The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.2.307
27(2), 307–315

A Test on a Specific Set of Outlier Candidates in a Linear Model

Han Son Seo[a] · Min Yoon[b,1]

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

(Received January 2, 2014; Revised February 21, 2014; Accepted March 20, 2014)

Abstract

An exact distribution of the test statistic to test for multiple outlier candidates does not generally exist; therefore, tests of individual outliers (or tests using simulated critical-values) are usually conducted instead of testing for groups of outliers. This article is on procedures to test outlying observations. We suggest a method that can be applied to arbitrary observations or multiple outlier candidates detected by an outlier detecting method. A Monte Carlo study performance is used to compare the proposed method with others.

Keywords: Linear regression model, outlier test, robust method.

1. 서론

선형회귀모형에서 이상치 탐지법은 활발하게 연구되어 다양한 접근법이 제시되어 있다. 이상치 탐지법은 이상치로 의심되는 관찰치들을 찾아내고, 그 값들이 실제로 이상치인지 아닌지 판별하는데 목적이 있다. 이상치 탐지법은 직접적 방법과 간접적 방법으로 분류할 수 있으며 직접적 방법은 이상치를 탐지, 제거하고 간접적 방법은 이상치에 영향을 작게 받는 강건 추정량을 사용하거나 이상치에 영향을 적게 받는 기준을 적용하여 이상치 문제를 해결하는 방식이다. 직접적 방법은 이상치를 탐지할 때 RSS 최소축소법 (Gentleman과 Wilk, 1975), 다단계 RSS 최소축소법 (Marasinghe, 1985) 등과 같이 이상치 개수가 미리 정해진 상황에서 수행되는 절차와 스튜던트화 잔차나 Cook's Distance와 같은 회귀진단 기준에 따라 관측값을 순서대로 정렬한 뒤에 잔차를 계산하여 이상치를 탐지하거나 (Kianifard와 Swallow, 1989) 잔차의 극단치를 사용하여 단계적으로 이상치를 탐지, 검정하는 (Hadi와 Simonoff, 1993) 순차적 절차(sequential process)가 있다.

간접적 이상치 탐지방법으로는 최소중위수제곱(LMS, Rousseeuw, 1984), 최소절사제곱(LTS; Rousseeuw, 1984, 1985), M-추정량 (Huber, 1973), MM-추정량 (Yohai, 1987) 등을 사용한 방법 등이 제시되어 있다. 이상치 탐지법은 이상치로 의심되는 관찰치를 탐지하는 측면과 탐지된 관찰치의 이상치 여부를 검정하는 측면으로 구분할 수 있다. 이 중 본 논문에서는 이상치 후보군에 대한 검정문제를 고려

This work was supported by the Konkuk University 2013.
[1]Corresponding author: Department of Statistics, Pukyong National University, 45, Yongso-ro, Nam-Gu, Busan 608-737, Korea. E-mail: myoon@pknu.ac.kr

하고자 한다. 다수의 관찰치로 구성된 특정 이상치 후보군을 검정하고자 할 때 일반적으로 F 분포등과 같은 정확한 검정 통계량의 분포가 존재하지 않는다 (Peña와 Yohai, 1995, p.150). 간접적 이상치 탐지 방법은 절차상 정확한 검정통계량 분포를 알 수 없으므로 이상치 가설검정에서 실험에 의해 계산된 유의값(simulated p-value)을 사용한다. 또한 직접적 이상치 탐지 방법에서도 일부를 제외하면, 전체 관찰치군에 대한 검정대신 개별 관찰치에 대한 t-검정을 수행하거나 (Peña와 Yohai, 1995) 검정절차에서 사용된 통계량을 검정통계량으로 사용할 때 실험에 의해 계산된 유의값으로 임계치를 추정하며 (Swallow와 Kianifard, 1996) 검정과정 없이 이상치 여부를 임의적으로 결정하기도 한다 (Sebert 등, 1998). 실험에 의해 계산된 유의값을 사용하는 경우 제한된 범위 내에서만 검정이 가능하며 실험의 변동성에 따른 유의값의 불확실성을 감수하여야 한다. 본 연구에서는 이상치 탐지절차와 상관없이 특정 관찰치군이 이상치로 탐지된 경우 근사적인 분포함수를 이용하여 검정절차를 수행하는 방법을 제시하고자 한다. 제시된 방법의 검정력과 실험에 의해 계산된 유의값을 사용하는 이상치 탐지기법의 검정력은 모의실험을 통해 비교된다. 2장에서는 본 연구에서 제안하는 스튜던트화 잔차를 이용한 근사 t 분포 검정 방법을 설명하며 이 방법과 비교할 반복잔차(recursive residual)에 의한 이상치 탐지법 (Swallow와 Kianifard, 1996)을 소개한다. 3장에서는 모의실험을 통하여 제안된 방법과 반복절차에 의한방법의 검정력을 비교하고 마지막으로 4장에서 연구결과를 요약한다.

2. 이상치 후보군에 대한 이상치 검정

여러 연구에서 제안된 이상치 탐지법들의 검정통계량은 정확한 분포 함수를 갖지 않아서 근사 분포를 이용하거나 모의 분포에 의하여 이상치 검정을 수행하는 경우가 많다. 만약 특정 관찰치 군이 서로 상이한 탐지법에 의하여 공통적으로 이상치로 간주되는 경우 이론적으로 정확한 분포가 확보된 검정통계량으로 검정하는 것이 효율적이다. 근사 분포가 알려진 대표적 검정통계량은 최소제곱추정법(OLS)에 의한 잔차를 이용한 탐지법이며 OLS 잔차를 통해 이상치 탐지를 시도하는 경우 우도비 검정과 누락 잔차 검정은 서로 상응한 관계가 성립된다. 최소제곱추정법에 따른 각종 잔차와 잔차의 분포에 관련된 특징은 다음과 같다.

설명변수와 종속변수 사이에 다음과 같은 관계식으로 표현되는 선형회귀 모형을 고려하자.

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

여기서 Y는 $n \times 1$ 반응변수 벡터, $\boldsymbol{\beta}$는 $p \times 1$회귀계수벡터, X는 p개의 설명변수를 나타내는 $n \times p$행렬이며 $\boldsymbol{\epsilon}$은 평균이 0이고 분산행렬이 $\sigma^2 I_n$인 $n \times 1$오차벡터이다. 최소제곱추정법에 의해 추정된 y의 추정치 $\hat{y}$이라고 할 때 선형회귀모형에서 모형의 진단에 사용되는 잔차 $\hat{\epsilon}_i$는 $\hat{\epsilon}_i = y - \hat{y}$으로 정의되며 잔차 $\hat{\epsilon}_i$를 표준편차로 나누어 표준화 시킨 표준화잔차(standardized residual) $\tilde{\epsilon}_i$는 다음과 같다.

$$\tilde{\epsilon}_i = \frac{\hat{\epsilon}_i}{S\sqrt{1-h_{ii}}}, \tag{2.2}$$

여기서 S는 σ의 추정치, h_{ii}는 해트행렬(hat matrix)의 i번째 대각원소(diagonal element)이다. 표준화잔차 $\tilde{\epsilon}_i$는 분자와 분모가 서로 독립적이지 않아서 정확하게 t 분포를 따르지는 않지만 n이 p보다 충분히 크면 자유도 $(n-p)$의 t 분포를 근사적으로 따르게 된다. i번째 관찰치가 이상치 이거나 다른 이유로 인해 자료에서 삭제되고 그 나머지 자료로부터 추정된 β와 σ의 추정치를 각각 $\hat{\beta}_{(i)}$, $S_{(i)}$라고 할 때 이 모형에 의해 y_i를 예측한 $\hat{y}_{(i)} = y_i - x_i'\hat{\beta}_{(i)}$와 y_i의 차이인 $\hat{\epsilon}_{(i)} = y_i - \hat{y}_{(i)}$를 삭제잔차(deleted residual) 또는 예측잔차(predicted error)라고 하며 예측잔차를 표준화시킨 다음과 같은 잔차 t_i를 외적스튜던트

화 잔차(externally studentized residual)라고 한다.

$$t_i = \frac{\hat{\epsilon}_{(i)}}{S_{(i)}\sqrt{1+h_{ii}}}. \tag{2.3}$$

식 (2.3)는 식 (2.2)의 표준화잔차 $\tilde{\epsilon}_i$에서 분산 σ^2의 추정치 S대신 해당 관찰치를 제외하고 계산된 $S_{(i)}$를 대입한, 아래와 같은 잔차 $\tilde{\epsilon}_{(i)}$와 같으므로 분자 분모의 독립관계가 성립하여 t_i는 자유도 $(n-p-1)$인 t 분포를 따르게 된다.

$$\tilde{\epsilon}_{(i)} = \frac{\hat{\epsilon}_i}{S_{(i)}\sqrt{1-h_{ii}}}.$$

또한 외적스튜던트화 잔차는 평균이동 이상치모형(mean shift outlier model)을 가정한 i번째 관찰치의 이상치 여부를 검정하는 검정통계량과 일치한다. 자료에 한 개의 이상치가 있는 경우 외적스튜던트화 잔차 t_i는 이상치를 더욱 부각시키지만 여러 개의 이상치가 있는 경우 t_i는 가면화 현상(masking)에 취약할 수 있다. 이 경우 이상치로 의심되는 후보군을 결정하고 이를 제외한 후 계산된 잔차를 검정통계량으로 사용하면 가면화 현상에 대비 할 수 있다. 따라서 순차적으로 일정한 크기의 관찰치를 제외한 후 계산된 이상치 후보군의 t_i를 t 분포에 비교하거나 t_i의 최소값에 Bonferroni 부등식을 적용하여 이상치를 탐지, 검정할 수 있다. 그러나 이상치 탐지와 검정을 위한 통계량이 이상치라고 생각되는 관찰치만 고려하는 경우 이에 속하지 않는 관찰치군에 의한 가면화 현상을 방지할 수 없다. 이에 따라 일정한 크기의 이상치 후보군을 결정할 때 잠정적 이상치군만 아니라 모형추정에 참여한 잠정적 양호치군도 함께 고려하는 것이 효과적이다. 잠정적 양호치군의 외적스튜던트화 잔차는 결국 표준화잔차와 일치하므로 모형추정 참여 여부에 따라 각 관찰치의 표준화잔차와 외적스튜던트화 잔차를 비교하여 최종적인 이상치 후보군을 결정 할 수 있다. 이와 같은 접근법을 이용하여 Hadi와 Simonoff (1993)는 일정 크기의 잠정적 양호치군으로부터 모형을 추정한 후, 잠정적 양호치군과 잠정적 이상치군에서 계산된 표준화잔차와 외적스튜던트화잔차의 절대값 순서 통계량에 대한 t 검정 결과에 따라 최종 이상치군을 결정하거나 잠정적 양호치군의 크기를 한 개 늘려서 반복된 절차를 수행하는 순차적 이상치 탐지법을 제안하였다.

본 연구에서는 위에서 설명한 과정을 응용하여 특정 관찰치군으로 이루어진 검정 대상군에 대한 이상치 여부를 판정하는 검정절차를 제안하고자 한다. 위의 과정을 적용하여 검정 대상군의 이상치 여부를 판단하기 위해서는 표준화잔차와 외적스튜던트화 잔차의 비교에 의해 검정 대상군이 최종 이상치후보군으로 지정되어야만 한다. 따라서 이와 같은 결과를 유도하는 잠정적 양호치군을 찾는 것이 필요하며 이를 위해 검정 대상군의 여집합을 잠정적 양호군으로 지정하여 모형추정을 수행하며 필요에 따라 검정 대상군 중 적정 개수 만큼을 교체하여 잠정적 양호군을 재 지정한다. 표준화잔차와 외적스튜던트화 잔차에 의해 검정 대상군이 최종 이상치군으로 판정되면 이상치 검정을 수행하고 그 결과에 의해 검정 대상치군의 이상치 여부를 판단하며 만약 다양한 잠정적 양호치군에 의해서도 검정 대상군이 최종 이상치군으로 판정되지 않으면 검정 대상군은 이상치가 아닌 것으로 판정한다. 본 연구에서 제안하는 특정 관찰치군에 대한 이상치 검정 과정을 구체적으로 설명하면 다음과 같다. 전체 관찰치 n개 중에서 검정에 포함되지 않는 관찰치의 크기는 s라고 할때 이상치 여부의 대상인 크기 $(n-s)$의 검정 대상군을 O라고 표기하자. 회귀식 추정에 참여하는 잠정적 양호치군을 M이라고 할때, 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 추정치라고 하자.

(0) 검정 대상군 O의 여집합 O^c를 회귀식추정에 참여하는 잠정적 양호치군 M으로 지정하여 회귀식을 추정한다.

(1) 회귀식 추정에 참여하는 잠정적 양호치군 M과 참여하지 않는 잠정적 이상치군 M^c로 나누어진 자

료에서 각각 표준화잔차와 외적스튜던트화 잔차인 d_i를 아래와 계산한다.

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

(2) d_i의 절대값 $|d_i|$의 크기가 작은 순서대로 전체 데이터를 정렬할 때의 순서통계량을 $d_{(i)}$라고 하자. $\{d_{(s+1)}, \ldots, d_{(n)}\}$에 해당하는 관찰치가 O와 일치하면 다음의 검정을 수행하고 O와 일치하지 않으면 단계 (3)을 수행한다.

a. 만약 $d_{(s+1)} \geq t_{(\alpha/2(s+1)),s-k}$이면, 검정 대상군 O를 이상치로 간주한다.

b. $d_{(s+1)} < t_{(\alpha/2(s+1)),s-k}$이면, 검정 대상군 O를 이상치가 아닌 것으로 판단한다.

(3) M의 관찰치중 한 개 (또는 적정갯수)를 M^c에 속한 것과 교체한 새로운 M으로 단계 (1)에서 부터 위 과정을 재 실행한다. 만약 일정 횟수의 반복 시도에서도 O가 최종 이상치 후보군으로 판정되지 않으면 검정 대상군 O는 이상치가 아닌 것으로 판단한다.

위의 과정에서 반복 시도의 횟수는 자료의 크기를 고려하여 결정하지만 경험적으로, 또한 계산량을 고려할 때 O^c의 관찰치중 한 개씩만 바꾸거나 최대한 두 개씩 바꾸는 경우까지만 시도한다.

본 연구에서 제안된 방법의 효율성을 검증하기 위하여 비교할 이상치 탐지법은 반복잔차를 이용한 고정적, 순차적 탐지법이다. 앞서 정의된 잔차는 이상치를 탐지하는데 있어서 저마다의 단점을 갖고 있으며 유용성이 부족하다는 것이 알려져있다 (Kianikfard와 Swallow, 1989; Barnett와 Lewis, 1994; Hawkins, 1980). 이에 따라 Kianifard와 Swallow (1989)는 반복잔차에 기반하여 이상치를 탐지할 것을 제안하였다. 식 (2.1)의 선형모형에서 반복잔차는 다음과 같이 정의되며 반복잔차를 효율적으로 계산할 수 있는 다양한 방법들 (Plackett, 1950; Brown 등, 1975)이 제안되어 있다.

$$w_j = \frac{y_j - x_j^T \widehat{\beta_{j-1}}}{\left[1 + x_j^T (X_{j-1}^T X_{j-1})^{-1} x_j\right]^{\frac{1}{2}}}, \quad j = p+1, \ldots, n.$$

Kianifard와 Swallow (1989)는 반복잔차를 이용한 개별적인 이상치 탐지과정을 다음과 같이 제안하였다.

(1) 회귀 모형을 데이터에 적합시킨후 n개의 관측값들 각각에 대해 적절한 회귀 진단값을 계산한다. (예를 들면, 스튜던트화 잔차 또는 Cook의 거리)

(2) 선택한 진단 측정값에 따라 관측값들을 정렬시키고 정렬된 데이터에서 처음 p개의 데이터를 사용하여 남은 $(n-p)$개의 관측값들에 대하여 반복된 잔차 w_j를 계산한다.

(3) 통계량 $|w_j/\hat{\sigma}|$ $(j = p+1, \ldots, n)$를 계산하고, 각각의 $|w_j/\hat{\sigma}|$값을 임계치와 비교하여 j번째 관찰치가 이상치가 아니다 라는 귀무가설을 기각하는 관찰치군을 이상치군으로 판정한다.

Kianifard와 Swallow (1989)는 위에서 제시한 각각의 t 통계량을 사용하는 과정외에 $\max|t_i|$의 순차적인 사용을 기반으로 한 순차적 반복 검정과정을 제안하였으며 순차적 반복검정과정은 위의 (1), (2) 과정 후 다음의 (3-1) 과정을 수행한다.

(3-1) 통계량 $\max|w_j/\hat{\sigma}|$ $(j=p+1,\ldots,n)$를 계산하고, 계산된 값을 임계치와 비교한다. 만약 귀무가설이 기각되지 않으면 과정은 끝마치게 되고, 귀무가설이 기각되어 최대값에 해당하는 관찰치가 이상치라고 판단되면 $(n-1)$개의 데이터로 다시 모형을 적합한 후 귀무가설을 기각하지 못할 때까지 과정을 반복한다.

반복잔차는 처음 p개의 관측값에 대해서는 계산되지 않으며 $w_{p+1},\ldots,w_n$는 독립이고 $N(0,\sigma^2)$을 따르는 확률변수이기 때문에 σ^2의 추정값으로 식 (2.3)에서 사용된 $s^2_{(j)}$를 이용하면 $|w_j/\hat{\sigma}_j|$를 t 분포의 임계치와 비교할 수 있다. 반복잔차를 이용한 이상치 검정에서 가면화 현상을 피하기 위하여 Swallow와 Kianifard (1996)는 σ^2의 추정값으로 $s^2_{(j)}$대신 강건 통계량인 사분위범위(IR; inter quartile range)와 중위절대편차(MAD; median absolute deviation from the median)를 사용할 것을 제안하였다. 사분위범위나 중위절대편차에 의한 검정통계량은 근사적인 확률분포를 계산하기 어렵기 때문에 실험에 의하여 계산된 임계치를 사용한다. 본 연구에서 제안된 검정과정은 반복절차를 이용한 두 가지 과정인 고정적 이상치 탐지 과정및 순차적 이상치 탐지과정과 효율성을 각각 비교하며 반복잔차 검정통계량에서 사분위범위와 중위절대편차를 σ^2의 추정치로 사용한다. 사분위범위와 중위절대편차를 사용한 검정통계량의 임계값은 Swallow와 Kianifard (1996, p.550, Table 2)가 실험에 의해 계산한 수치를 사용하기로 한다.

3. 모의실험

2절에서 제안된 특정 이상치군 검정방법의 효율성을 반복절차를 이용한 검정법과 비교하기 위하여 모의실험을 실시한다. 가상의 데이터는 Kianifard와 Swallow (1990)와 Swallow와 Kianifard (1996)와 유사한 방식으로 생성된다. 자료생성에 사용되는 모형은 $y_i=\beta_0+\beta_1 x_i+\epsilon_i$이며 잔차는 β_0와 β_1의 특정값에 영향을 받지 않으므로 $\beta_0=0,\beta_1=1$을 지정한다. 설명변수 x는 $U(0,15)$에서 생성되었고 오차 ϵ은 표준정규분포로부터 생성되었다. 이상치 집단의 관찰치는 실제 회귀선과 δ_i만큼 떨어진 $y_i=\beta_0+\beta_1 x_i+\delta_i$에 위치한다. 이상치는 한 개에서 세 개까지 지정되며, 이상치의 개수와 위치에 따른 다음과 같은 7개의 이상치 유형의 모의자료를 생성한다.

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

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

이상치를 포함한 전체 관찰치의 크기는 $n=25$로 고정하며 동일한 x값이 10번 반복하여 사용되었고 전체적으로 1000번의 실험이 수행되었다. 각 방법에 대한 검정력의 척도는 p_1, p_2, p_3로 표기하며 p_1은 이상치를 정확하게 전부 찾아낸 경우의 비율이고, p_2는 적어도 한 개 이상의 이상치를 찾아낸 경우의 비율이며, p_3는 탐지된 이상치 중에 정상관찰치가 포함되어 잘못 탐지된 경우의 비율이다. 따라서 이상치를 정상치로 탐지하게 되는 가면효과가 발생할 비율은 $(1-p_2)$이고, 정상치를 이상치로 탐지하게 되는 수렁현상(swamping phenomenon)이 발생할 비율은 p_3이다. p_1, p_2, p_3를 계산할 때 반복잔차에 의한 방법들은 이상치 탐지과정이 병행되어 검정이 수행되므로 전체 1000번의 실험이 기준이 된다. 하지만 본

Table 3.1. 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 four procedures and new procedure for each procedure under seven outlier patterns (a-g) having outliers planted at vertical distances from the true regression line.

Type	$\delta_i's$		Static-IR			Static-MAD			Seq-IR			Seq-MAD		
			p_1	p_2	p_3	p_1	p_2	p_3	p_1	p_2	p_3	p_1	p_2	p_3
a	3.0	Ori	0.24	1.00	0.76	0.36	1.00	0.64	0.84	1.00	0.16	0.87	0.97	0.10
		New	0.93	1.00	0.07	0.94	1.00	0.06	0.96	1.00	0.04	0.95	0.98	0.03
	3.5	Ori	0.27	1.00	0.74	0.35	1.00	0.65	0.84	1.00	0.16	0.89	1.00	0.11
		New	0.95	1.00	0.05	0.94	1.00	0.06	0.96	1.00	0.04	0.97	1.00	0.03
b	3.0	Ori	0.30	1.00	0.71	0.40	1.00	0.60	0.82	0.98	0.16	0.85	0.95	0.10
		New	0.94	1.00	0.06	0.93	1.00	0.07	0.94	0.98	0.04	0.92	0.95	0.03
	3.5	Ori	0.32	1.00	0.68	0.43	1.00	0.57	0.85	1.00	0.15	0.89	0.99	0.10
		New	0.93	1.00	0.07	0.94	1.00	0.06	0.96	1.00	0.04	0.97	1.00	0.03
c	3.0, −3.0	Ori	0.30	1.00	0.70	0.42	1.00	0.58	0.83	1.00	0.15	0.82	1.00	0.11
		New	0.90	1.00	0.10	0.90	1.00	0.10	0.92	1.00	0.05	0.89	1.00	0.03
	3.5, −3.5	Ori	0.29	1.00	0.71	0.44	1.00	0.56	0.83	1.00	0.17	0.86	1.00	0.13
		New	0.92	1.00	0.09	0.92	1.00	0.08	0.94	1.00	0.06	0.94	1.00	0.04
d	3.0, 3.0	Ori	0.42	1.00	0.58	0.52	1.00	0.48	0.82	0.97	0.15	0.78	0.90	0.11
		New	0.93	1.00	0.07	0.93	1.00	0.07	0.92	0.97	0.05	0.86	0.89	0.03
	3.5, 3.5	Ori	0.43	1.00	0.57	0.52	1.00	0.48	0.83	1.00	0.17	0.84	0.98	0.14
		New	0.94	1.00	0.06	0.93	1.00	0.07	0.94	0.99	0.05	0.94	0.97	0.04
e	3.0, 5.0	Ori	0.39	1.00	0.61	0.49	1.00	0.51	0.84	1.00	0.16	0.84	1.00	0.11
		New	0.94	1.00	0.06	0.94	1.00	0.05	0.94	1.00	0.05	0.92	1.00	0.03
	3.5, 5.5	Ori	0.38	1.00	0.62	0.46	1.00	0.54	0.83	1.00	0.17	0.88	1.00	0.11
		New	0.94	1.00	0.06	0.95	1.00	0.05	0.95	1.00	0.05	0.96	1.00	0.04
f	3.0, 4.0, 5.0	Ori	0.36	1.00	0.62	0.43	1.00	0.54	0.83	0.99	0.15	0.78	0.97	0.12
		New	0.92	1.00	0.05	0.89	1.00	0.05	0.93	0.99	0.04	0.88	0.97	0.03
	3.5, 4.5, 5.5	Ori	0.42	1.00	0.57	0.49	1.00	0.51	0.83	1.00	0.16	0.87	0.99	0.11
		New	0.94	1.00	0.06	0.95	1.00	0.05	0.95	1.00	0.05	0.95	0.99	0.04
g	3.0, 3.0, 3.0	Ori	0.47	1.00	0.53	0.53	1.00	0.46	0.74	0.90	0.16	0.69	0.81	0.11
		New	0.94	1.00	0.06	0.93	1.00	0.07	0.84	0.89	0.05	0.76	0.80	0.03
	3.5, 3.5, 3.5	Ori	0.53	1.00	0.47	0.57	1.00	0.43	0.81	0.98	0.17	0.84	0.95	0.11
		New	0.94	1.00	0.06	0.94	1.00	0.06	0.93	0.98	0.05	0.91	0.95	0.04

연구에서 제안한 방법은 단지 이상치 여부만 판단하기 때문에 반복잔차 방법에 의하여 결정된 이상치군에 대하여 검정을 적용하여 이상치로 판정된 경우들의 횟수가 기준이 된다. 실험의 결과에 대한 표기법으로, 본 연구에서 제안한 방법을 New, 기존의 방법을 Ori라고 표기한다. 4개의 기존 방법 중 반복잔차를 이용한 고정적 방법을 Static, 순차적 방법을 Seq라고 표기하고 사용된 강건통계량의 종류에 각각 IR(사분위범위), MAD(중위절대편차)를 Static 또는 Seq 뒤에 하이픈(-)으로 첨가시켜 표기한다. 앞에서 설명한 7개의 이상치군의 유형(a)–(e)에서 δ_i값들을 달리 지정하여 생성된 자료에 각 방법들을 적용한 실험 결과는 Table 3.1과 같다.

실험의 결과를 보면, 기존의 반복잔차를 이용한 방법들에서 가면화 현상이 발생하고 있지 않으나 p_1과 p_3의 측면에서 비교해 보면, 고정적 방법보다 순차적방법이, 또한 사분위범위를 사용하는 것 보다 중위절대편차를 사용하는 것이 더 효과적인 것을 알 수 있다. 본 논문에서 제안한 방법은 기존의 네 가지 방법 보다 p_1, p_2, p_3의 관점에서 더 효율적이다. 본 논문에서 제안한 방법에서도 가면화 현상이 발생하지

않으며, 이상치를 정확하게 모두 탐지하는 비율 (p_1)과 수렁현상 (p_3)을 방지하는 측면에서 향상된 결과를 보여준다. 특히 반복잔차를 이용하는 고정적 방법과 비교할 때 p_1과 p_3의 비율이 급격하게 개선되며 순차적방법과의 비교에서도 명확하게 개선된 차이를 확인할 수 있다. 예를 들면 $\delta = 3$의 이상치 유형 (a)에서 사분위 범위를 사용한 고정적 방법의 p_1, p_2, p_3는 각각 0.24, 1.00, 0.76이지만 새롭게 제안된 방법은 $p_2 = 1$로써 가면화 현상이 잘 방지되며 p_1은 0.93으로 향상되어 정확하게 이상치를 탐지하는 비율이 높아지고 p_3는 0.07로 감소되어 수렁효과가 낮아지게 된다. 또한 중위절대편차를 사용한 순차적 방법과 비교하면 p_1, p_2, p_3가 각각 0.87, 0.97, 0.1에서 0.95, 0.98, 0.03으로 개선된 결과를 보여준다. 이상치 집단의 관찰치와 실제 회귀선과의 차이인 δ_i를 Table 3.1에서 지정한 3 또는 3.5보다 더 크게 5, 7등으로 지정하여 모의 실험한 결과에서도 비슷한 결과를 얻을 수 있었다.

일반적으로 이상치가 없는 경우의 실험을 통하여 검정방법의 유의수준을 확인하는 것이 필요하지만 본 연구에서는 특정군이 이상치로 탐지된 경우에 한해서만 제안된 검정방법이 수행되므로 이상치가 없는 경우에는 p_1과 p_2의 계산이 불가능하여 실험을 수행하지 않았다. 또한 다양한 표본크기에 대한 실험도 비교대상이 되는 반복절차에 의한 이상치 탐지법의 임계치가 몇 가지 경우에만 계산되어 있어 대표적인 크기의 모의실험만을 수행하였다.

4. 결론

본 연구에서는 자료의 일부분을 이용한 회귀분석에서 표준화잔차와 외적스튜던트화잔차를 계산하여 특정 관찰치군의 이상치 여부를 판단하는 검정과정을 제안하였고 반복잔차에 의한 이상치 탐지, 검정 방법과 검정력을 비교하였다. 다양한 이상치 유형을 적용한 모의실험의 결과에 의하면, 제안된 방법이 실험에 의한 유의값을 사용하는 이상치 검정 방법보다 더 효율적인 것을 확인할 수 있었다. 다만 제안된 방법은 표준화잔차와 외적스튜던트화잔차의 순서통계량을 검정통계량으로 사용할 때 잔차들 간 독립성이 보장되지 않아 근사적인 t 분포의 임계치와 차이가 날 수가 있으며 이러한 부분에 대한 보정 작업이 연구 중에 있다. 본 연구에서 제시하는 방법은 어떠한 이상치 탐지법의 검정과정에도 적용될 수 있을 뿐만 아니라 임의의 관찰치군에 대해서도 이상치 여부를 검정할 수 있다. 이에 따라 검정절차가 없는 이상치 탐지법은 본 연구에서 제안한 방법을 적용하여 순차적 방법으로 수정, 개선 될 수 있다.

References

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data*, 3rd Edition, Wiley, Chichester, UK.

Brown, R. L., Durbin, J. and Evans, J. M. (1975). Techniques for testing the consistency of regression relations over time, *Journal of the Royal Statistical Society, Series B*, **37**, 149–163.

Gentleman, J. F. and Wilk, M. B. (1975). Detecting outliers. Ⅱ .Supplementing The Direct Analysis of Residuals, *Biometrics*, **31**, 387–410.

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.

Hawkins, D. M. (1980). *Identification of Outliers*, Chapman and Hall, New York.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo, *Annals of Statistics*, **1**, 799–821.

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 - Theory and Methods*, **19**, 1913–1938.

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

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.

Plackett, R. L. (1950). Some theorems in least squares, *Biometrika*, **37**, 149–157.

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

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. In: Grossmann, W., Pflug, G., Vincze, I., Wertz, W. (Eds.), *Mathematical Statistics and Applications*, **B**, Reidel, Dordrecht, 283–297.

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.

Swallow, W. H. and Kianifard, F. (1996). Using robust scale estimates in detecting multiple outliers in linear regression, *Biometrics*, **52**, 545–556.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression, *The Annals of Statistics*, **15**, 642–656.

선형모형에서 특정 이상치 후보군에 대한 검정

서한손[a] · 윤민[b,1]

[a]건국대학교 응용통계학과, [b]부경대학교 통계학과

(2014년 1월 2일 접수, 2014년 2월 21일 수정, 2014년 3월 20일 채택)

요 약

이상치 후보군을 검정할 때 일반적으로 정확한 검정 통계량의 분포가 존재하지 않는다. 이에 따라 전체 관찰치군에 대한 검정대신 개별 관찰치에 대한 검정을 수행하거나 실험에 의해 계산된 유의값을 사용하여 이상치 가설검정을 수행한다. 본 연구에서는 임의의 관찰치 집단 또는 이상치 탐지절차에 따라 이상치 후보로 탐지된 특정 관찰치 집단의 이상치 여부를 검정하는 방법을 제시한다. 제시된 방법은 기존의 이상치 탐지기법에서 사용되는 검정방법과 모의실험을 통해 검정력을 비교한다.

주요용어: 강건 방법, 선형회귀모형, 이상치 검정.

이 논문은 2013학년도 건국대학교의 지원에 의하여 연구되었음.

[1]교신저자: (608-737) 부산 남구 용서로 45, 부경대학교 통계학과. E-mail: myoon@pknu.ar.kr