The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.4.699
29(4), 699–706

A sequential outlier detecting method using a clustering algorithm

Han Son Seo[a] · Min Yoon[b,1]

[a]Department of Applied Statistics, Konkuk University;
[b]Department of Statistics, Pukyong National University

(Received March 4, 2016; Revised April 9, 2016; Accepted April 16, 2016)

Abstract

Outlier detection methods without performing a test often do not succeed in detecting multiple outliers because they are structurally vulnerable to a masking effect or a swamping effect. This paper considers testing procedures supplemented to a clustering-based method of identifying the group with a minority of the observations as outliers. One of general steps is performing a variety of t-test on individual outlier-candidates. This paper proposes a sequential procedure for searching for outliers by changing cutoff values on a cluster tree and performing a test on a set of outlier-candidates. The proposed method is illustrated and compared to existing methods by an example and Monte Carlo studies.

Keywords: clustering, linear regression model, outlier test, sequential procedure

1. 서론

선형회귀모형에서 모수 추정을 위해 일반적으로 사용되는 최소제곱추정법은 자료에 이상치가 존재할 때 왜곡된 결과를 초래할 수 있다. 이상치 문제를 해결하는 방식에는 직접 이상치를 탐지하여 제거하는 방식과 이상치에 대하여 강건 추정량이나 강건 기준을 적용하는 방식이 있다. 이상치를 탐지하는 과정은 일단 이상치 후보군을 탐지한 후 이상치 후보군 전체 혹은 일부에 대하여 검정을 수행하여 최종적인 이상치 여부를 결정한다. 이상치 후보군 전체에 대한 검정은 검정통계량의 정확한 분포가 알려져 있지 않은 경우가 많아서 실험에 의해 도출된 유의값이나 검정통계량의 근사분포에 의존해 수행된다. 반면에 이상치 후보군의 개별 관찰치에 대한 이상치 여부는 여러 유형의 t-검정에 의해 수행된다 (Peña와 Yohai, 1995).

본 연구에서는 탐지된 이상치 후보군에 대한 검정을 수행하여 그 결과에 따라 이상치 후보군을 재탐지하여 순차적으로 최종적인 이상치군을 판정하고자 한다. 이러한 과정이 적용될 기존의 이상치 탐지법은 군집화에 의한 이상치 탐지법이다. Sebert 등 (1998)은 단순결합에 의한 군집화를 수행하고 이에 따라 작성되는 군집나무에 적정기준을 적용하여 이상치를 탐지하는 방법을 제안하였으나 이상치군에 대한 검정과정이 생략되어 군집화의 오류에 따른 수렁화나 가면효과에 취약하다. 이에 대한 보완으로 본 연

This paper was supported by Konkuk University in 2015.
[1]Corresponding author: Department of Statistics, Pukyong National University, 45, Yongso-ro, Nam-Gu, Busan 48513, Korea. E-mail: myoon@pknu.ac.kr

구에서는 군집화 방법에서 결정된 기본적인 이상치군 전체에 대한 검정을 수행하여 이상치 여부를 판정한 후 그 결과에 따라 군집나무의 절단기준을 변경시켜 새로운 이상치군을 탐색하는 순차적 방법을 제안한다. 제안된 방법은 예제와 모의실험을 통해 검정절차가 없는 군집화 방법 (Sebert 등, 1998)과 군집화에 의해 결정된 이상치 후보군에 대하여 개별적 t-검정을 수행하는 방법 (Kim과 Krzanowski, 2007) 등과 효율성을 비교한다. 2장에서는 본 연구에서 제안된 방법에서 적용되는 이상치 검정법을 소개하며 3장에서는 군집화 방법과 본 연구에서 제안하는 순차적인 이상치 탐지법을 설명한다. 4장에서는 예제와 모의실험의 결과를 소개하고 5장에서는 연구결과를 요약한다.

2. 이상치후보군 검정법

다음과 같은 선형회귀모형을 고려하자

$$Y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

여기서 Y는 $n \times 1$ 반응변수 벡터이고 X는 $n \times k$ 설명변수 행렬이며 $\boldsymbol{\beta}$는 회귀계수 벡터, $\boldsymbol{\varepsilon}$은 $N(\mathbf{0}, \sigma^2 I_n)$을 따르는 오차를 표시한다.

선형회귀모형의 추정에서 이상치 후보군에 대한 최종적인 이상치 여부는 검정을 통해 결정된다. 일반적으로 후보군에 속한 각 관찰치에 본페로니(Bonferroni) 부등식을 적용한 개별적인 t-검정이 이용되지만 본 연구에서는 이상치 후보군의 부분집합을 대상으로 이상치 여부를 판정하는 절차를 사용한다. 임의의 관찰치군에 대한 이상치 검정방법 중 Seo와 Yoon (2014)은 Hadi와 Simonoff (1993)의 이상치 탐지과정을 응용하여 외적스튜던트화잔차와 표준화잔차의 순서통계량에 의해 검정대상군에 대한 이상치 여부를 판단한다. 검정대상군에 대한 이상치 판정은 해당 관찰치군의 최종 이상치 후보군 지정 여부와 지정될 경우 이상치 검정의 결과에 따라 결정된다. 즉 표준화잔차와 외적스튜던트화잔차에 의해 검정대상군이 최종 이상치 후보군으로 지정되지 않으면 검정대상군은 이상치가 아닌 것으로 판정하며, 최종 이상치후보군으로 지정되면 이상치 검정을 수행하여 그 결과에 따라 이상치 여부를 판정한다. 검정대상군에 대한 이상치 판정의 구체적인 절차는 다음과 같다.

(0) 이상치 여부를 판정할 검정대상군을 O라고 표기하고 크기는 r이라고 하자. O의 여집합 O^c를 잠정적 양호치군 M으로 지정하여 선형회귀식 (2.1)을 추정한다.

(1) 추정된 회귀모형으로 부터 다음과 같이 잠정적 양호치군 M에 속하는 관찰치는 표준화잔차를 계산하고 잠정적 이상치군 M^c에 속하는 관찰치는 외적스튜던트화잔차를 계산하여 이를 d_i라고 표기한다.

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

이때 X_M은 집합 M에 해당하는 X의 부분행렬, $\hat{\beta}_M$은 집합 M에 의해 추정된 회귀계수, $\hat{\sigma}_M$은 집합 M에 의해 계산된 σ의 추정치이다.

(2) $d_{(i)}$를 d_i의 절대값 $|d_i|$에 대한 오름차순 순서통계량이라고 할 때 $\{d_{(n-r+1)}, \ldots, d_{(n)}\}$에 해당하는 관찰치가 검정대상군 O와 일치하면 다음의 검정 절차를 수행하고 일치하지 않으면 단계 (3)을 수행한다.

1. 만약 $d_{(n-r+1)} \geq t_{(\alpha/2;(n-r+1),(n-r-k))}$이면, 검정 대상군 O를 이상치로 판단한다.
2. $d_{(n-r+1)} < t_{(\alpha/2;(n-r+1),(n-r-k))}$이면, 검정 대상군 O를 이상치가 아닌 것으로 판단한다.

(3) M과 M^c에 속하는 관찰치중 한 개 또는 일부를 서로 교체하여 새로운 잠정치 양호군 M에 의해 선형회귀식을 추정하고 단계 (1)에서 부터 위 과정을 재실행하며 O가 최종 이상치 후보군으로 선정될 때 까지 일정 횟수 내에서 새로운 M으로 반복 시도한다. 일정 횟수의 반복 시도에서도 O가 최종 이상치 후보군으로 선정되지 않으면 검정대상군 O는 이상치가 아닌 것으로 판정한다.

Seo와 Yoon (2014)에 따르면 위의 과정에서 반복 시도의 횟수는 자료의 크기를 고려하여 결정하지만 경험적으로 O^c의 관찰치중 한 개 또는 두 개까지 교환하는 것을 추천하고 있다.

3. 군집방법을 이용한 순차적 이상치 탐지

모형의 적합도를 판단하기 위해 최소자승추정법으로 추정된 예측치와 잔차의 산점도에서 수평 밴드 형태의 모양 여부를 확인하는 것은 단순연결군집화(single linkage clustering)의 관점에서 체인 형태의 긴 수평 군집을 찾는 것과 같으므로 예측치와 잔차에 의해 군집화 과정을 수행하여 소수의 집단에 속하는 관찰치군을 이상치로 판단할 수 있다. 이러한 근거에 따라 Sebert 등 (1998)이 제안한 데이터의 군집화에 기반한 이상치 탐지법의 절차는 다음과 같다.

(1) 각 관찰치의 예측치와 잔차를 평균과 표준편차로 표준화 한다.

(2) 표준화된 예측치와 잔차에 의해 계산된 각 관찰치간의 유클리디언 거리를 유사성의 측도로 사용하여 단일연결군집화를 수행하여 군집나무를 생성한다.

(3) 관찰치들의 군집을 결정하기 위하여 군집나무 절단에 사용되는 기준은 Mojena (1977)의 정지규칙을 사용한다. 즉 $\bar{h}$와 s_h를 각각 $N-1$개 군집나무 높이들의 평균과 표준편차라고 할 때 $\bar{h} + 1.25s_h$에서 군집나무를 절단하여 군집을 구성한다.

(4) 완성된 군집 구성에서 소수 집단에 속하는 군집의 관찰치를 이상치로 간주한다.

그러나 단일연결군집화에 기반을 둔 이상치 탐지법은 일반적으로 이상치 탐지과정의 수렁 효과, 즉 정상치를 이상치로 잘못 판정하게 되는 현상에 취약하다는 것이 예제 및 모의실험을 통해 확인된다. 이를 보완하기 위하여 군집화 방법에서 판정된 이상치의 각 관찰치에 대하여 t-검정을 수행하고 최종적으로 이상치를 판단할 때 각종 잔차그림에 minimal spanning tree(MST)를 중복시킴으로써 시각적으로 파악되는 각 관찰치의 유형을 고려하는 방법이 제안 되었다 (Kim과 Krzanowski, 2007). 본 연구에서는 단일연결군집화 방법을 보완하기 위하여 이상치 후보로 판정된 군집에 대하여 이상치 검정을 수행하고 그 결과에 따라 군집나무의 절단기준을 변경시켜 군집의 크기를 축소 또는 확장하여 새로운 이상치 후보군을 찾는 방법을 제안한다. 기존의 개별적 t-검정 등은 Mojena (1977)의 정지규칙에 의하여 탐지된 소수 군집을 이상치 후보군으로 확정하여 여기에 속하지 않는 관찰치의 이상치 여부를 고려하지 않는데 반해 본 연구에서 제안한 방법은 군집나무에서 절단기준에 따라 다르게 결정되는 이상치 후보군에 대해 군 단위로 이상치 여부를 판정하는 것이다. 본 연구에서 제안하는 방법의 절차는 다음과 같다.

단계 1: 단일연결군집화에 의해 군집나무를 생성하고 Mojena (1977)의 정지규칙을 적용하여 이상치 후보군을 정한다.

단계 2: 이상치 후보군에 대하여 Seo와 Yoon (2014)의 방법을 적용하여 이상치 여부를 검정한다. 검정결과 이상치가 아닌 것으로 판정되면 이상치 규모를 줄이는 단계 3-a를 수행하고, 이상치인 경우는 이상치 규모를 늘이는 단계 3-b를 수행한다.

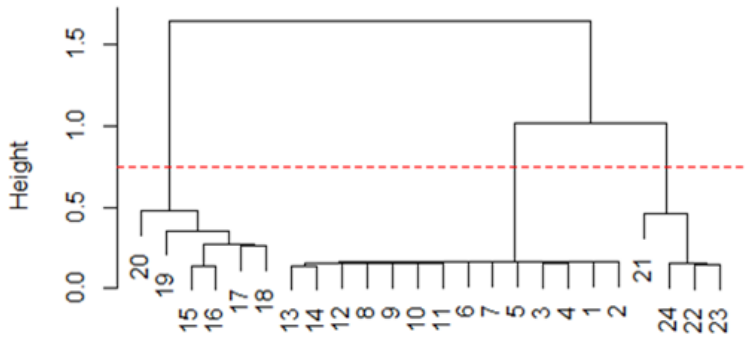

Figure 4.1. Cluster dendrogram for Telephone data.

단계 3-a: (1) 앞선 단계에서 사용한 절단 기준값보다 큰 절단 기준값을 적용하여 더 작은 크기의 이상치 후보군을 결정한다.

(2) 새로운 이상치 후보군에 대하여 검정을 수행하여 그 결과 이상치군으로 판정되면 탐지과정을 종료하며 이상치군이 아닌 것으로 판정되면 단계 3-a (1) 과정부터 반복 수행한다. 반복 수행에서 이상치 후보군이 공집합 일 때 이상치가 없는 것으로 판정한다.

단계 3-b: (1) 전 단계에서 사용했던 절단기준 보다 더 작은 값으로 군집나무를 절단하여 더 큰 크기의 새로운 이상치 후보군을 결정한다.

(2) 새 이상치 후보군에 대하여 검정을 수행하여 이상치군이 아닌 것으로 판정되면 앞선 단계에서 이상치군으로 판정된 이상치후보군을 최종 이상치군으로 확정하고 탐지과정을 종료하며, 검정결과 이상치군으로 판정되면 단계 3-b (1) 과정을 재차 실행하여 새로운 이상치 후보군이 이상치가 아닌 것으로 판정될 때 까지 반복 수행한다.

4. 모의실험 및 예제

기존의 군집화에 의하여 이상치군을 확정하는 방법과 이를 보완한 개별적 t-검정법 그리고 본 연구에서 제안한 방법을 Telephone data (Rousseeuw와 Leroy, 1987, p.26)에 적용한 과정 및 결과를 소개한다. Telephone data는 x, y의 24개 관찰치로 구성되어 있으며 이중 관찰치 15–20은 이상치로 간주되며 14와 21은 이상치와 정상치의 경계에 있다고 간주된다. Figure 4.1은 Telephone data에 단일연결군집화를 수행하여 완성된 군집나무이며 그림 중앙의 점선은 Mojena (1977)의 정지규칙에 의한 절단 기준값이다. Sebert 등의 방법에 따르면 절단 기준값에 의하여 구분된 군집 중 소수 집단인 15–24번째 관찰치를 이상치로 확정한다. 이러한 결과에 검정절차를 적용하여 개별적 t-검정을 15–24번째 관찰치에 수행하면 이중 15–21번째 관찰치들이 이상치로 판정된다. 본 연구에서 제안한 순차적 방법을 단계별로 적용한 과정은 다음과 같다.

첫 번째 단계에서 최초의 이상치 후보군인 15–24번째 관찰치에 대하여 Seo와 Yoon (2014) 검정을 수행한 결과 이상치가 아닌 것으로 판정된다. 따라서 Figure 4.1의 절단기준을 상향조정하면 구별되는 두 집단중 소수집단인 15–20번째 관찰치가 새로운 이상치 후보군으로 지정된다. 새로운 이상치 후보군에 대해 Seo와 Yoon (2014) 검정을 수행한 결과 대상군이 이상치로 판정되어 최종적으로 15–20번째 관찰치군을 이상치로 판정한다.

군집화에 의하여 이상치군을 확정하는 방법이 수렁효과 등으로 인해 정확하게 이상치를 탐지하지 못하는 반면 t-검정이나 본 연구에서 제안한 방법이 성공적으로 이상치군을 탐지하는 경우는 Telephone data 뿐만 아니라 Hertzsprung-Russell Star data (Rousseeuw와 Leroy, 1987, p.27), Hawkins, Bradu와 Kass data (Rousseeuw와 Leroy, 1987, p.94) Hadi와 Simonoff data (Hadi와 Simonoff, 1993, p.1269) 등 이상치에 관련된 여러 자료에서도 볼 수 있다. 이상치 후보군을 검정하지 않는 방법의 경우, Hertzsprung-Russell Star data에서는 2개의 관찰치(7번째, 14번째)를, Hawkins, Bradu와 Kass data에서는 네 개의 관찰치(11–14번째)를, Hadi와 Simonoff data에서는 1개의 관찰치(4번째)를 수렁효과에 의하여 이상치로 잘못 탐지 되는 반면 t-검정과 본 연구에서 제안된 순차적 검정법은 제대로 이상치군을 탐지한다.

본 연구에서 제안된 방법을 군집화만 수행하는 방법, t-검정이 추가된 방법 등과 비교하기 위하여 모의실험을 실시한다. 모의실험에 사용되는 자료의 생성 모형은 단순선형회귀모형 $y_i = X_i + \varepsilon_i$이며 x_i는 균등분포 unif(0, 15), 오차 ε_i는 표준정규분포에서 생성되었다. 이상치는 실제 회귀선과 δ_i만큼 차이가 있는 $y_i = X_i + \delta_i$값으로 지정되며 이상치의 위치와 개수에 따라 다음과 같은 일곱 종류의 이상치 모형 (Kianifard와 Swallow, 1989, 1996)으로 부터 생성된다.

(a) $y_1 = 7.5 + \delta_1$ $(\delta_1 > 0)$.

(b) $y_1 = 15 + \delta_1$ $(\delta_1 > 0)$.

(c) $y_1 = 15 + \delta_1, \quad y_2 = 15 + \delta_2$ $(\delta_1 > 0, \ \delta_2 < 0)$.

(d) $y_1 = 15 + \delta_1, \quad y_2 = 14.95 + \delta_2$ $(\delta_1, \delta_2 > 0)$.

(e) $y_1 = 15 + \delta_1, \quad y_2 = 15 + \delta_2$ $(\delta_1 > 0, \ \delta_2 > 0)$.

(f) $y_1 = 15 + \delta_1, \quad y_2 = 15 + \delta_2, \quad y_3 = 15 + \delta_3$ $(\delta_1, \delta_2, \delta_3 > 0)$.

(g) $y_1 = 15 + \delta_1, \quad y_2 = 14.95 + \delta_2, \quad y_3 = 14.90 + \delta_3$ $(\delta_1, \delta_2, \delta_3 > 0)$.

총 실험의 횟수는 1,000번이며 추출된 한 조의 설명변수 자료는 10번 반복하여 사용한다. 실험의 다양성을 위해 표본의 크기는 $n = 20, 30, 50$으로 하며, δ_i값은 서로 다른 두 개의 값을 각각 적용한다. 각 방법의 효율성은 가면효과와 수렁효과 등을 고려하여 세 가지 비율 p_1, p_2, p_3을 계산하여 비교한다. p_1은 이상치를 정확하게 전부 찾아낸 비율이고 p_2는 적어도 한 개 이상의 이상치를 찾아낸 비율이며 p_3는 탐지한 이상치 중 정상치가 포함된 비율이다. 따라서 가면효과가 발생한 비율은 $1 - p_2$이고, 수렁효과가 발생한 비율은 p_3이 된다. 실험의 결과를 요약한 Table 4.1에서 군집화만 수행하는 방법은 CL, t-검정이 추가된 방법은 TT, 본 연구에서 제안된 순차적 방법은 SQ로 표기한다.

모의실험 결과, 전반적으로 군집화만 수행하는 방법보다 t-검정이 추가된 방법과 본 논문에서 제안한 순차적 방법이 더 효율적이며 보완된 방법 중 t-검정 보다 순차적 방법이 조금 더 효과적임을 알 수 있다. 세 방법은 가면효과를 반영하는 p_2값 비교에서는 대체로 대등한 결과를 보이지만 수렁현상을 반영하는 p_3값과 이상치를 정확하게 탐지하는 비율인 p_1값에서 차이가 있다. 예를 들면 $\delta = 2.5, 2.5$, $n = 50$의 이상치 유형 (d)에서 군집화만 수행하는 방법의 p_1, p_2, p_3는 각각 0.295, 1.00, 0.705이고 t-검정이 추가된 방법의 p_1, p_2, p_3는 0.875, 0.993, 0.118이지만 본 연구에서 제안된 순차적 방법의 p_2는 0.992로 가면효과가 잘 방지되며 p_1은 0.938로 정확하게 이상치를 탐지하는 비율이 높고 $p_3 = 0.054$로 수렁효과의 빈도가 낮다. 특히 이상치모형 (g)에서 군집화만 수행하는 방법과 t-검정이 추가된 방법은 가면효과와 수렁효과에는 강건하지만 이상치 정도가 작을 때 $(\delta_1, \delta_2, \delta_3 > 0)$ 이상치를 정확하게 탐지하는 비율 p_1이 0을 나타내어 이상치 탐지기능을 전혀 수행하지 못하나 순차적 방법은 자료 크기가 커짐에 따라 p_1값이 0.134, 0.500, 0.924가 되어 상대적으로 이상치를 정확하게 탐지하고 있다 (Table 4.1).

Table 4.1. Proportion of all planted outliers correctly identified (p_1), at least one planted outlier correctly identified (p_2) and other observations incorrectly detected to be outliers (p_3) by three methods under seven outlier patterns (a-g) having outliers planted at vertical distances from the true regression line

Type	$\delta_i's$		$n=20$			$n=30$			$n=50$		
			p_1	p_2	p_3	p_1	p_2	p_3	p_1	p_2	p_3
(a)	2.5	CL	0.525	0.994	0.475	0.335	1.000	0.665	0.110	1.000	0.890
		TT	0.755	0.851	0.101	0.802	0.949	0.148	0.802	0.852	0.051
		SQ	0.766	0.833	0.067	0.771	0.829	0.058	0.817	0.877	0.060
	3	CL	0.900	1.000	0.100	0.775	1.000	0.225	0.568	1.000	0.432
		TT	0.935	1.000	0.065	0.950	1.000	0.050	0.913	1.000	0.087
		SQ	0.963	1.000	0.037	0.954	1.000	0.046	0.949	1.000	0.051
(b)	2.5	CL	0.765	1.000	0.235	0.593	1.000	0.407	0.348	1.000	0.652
		TT	0.850	0.908	0.058	0.900	0.974	0.074	0.894	0.999	0.105
		SQ	0.874	0.922	0.050	0.912	0.966	0.054	0.934	0.994	0.060
	3	CL	0.900	1.000	0.100	0.796	1.000	0.204	0.561	1.000	0.439
		TT	0.947	0.993	0.046	0.947	0.999	0.052	0.943	1.000	0.057
		SQ	0.969	0.987	0.018	0.950	0.999	0.049	0.967	1.000	0.033
(c)	2.5, −2.5	CL	0.771	0.955	0.041	0.898	1.000	0.102	0.682	1.000	0.318
		TT	0.793	0.980	0.018	0.919	0.993	0.036	0.933	1.000	0.062
		SQ	0.935	1.000	0.044	0.937	1.000	0.024	0.937	1.000	0.052
	3, −3	CL	0.934	0.998	0.049	0.965	1.000	0.035	0.863	1.000	0.137
		TT	0.986	1.000	0.011	0.947	1.000	0.052	0.963	1.000	0.037
		SQ	0.978	1.000	0.001	0.983	1.000	0.016	0.968	1.000	0.032
(d)	2.5, 2.5	CL	0.722	1.000	0.278	0.571	1.000	0.429	0.295	1.000	0.705
		TT	0.848	0.888	0.042	0.901	0.971	0.071	0.875	0.993	0.118
		SQ	0.853	0.897	0.046	0.925	0.973	0.048	0.938	0.992	0.054
	3, 3	CL	0.863	1.000	0.137	0.746	1.000	0.254	0.541	1.000	0.459
		TT	0.942	0.994	0.052	0.946	0.999	0.053	0.912	1.000	0.088
		SQ	0.955	0.987	0.033	0.953	0.999	0.046	0.944	0.999	0.055
(e)	2.5, 4.5	CL	0.860	0.999	0.035	0.885	1.000	0.114	0.732	1.000	0.268
		TT	0.901	1.000	0.079	0.918	1.000	0.051	0.952	1.000	0.048
		SQ	0.873	0.999	0.014	0.939	1.000	0.037	0.950	1.000	0.046
	3, 5	CL	0.941	1.000	0.059	0.918	1.000	0.082	0.786	1.000	0.214
		TT	0.946	0.999	0.036	0.964	1.000	0.035	0.945	1.000	0.055
		SQ	0.971	0.998	0.007	0.983	1.000	0.016	0.952	1.000	0.048
(f)	2.5, 3.5, 4.5	CL	0.695	0.995	0.303	0.671	1.000	0.329	0.494	1.000	0.506
		TT	0.774	0.976	0.035	0.900	0.998	0.057	0.920	1.000	0.074
		SQ	0.831	0.859	0.028	0.911	0.963	0.049	0.947	0.993	0.043
	3, 4, 5	CL	0.722	1.000	0.278	0.751	1.000	0.249	0.659	1.000	0.341
		TT	0.907	0.998	0.022	0.959	0.998	0.025	0.938	1.000	0.061
		SQ	0.957	0.978	0.021	0.965	0.997	0.032	0.953	1.000	0.047
(g)	2.5, 2.5, 2.5	CL	0.000	1.000	0.000	0.000	1.000	0.000	0.000	1.000	0.000
		TT	0.000	1.000	0.000	0.000	1.000	0.000	0.000	1.000	0.000
		SQ	0.134	1.000	0.001	0.500	1.000	0.002	0.924	1.000	0.027
	3, 3, 3	CL	0.702	1.000	0.298	0.693	1.000	0.307	0.511	1.000	0.489
		TT	0.907	0.935	0.032	0.940	0.980	0.040	0.923	1.000	0.077
		SQ	0.945	0.982	0.038	0.964	1.000	0.036	0.950	1.000	0.050

CL = 군집화만 수행하는 방법; TT = t-검정이 추가된 방법; SQ = 본 연구에서 제안된 순차적 방법.

5. 결론

Sebert 등 (1998)이 제안한 군집화에 기반한 이상치 탐지방법은 이상치 검정절차가 생략되어 가면효과나 수렁효과에 취약하다. 이에 대한 보완방법으로 본 연구에서는 이상치 후보군에 대한 검정을 통하여 그 결과에 따라 군집나무 절단기준을 변경하는 순차적 방법을 제안하였다. 본 연구에서 제안한 방법은 변경된 절단기준에 의해 순차적으로 탐지되는 새로운 이상치 후보군들이 실제 이상치군과 상이할 경우에는 최초 이상치 후보군에서 실질적인 이상치를 확정하는 접근법보다 탐지력이 낮을 수 있으나 여러 유형의 이상치 모형을 적용한 모의실험 결과, 제안된 방법은 군집화에 의해 구분된 소수의 관찰치군을 확정된 이상치로 간주하는 방법이나 이상치 후보군의 각 관찰치에 대하여 t-검정을 수행하는 방법보다 더 효과적임을 확인할 수 있었다. 제안된 검정절차는 일정 기준에 의해 이상치 후보군의 크기가 정해지는 어떠한 이상치 탐지법에도 적용될 수 있다.

References

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.

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

Kianifard, F. and Swallow, W. H. (1996). A review of the development and application of recursive residuals in linear models, *Journal of the American Statistical Association*, **91**, 391–400.

Kim, S. S. and Krzanowski, W. J. (2007). Detecting multiple outliers in linear regression using a cluster method combined with graphical visualization, *Computational Statistics*, **22**, 109–119.

Mojena, R. (1977). Hierarchical grouping methods and stopping rules: an evaluation, *The Computer Journal*, **20**, 359–363.

Peña, D. and Yohai, V. J. (1995). The detection of influential subsets in linear regression by using an influence matrix, *Journal of the Royal Statistical Society, Series B*, **57**, 145–156.

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

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

Seo, H. S. and Yoon, M. (2014). A test on a specific set of outlier candidates in a linear model, *The Korean Journal of Applied Statistics*, **27**, 307–315.

군집 알고리즘을 이용한 순차적 이상치 탐지법

서한손[a] · 윤민[b,1]

[a]건국대학교 응용통계학과, [b]부경대학교 통계학과

(2016년 3월 4일 접수, 2016년 4월 9일 수정, 2016년 4월 16일 채택)

요 약

검정절차가 생략된 이상치 탐지법은 구조적으로 수렁효과나 가면효과에 취약하기 때문에 다수의 이상치를 제대로 탐지하지 못할 때가 있다. 본 연구에서는 군집화에 의하여 구분된 소수 관찰치군을 이상치로 판정하는 방법에 보완될 검정절차를 다룬다. 이에 관련된 일반적인 방법은 탐지된 이상치 후보군의 개별적인 관찰치에 대해 다양한 종류의 t-검정을 수행하는 것이다. 본 연구에서는 이상치 후보군에 대한 검정을 수행하고 군집나무의 절단기준을 변경시켜 새로운 이상치군을 탐색해 나가는 순차적인 방법을 제안한다. 예제와 모의실험을 통해 제시된 방법과 기존의 방법들을 비교한다.

주요용어: 군집화, 선형회귀모형, 순차적방법, 이상치 검정

이 논문은 2015학년도 건국대학교의 지원에 의하여 연구되었음.

[1]교신저자: (48513) 부산시 남구 용소로 45, 부경대학교 통계학과. E-mail: myoon@pknu.ar.kr