응용통계연구 (2011)
24(5), 979–986
DOI: http://dx.doi.org/10.5351/KJAS.2011.24.5.979

동적 그림을 이용한 이상치 검색

안병진[1] · 서한손[2]

[1]건국대학교 응용통계학과, [2]건국대학교 응용통계학과

(2011년 8월 접수, 2011년 9월 채택)

요 약

선형회귀모형분석은 방법의 간편성과 높은 적용성에 의해 다양한 종류의 자료 분석에 활용되고 있다. 하지만 자료에 이상치가 포함되는 경우 이에 민감하게 영향을 받게 되므로 의심되는 관찰치를 찾아서 이상치 여부를 검토하는 것이 중요하다. 그러나 이상치를 탐지하는 방법의 대부분은 가면화 효과 등 이상치로부터 영향을 받아 정확하게 이상치를 발견하지 못하는 경우가 있다. 본 연구에서는 이를 개선하기 위하여 동적 잔차도를 활용한 방법을 제안한다. 제안된 방법은 종속적 이상치탐지방법을 사용할 때 다양한 기초군을 제공하는데 유용하며 결과적으로 정확한 이상치군을 탐지하게 되는 것을 예를 통해 검증한다.

주요용어: 동적화, 이상치, 선형회귀모형, 잔차도.

1. 서론

통계분석의 대표적인 기법인 선형회귀모형분석은 설명변수와 종속변수 간에 다음과 같은 관계식을 가정한다.

$$Y = X\beta + \epsilon,$$

여기서 Y는 $n\times1$ 설명변수 벡터, β는 $p\times1$ 회귀계수벡터, X는 p개의 설명변수를 나타내는 $n\times p$ 행렬이며 ϵ는 평균이 0이고 분산행렬이 $\sigma^2 I_n$인 $n \times 1$ 오차벡터이다. 다양한 종류의 자료 분석에 활용되는 선형회귀모형분석은 이상치에 민감하게 영향을 받는다는 것이 잘 알려져 있다. 선형회귀모형분석에서 이상치를 탐색하는 방법은 다양한 측면에서 연구되어 왔다. Gentleman과 Wilk (1975)는 주어진 크기의 관찰치군들 중에서 전체 자료에 의한 분석에서 계산되는 잔차 합과 비교하였을 때 해당 관찰치군을 제외한 분석에 의한 잔차 합 축소가 가장 큰 관찰치군을 이상치 후보로 선정하였다. Marasinghe (1985)은 이상치군을 한꺼번에 찾는 경우 발생하는 계산의 복잡함을 개선하기 위하여 다단계 방법으로 정해진 크기의 이상치 후보군을 구성한 후 하나씩 이상치 여부를 검정하는 방법을 제안하였다. Paul과 Fung (1991)은 이른바 일반적 극단스튜던트화 잔차(generalized extreme studentized residual; GESR)를 이용하여 정해진 크기의 이상치후보군을 탐지하는 방법을 제안하였으며 GESR과 영향치의 측도를 모두 고려하여 이상치 후보를 파악한 후 이를 차례로 검정하는 이 단계 탐지법도 제안하였다. Kianifard와 Swallow (1989, 1990)는 반복잔차(recursive residual)에 의하여 이상치후보군을 형성하고 후보군의 각

이 논문은 2011학년도 건국대학교의 지원에 의하여 수행되었음.
[2]교신저자: (143-701) 서울시 광진구 화양동 1번지, 건국대학교 응용통계학과, 교수.
E-mail: hsseo@konkuk.ac.kr

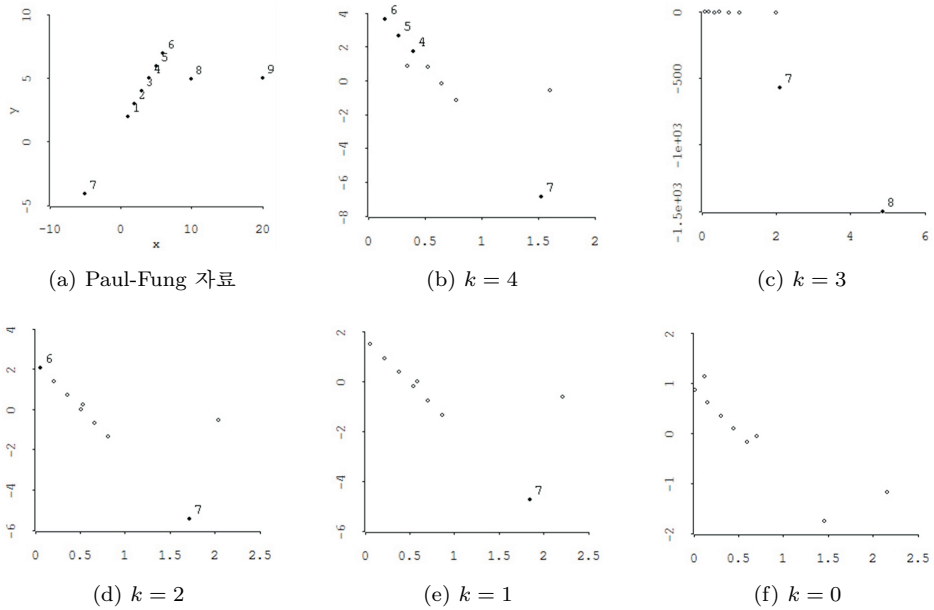
관찰치에 대하여 순차적 검정을 수행하는 방법을 제시하였다. Hadi와 Simonoff (1933)은 기초적인 순수관찰치군으로부터 이상치군의 크기를 순차적으로 줄여나가는 방법으로써 각 단계별로 잔차가 가장 큰 관찰치에 대하여 t-검정을 수행하는 방법을 제안하였다. 이상치군 탐색에서 중요한 것 중 하나는 계산량의 크기이다. 이와 관련하여 Atkinson (1994)은 전진탐색법에 의해 계산량을 감소시켜 이상치군을 빠르게 찾을 수 있는 방법을 제시하였으며 Pena와 Yohai (1999)도 설명변수의 개수가 많은 경우에도 적절한 계산량에 의해 이상치 탐색이 가능한 방법을 제안하였다. 최근 Jajo (2005)는 잔차에 기반을 둔 여러 가지 이상치군 탐지 방법을 소개하고 그들 간의 장단점을 비교하였다. 이상치군 탐지를 위해 제안된 여러가지 방법들 중 초기에 제안된 방법들 (Gentleman과 Wilk, 1975; Marasinghe, 1985)은 이상치의 크기가 미리 정해져야하는 단점이 있는 반면 최근에는 이상치군의 크기를 순차적인 과정을 통해 결정하는 방법이 제안되고 있다. 두 가지 방법에서 공통적으로 중요한 절차는 이상치군 크기에 따라 전체 관찰치를 이상치 한계를 벗어나는 이상치군과 그렇지 않은 양호치군으로 나누는 작업이다. 이 과정에서 사용되는 접근법은 두 가지로 구분된다. 이상치군의 크기에 따른 각 단계별로 이상치군을 탐지할 때 그 이전 단계에서 발견한 이상치군에 기반을 둔 방법과 각 단계별로 독립적으로 이상치군을 탐지하는 방법이다. 전자에 속하는 방법의 예로는 Gentleman과 Wilk (1975)의 방법이 있고 후자에 속하는 방법으로는 Hadi와 Simonoff (1933)의 방법이 있다. 두 종류의 방법은 모두 각 단계별로 왜곡되게 이상치군을 선정할 가능성이 있으며 특히 방법의 전개상 전자에 속하는 방법들은 이전 단계에서 이상치군을 잘못 결정하였을 때 지속적인 오류를 범할 가능성이 크다. 이러한 위험의 중요 원인은 이상치 탐색법 자체가 이상치들에 의하여 영향을 받아 왜곡된 결과를 초래케 하는 가면화 효과(masking effect)이다. Hadi와 Simonoff (1993)는 앞서 소개한 방법들보다 가면화 효과의 위험성이 개선된 절차를 제안하고 있으나 본 논문에서는 이를 더욱 보완할 절차를 제시하고자 한다. 본 논문에서는 각 단계에서 찾은 이상치군을 기반으로 적절한 잔차도를 작성하고 동적기능을 통해 이를 연속적으로 관찰함으로써 왜곡된 이상치군으로 인해 뒤따르는 오류를 방지하는 과정을 제안하고자 한다. 2장에서는 동적화의 대상이 되는 잔차도와 동적그림 방법의 과정을 설명하며 이러한 과정이 이상치군 파악의 오류를 방지하는데 효과적으로 사용되는 예를 제시하기로 한다. 3장에서는 본 논문에서 제시하는 과정의 확장성에 대하여 언급하기로 한다.

2. 동적 잔차도

잔차도는 Y축에 잔차를 X축에 반응변수의 예측치 등 관련 수치를 지정하여 작성된다. 이상치를 발견할 목적으로 잔차도를 사용할 때 X축에 지정될 값은 영향치의 정도를 나타내는 측도가 가장 적합하다. 영향치는 분석에서 해당 관찰치를 제외하였을 때 추정된 모형식의 변화가 큰 관찰치를 의미하여 잔차와 함께 그려졌을 때 이상치의 중요도 여부를 판단할 수 있다. 영향치의 측도로 일반적으로 사용되는 해트행렬 $X(X^TX)^{-1}X^T$의 대각항원소 h_{ii}는 다음과 같다.

$$h_{ii} = \frac{(x_i - M(X))V(X)^{-1}(x_i - M(X))^T}{(n-1)} + \frac{1}{n},$$

여기서 $M(X)$는 X의 평균벡터이고 $V(X)$는 X의 분산행렬이다. 따라서 h_{ii}는 $(x_i - M(X))V(X)^{-1}(x_i - M(X))$에 전적으로 의존하게 되며 이 값은 마할라노비스 거리(Mahalanobis distance)의 제곱값과 일치한다. 따라서 잔차도의 X축에는 각 관찰치의 마할노비스 거리를 표시하기로 하며 잔차도의 Y축에는 표준화 잔차인 r_i/σ를 표시한다. 이와 같이 정의된 잔차도는 이상치군의 크기에 따른 각 단계별로 작성된다. 양호치군과 이상치군으로 나누어진 자료중에서 양호치군 만으로 모형을 추정하여 관련 통계량을 계산하고 이를 모든 관찰치에 적용하여 잔차도를 작성한다. 이때 잔차도의 각 관찰치는 양호치군과 이상치군으로 구분하여 표시된다. 동적 잔차도를 작성하는 구체적인 절차는 다음과 같으며 이상치의 크기에 따른 잔차도의 동적화는 Lisp-Stat (Tierney, 1990)을 이용하여 프로그래밍한다.



(a) Paul-Fung 자료 (b) $k = 4$ (c) $k = 3$

(d) $k = 2$ (e) $k = 1$ (f) $k = 0$

그림 2.1. Paul-Fung 자료와 이상치군 크기(k)에 따른 Hadi-Simonoff 방법에 의한 잔차도

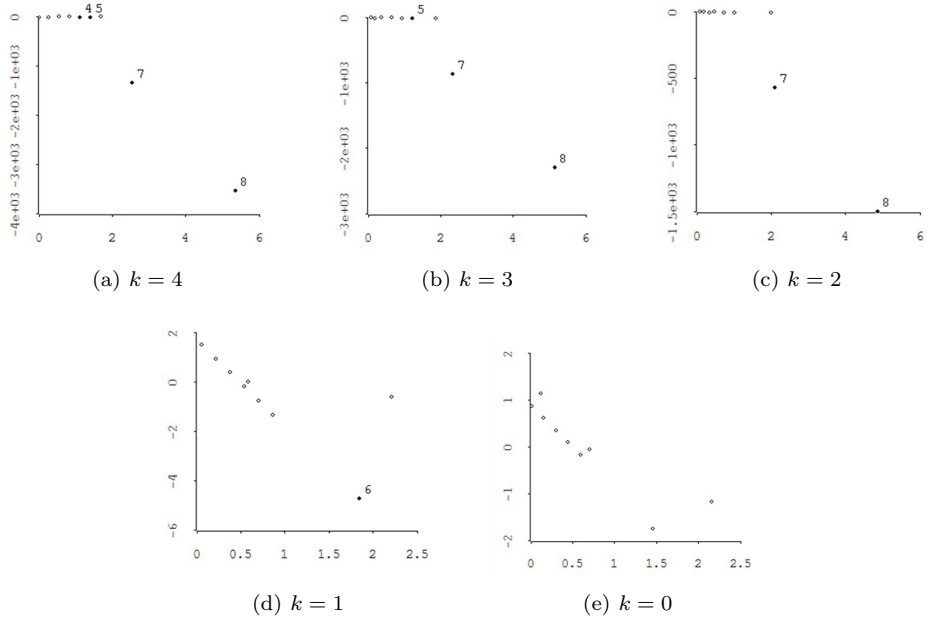
1. 특정 이상치탐지법에 따라 이상치군의 크기별로 이상치군을 구성한다 (이상치의 크기를 k라고 할 때 이상치를 제외한 양호치군으로 구성된 설명변수 행렬과 반응변수 벡터를 각각 $X^{(k)}$와 $Y^{(k)}$라고 하자).
2. $X^{(k)}$에 의해 설명변수의 평균과 공분산을 계산하고 (이를 각각 $M(X^{(k)})$와 $V(X^{(k)})$라고 하자) $X^{(k)}$와 $Y^{(k)}$에 의해 모형을 추정한다 (추정된 모수와 추정된 오차의 표준편차를 각각 $\hat{\beta}^{(k)}$, $\hat{\sigma}^{(k)}$라고 하자).
3. 각 관찰치 $i = 1, \ldots, n$ 별로 다음과 같이 $D_i^{(k)}$와 $SR_i^{(k)}$를 계산한다.

$$D_i^{(k)} = \sqrt{(x_i - M(X^{(k)}))V(X^{(k)})^{-1}(x_i - M(X^{(k)}))^T},$$
$$SR_i^{(k)} = \frac{r_i^{(k)}}{\hat{\sigma}^{(k)}}$$

   이때 $r_i^{(k)} = y_i - x_i\hat{\beta}^{(k)}$이다.
4. X축에 $D_i^{(k)}$, Y축에 $SR_i^{(k)}$를 지정한 잔차도를 작성하고 모든 관찰치는 이상치군과 양호치군으로 구분하여 표시한다.
5. k값의 변화에 따른 잔차도의 추세와 이상치군의 구성에 급격한 변화가 있는지 관찰한다.

서론에서 언급했듯이 이상치군을 탐지하는 방법들은 이상치군의 크기에 따른 각 단계별로 그 이전 단계에서 발견한 이상치군에 기반을 둔 방법과 각 단계별로 독립적으로 이상치군을 탐지하는 두 가지 종류로 구분될 수 있으며 전자를 종속적방법, 후자를 독립적방법으로 부르기로 하자.

종속적 방법은 독립적방법보다 일반적으로 계산량이 적다는 장점이 있는 반면에 그 이전 단계의 결과에 의존함으로써 어느 단계에서 가면화현상 등으로 오류가 발생할 경우 그 이후 단계에서도 이상치군을 제



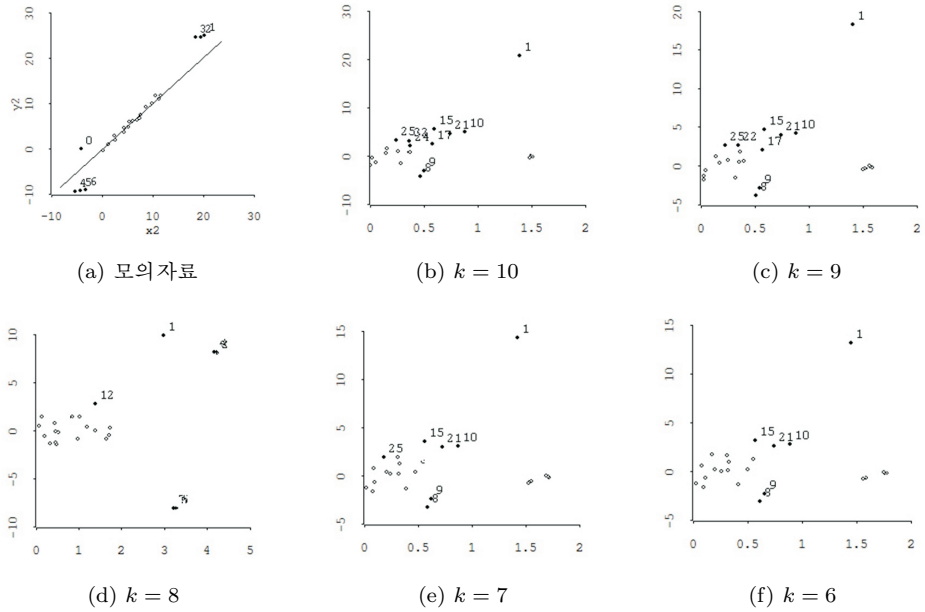
(a) $k = 4$ (b) $k = 3$ (c) $k = 2$

(d) $k = 1$ (e) $k = 0$

그림 **2.2.** Paul-Fung 자료와 Gentleman-Wilk 방법에 의한 이상치군 크기(k)에 따른 잔차도

대로 탐지하지 못할 가능성이 크다는 단점이 있다. 반면에 독립적 방법은 이전 단계의 결과와 상관없이 이상치군을 탐지함으로써 이전 단계에서의 오류에 영향을 받지 않지만 이 방법도 여전히 가면화현상 등으로 부터 자유롭지 못하다. 종속적 방법에 의해 탐지된 일련의 이상치군을 가지고 동적 잔차도를 구현했을 때 이상치군의 패턴에 일관성이 있다면 두 가지 가능성을 예측해 볼 수 있다. 첫 번째는 이상치군들이 오류 없이 성공적으로 탐지된 경우이고 두 번째는 최초 단계에서 발견된 기초 양호치군(혹은 이상치군)의 오류로 인해 결과적으로 이상치군들이 잘못 탐지된 경우이다. 이와 관련하여 Paul과 Fung (1991)의 자료에 대하여 종속적 방법인 Hadi-Simonoff 방법 (Hadi와 Simonoff, 1993)과 독립적 방법인 Gentleman-Wilk 방법 (Gentleman와 Wilk, 1975)에 의해 각각 탐지된 이상치군들의 동적 잔차도를 보기로 하자. 그림 2.1의 (a)는 자료의 산점도이며 관찰치 8과 9가 이상치군을 형성하고 있는 것을 알 수 있다. 그림 2.1의 (b)에서 (f)까지는 Hadi-Simonoff 방법에 의해 탐지된 이상치군의 잔차도들이며 그림에서 검은 점은 이상치군을 표시한다. 이상치군의 크기에 따라 생성되는 일련의 잔차도를 보면 이상치군의 구성과 잔차 및 마할노비스 거리 값 등에서 일관성을 유지하는 것을 알 수 있다. 그러나 이 방법은 어느 단계에서도 제대로 된 이상치군을 탐지하지 못하고 있으며 이는 최초 기초 양호치군을 잘못 선정함에 원인이 있다. 반면에 그림 2.2의 Gentleman-Wilk 방법의 경우에서는 동적 산점도의 패턴이 일관성을 유지하면서 성공적으로 이상치를 탐지하고 있다.

앞의 예에서 보듯이 종속적인 방법의 효율성은 이상치군 탐색의 시작 시점에서 사용하는 기초군의 우량 여부에 의해 결정되며 이와 관련하여 Hadi와 Simonoff도 두 가지의 기초군선정 방법을 제시하고 있다. 따라서 상대적인 계산량 등을 고려해 볼 때 독립적인 방법에 의해 이상치군을 탐지해 가는 과정에서 일관성이 결여되는 패턴이 나타나면 해당되는 두 단계의 결과를 기초군으로 삼아 종속적인 방법을 적용하는 것이 효과적인 전략이 될 수 있다. 이러한 전략을 적용하여 성공적으로 이상치군을 탐지한 예를 보기로 하자. 그림 2.3의 (a)는 모의실험에 의해 생성된 자료를 나타내는 산점도이다. 크기 $n = 25$개의 자료 중 28개의 자료는 $Y_i = X_i + \epsilon_i$, $i = 8, \ldots, 25$에서 생성되었으며 $X_i \sim U(0, 15)$와 $\epsilon_i = N(0, 0.5^2)$는



(a) 모의자료 (b) $k = 10$ (c) $k = 9$

(d) $k = 8$ (e) $k = 7$ (f) $k = 6$

그림 2.3. 모의자료 산점도와 Gentleman-Wilk 방법에 의한 이상치크기(k)에 따른 잔차도 일부

각각 균등분포와 정규분포에서 생성되었다. 나머지 7개 자료 $i = 1, \ldots, 7$는 추세선에서 4 또는 6 만큼 벗어나 이상치군을 형성하도록 생성되었다. 그림상의 직선은 28개의 정상관찰치 만으로 추정된 회귀선이다. 그림 2.3의 (b)에서 (f)까지는 모의자료에 대하여 독립적방법인 Gentleman-Wilk 방법에 따라 이상치군을 탐지하여 작성한 잔차도의 일부이다. $k = 10$에서 $k = 9$까지의 잔차도는 일관성 있게 변하지만 $k = 8$에서는 그 이전과 다른 패턴을 보여주고 결과적으로 $k = 7$에서 이상치군을 정확히 탐지하는데 실패하고 있다. 여기서 이상치군의 패턴변화가 발생한 $k = 8$과 $k = 9$를 기초군으로 Hadi-Simonoff 방법을 시도해 보기로 한다.

Hadi-Simonoff 방법은 기초양호군에서 시작하여 각 단계별로 이상치군의 크기를 줄여가면서 이상치군에 대한 최종적인 이상치 여부를 순서통계량에 대한 t-검정의 결과를 적용하여 결정한다. 만약 현재 단계의 이상치군 크기가 k라고 한다면 다음과 같이 양호치군에 속하는 관찰치에 대해서는 내부 스튜던트 잔차(internally studentized residual)를 계산하고 이상치군에 속하는 관찰치에 대해서는 예측오류를 계산하여 d_i라고 표시한다.

$$d_i = \begin{cases} \dfrac{y_i - x_i^T \hat{\beta}^{(k)}}{\hat{\sigma}^{(k)} \sqrt{1 - x_i^T (X^{(k)^T} X^{(k)})^{-1} x_i}}, & \text{관찰치가 양호치군에 속할때,} \\ \dfrac{y_i - x_i^T \hat{\beta}^{(k)}}{\hat{\sigma}^{(k)} \sqrt{1 + x_i^T (X^{(k)^T} X^{(k)})^{-1} x_i}}, & \text{관찰치가 이상치군에 속할 때.} \end{cases}$$

이상치군에 대한 이상치 여부는 d_i의 절대값을 오름차순으로 정리한 후 $(k+1)$번째 값에 대한 t-검정으로 판단한다. 즉 $|d|_{(k+1)}$을 $|d|$의 $(k+1)$번째 순서통계량이라고 할 때 $|d|_{(k+1)} \geq t_{(\alpha/\{2(k+1)\}, k-p)}$이면 $(n-k)$개의 큰 순서관찰치를 이상치라고 판단하고 그렇지 않을 경우에는 이상치군의 크기를 하나 줄여서 다음 단계의 이상치 탐색을 시도한다.

표 **2.1.** 고유방법에 의해 채택된 관찰치군을 기초군으로 이용하여 수행한 Hadi-Simonoff 결과

k	이상치군	검정
12	(1 5 6 7 8 9 10 14 15 16 20 21)	N
11	(1 5 6 7 8 9 14 15 16 20 21)	N
10	(1 5 6 7 8 9 14 15 16 20)	N
9	(1 5 6 7 8 9 14 16 20)	N
8	(1 5 6 7 8 9 16 20)	N
7	(1 5 6 7 8 9 16)	N
6	(1 5 6 7 8 9)	N
5	(1 5 6 7 8)	N
4	(1 5 8 15)	N
3	(1 8 15)	N
2	(1 8)	N
1	(1)	Y

표 **2.2.** 그림 2.3의 $k = 9$와 $k = 8$을 기초군으로 이용하여 수행한 Hadi-Simonoff 결과

k	이상치군	검정	k	이상치군	검정
9	(1 8 9 10 15 17 21 22 25)	N			
8	(1 8 9 10 15 21 22 25)	N	8	(1 2 3 4 5 6 7 12)	N
7	(1 8 9 10 15 21 22)	N	7	(1 2 3 4 5 6 7)	Y
6	(1 8 9 10 15 21)	N			
5	(1 8 10 15 21)	N			
4	(1 8 15 21)	N			
3	(1 8 15)	N			
2	(1 8)	N			
1	(1)	Y			

표 2.1은 Hadi-Simonoff 이 제안한 기초양호치군에 의하여 Hadi-Simonoff 과정을 적용할 결과이며 표 2.2는 Gentleman-Wilk 방법에서 $k = 9$와 $k = 8$의 결과(그림 2.3(c), (d))를 기초양호치군으로 삼아 Hadi-Simonoff 방법을 수행한 결과이다. 표 2.1에서 보듯이 원래의 Hadi-Simonoff 방법이 이상치군을 탐지하는데 실패하는데 반해 동적 잔차도에 의하여 패턴의 이상 단계를 파악한 $k = 8$을 기초양호군으로 적용한 결과 이상치군(1 2 3 4 5 6 7)을 정확하게 탐지하는 것을 알 수 있다.

3. 결론

본 연구에서 제안하는 동적그림에 의한 이상치군 탐색방법은 단계별로 나타나는 동적그림의 이상치 패턴을 분석하여 가면화현상 등을 파악하는데 목적이 있다. 단계별로 이상치를 탐지하는 방법은 어떤 방법도 가능하며 동적화의 대상인 잔차도도 다양한 종류가 사용될 수 있다. 본 논문에서는 마할로비스 거리와 표준화잔차를 이용한 일반적인 잔차도를 사용하였지만 이상치에 강건한 잔차도 (Rousseeuw와 Zomeren, 1990)를 이용할 수도 있다. 예를 들어 마할라노비스 거리는 $M(X)$와 $V(X)$의 비강건성에 의해 자료의 극단치를 제대로 측정할 수 없는 단점이 있으므로 최소부피타원형추정량(Minimum Volume Ellipsoid Estimator) 등 $M(X)$와 $V(X)$를 보다 강건한 추정량으로 대체할 수 있다. 잔차도의 Y축에도 표준화 잔차인 r_i/σ대신 강건통계량으로 대체될 수 있다. 일반적으로 사용되는 최소제곱잔차(least squared residual) 대신 최소중위수제곱 잔차(least median squared residual)를 사용하거나 σ 또한 잔차제곱의 중위수로 추정하여 사용할 수 있다.

참고문헌

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

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.

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

Kianifard, F. and Swallow, W. H. (1989). Using recursive residuals, calculated on adaptive-ordered observations, to identify outliers in linear regression, *Biometrics*, **45**, 571–885.

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.

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. and Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points, *Journal of the American Statistical Association*, **85**, 633–639.

Tierney, L. (1990). *Lisp-Stat*, John Wiley & Sons, New York.

Outlier Detection Using Dynamic Plots

Byung Jin Ahn[1] · Han Son Seo[2]

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

(Received August 2011; accepted September 2011)

Abstract

A linear regression method is commonly used to analyze data because of its simplicity and applicability; however, it is well known that data may contain some outliers and influential cases that may have a harmful effect on a statistical analysis. Thus detection and examination of outliers or influential cases are important parts of data analysis. In detecting multiple outliers, masking effects usually occur and make it difficult to identify the true outliers. We propose to use dynamic plots as a method resistant to masking effect. The procedure using dynamic plots is useful to find appropriate basic sets with which a dependent outliers detection method start and detect a true outliers set. Examples are given to demonstrate the effectiveness of the suggested idea.

Keywords: Dynamic graphics, linear regression model, outliers, residual plots.

This work was supported by the Konkuk University in 2011.
[2]Corresponding author: Professor, Department of Applied Statistics, Konkuk University, Hwayang-dong 1, Kwangjin-gu, Seoul 143-701, Korea. E-mail: hsseo@konkuk.ac.kr