The Korean Journal of Applied Statistics (2017) DOI: https://doi.org/10.5351/KJAS.2017.30.1.159
30(1), 159–167

Outlier tests on potential outliers

Han Son Seo[a,1]

[a]Department of Applied Statistics, Konkuk University

(Received December 14, 2016; Revised January 30, 2017; Accepted February 1, 2017)

Abstract

Observations identified as potential outliers are usually tested for real outliers; however, some outlier detection methods skip a formal test or perform a test using simulated p-values. We introduce test procedures for outliers by testing subsets of potential outliers rather than by testing individual observations of potential outliers to avoid masking or swamping effects. Examples to illustrate methods and a Monte Carlo study to compare the power of the various methods are presented.

Keywords: diagnostics, linear model, masking, outliers, swamping

1. 서론

선형회귀모형에서 이상치에 관련된 연구는 광범위하게 진행되어 왔으며 아직도 중요한 연구주제의 하나로 간주되고 있다. 이상치 탐지법은 잠재적인 이상치군을 찾는 과정과 탐지된 이상치 후보군에 대한 검정과정으로 구성되며 제시된 여러 방법 중 일부는 검정과정이 생략된 방법들도 있다 (Seber 등, 1998). 검정과정없이 일정한 이상치 탐지 절차를 통해 탐지된 관찰치군을 이상치로 간주하는 것은 통계적으로 유효한 결정이라고 할 수 없다. 본 연구에서는 잠재적 이상치로 간주되는 관찰치 군에서 최종적인 이상치를 결정하기 위한 검정방법을 다룬다. 임의 관찰치군에 대한 이상치 여부를 검정하는 절차 (Seo와 Yoon, 2014)와 달리 본 연구에서는 잠재적 이상치군에서 최종적으로 이상치로 판단되는 군의 일부 또는 전부를 찾는다.

선형회귀모형에서 이상치 탐지를 위해 제시된 여러 방법들은 다양한 측면에서 비교되었다 (Peña와 Yohai, 1995). 이중 본 연구에서 활용되는 기본적인 이상치 탐지법은 탐지력과 아울러 절차의 간결성, 계산의 단순성 측면에서 유용하다고 평가되는 Hadi와 Siminoff (1993)의 순차적 탐지과정이다. 잠재적 이상치군이 정해진 상황에서 각 관찰치의 이상치 여부를 판정하는데 사용되는 가장 보편적인 방법은 Bonferroni 검정법이다. Bonferroni 부등식에 기반을 둔 검정은 개별 관찰치의 이상치 여부를 검정할 때 검정대상군의 크기에 따른 유의수준을 조절한다. 이상치 검정에서 Bonferroni 검정은 주로 t 분포 검정에 적용되며 일반적으로 외적스튜던트화 잔차(externally studentized residual)를 검정통계량으로 사용할 수 있지만 다양한 검정통계량들이 사용될 수 있다 (Kim과 Krzanowski, 2007).

잠재적 이상치군을 대상으로 수행되는 이상치 검정에서 검정통계량의 분포를 찾기 어려운 경우 모의실험에 의한 유의확률을 검정에 적용하며 그 대표적인 예로 Paul과 Fung (1991)이 제안한 generalized

This paper was supported by Konkuk University in 2016.
[1]Department of Applied Statistics, Konkuk University, 120, Neungdong-ro, Gwangjin-gu, Seoul 05029, Korea. E-mail: hsseo@konkuk.ac.kr

extreme studentized residual(GESR)에 의한 이상치 탐지법, Kianifard와 Swallow (1990, 1996)의 순차적 회귀잔차법 등이 있다. 본 연구에서는 대부분의 Bonferroni 검정에서 적용되는 잠재적 이상치군에 속한 관찰치를 개별적으로 검정하는 절차를 사용하지 않고 Hadi와 Siminoff의 순차적 이상치 탐지법 또는 Seo와 Yoon의 집단 이상치 검정을 활용하여 잠재적 이상치군의 부분집합에 대해 최종적인 이상치군을 판정하는 방법을 제안한다. 제시된 방법은 예제와 모의실험을 통해 기존의 개별적 검정인 Bonferroni 방법, Paul-Fung 방법 등과 검정력을 비교한다. 잠재적 이상치군의 선정은 Paul-Fung이 제안한 GESR 방법과 검정절차없이 군집화에 의하여 이상치후보군을 선정하는 Seber, Montgomery, Rollier 방법을 사용한다. 제 2장에서는 잠재적 이상치군 탐지와 검정을 위한 기본적 방법인 Hadi와 Siminoff 방법, 임의의 집단에 대한 이상치 여부를 판정할 수 있는 Seo와 Yoon 방법, 본 연구에서 제안한 방법과 비교될 다양한 통계량의 Bonferroni 방법을 소개하고 본 연구에서 제안하는 잠재적 이상치군에 대한 이상치 검정절차를 설명한다. 제 3장에서는 예제와 모의실험을 통하여 제안된 방법과 기존방법의 검정력을 비교하며 4장에서는 연구결과와 추후 연구 방향을 요약, 제시한다.

2. 이상치탐지법과 이상치후보군에 대한 검정법

Hadi와 Siminoff (1993)의 방법은 일정한 크기의 기초 양호치군에서 점차적으로 이상치군의 크기를 줄여가는 순차적 탐지법이며 각 단계의 양호치군만으로 회귀모형을 추정하여 계산된 잔차로 잠재적이상치군을 탐지하고, 잔차의 순서통계량을 이용하여 잠재적 이상치군에 대한 최종적인 이상치 여부를 결정한다. Hadi와 Siminoff 방법의 대략적인 절차는 다음과 같다.

(1) 현 단계에서 회귀식 추정에 사용할 양호치군을 M이라고 하고 크기를 s라고 하자.

(2) 양호치군 M만으로 회귀식을 추정한 후 양호치군 M과 비양호치군 M^c에 해당하는 관찰치에 내적 스튜던트화 잔차(internally studentized residual) d_i를 다음과 같이 계산한다.

$$d_i = \begin{cases} \dfrac{y_i - x_i^T \hat{\beta}_M}{\hat{\sigma}_M \sqrt{1 - x_i^T (X_M^T X_M)^{-1} x_i}}, & \text{if } i \in M, \\ \dfrac{y_i - x_i^T \hat{\beta}_M}{\hat{\sigma}_M \sqrt{1 + x_i^T (X_M^T X_M)^{-1} x_i}}, & \text{if } i \notin M, \end{cases} \tag{2.1}$$

여기서 X_M, $\hat{\beta}_M$, $\hat{\sigma}_M$은 각각 집합 M으로 구성된 X의 부분행렬, 추정된 회귀계수, σ의 추정치이다.

(3) $|d|_{(j)}$는 $|d_i|$의 오름차순에 따른 j번째 순서통계량이고 $t_{(a;b)}$는 X가 자유도 b인 t분포를 따를 때 $P(X \geq x) = a$가 되는 x값이라고 할 때, $|d|_{(s+1)} \geq t_{(\alpha/2(s+1);(s-p))}$이면, $(n-s)$개 후 순위 순서통계량에 속하는 관찰치를 최종적인 이상치군으로 판단하며 그렇지 않을 경우 $|d|_{(1)}$, $|d|_{(2)}, \ldots,$ $|d|_{(s+1)}$에 해당하는 관찰치를 새로운 양호치군 M으로 간주하고 위의 절차를 반복한다.

초기 양호치의 크기는 대략 전체자료의 절반 정도가 적절하며 Hadi와 Siminoff (1993)는 두 가지의 기초군 선정 방법을 제시하고 있다. 본 연구에서 활용하는 Seo와 Yoon (2014)의 이상치군 집단적 검정법은 Hadi와 Siminoff 방법과 같이 내적 스튜던트화 잔차(internally studentized residual)의 순서통계량을 이용하여 임의의 관찰치군에 대한 이상치 여부를 판정하는 절차이다. Hadi와 Siminoff 방법에서 임의의 관찰치군의 이상치 여부를 판단하기 위해서는 내적스튜던트화잔차에 의해 검정 대상군이 일단 잠재적이상치군으로 지정되어야 한다. 이를 위해 다양한 양호치군으로 회귀모형을 추정하여 검정 대상군이 잠재적이상치군으로 판정되는지 확인하고 만약 잠재적 이상치군으로 확인되면 이에 대한 이상치 검정을 수행하며, 만약 여러 시도에도 검정대상군이 잠재적 이상치후보군으로 지정되지 않으면 검

정대상군은 이상치가 아닌 것으로 판정한다. 임의의 관찰치군에 대한 이상치 검정방법 중 Seo와 Yoon (2014)은 Hadi와 Simonoff의 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.

Seo와 Yoon (2014)이 제안한 임의의 관찰치군에 대한 이상치 검정 과정을 요약하면 다음과 같다.

(1) 검정대상군의 크기를 k라고 하자. 검정대상군의 여집합을 양호치군 M으로 지정하여 회귀식을 추정하고 식 (2.1)과 같은 내적 스튜던트화 잔차를 계산한다.

(2) 내적 스튜던트화 잔차 d_i의 절대값이 가장 큰 k개의 관찰치가 검정대상군과 일치하면 아래의 검정을 수행하고, 검정대상군과 일치하지 않으면 집합 M에 속한 관찰치중 한 개(또는 적정 갯수)를 교체하여 위 과정을 반복한다.

   a) 만약 $d_{(n-k+1)} \geq t_{(\alpha/2(n-k+1);n-k-p)}$이면, 검정대상군을 이상치로 판정한다.

   b) $d_{(n-k+1)} < t_{(\alpha/2(n-k+1);n-k-p)}$이면, 검정대상군을 이상치가 아닌 것으로 판단한다.

(3) 만약 일정 횟수의 반복 시도에서도 검정대상군이 이상치 후보군으로 지정되지 않으면 검정대상군은 이상치가 아닌 것으로 판정한다.

위의 과정에서 반복 시도의 횟수는 자료의 크기에 따라 결정하지만 대부분의 경우 검정대상군중 한 개씩 교환하는 것으로도 충분하다.

본 연구에서는 특정한 절차에 따라 탐지된 잠재적 이상치군에서 최종적으로 이상치로 판단되는 관찰치를 찾는 방법을 제시한다. 이러한 목적으로 사용되는 전통적인 방법은 Bonferonni 부등식을 이용한 검정이다. Bonferonni 검정은 이상치 검정에서 단일 관찰치의 이상치 여부에 주로 사용되는 검정통계량인 외적스튜던트화 잔차(externally studentized residual)에 잠재적이상치군의 크기를 고려한 유의수준을 적용하는 것이다. 본 연구에서는 일반적인 외적스튜던트화 잔차를 이용한 Bonferroni 검정외에 Hadi와 Siminoff 검정에서 사용하는 방식의 외적스튜던트화 잔차를 Bonferroni 검정에 적용한다. 이 방법은 군집화에 의한 이상치 탐지법에서 가면화 효과(masking effect)와 수렁화 효과(swamping effect)에 대한 보완책으로 제시된 시각화 기법과 더불어 이상치후보군에 대한 검정 절차로써 사용되었다 (Kim과 Krzanowski, 2007). Hadi와 Siminoff 통계량을 사용한 Bonferroni 검정절차는 외적스튜던트화 잔차를 계산할 때 해당 관찰치만 제외하는 일반적인 Bonferroni 검정과는 달리 식 (2.1)의 잔차처럼 잠재적이상치군 전체를 제외하여 외적스튜던트화 잔차를 계산한다. 이와 같이 검정통계량인 외적스튜던트화 잔차 계산에서 제외되는 대상이 개별 관찰치(individual)와 관찰치 군(group)의 여부에 따라 수행되는 두 종류의 Bonferroni 검정을 각각 "Bonferroni-i 검정", "Bonferroni-g 검정"이라고 부르기로 한다. Bonferroni-g 검정이 Hadi와 Siminoff 검정과 다른 점은 잔차 절대값의 순서통계량에 대하여 검정하지 않고 이상치후보군의 관찰치에 대하여 개별적인 t 검정을 수행하는 것이다. 즉 이상치후보군의 외적스튜던트화잔차 절대값에 대하여 $t_i \sim t_{(\alpha/2\times(s+1);s-p)}$, (이때 s = clean set 개수, p = 독립변수 개수)인 검정을 수행하여 이상치후보군에 속한 각 관찰치에 대한 이상치 여부를 판단한다.

본 연구에서 제안하는 잠재적 이상치군에 대한 두 가지 검정 방법을 각각 S1-검정, S2-검정 이라고 하자. 첫 번째 검정절차인 S1-검정에서는 순차적 검정절차를 수행할 때 잠재적 이상치군을 고려하여 각 단계별 검정대상을 지정한다. 즉 Hadi와 Siminoff 과정에 의해 탐지된 이상치후보군이 잠재적 이상치

군의 부분집합일 때, 검정대상인 잠정적 이상치군에 대한 검정을 수행하며 그렇지 않은 경우 검정을 생략하고 다음 단계의 탐지과정을 수행한다. S1-검정의 구체적 절차는 다음과 같다.

검정대상인 잠재적 이상치군을 L, 그 크기를 k라고 하고 현 단계에서의 잠정적 양호치군을 M, 그 크기를 s라고 하자.

[S1-검정]:

(1) L의 여집합인 L^c를 양호군 M으로 간주하여 회귀모형을 추정한 후 식 (2.1)의 잔차를 계산한다.

(2) 잔차의 절대값인 $|d_i|$의 오름차순에 의한 $(s+1)$번째의 순서통계량을 $d_{(s+1)}$라고 할 때 $d_{(s+1)}$, $d_{(s+2)}, \ldots, d_{(n)}$에 해당하는 관찰치 군을 R_s^c라고 표기하고 이에 대하여 다음과 같은 검정을 수행한다.

   a) 만약 $R_s^c \subseteq L$이고 $d_{(s+1)} \geq t_{(\alpha/2(s+1);s-p)}$일 때 R_s^c를 이상치로 판정한다.

   b) 위의 조건을 만족하지 않은 경우, 만약 $s < n$이면 Hadi와 Siminoff 방법에서처럼 $d_{(i)}$를 기준으로 $(s+1)$개의 관찰치를 새로운 M으로 지정하여 탐지과정을 반복하며 $s = n$이면 잠재적 이상치군에는 어떠한 이상치도 없다고 판정한다.

본 연구에서 제안하는 두 번째 검정절차인 S2-검정은 잠재적 이상치군 L의 부분집합에 대하여 단계적 검정을 수행하며 부분집합의 이상치 여부는 Seo와 Yoon 방법을 적용하여 판단한다. 단계별 검정대상과 검정방법에 관한 S2-검정의 절차는 다음과 같다.

[S2-검정]:

(1) 크기 s의 잠정적 양호치군 M을 이용하여 회귀모형을 추정한 후 식 (2.1)의 잔차를 계산한다.

(2) 잠재적 이상치군 L 중에서 잔차의 절대값인 $|d_i|$를 기준으로 가장 큰 $(n-s)$개의 관찰치를 검정대상으로 선정하고 Seo와 Yoon 검정을 수행한다.

(3) 검정 결과 이상치로 판정되면 검정대상이었던 L의 부분집합을 이상치로 판정하고 만약 검정결과 이상치가 아니면 전체 관찰치 중에서 잔차의 절대값 기준으로 가장 작은 $(s+1)$개에 해당하는 관찰치들인 R_{s+1} 관찰치군을 새로운 양호군 M으로 지정하여 위의 과정을 반복한다.

위 절차의 최초 과정은 $M = L^c$로 지정하여 이상치 후보군 L을 검정한다.

3. 예제와 모의실험

예제에서 사용될 세 개의 자료는 Hadi와 Simonoff (1993, p.1269) 자료, Hertzsprung와 Russell의 stars 자료 (Rousseeuw와 Leroy, 1987, p.27), stack loss 자료 (Brownlee, 1965, p.454)이다. 검정대상군인 잠재적 이상치군은 Kim과 Krzanowski (2007)와 유사하게 Rousseeuw와 Van Zomeren (1990)이 제안한 강건그림 방법으로 탐지하고 탐지된 잠재적 이상치군에 대하여 두 종류의 Bonferroni 검정과 본 연구에서 제안된 두 가지 방법을 적용 비교하기로 한다. 예제의 결과는 Table 3.1에 요약되어 있다.

Hadi와 Simonoff 자료는 $n = 25$이며 기본적으로 $Y = x_1 + x_2 + \epsilon$의 관계로부터 생성된 임의의 자료이다. 단, 관찰치 1, 2, 3은 이상치군으로써 $Y = x_1 + x_2 + 4$의 모형에서 생성되었다. Hadi와 Simonoff 자료에 대해 Rousseeuw와 Van Zomeren 방법은 관찰치 1, 2, 3, 6, 11, 13, 17, 19, 20, 24를 잠재적 이상치군으로 판정하였으며 여기에 네 가지 검정 방법을 적용한 결과, Bonferroni 검정들은 관찰치 20을 제외한 9개의 관찰치를 이상치로 판정하여 수렁화 효과의 영향을 받는 반면 본 연구에서 제안한 두 가지 방법은 관찰치 1, 2, 3을 이상치로 판정하였다.

Table 3.1. Three examples: potential outliers are detected by using robust plots (Rousseeuw and Van Zomeren, 1990)

	Hadi and Simonoff data	Star data	Stack loss data
Outliers	1, 2, 3	11, 20, 30, 34	1, (2), 3, 4, 21
Potential outliers	1, 2, 3, 6, 11, 13, 17, 19, 20, 24	7, 9, 11, 20, 30, 34	1, 2, 3, 4, 13, 14, 20, 21
Bonferroni-i	1, 2, 3, 6, 11, 13, 17, 19, 24	7, 9, 11, 20, 30, 34	1, 3, 4, 13, 21
Bonferroni-g	1, 2, 3, 6, 11, 13, 17, 19, 24	11, 20, 30, 34	1, 3, 4, 13, 21
S1	1, 2, 3	11, 20, 30, 34	1, 3, 4, 21
S2	1, 2, 3	11, 20, 30, 34	1, 3, 4, 21

In stack loss data case 2 is marginal.

Star 자료는 별의 표면적온도(temperature at the surface)와 밝기(light intensity)에 관한 $n = 47$개의 자료이다. 두 변수 간에 직접적인 연관 관계는 없지만 네 개의 관찰치(11, 20, 30, 34)는 낮은 온도와 높은 밝기로 인해 나머지 관찰치들과는 다른 관계성을 갖는 이상치로 간주된다. Star 자료에 대하여 Rousseeuw와 Van Zomeren 방법은 관찰치 7, 9, 11, 20, 30, 34를 잠재적 이상치로 탐지하였으며 이에 대하여 Bonferroni-i 검정은 잠재적 이상치군 전부를 최종 이상치로 판정하였고 나머지 방법인 Bonferroni-g 검정, S1 검정, S2 검정은 관찰치 11, 20, 30, 34를 최종적인 이상치로 판정하였다. Stack loss 자료는 암모니아 산화법에 의한 질산 생산에 대한 자료이며 $n = 21$일 동안 세 가지 요인(공기유입, 냉각수 온도, 산성도)에 따라 발생하는 암모니아 손실을 측정한 자료이다. 이 자료에 대한 다양한 분석 결과가 제시되었지만 (Atkinson, 1985, p.266) 관찰치 1, 3, 4, 21은 이상치이고 관찰치 2는 경계치로 간주되고 있다. Rousseeuw와 Van Zomeren 방법은 stack loss 자료에 대하여 관찰치 1, 2, 3, 4, 13, 14, 20, 21이 잠재적 이상치군이라고 진단한다. 8개의 잠재적 이상치군에 대한 두 종류의 Bonferroni 검정은 1, 3, 4, 13, 21을 이상치로 판정하는 반면 본 연구에서 제안한 두 가지 방법들은 1 3, 4, 21을 이상치로 판정한다. 단 Bonferroni-g 검정은 잠재적 이상치군이 1, 2, 3, 4, 21인 경우 최종적으로 1, 3, 4, 21이 이상치로 판정된다 (Kim과 Krzanowski, 2007, p.117).

여러 검정 방법의 검정력을 비교하기 위하여 모의실험을 수행한다. 모의실험에 사용될 가상자료의 이상치는 이상치의 개수와 위치에 따라 아래와 같은 7개의 평균이동모형 (Kianifard와 Swallow, 1990, 1996)에 의해 생성된다. 정상적인 자료는 모형 $y_i = x_i + \epsilon_i$에 의해 생성되며 설명변수 x는 unif(0, 15)에서, 오차항 ϵ은 표준정규분포로부터 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) $y_1 = 15 + \delta_1, \quad y_2 = 14.95 + \delta_2, \quad y_3 = 14.90 + \delta_3 \ (\delta_1, \delta_2, \delta_3 > 0)$

가상자료 생성을 위해 위의 각 이상치 모형에 실제로 적용된 δ_i값은 다음과 같다. 모형 (a)와 (b), $\delta_1 = 3.5$; 모형 (c), $\delta_1 = 3.5$, $\delta_2 = -3.5$; 모형 (d), $\delta_1 = 3.5$, $\delta_2 = 3.5$; 모형 (e), $\delta_1 = 3.5$, $\delta_2 = 5.5$; 모형 (f), $\delta_1 = 3.5$, $\delta_2 = 4.5$, $\delta_3 = 5.5$; 모형 (g), $\delta_1 = 3.5$, $\delta_2 = 3.5$, $\delta_3 = 3.5$. 가상 자료의 크기는 이상치를 포함하여 $n = 25$로 고정하며 이것은 Paul과 Fung의 GESR 방법에서 제시된 가상 p-값을 고려한 것이다.

Table 3.2. Summary of simulation results: potential outliers are detected by using clustering method (Seber *et al.*, 1998)

Model	(a)			(b)			(c)			(d)			(e)			(f)			(g)		
	P_1	P_2	P_3	P_1	P_2	P_3	P_1	P_2	P_3	P_1	P_2	P_3	P_1	P_2	P_3	P_1	P_2	P_3	P_1	P_2	P_3
Bonferroni-i	0.264	1	0.736	0.277	1	0.723	0.413	1	0.587	0.406	1	0.594	0.452	1	0.548	0.614	1	0.386	0.557	0.96	0.442
Bonferroni-g	0.646	1	0.354	0.605	1	0.395	0.751	1	0.249	0.750	1	0.250	0.750	1	0.250	0.880	1	0.120	0.825	0.96	0.169
S1	0.942	1	0.058	0.931	1	0.069	0.942	1	0.058	0.932	1	0.068	0.936	1	0.064	0.950	1	0.050	0.904	0.96	0.054
S2	0.947	1	0.053	0.932	1	0.068	0.943	1	0.057	0.932	1	0.068	0.937	1	0.063	0.952	1	0.048	0.904	0.96	0.054

Table 3.3. Summary of simulation results: potential outliers are detected by using GESR (Paul and Fung, 1991)

Model	(a)			(b)			(c)			(d)			(e)			(f)			(g)		
	P_1	P_2	P_3	P_1	P_2	P_3	P_1	P_2	P_3	P_1	P_2	P_3	P_1	P_2	P_3	P_1	P_2	P_3	P_1	P_2	P_3
GESR	0.936	1	0.064	0.918	1	0.082	0.984	1	0.016	0.883	1	0.117	0.963	1	0.037	0.853	1	0.147	0.824	0.96	0.176
Bonferroni-i	0.947	1	0.053	0.929	1	0.071	0.993	1	0.007	0.937	1	0.062	0.986	1	0.014	0.949	1	0.045	0.925	0.96	0.067
Bonferroni-g	0.977	1	0.023	0.942	1	0.058	0.996	1	0.004	0.968	1	0.030	0.992	1	0.008	0.970	1	0.024	0.958	0.96	0.032
S1	0.978	1	0.022	0.973	1	0.027	0.996	1	0.004	0.969	1	0.030	0.993	1	0.007	0.978	1	0.022	0.976	0.96	0.024
S2	0.978	1	0.022	0.972	1	0.028	0.996	1	0.004	0.973	1	0.027	0.993	1	0.007	0.985	1	0.015	0.976	0.96	0.024

실험의 횟수는 총 1,000번이며 동일한 x값이 10번 반복 사용되었다. 각 방법의 검정력은 세 개의 척도 P_1, P_2, P_3에 의해 계산된다. P_1은 이상치를 정확하게 찾은 비율이고, P_2는 적어도 한 개 이상의 이상치를 찾은 비율이어서 가면현상(masking phenomenon)이 발생한 비율은 $(1 - P_2)$가 된다. P_3는 탐지된 이상치 중에 정상관찰치가 포함된 비율, 즉 수렁현상(swamping phenomenon)이 발생한 비율이다.

한 개의 가상자료에 대해 잠재적 이상치군은 Seber, Montgomery, Rollier방법의 군집화 방법과 Paul과 Fung (1991)이 제안한 GESR 방법을 각각 적용하여 선정한다. 군집화방법에 의한 잠재적 이상치군 선정의 모의실험에서는 네 가지 검정방법이 비교되고 GESR 방법에 의해 잠재적 이상치군이 선정된 모의실험에서는 네 가지 방법과 더불어 Paul과 Fung이 제시한 모의 p-값에 의한 검정방법을 추가하여 비교한다.

Table 3.2는 군집화에 의하여 잠재적 이상치군을 선정한 경우 이에 대한 이상치 검정방법들 간의 비교결과이다. 이상치 탐지법을 거친 관찰치들을 검정대상군으로 지정하였으므로 모든 방법에 있어서 P_1값, 즉 정확하게 이상치를 판정하는 비율이 높고, 특히 P_2값은 1 또는 1에 근접하여 적어도 한 개 이상의 실제 이상치를 최종 이상치로 판정하는 비율이 높으며 수렁현상 비율이 낮은 것을 알 수 있다. 본 연구에서 제안된 검정방법들인 S1 검정과 S2 검정은 비슷한 수준의 검정력을 보이고 있으며 이상치 생성모형에 상관없이 기존 방법들보다 개선된 결과를 보이고 있다.

Table 3.3은 GESR에 의하여 잠재적 이상치군을 탐지하고 이에 대해 이상치 검정방법들을 적용한 결과이다. 군집화에 의하여 잠재적 이상치군을 탐지했을 때와 비교하면, S1 검정과 S2 검정은 비슷한 수준의 검정력을 보이고 있으나 Bonferroni 검정들은 상대적으로 낮은 검정력을 보인다. 검정방법들 간 검정력을 비교하면, 군집화에 의해 잠재적 이상치군을 탐지한 경우와 마찬가지로 새로 제안된 검정방법은 이상치 생성모형에 상관없이 전반적으로 Bonferroni 검정들 보다 우수한 결과를 보이고 있으며 잠재적 이상치군 탐지과정을 기반으로 계산된 GESR 검정통계량의 모의 p-값에 의한 검정과도 대등한 검정력을 보이고 있다. Table 3.2와 Table 3.3에서 보듯이 이상치검정법의 검정력은 잠재적 이상치군 선정방법에 따라 차이가 발생할 수도 있다. 이런 가능성을 점검하기 위해 이상치탐지법을 거치지 않고 임의로 이상치 후보군을 선정한 후 이를 대상으로 이상치 검정방법들의 검정력을 알아보기로 한다. 모형 (f)에서 $\delta_1 = 2.0$, $\delta_2 = -2.5$, $\delta_3 = 2.5$에 의해 세 개의 이상치를 갖는 $n = 25$의 자료를 생성한 후 지정된 세 개의 이상치 $(23, 24, 25)$와 임의로 선택된 세 개의 관찰치로 구성된 6개의 관찰치를 잠재적 이상치군으로 간주하여 이상치 검정을 수행한다. Table 3.4는 생성된 모의자료이며 Table 3.5는 이와 같은 임의의 잠재적 이상치군에 대해 이상치검정을 수행한 결과이다. Bonferroni 검정들은 이상치 후보군에 따라 가면화 또는 수렁화현상이 발생하지만 본 연구에서 제안된 방법들은 이상치군을 올바르게 판정하는 것을

Table 3.4. Artificial data

ID	y	x	ID	y	x
1	7.15	7.12	14	1.46	1.29
2	8.19	8.81	15	7.73	7.92
3	10.49	10.26	16	3.05	3.43
4	4.09	3.81	17	2.88	2.93
5	3.45	3.65	18	12.08	12.10
6	2.82	3.37	19	12.31	12.54
7	10.60	10.37	20	12.61	13.55
8	13.69	13.37	21	9.73	9.70
9	15.61	14.52	22	1.63	2.39
10	5.40	5.47	23	17.00	15.00
11	6.76	6.54	24	13.00	15.00
12	8.14	8.45	25	17.50	15.00
13	11.24	10.82			

Table 3.5. Several examples with randomly selected potential outliers

Potential outliers	Bonferroni-i	Bonferroni-g	S1	S2
10, 18, 21, 23, 24, 25	23, 24, 25	24, 25	23, 24, 25	23, 24, 25
9, 16, 22, 23, 24, 25	9, 23, 24, 25	23, 24, 25	23, 24, 25	23, 24, 25
2, 18, 20, 23, 24, 25	20, 23, 24, 25	20, 23, 24, 25	23, 24, 25	23, 24, 25
3, 8, 16, 23, 24, 25	23, 24, 25	23, 24, 25	23, 24, 25	23, 24, 25

Table 3.6. Summary of simulation results: potential outliers are selected randomly

Methods	P_1	P_2	P_3
Bonferroni-i	0.898	1	0.102
Bonferroni-g	0.864	1	0.008
S1	1.000	1	0.000
S2	1.000	1	0.000

알 수 있다. 잠재적 이상치군을 임의로 1,000개 추출한 후 각 검정절차를 적용한 결과인 Table 3.6에서도 동일한 결론을 얻을 수 있다.

4. 결론

잠재적 이상치군이 선별되면 기존의 방법들은 가면화 또는 수렁화의 가능성이 높은 개별 관찰치에 대한 검정을 수행하거나 유의값의 불확실성이 높은 실험에 의해 계산된 유의값으로 최종 이상치 여부를 판단한다. 본 연구에서 제시된 방법은 잠재적 이상치군에 대해 부분집합 단위로 단계적인 이상치 검정을 수행하며 예제와 모의실험을 통해 기존의 방법들보다 개선된 결과를 보여준다.

잠재적 이상치군을 부분집합 단위로 검정할 때 단계별 검정대상을 정하는 방식에 다양한 방법을 적용할 수 있다. 일반적인 순차검정에서는 단계별 잠정적 양호군을 잔차에 의하여 결정하지만 본 연구에서는 잔차계산을 위한 회귀모형의 구현과정에서 모형구현에 참여하는 관찰치군을 선정할 때 검정대상군인 잠재적이상치군을 고려한다. 이때에도 전 단계에서의 잠정적양호군을 고려하거나 아니면 독립적으로 결정하는 방법이 있을 수 있으며 일정한 기준에 따라 잠재적 이상치군에서 한 개씩 관찰치를 제외하여 단계별 검정대상으로 지정하는 방법도 고려할 수 있다. 잠재적 이상치군의 크기가 크지 않을 경우 잠재적 이상치군의 모든 부분집합에 대하여 검정을 수행할 수도 있다.

References

Atkinson, A. C. (1985). *Plots, Transformations and Regression: An Introduction to Graphical Method of Diagnostic Regression Analysis*, Oxford University Press, Oxford.

Brownlee, K. A. (1965). *Statistical Theory and Methodology in Science and Engineering*, John Wiley, New York.

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. (1990). A Monte Carlo comparison of five procedures for identifying outliers in linear regression, *Communications in Statistics - Theory and Methods*, **19**, 1913–1938.

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.

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

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 (Methodological)*, **57**, 145–156.

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

Rousseeuw, P. J. and Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points(with comments), *Journal of the American Statistical Association*, **85**, 633–651.

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.

잠재적 이상치군에 대한 검정

서한손[a,1]

[a]건국대학교 응용통계학과

(2016년 12월 14일 접수, 2017년 1월 30일 수정, 2017년 2월 1일 채택)

요 약

일반적으로 잠재적 이상치군은 검정과정을 통해 최종적으로 이상치 여부를 판단하지만 검정절차를 생략하거나 모의실험에 의해 계산된 유의값을 기반으로 검정을 수행하는 이상치 탐지법들도 있다. 본 논문에서는 가면화나 수렁화현상을 피하기 위하여 이상치후보군에 속한 개별 관찰치를 검정하지 않고 이상치후보군의 부분집합들을 검정하는 절차를 제안한다. 제안된 방법의 활용을 보여주는 예제와 다른 방법과의 검정력 비교를 위한 모의실험 결과가 제시된다.

주요용어: 가면화, 선형모형, 수렁화, 이상치, 진단

이 논문은 2016학년도 건국대학교 KU학술연구비 지원에 의한 논문임.

[1](05029) 서울시 광진구 능동로 120, 건국대학교 응용통계학과. E-mail: hsseo@konkuk.ac.kr