon$에서 계산된 부분잔차를 이용하여 이상치를 탐지한 결과를 보여주고 있다. 세 방법 모두 관찰치 5가 수렁효과에 영향을 받아 이상치를 올바르게 탐지하지 못하는 것을 알 수 있다.

Table 2.6. Results of applying Gentleman-Wilk method, Hadi-Simonoff method, reverse Hadi-Simonoff method for partial linear model fitted Table 2.5 data

크기	G-W 이상치	Hadi-Simonoff 이상치	검정	역 Hadi-Simonoff 이상치	검정
5	1 4 5 9 10	4 5 8 9 10	N	1 4 5 9 10	N
4	4 5 9 10	4 5 9 10	N	4 5 9 10	N
3	5 9 10	5 9 10	Y	5 9 10	Y
2	9 10	9 10	Y	5 10	N
1	5	10	N	5	N

Table 2.7. Artificial data with 5 explanatory variables

ID	X_1	X_2	X_3	X_4	X_5	Y
1	7.54	8.60	8.29	9.93	1.00	35.03
2	5.65	7.21	8.17	6.78	1.50	28.76
3	9.78	8.42	5.47	9.14	2.00	34.11
4	7.89	9.67	9.95	5.67	2.50	34.99
5	5.50	9.76	9.09	7.70	3.00	35.14
6	9.31	8.54	9.33	7.44	3.50	37.69
7	8.96	5.96	8.77	5.87	4.00	33.72
8	5.16	8.86	6.15	5.48	4.50	29.67
9	6.18	7.44	6.02	5.78	5.00	29.87
10	7.34	7.74	7.27	8.26	5.50	37.74
11	9.25	9.79	9.57	5.98	6.00	40.31
12	6.32	9.43	8.44	5.93	−5.00	24.96
13	9.16	6.74	7.91	7.48	−4.90	26.13
14	5.12	8.72	6.84	5.89	10.00	30.70
15	5.44	8.48	7.16	6.18	9.90	31.43

Table 2.8. Results of applying Gentleman-Wilk method, Hadi-Simonoff method, reverse Hadi-Simonoff method for Table 2.7 data

크기	G-W 이상치	검정	Hadi-Simonoff 이상치	검정	역 Hadi-Simonoff 이상치	검정
5	4 7 10 14 15	N	5 10 12 14 15	N	4 7 10 14 15	N
4	7 10 14 15	N	5 10 14 15	N	7 10 14 15	N
3	10 14 15	Y	10 14 15	Y	10 14 15	Y
2	14 15	N	14 15	Y	10 15	N
1	10	N	14	N	10	N

예제 2.4: 설명변수의 개수가 많은 경우 이상치 탐지법을 적용하기 위하여 다음과 같이 5개의 설명변수 X_1, X_2, X_3, X_4, X_5와 반응변수 Y를 임의로 생성한다 (Table 2.7). 설명변수 X_1, X_2, X_3, X_4는 $X_i \sim 5 + 5 \times U(0,1)$ $i = 1, 2, 3, 4$에서 생성되었으며 X_5는 1에서 6 사이에 동일간격으로 열 한개의 값을 생성하였고 양극단으로 네 개의 값을 추가하였다. 관찰치 1에서 13까지의 반응변수 Y는 표준정규분포를 따르는 ε을 포함하여 모형 $Y = X_1 + X_2 + X_3 + X_4 + X_5 + \varepsilon$에서 생성되었으며 관찰치 14와 15의 반응변수 Y는 모형 $Y = X_1 + X_2 + X_3 + X_4 + X_5 - 6 + \varepsilon$에서 생성되었다. 따라서 관찰치 14, 15가 이상치로 탐지되어야 한다.

이 자료에 대하여 세 가지 방법을 적용하여 이상치를 탐지한 결과 Table 2.8에서 보듯이 세 방법은 모두 수렁효과에 영향을 받아 정상치인 관찰치 10을 이상치로 탐지한다.

Table 3.1. Clustered Table 2.1 data

New ID	Old ID	X	Y	New ID	Old ID	X	Y
1	2 4 3	19.90	23.900	11	14	9.87	10.110
2	7 6 5	−4.90	−8.900	12	15	2.55	3.030
3	24 22	4.42	4.675	13	16	7.51	6.860
4	1	−4.00	0.000	14	17	2.67	2.100
5	8	11.36	11.100	15	18	4.40	3.740
6	9	11.66	11.920	16	19	7.65	7.570
7	10	0.20	−0.270	17	20	7.01	6.400
8	11	5.27	4.950	18	21	1.28	1.050
9	12	10.52	11.830	19	23	8.73	9.390
10	13	6.16	6.340	20	25	5.47	6.040

3. 군집화를 통한 이상치 탐지

앞 절에서 보듯이 단계별 독립적 이상치 탐지방법이나 순차적 이상치 탐지방법은 모두 가면화 효과나 수렁효과에 취약할 가능성이 있으며 이는 대부분의 경우 인접한 관찰치들을 상반되게 평가하는 경우에 발생된다. 따라서 관찰치군을 일정기준에 따라 군집화하여 인접한 관찰치들을 동일한 군으로 자료를 재편성하면 이상치 탐지의 오류를 방지할 수 있다. 군집화는 주로 다변량 자료에서 직접적으로 이상치군을 탐지할 때 사용되지만 (Atkinson 등, 2004) 군집화를 매개로 모형분석 진단을 수행하는 방법이 Gray와 Ling (1984) 등에 의하여 시도된 바 있다. Gray와 Ling의 경우에는 영향치군을 탐색함에 있어서 설명변수 행렬 X뿐만 아니라 반응변수 벡터 Y가 추가된 행렬을 이용하여 수정된 해트행렬(hat matrix)을 계산하고 수정된 해트행렬이 집단 대각(block-diagonal) 구조를 가지도록 군집화를 통하여 관찰치를 순열화한다. 이와 같은 군집화에 의하여 집단화된 자료에 대하여 영향치 여부를 판단하면 가면화 효과의 가능성이 낮아지게 된다. 일반적으로 사용되는 이상치의 측도는 잔차등과 같은 단변량 값이므로 Gray와 Ling의 접근법을 이상치 탐지과정에 직접적으로 적용하는 것은 불가능하지만 관찰치의 위치를 기준으로 자료의 집단화를 수행하면 동일한 효과를 기대할 수 있다.

군집화의 방법은 다양하게 제안되어 있으며 (Cormack, 1971) Gray와 Ling은 인접 이웃 알고리즘(nearest-neighbor algorithm)을 확장시킨 k-군집화 알고리즘(k-clustering algorithm)을 사용하고 있다 (Ling, 1972). 본 논문에서는 자료의 군집화를 위하여 병합적 방법(agglomerative nesting) (Kaufman과 Rousseeuw, 1990)을 사용하며 군집간 연결 기준은 중심연결 방법(centroid linkage method)이고 최종군집을 결정하는 군집간 거리는 0.5 이하로 하고 군집화된 자료의 값은 집단의 평균값을 사용한다. 군집간 거리에 대한 기준은 cross-validation 등 보다 더 정교한 방법을 적용하는 것도 가능하지만 인위적으로 무리하게 관찰치들이 군으로 분류되지 않도록 작은 값을 적용하였다.

본 연구에서 제안된 방법의 효과를 검증하기 위하여 2절에서 수행된 예제의 자료를 이용하여 군집화를 통한 이상치군 탐지를 수행한다.

예제 3.1: Table 2.1 자료를 X, Y에 의하여 군집화 한 결과 8개의 관찰치가 각각 3개, 3개, 2개씩 모여 하나의 집단을 형성하게 되어 Table 3.1과 같은 크기 20개의 군집화된 자료가 생성된다. 이 자료를 이용하여 이상치 탐지방법들을 수행하게 되면 Table 3.2에서 볼 수 있듯이 Gentleman-Wilk 방법과 Hadi-Simonoff 방법, 역 Hadi-Simonoff 방법 모두 군집자료상의 관찰치 1, 2, 4를 이상치로 탐지하여 이에 해당하는 원 자료인 Table 2.1의 관찰치 1, 2, 3, 4, 5, 6, 7이 이상치로 탐지된다.

Table 3.2. Results of applying Gentleman-Wilk method, Hadi-Simonoff method, reverse Hadi-Simonoff method for Table 3.1 data

크기	G-W 이상치	Hadi-Simonoff 이상치	검정	역 Hadi-Simonoff 이상치	검정
7	1 2 4 9 12 19 20	1 2 4 9 12 19 20	N	1 2 4 9 12 19 20	N
6	1 2 4 9 12 20	1 2 4 9 12 19	N	1 2 4 9 12 19	N
5	1 2 4 9 12	1 2 4 9 20	N	1 2 4 9 20	N
4	1 2 4 9	1 2 4 9	N	1 2 4 9	N
3	1 2 4	1 2 4	Y	1 2 4	Y
2	1 4	1 4	N	1 4	N
1	4	4	Y	1	Y

Table 3.3. Clustered Table 2.3 data

New ID	Old ID	X	Y	New ID	Old ID	X	Y
1	10 9	9.950	5.010	5	3	3.000	4.030
2	8 7	−4.950	−4.005	6	4	4.000	5.030
3	1	1.000	2.010	7	5	5.000	6.020
4	2	2.000	3.010	8	6	6.000	7.010

Table 3.4. Results of applying Gentleman-Wilk method, Hadi-Simonoff method, reverse Hadi-Simonoff method for Table 3.3 data

크기	G-W 이상치	Hadi-Simonoff 이상치	검정	역 Hadi-Simonoff 이상치	검정
4	1 2 3 4	1 2 7 8	N	1 4 7 8	N
3	1 7 8	1 2 8	N	1 2 8	N
2	1 2	1 8	N	1 2	N
1	1	1	Y	1	Y

Table 3.5. Clustered Table 2.5 data by partial residual and X_2

New ID	1	2	3	4	5	6	7	8
Old ID	8 7	10 9	1	2	3	4	5	6
X_2	−4.950	9.950	1.000	2.000	3.000	4.000	5.000	6.000
부분잔차	−6.946140	2.329360	0.101392	0.835595	1.817340	2.196070	5.980430	4.775570

예제 3.2: Table 2.3의 수정된 Paul-Fung 자료를 군집화 통해 재정리된 자료는 Table 3.3과 같다. 이 자료에 의해 이상치 탐지법을 적용한 결과 원 자료가 수렁효과에 영향을 받는 것과는 달리 Table 3.4에서 보듯이 원 자료의 관찰치 9, 10에 해당하는 군집화 자료의 관찰치 1을 이상치로 올바르게 탐지한다.

예제 3.3: 자료의 군집화는 관련된 변수값을 대상으로 수행할 수 있지만 모형의 추정과정에 따라 군집대상을 적절히 선택할 수 있다. 부분선형모형에서는 관찰값 (Y, X_1, X_2)을 사용하기 보다는 곡선함수 $f(X_2)$를 추정하는데 사용된 부분잔차와 해당 변수 X_2에 의해 군집화를 하는 것이 더 유용하다. Table 3.5는 X_2와 부분잔차에 의하여 원 자료인 Table 2.5의 자료를 군집화한 결과를 보여준다.

Table 3.5의 자료를 이용하여 부분잔차와 X_2에 대한 세 가지 추정법을 적용한 결과 Table 3.6에서 보듯이 원 자료의 관찰치 9, 10에 해당하는 군집자료의 관찰치 2가 이상치로 탐지된다.

예제 3.4: 설명변수가 다섯 개인 Table 2.7의 자료를 군집화한 결과는 Table 3.7과 같다. 군집화된 자료에 대한 이상치 탐지 결과인 Table 3.8은 원 자료를 사용한 경우와 달리 원 자료의 관찰치 14, 15에 해

Table 3.6. Results of applying Gentleman-Wilk method, Hadi-Simonoff method, reverse Hadi-Simonoff method for Table 3.5 data

크기	G-W 이상치	Hadi-Simonoff 이상치	검정	역 Hadi-Simonoff 이상치	검정
4	1 2 6 7	1 2 7 8	N	1 2 6 7	N
3	1 2 6	1 2 7	N	2 6 7	N
2	2 7	2 7	N	2 7	N
1	2	2	Y	2	Y

Table 3.7. Clustered Table 2.7 data

New ID	Old ID	X_1	X_2	X_3	X_4	X_5	Y
1	8 9	5.67	8.15	6.08	5.63	4.75	29.77
2	14 15	5.28	8.60	7.00	6.03	9.95	31.07
3	1	7.54	8.60	8.29	9.93	1.00	35.03
4	2	5.65	7.21	8.17	6.78	1.50	28.76
5	3	9.78	8.42	5.47	9.14	2.00	34.11
6	4	7.89	9.67	9.95	5.67	2.50	34.99
7	5	5.50	9.76	9.09	7.70	3.00	35.14
8	6	9.31	8.54	9.33	7.44	3.50	37.69
9	7	8.96	5.96	8.77	5.87	4.00	33.72
10	10	7.34	7.74	7.27	8.26	5.50	37.74
11	11	9.25	9.79	9.57	5.98	6.00	40.31
12	12	6.32	9.43	8.44	5.93	−5.00	24.96
13	13	9.16	6.74	7.91	7.48	−4.90	26.13

Table 3.8. Results of applying Gentleman-Wilk method, Hadi-Simonoff method, reverse Hadi-Simonoff method for Table 3.7 data

크기	G-W 이상치	검정	Hadi-Simonoff 이상치	검정	역 Hadi-Simonoff 이상치	검정
4	1 2 4 5	N	2 7 10 12	N	2 6 9 10	N
3	2 9 10	N	2 7 10	N	2 9 10	N
2	2 10	N	2 10	N	2 10	N
1	2	Y	2	Y	2	Y

당하는 군집화 자료의 관찰치 2가 이상치로 올바르게 탐지되며 군집화는 설명변수가 많은 경우에도 수렁효과나 가면화 효과의 오류를 개선하는데 효과가 있음을 보여준다.

4. 결론

본 연구에서는 인접한 관찰치들을 이상치 관점에서 서로 다르게 평가하여 순차적인 이상치 탐지과정에서 가면화 효과나 수렁효과가 발생하는 것을 개선하고 부수적으로 탐지과정에서 필요한 계산량을 감소시키는 의도로써 군집화에 의한 이상치군 탐지방법을 제안한다. 군집화의 효과를 검증하기 위하여 적용된 방법은 Gentleman-Wilk 방법, Hadi-Simonoff 방법, 역 Hadi-Simonoff 방법이다. 그 중에서 역 Hadi-Simonoff 방법은 이상치군의 크기를 줄여가는 순차적 방법인 Hadi-Simoff 방법이 데이터의 크기가 크면 계산량의 부담이 커진다는 점을 감안하여 가산형 이상치 탐지법으로 제안된 방법이며 절차상 가면화 효과에 취약할 수 있지만 계산량이 현저히 줄어든다는 장점을 기대할 수 있다. 예제를 통해 세 가지 방법 모두에서 군집화는 가면화 효과와 수렁효과를 방지하는데 도움이 된다는 것을 검증할 수 있다. 데이터의 크기가 큰 경우 본 논문에서 제안하는 방법은 기초적인 이상치군의 탐색에도 도움이 될 것으로

기대되며 부분선형모형의 예에서 보듯이 군집화는 관찰치의 위치뿐만 아니라 모형과 환경에 따른 다양한 값을 기반으로 수행될 수 있다.

References

Ahn, B. J. and Seo, H. S. (2011). Outlier detection using dynamic plots, *The Korean Journal of Applied Statistics*, **24**, 979–986.

Atkinson, A. C. (1994). Fast very robust methods for the detection of multiple outliers, *Journal of the American Statistical Association*, **89**, 1329–1339.

Atkinson, A. C., Riani, M. and Cerioli, A. (2004). *Exploring Multivariate Data with The Forward Search*, Springer, New York.

Cormack, R. M. (1971). A review of classification, *Journal of the Royal Statistical Society, Series A*, **134**, 321–367.

Gentleman, J. F. and Wilk, M. B. (1975). Detecting outliers. Ⅱ. supplementing the direct analysis of residuals, *Biometrics*, **31**, 387–410.

Gray, J. B. and Ling, R. F. (1984). K-clustering as a detection tool for influential subsets in regression, *Technometrics*, **26**, 305–318.

Hadi, A. S. and Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models, *Journal of the American Statistical Association*, **88**, 1264–1272.

Jajo, N. K. (2005). A review of Robust regression an diagnostic procedures in linear regression, *Acta Mathematicae Applicatae Sinica*, **21**, 209–224.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York.

Kianifard, F. and Swallow, W. H. (1989). Using recursive residuals, calculated on adaptively-ordered observations, to identify outliers in linear regression, *Biometrics*, **45**, 571–585.

Kianifard, F. and Swallow, W. H. (1990). A Monte Carlo comparison of five procedures for identifying outliers in linear regression, *Communications in Statistics*, **19**, 1913–1938.

Ling, R. F. (1972). On the theory and construction of k-clusters, *Computer Journal*, **15**, 326–332.

Marasinghe, M. G. (1985). A multistage procedure for detecting several outliers in linear regression, *Technometrics*, **27**, 395–399.

Paul, S. R. and Fung, K. Y. (1991). A generalized extreme studentized residual multiple-outlier-detection procedure in linear regression, *Technometrics*, **33**, 339–348.

Pena, D. and Yohai, V. J. (1999). A fast procedure for outlier diagnostics in linear regression problems, *Journal of the American Statistical Association*, **94**, 434–445.

Rousseeuw, P. J. (1984). Least median of squares regression, *Journal of the American Statistical Association*, **79**, 871–880.