한국통계학회논문집 제3권 3호
1996년 12월 pp. 11-23

Hadi와 Simonoff의 다중이상점 식별방법의 개선과 여러 다중이상점 식별방법의 효율성 비교

유 종 영[1], 김 현 철[2]

요 약

본 연구에서는 선형회귀분석에서 Hadi와 Simonoff의 다중이상점 식별방법을 수정하여 새로운 알고리즘을 제시하였다. Hadi와 Simonoff의 알고리즘 첫 단계에서 이상점일 가능성이 없는 점들의 집합을 추출할 때 가장효과와 편승효과에 영향을 받을 수 있음으로, 이 첫 단계를 수정하였다. 우리는 잔차가 일정한 분산을 갖는 정규분포에 따르다는 가정하에서 잔차의 신뢰구간을 생각하고, 이 구간안에서 잔차의 MAD가 최소인 새로운 모형을 탐색하고, 이를 이상점일 가능성이 없는 점들의 집합을 추출하는데 이용하는 새로운 알고리즘을 제시하였다. 제시된 방법은 실제자료에서 다른 방법에 비해 효율적으로 이상점을 식별할 수 있었다.

1. 서론

우리는 선형회귀모형에서 다중이상점을 찾는 방법에 대해 다루고자 한다. 선형회귀모형은 많은 분야에서 자료를 분석할 때 폭 넓게 사용되고 있으나, 분석하고자 하는 자료들은 종종 이상점(outlier)을 포함하기도 한다. 이런 이상점들은 분석결과를 크게 왜곡시킬 수도 있고 경우에 따라서는 가정한 모형이 잘못되었다는 중요한 정보를 갖고 있을 수도 있기 때문에 최종적인 분석에 앞서 이러한 점들을 구분해 낼 필요가 있다.

만약 자료에 하나의 이상점이 있다면 이 점을 식별하는 데는 이론적으로나 계산상에 별 문제가 없다. 그러나 두 개 이상의 이상점이 있을 때 단일이상점 검정(single outlier testing) 방법을 사용하면 가장효과(masking effect)와 편승효과(swamping effect) 때문에 식별이 곤란해진다. 가장효과란 두 개 이상의 이상점이 근접한 곳에 위치하여 단일이상점 검정방법에서 서로 상대방이 이상점으로 식별되지 못하도록 방해하는 것을 말한다. 이 때 처음부터 두 개 이상의 이상점이 있다는 것을 알고 있다면 가장효과를 피할 수 있도록 검정방법을 수정할 수 있는데, 이렇게 수정된 검정방법을 블록검정(block test)이라고 한다. 그러나 블록검정에서는 이상점이 아닌 점을 이상점으로 판정하는 경향이 있는데 이를 편승효과라고 한다. 따라서 블록검정에서는 이상점의 수를 미리 정확히 알고 있는 것이 매우 중요하다. 이처럼 단일이상점 검정이나 블록검정이 가장효과나 편승효과의 영향을 받는 까닭에 많은 대안들이 제시되어 왔다.

1) (449-714) 경기도 용인시 삼가동 117-6 용인대학교 전산통계학과 전임강사.
2) (573-701) 전라북도 군산시 미룡동 산 68번지 군산대학교 계산통계학과 조교수.

일반적으로 선형회귀모형에서 이상점을 식별하는 방법들은 k개의 이상점이 있다고 가정했을 때, 이상점일 가능성이 있는 부분집합과 그렇지 않은 나머지 n-k개의 관측점의 집합을 분리하여 식별을 시작한다.

Gentleman과 Wilk (1975)는 총 ${}_nC_k$개의 모든 부분집합에 대해 각 부분집합에 속한 점들을 제거시켰을 때 잔차의 제곱합이 가장 많이 줄어드는 k개의 관측점 집합을 이상점일 가능성이 가장 높은 집합으로 정의했다. 그리고 이들에 대해 좀 더 상세한 검정을 통해 이들이 정말 이상점들인지를 판정하고 만약 이상점들이 아니라면 이상점의 수를 하나씩 줄여 가면서(k-1, k-2, ...) 유의한 집합이 나타날 때까지 반복하는 연속적인 방법을 제안했다. 이 방법은 논리가 매우 간단하고 명확하다는 장점이 있으나 매우 많은 계산이 필요한 지루한 방법이다.

Marasinghe (1985), Paul과 Fung (1991)은 단일 이상점 검정에 사용되는 여러 가지 회귀진단값 중 하나를 선택하여 회귀진단값이 제일 큰 점을 찾고 이 점을 제외한 나머지 자료 군에 대해서 다시 분석된 회귀진단값이 제일 큰 점을 찾는 작업을 반복해서, 이상점일 가능성이 있는 k개의 점을 결정했다. 이렇게 결정된 최초의 이상점 부분집합에서 다시 더욱 상세한 검정을 수행하는 방법을 제안했다. 이런 방법들은 Gentleman과 Wilk의 방법에 비해 계산 측면에서는 크게 개선된 것이나, 일반적으로 미리 알 수 없는 이상점의 개수 k를 알고 있어야 한다는 단점을 갖고 있다. 또한 k를 모를 때 Marasinghe가 권장하듯이 k의 수를 안전하게 크게 잡는 방법을 사용할 경우 가장효과가 개입되는 것으로 알려져 있다(Barnett와 Lewis 1994, p.345).

이상의 방법들이 이상점일 가능성이 높은 점들의 집합을 먼저 생각하고 이들이 과연 이상점들인가를 확인하는 절차를 취한 데 비해, Hadi와 Simonoff (1993)는 반대로 이상점일 가능성이 없는 관측점들의 집합을 먼저 정하고 여기에 기초해서 일종의 거리를 계산한 뒤 이 값으로 이상점일 가능성이 없는 점들을 하나씩 처음에 정한 집합에 추가해 가는 방법을 제안했다. 그러나 이 방법도 처음에 이상점일 가능성이 없는 집합을 찾는 과정에서 Marasinghe나 Paul과 Fung처럼 기존의 통계량들을 사용함으로써 가장효과나 편승효과의 영향을 받는 것으로 분석된다.

앞에서 소개한 여러 가지 방법들은 만약 이상점일 가능성이 있는 집합, 또는 이상점일 가능성이 없는 집합을 적절하게 찾을 수만 있다면 좋은 이상점 식별방법이 될 수 있다. 그러나 처음 집합을 찾는 과정에서 가장효과나 편승효과가 개입되어 선형회귀모형에서 다중이상점을 찾는 문제는 사실상 이런 부분집합을 찾는 문제로 좁혀 볼 수 있다.

따라서 본 연구에서는 이상점일 가능성이 없는 점들의 집합을 찾는 방법에 대해 제의하고, 이 방법을 가장효과와 편승효과의 영향을 받을 수 있는 Hadi와 Simonoff의 알고리즘 첫 단계에 적용시켜 Hadi와 Simonoff의 다중이상값 식별방법을 수정하여 새로운 방법을 제시하였다.

이 방법은, 여러 연구자들에 의해 이상점 식별방법을 비교하는데 사용되었던 자료들에 적용시켜 본 결과, 가장효과나 편승효과에 영향을 덜 받는 것으로 나타났으며, 특히 동적 그래픽스에 의한 회귀진단 방법에 손쉽게 적용시킬 수 있다는 장점을 갖고 있는 것으로 분석된다.

2. 추정회귀선의 주변에서 새로운 회귀식의 탐색

2장에서는 추정회귀선의 주변에서 새로운 회귀식의 탐색방법에 대해 논의하고자 한다. 본 논문에서 제안하는 새 방법은 잔차가 일정한 분산을 갖는 정규분포에 따르다는 가정하에서 잔차의 신

뢰구간을 생각하고, 특정한 한 자료를 선택하여 그 잔차의 99% 신뢰구간내에서 잔차의 크기를 변동시켜가면서 새로운 회귀식과 잔차들을 생성한 후 이들 새로운 잔차중 잔차의 MAD(median absolute deviation)가 최소가 되는 모형을 최종적으로 선택한다. 그 다음 최종적인 모형의 잔차에서 잔차의 중앙값을 뺀 후 절대값을 취하여 작은 순서대로 배열한 후 이상점일 가능성이 없는 점들을 선택한다. 따라서 우리는 먼저 특정한 잔차의 크기를 변동시키면서 새로운 회귀식과 잔차를 생성시키는 방법을 제안하고자 한다.

2장에서는 3개의 Lemma를 제시하였는 데 이중 Lemma 1은 특정잔차를 신뢰구간내의 특정한 위치로 변동시켰을 때의 새로운 추정식과 잔차를 구하는 방법을 보여주고 있으며, Lemma 2와 3은 특정잔차를 선택하는 기준을 제공하고 있다.

우리는 다음과 같은 일반적인 선형회귀모형에 관심을 갖고 있다.

$$\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

여기서 $\mathbf{y}$는 $n\times 1$의 관측점 벡터, $\boldsymbol{\beta}$는 $p\times 1$의 모수벡터, $\boldsymbol{X}$는 $n\times p$인 알려진 상수들의 행렬이다. 또 $\boldsymbol{\varepsilon}$은 $n\times 1$인 오차항들의 벡터로 다음과 같은 분포를 갖는 것으로 가정하자.

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n),$$

여기서 $\boldsymbol{I}_n$은 차수 n인 단위행렬을 의미한다. 모형 (1)에서 모수 $\boldsymbol{\beta}$에 대한 최소제곱추정량과 그 분포는 각각 식 (2) 및 (3)과 같이 나타낼 수 있다.

$$\mathbf{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{y} \tag{2}$$

$$\mathbf{b} \sim N(\boldsymbol{\beta}, (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2). \tag{3}$$

또 잔차벡터와 그 분포는 식 (4) 및 (5)와 같다. 여기서 $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$이다.

$$\mathbf{e} = (\boldsymbol{I}_n - \boldsymbol{P})\mathbf{y} \tag{4}$$

$$\mathbf{e} \sim N(\mathbf{0}, (\boldsymbol{I}_n - \boldsymbol{P})\sigma^2). \tag{5}$$

따라서 추정된 회귀식은 다음과 같이 표현된다.

$$\mathbf{y} = \boldsymbol{X}\mathbf{b} + \mathbf{e}. \tag{6}$$

Easton (1994)은 선형회귀모형에서 최소제곱법으로 회귀계수 $\mathbf{b}$를 추정한 다음, 그 추정된 회귀계수를 99% 신뢰구간에서 변화시켜 원래 추정된 회귀식의 주변에서 흥미있는 새로운 회귀식을 추정하여 이상값을 식별하는 방법을 제안했다. 이 방법은 적절히 선택된 값 $\delta_1, \delta_2, \cdots, \delta_p$가 있을 때 p개의 회귀계수에 대해서 원래 추정된 최소제곱추정값의 주변에서 2^p 요인실험 $(b_1 \pm \delta_1, b_2 \pm \delta_2, \cdots, b_p \pm \delta_p)$을 하고 이 결과를 이용한 반응표면분석을 통해 콘트롤러 박스(controller box)를 이용하여 적절한 회귀식을 찾도록 되어 있다. 그는 δ_i로 $3s_{b_i}$을 사용하고 있는데, 이는 $\mathbf{b}$가 (3)과 같은 분포를 따르는 확률변수라는 점으로부터 정당화될 수 있다.

이런 분석은 잔차에 대해서도 비슷하게 수행할 수 있다. 우리가 회귀모형을 추정함으로써 얻게 되는 잔차 $\mathbf{e}$는 (5)와 같은 분포를 갖는다고 가정할 수 있다. 즉, $\mathbf{e}$는 일정한 표본오차를 갖기

때문에 현재 표본에서 계산된 잔차는 일정한 범위 내에서 오차를 수반할 수 있다. 이 점은 우리가 현재 계산된 잔차를 중심으로 일정한 범위 내에서 잔차값이 달라지는 흥미 있는 다른 회귀식을 찾을 수 있다는 의미가 된다. 여기에서 흥미 있는 추정식이란 MAD가 최소인 추정식을 말한다.

이제 (6)과 같이 추정된 회귀식에서 특정한 하나의 관측점에 대응하는 잔차의 크기가 최소제곱 추정 결과와 다른 값이 되는 새로운 회귀계수와 새로 구한 회귀식에서의 잔차를 구해 보자. 잔차는 오차와 달리 서로 독립이 아니기 때문에 하나의 관측점에 대응하는 잔차만 달라져도 다른 잔차들은 모두 함께 달라지게 된다. 먼저 다음과 같이 i번째 요소만 t이고 나머지는 모두 0인 $n\times 1$ 열 벡터 $\mathbf{d}_i(t)$를 정의하자.

$$\mathbf{d}_i'(t)=[0,\cdots,0,t,0,\cdots,0].$$

<Lemma 1> i번째 관측점에 대응하는 잔차가 $e_i^*=e_i+\delta_i$라는 제약 하에서의 최소제곱추정량 $\mathbf{b}_i$와 잔차벡터 $\mathbf{e}_i$는 각각 다음과 같다.

$$\mathbf{b}_i = \mathbf{b}-(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\ \mathbf{d}_i(t) \tag{7}$$

$$\mathbf{e}_i = \mathbf{e}+\boldsymbol{P}\ \mathbf{d}_i(t), \tag{8}$$

여기서 $\mathbf{b}$와 $\mathbf{e}$는 식 (2) 및 (4)와 같고, t는 δ_i/p_{ii}이고, p_{ii}는 $\boldsymbol{P}$의 i번째 대각요소이다.

<증 명> i번째 관측점에 대응하는 잔차가 $e_i^*=e_i+\delta_i$가 되는 회귀모형을

$$\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta}_i+\boldsymbol{\varepsilon}_i \tag{9}$$

라 하자. 또 $\boldsymbol{\varepsilon}_i$를 다음과 같이 분해했다고 하자.

$$\boldsymbol{\varepsilon}_i = \boldsymbol{\varepsilon}^* + \mathbf{e} + \mathbf{d}_i(t). \tag{10}$$

그러면 모형 (9)는 다음과 같이 고쳐 쓸 수 있다.

$$\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}^* + \mathbf{e} + \mathbf{d}_i(t)$$

위 식에서 우변의 끝 두 항은 우리가 알고 있는 벡터이므로 좌변으로 이항시키면 식 (11)과 같은 변형된 모형을 얻는다.

$$\mathbf{y} - \mathbf{e} - \mathbf{d}_i(t) = \boldsymbol{X}\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}^*. \tag{11}$$

모형 (9)에서 최소제곱추정량을 구하기 위해 편의상 변형된 모형 (11)에서 $\boldsymbol{\beta}_i$을 최소제곱법으로 추정하면 다음과 같다.

$$\begin{aligned}\mathbf{b}_i &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\ \mathbf{y}-\mathbf{e}-\mathbf{d}_i(t))\\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\ \mathbf{y}-(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\ \mathbf{e}-(\boldsymbol{X}'\boldsymbol{X})^{-1}\ \boldsymbol{X}'\mathbf{d}_i(t)\\ &= \mathbf{b}-(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\ \mathbf{d}_i(t)\end{aligned}$$

또 모형 (9)에서의 잔차 $\mathbf{e}_i$는 다음과 같다.

$$\begin{aligned}\mathbf{e}_i &= \mathbf{y}-\boldsymbol{X}\ \mathbf{b}_i \\ &= \mathbf{y}-\boldsymbol{X}(\ \mathbf{b}-(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\ \mathbf{d}_i(t)) \\ &= \mathbf{e}+\boldsymbol{P}\ \mathbf{d}_i(t).\end{aligned}$$

즉 특정한 관측점에 대한 잔차가 우리가 원하는 어떤 값($e_i^*=e_i+\delta_i$)이 되는 회귀식의 계수는 $\mathbf{y}$ 대신 $\mathbf{y}-\ \mathbf{e}-\mathbf{d}_i(t)$를 종속변수로 하는 최소제곱추정으로부터 구할 수 있다. 이 말은 최소제곱추정의 결과로 얻은 잔차중 e_i를 δ_i만큼 이동시키고자 할 때 (즉 $e_i^*=e_i+\delta_i$), 새로운 회귀식과 잔차는 식 (7), (8)로부터 손쉽게 구할 수 있음을 의미한다.

한편 $\mathbf{e}_i$는 (4)와 같이 계산된 잔차벡터에 다음과 같이 표현되는 $\boldsymbol{P}\ \mathbf{d}_i(t)$가 더해진 벡터이다.

$$\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & & \vdots \\ p_{i1} & p_{i2} & \cdots & p_{in} \\ \vdots & \vdots & & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ t \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} p_{1i}t \\ p_{2i}t \\ \vdots \\ p_{ii}t \\ \vdots \\ p_{ni}t \end{bmatrix} \qquad (12)$$

따라서 i번째 관측점에 대응하는 잔차를 e_i^*가 되게 하려면 $p_{ii}t=\delta_i$가 되어 t는 δ_i/p_{ii}이어야 한다. ■

식 (12)는 특정한 잔차를 어떤 값으로 미리 정해 주면 다른 잔차들도 모두 제약이 없을 때의 잔차와 달라지는 것을 보여 준다. 따라서 i번째 관측점에 대응하는 잔차를 e_i^*가 되게 하는 방법은 직접 그 잔차를 e_i^*라고 정해 주는 방법만 가능한 것이 아니라 다른 점에 대응하는 잔차를 어떤 값으로 정해 줌으로써 간접적으로 i번째 관측점에 대응하는 잔차를 e_i^*가 되게 할 수도 있다. 따라서 하나의 관측점에 대응하는 잔차를 움직여서 e_i를 e_i^*가 되게 하는 방법은 총 n가지가 있을 수 있다. 이 중에서 잔차제곱합을 가장 적게 해 주는 추정 방법은 i번째 관측점에 대응하는 잔차를 직접 e_i^*라고 제약하는 방법이다. 다음의 Lemma 2와 3은 이를 보여주고 있다.

<Lemma 2> i번째 관측점에 대응하는 잔차를 e_i^*로 하는 회귀식을 Lemma 1에서처럼 추정했을 때 잔차제곱합(RSS_i)은 다음과 같다.

$$RSS_i = \mathbf{e}'\mathbf{e} + \mathbf{d}_i'(t)\ \boldsymbol{P}\ \mathbf{d}_i(t).$$

<증 명>

$$\begin{aligned}RSS_i &= (\ \mathbf{e}+\boldsymbol{P}\ \mathbf{d}_i(t))'(\ \mathbf{e}+\boldsymbol{P}\ \mathbf{d}_i(t)) \\ &= \mathbf{e}'\mathbf{e}+\ \mathbf{d}_i'(t)\boldsymbol{P}\ \mathbf{d}_i(t) \quad (\because \boldsymbol{X}'\ \mathbf{e}=0).\end{aligned}$$

■

< **Lemma 3**> 임의의 관측점 하나에 대응하는 잔차를 어떤 값으로 지정하여 i번째 관측점에 대응하는 잔차를 e_i^*가 되게 만드는 총 n가지의 방법 중에서 잔차의 제곱합이 최소인 방법은 식 (10)의 분해식에서 $\mathbf{d}_i(t)$에 다음과 같은 $\mathbf{d}_i(\delta_i/p_{ii})$를 사용하는 것이다.

$$\mathbf{d}_i'(\delta_i/p_{ii})=[0,\cdots,0,\delta_i/p_{ii},0,\cdots,0].$$

<**증 명**> k번째 잔차를 어떤 값으로 미리 정해 놓고 최소제곱추정을 하면 잔차벡터는 원래의 잔차벡터에 다음과 같은 벡터가 더해진 것과 같다.

$$[p_{1k}t,\ p_{2k}t,\ \cdots,\ p_{kk}t,\ \cdots,\ p_{nk}t\ ]'.$$

따라서 하나의 잔차를 움직여서 i번째 관측점에 대응하는 잔차를 e_i^*가 되게 만드는 방법은, 앞의 식 (10)의 오차벡터 $\boldsymbol{\varepsilon}_i$의 분해식에서 $\mathbf{d}_i(t)$를 다음과 같이 n가지 벡터중 하나의 벡터로 지정해 주면 되기 때문에 총 n가지가 존재한다.

$$\mathbf{d}_1(\delta_i/p_{1i}),\ \mathbf{d}_2(\delta_i/p_{2i}),\ \cdots,\ \mathbf{d}_i(\delta_i/p_{ii}),\ \cdots,\ \mathbf{d}_n(\delta_i/p_{ni}).$$

이때 총 n가지의 방법중 $\mathbf{d}_i(\delta_i/p_{ii})$를 사용하는 방법과 $\mathbf{d}_j(\delta_i/p_{ji})$ $(i\neq j)$를 사용하는 방법을 생각해 보자. $\mathbf{d}_i(\delta_i/p_{ii})$를 사용하였을 때의 잔차제곱합을 RSS_i, $\mathbf{d}_j(\delta_i/p_{ji})$를 사용하였을 때의 잔차제곱합을 RSS_j라 하면 RSS_i와 RSS_j는 각각 다음과 같다.

$$\begin{aligned} RSS_i &= \mathbf{e}'\mathbf{e} + \mathbf{d}_i(\delta_i/p_{ii})\boldsymbol{P}\ \mathbf{d}_i(\delta_i/p_{ii}) \\ &= \mathbf{e}'\mathbf{e} + \delta_i^2/p_{ii}, \end{aligned}$$

$$\begin{aligned} RSS_j &= \mathbf{e}'\mathbf{e} + \mathbf{d}_j(\delta_i/p_{ji})\boldsymbol{P}\ \mathbf{d}_j(\delta_i/p_{ji}) \\ &= \mathbf{e}'\mathbf{e} + \delta_i^2\frac{p_{jj}}{p_{ji}^2}. \end{aligned}$$

따라서 RSS_j에서 RSS_i을 빼면 다음과 같은 결과를 얻는다.

$$RSS_j-RSS_i=\delta_i^2\frac{p_{jj}}{p_{ij}^2}-\frac{\delta_i^2}{p_{ii}}=\frac{\delta_i^2}{p_{ji}^2p_{ii}}\begin{vmatrix} p_{ii} & p_{ji} \\ p_{ji} & p_{jj} \end{vmatrix} > 0 \qquad (\because \boldsymbol{P}\text{는 양정치 행렬}).$$

그러므로 i번째 관측점에 대응하는 잔차를 e_i^*로 만들고자 할 경우 $\mathbf{d}_i(\delta_i/p_{ii})$를 사용하는 것이 다른 어떤 $\mathbf{d}_j(\delta_i/p_{ji})$ $(i\neq j)$를 사용하는 것보다 잔차제곱합이 적어진다. ■

이상의 Lemma들은 Hadi와 Simonoff의 방법에서 이상점일 가능성이 없는 점들의 집합을 찾는 단계를 수정하는 데 이용된다.

3. 수정된 Hadi와 Simonoff의 다중이상점 식별방법

Hadi와 Simonoff는 M을 이상점일 가능성이 없는 점들의 집합을 나타내는 지표, $\mathbf{y}_M$과 $\boldsymbol{X}_M$을 지표 M에 해당되는 $\mathbf{y}$와 $\boldsymbol{X}$의 부분행렬, $\mathbf{b}_M$과 $\hat{\sigma}_M^2$은 지표 M에 의해 추정된 회귀계수와 표본분산이라 정의하고 지표 M을 추출하는 두 가지의 방법을 제시하였다. 본고에서는 두 가지의 방법중 첫 번째의 방법을 소개하고 실증분석으로 비교하였다.

3.1 Hadi와 Simonoff의 다중이상점 식별방법(이하 H&S)

<단계 1> 크기가 $h((n+p-1)/2$의 정수부분)인 이상점일 가능성이 없는 점들의 부분집합 M을 선택한다.

(1) 전체 자료를 가지고 회귀모형을 적합시킨 뒤 적절한 회귀진단값(예를 들면 표준화잔차)의 절대값의 크기가 작은 점 $p+1$개로 구성된 초기집합 B를 선택한다.

(2) B에 대해 다시 회귀모형을 적합시킨 뒤 다음 값을 계산하여 오름차순으로 배열한다.

$$\frac{|y_i - \mathbf{x}_i' \mathbf{b}_B|}{\sqrt{1 - \mathbf{x}_i'(\boldsymbol{X}_B' \boldsymbol{X}_B)^{-1} \mathbf{x}_i}} \quad \text{if} \quad i \in B$$

$$\frac{|y_i - \mathbf{x}_i' \mathbf{b}_B|}{\sqrt{1 + \mathbf{x}_i'(\boldsymbol{X}_B' \boldsymbol{X}_B)^{-1} \mathbf{x}_i}} \quad \text{if} \quad i \notin B.$$

(3) s를 초기집합 B의 크기라 했을 때 $s=h$이면 크기 h인 초기집합을 부분집합 M으로 결정한다. 그러나 $s<h$이면 $s+1$번째의 관측점을 초기집합에 포함시키고 (2)의 과정을 지속한다.

<단계 2> 각각의 d_i를 계산한다.

$$d_i = \frac{y_i - \mathbf{x}_i' \mathbf{b}_M}{\hat{\sigma}_M \sqrt{1 - \mathbf{x}_i'(\boldsymbol{X}_M' \boldsymbol{X}_M)^{-1} \mathbf{x}_i}} \quad \text{if} \quad i \in M$$

$$= \frac{y_i - \mathbf{x}_i' \mathbf{b}_M}{\hat{\sigma}_M \sqrt{1 + \mathbf{x}_i'(\boldsymbol{X}_M' \boldsymbol{X}_M)^{-1} \mathbf{x}_i}} \quad \text{if} \quad i \notin M.$$

(위에서 $i \in M$인 경우 d_i는 표준화된 잔차를 의미하고 $i \notin M$인 경우는 부분집합 M에 기초한 예측오차임을 알 수 있다.)

<단계 3> d_i에 절대값을 취한 후 오름차순으로 나열하고, $d_{(s+1)}$을 $|d_i|$의 $(s+1)$번째의 순서통계량이라고 가정한다. 이때 s는 현재의 부분집합 M의 크기이다.

(1) 만약 $d_{(s+1)} \geq t_{(\alpha/2(s+1),s-k)}$ 가 성립하면 $|d_i| \geq t_{(\alpha/2(s+1),s-k)}$ 를 만족하는 모든 관측점들을 이상점이라 판정하고 종료한다.

(2) 위의 식이 성립하지 않는 경우 처음의 $s+1$개의 관측점을 선택하여 새로운 집합 M을 구성한다. 만약 $n=s+1$이 성립하면 자료에는 이상점이 없는 것으로 판정하고, $n>s+1$인 경우 <단계 2>로 돌아간다.

3.2 수정된 Hadi와 Simonoff의 다중이상점 식별방법(이하 MH&S)

위의 Hadi와 Simonoff의 다중이상점 식별방법은 기본집합 M을 추출하는 과정에서 가장효과와 편승효과의 영향을 많이 받는 기존의 단일이상점 검정 방법을 이용한다는 데에 문제가 있는 것으로 분석된다. 따라서 우리는 Hadi와 Simonoff의 다중이상점 식별방법의 <단계 1>을 다음과 같이 수정하여 새로운 식별방법을 제시하였다. <단계 2>부터는 Hadi와 Simonoff의 방법과 동일하다.

<단계 1> 크기가 $h((n+p-1)/2$의 정수부분)인 이상점일 가능성이 없는 점들의 부분집합 M을 선택한다.

(1) $\boldsymbol{P}$의 대각요소중 값이 가장 큰 p_{kk}에 해당하는 잔차를 선택한다.

(2) 선택된 잔차를 다음 범위 내에서 변화시키면서 가장 작은 MAD를 갖는 모형을 탐색한다.

$$(e_k - 3\,\hat{\sigma}_{e_k},\ e_k + 3\,\hat{\sigma}_{e_k}) .$$

(3) 앞에서 최종 선정된 모형의 잔차에서 그 잔차들의 중앙값을 뺀 후 절대값을 취하여 작은 것부터 크기 h인 부분집합 M을 선택한다.

위의 (1)에서 $\boldsymbol{P}$의 대각요소중 가장 큰 p_{kk}를 선택하는 것은 잔차를 δ_k 만큼 이동시켰을 때의 잔차제곱합은 $RSS_k = \mathbf{e}'\mathbf{e} + \delta_k^2 / p_{kk}$ 이므로 p_{kk}가 클수록 잔차제곱합의 크기가 적어지기 때문이다. 또 (2)에서 MAD가 가장 작은 모형을 탐색하는 것은 이상점의 영향을 받지 않는 모형을 선택하기 위함이다.

4. 여러 가지 다중이상점 식별방법

우리는 3장에서 제시된 개선된 Hadi와 Simonoff의 다중이상점 식별방법(MH&S)의 효율성을 검정하기 위해 많은 연구자들에 의하여 이상점의 식별연구에 사용되어 온 여러 가지 분석자료들의 이상점을 식별하여 다른 방법과의 효율성을 비교하고자 한다. 이상값 식별방법은 크게 직접적인 방법과 간접적인 방법이 있는데(Hadi와 Simonoff), 여기서는 이들 방법 중에서 다음과 같은 방법들을 선택하여 비교하고자 한다. .

4.1 직접적인 방법

(1) 표준화 잔차를 이용하는 방법

표준화(standardized) 잔차의 절대값 중 최대값을 Prescott(1975)이 제시한 기각값과 비교하여 더 크면 그 점을 이상점으로 판정하고 나머지 자료의 새로운 표준화 잔차로 이상점 판정을 지속한다(이하 SME).

$$E_n = \max | e_{si} | > [(n-p)F/(n-p-1+F)]^{1/2},$$

여기서 $F = F_{\alpha/n}(1, n-p-1)$ 이고, $e_{si} = \dfrac{e_i}{s(1-p_{ii})^{1/2}}$ 이며, $s^2 = \dfrac{\mathbf{e}'\mathbf{e}}{n-p}$ 이다.

(2) 스튜던트화 잔차를 이용하는 방법

스튜던트화(studentized) 잔차중 최대값을 스튜던트의 t분포에서 구한 기각값과 비교하여 더 크면 그 점을 이상점으로 판정하고 나머지 자료의 새로운 스튜던트 잔차로 이상점 판정을 지속한다(이하 SMT).

$$T_n = \max | t_i | > t_{\alpha/2n,\, n-p-1}.$$

여기서 $t_i = \dfrac{e_i}{s_{(i)}(1-p_{ii})^{1/2}}$ 이고, $s_{(i)}$는 i번째 관측점을 제외하고 구한 s이다.

(3) 수정된 순환잔차에 의한 방법

Kianifard와 Swallow(1989)는 p개의 이상점 가능성이 없는 집합을 선택하여 기본집합으로 삼은 후 이를 이용하여 다음과 같이 정의되는 순환잔차(recursive residual)를 계산하여 이상점을 식별하는 방법을 제안했다.

$$w_j = \frac{\mathbf{y}_j - \mathbf{x}_j\, \mathbf{b}_{j-1}}{\{1 + \mathbf{x}_j'(\mathbf{X}_{j-1}'\, \mathbf{X}_{j-1})^{-1}\, \mathbf{x}_j\}^{1/2}}, \quad j = p+1, p+2, \cdots, n$$

여기서 $\mathbf{b}_{j-1} = (\mathbf{X}_{j-1}'\, \mathbf{X}_{j-1})^{-1}\, \mathbf{X}_{j-1}'\, \mathbf{y}_{j-1}$ 이고, $\mathbf{X}_{j-1}$은 처음 j-1행으로 구성된 부분행렬, $\mathbf{y}_{j-1}$은 처음 j-1개의 원소로 구성된 벡터이다.

이 방법은 다음과 같이 Kianifard와 Swallow의 통계량을 t분포의 기각값과 비교하여 더 크면 이상점으로 판정한다.

$$\max | w_j / s_{(j)} | > t_{\alpha/2n,\,(n-p-1)}$$

위 검정에서 이상점이 없는 것으로 판정되면 중단하고, 만약 이상점이 있다고 판정되면 그 점을 제외하고 다시 순환잔차를 계산한 뒤 검정을 반복한다.

한편 Kianifard와 Swallow(1990)는 이 방법이 너무 많은 계산을 필요로 하기 때문에 이 방법을 수정한 새 방법을 제안했는데, 처음에 이상점이 있는지에 대한 검정은 원래 방법을 따르고 이상점이 있다고 판정되면 나머지 n-p-1개의 순환잔차를 $t_{\alpha/n,\,(n-p-1)}$ 와 비교하여 판정하는 방법이다(이하 MR).

4.2 간접적인 방법

간접적인 방법은 일반적으로 로버스트 추정을 통해 얻은 추정결과에서 그래프를 통해 직관적으로 이상점을 찾는 방법이다. 이중에서 특히 LMS(least median of squares regression), LTS(least trimmed of squares regression) 등의 방법이 이상점의 식별에 있어서 효율성이 매우 높은 것으로 알려져 있으나 계산이 매우 복잡하다는 단점도 안고 있다. 그러나 Hadi와 Simonoff는 자신들의 방법의 효율성을 검증하기 위하여 이런 방법들과의 비교를 시도한 바 있다. 따라서 본 연구에서도 L1-추정법, M-추정법, LMS(least median of squares regression), LTS(least trimmed of squares regression)와 비교해 보았다. 여기서 사용된 간접적인 방법은 모두 S-Plus 통계팩키지를 사용하여 계산했다.

5. 실증분석

실증분석에서는 Hadi와 Simonoff가 사용했던 자료들과 Easton이 동적 그래픽스를 이용한 회귀진단에 사용했던 스택로스(Stack Loss) 자료를 추가하여 사용했다. 이 자료 중에서 처음 네 가지는 회귀분석에서 이상점 검정과 관련하여 지표처럼 사용되고 있다(Hadi와 Simonoff).

5.1 전화자료(Telephone Data, Rousseeuw와 Leroy 1987에서 재인용))

이는 24년간 벨기에의 국제전화 사용량 자료로, LMS나 LTS에 의해 15-20번의 자료가 이상점인 것으로 알려져 있다. 이 자료에 대해서는 L1-추정법이나 M-추정법도 모두 15-20번 자료를 이상점으로 판정하고 있다. 그러나 H&S와 MH&S를 제외한 다른 직접적인 방법들은 모두 가장효과에 의해 이상점을 전혀 찾아내지 못하였다. 다만 H&S와 MH&S는 14, 21번 두 개의 점을 추가로 이상점으로 판정하고 있는데, Hadi와 Simonoff는 이 두 점이 거의 이상점(marginal outliers)이라고 주장하고 있다.

5.2 별자료(Hertzsprung-Russell의 Stars Data, Rousseeuw와 Leroy 1987에서 재인용)

이 자료는 47개 별의 표면 온도와 밝기의 대수값으로 11, 20, 30, 34번 등 4개의 별이 낮은 온도이나 높은 밝기를 나타내는 이상점으로 알려져 있다. 이 자료에 대해서는 LTS나 LMS와 마찬가지로 H&S와 MH&S는 모두 이상점들을 정확하게 찾아내고 있다. 여기에서 특이한 점은 L1-추정법이나 M-추정법은 이상점을 전혀 찾아내지 못하고 있다는 사실이다.

5.3 중력자료(Gravity Data, Rousseeuw 1984)

중력을 측정한 이 자료는 다섯 개의 설명변수를 갖는 자료인데 Rousseeuw(1984)가 인위적으로 4, 6, 8, 19번을 이상점으로 만든 자료이다. 우리의 분석에 의하면 LTS와 LMS만이 19번을 제외한 3점을 이상점으로 판정했고, L1-추정법과 M-추정법은 모두 엉뚱한 11번을 이상점으로 식별했다. 그리고 나머지 모든 직접적인 방법들은 이상점이 없는 것으로 판정했다.

5.4 인공자료(Artificial Data, Hawkins 등 1984, Rousseeuw와 Leroy 1987에서 재인용)

3개의 설명변수를 갖는 자료로 1-10번 자료가 이상점이 되도록 만들어졌다. 그러나 1-10번 점들은 가장효과에 의해 이상점으로 식별되지 않고 11-14번 점들은 편승효과에 의해 이상점으로 식별되도록 교묘히 만들어져 있다. 분석 결과 우리가 제안한 MH&S와 LTS, LMS만이 정확한 이상점을 찾아내고 나머지 방법들은 모두 11-14번 점들을 이상점으로 판정했다.

5.5 새로운 인공자료(New Artificial Data, Hadi와 Simonoff 1993)

이 자료는 Hadi와 Simonoff가 그들 방법의 효율성을 검증하기 위하여 인위적으로 1, 2, 3번 점을 이상점으로 만든 자료이다. 이 자료에 대한 분석 결과에서도 우리가 제안한 방법은 정확하게 이상점을 식별해 내었으며 H&S도 정확하게 이상점을 찾아낸 것은 물론이다. 그러나 LMS와 LTS의 방법은 1, 2, 3점 이외에 6, 11, 17번을 이상점으로 판정하고 있다.

5.6 스택로스 자료(StackLoss Data, Brownlee 1965, Rousseeuw와 Leroy 1987에서 재인용)

이 자료는 그 동안 많은 사람들에 의해 이상점 식별 방법을 보여주기 위해 인용되어온 자료이다. 3개의 설명변수에 21개의 관측점을 갖고 있는데 LTS, LMS, H&S와 마찬가지로 우리가 제안한 방법으로도 1, 3, 4, 21번 점들이 이상점으로 판별되었으며 Easton은 최소제곱법으로 추정한 회귀식의 주위에서 탐색한 추정식으로 1, 2, 4, 21점을 이상점으로 주장하였다.

<표 1> 여러 가지 자료의 이상점 식별 결과 비교

비교방법 \ 비교자료		전화 자료	별 자료	중력자료	인공자료	새로운 인공자료	스택로스 자료
직접적인 방법	SME	× [1]	×	×	11,12,13,14	×	×
	SMT	×	×	×	11,12,13,14	×	×
	MR	×	×	×	11,12,13,14	×	4,21
	H&S	14,15,16,17,18,19,20,21	11,20,30,34	×	11,12,13,14	1,2,3	1,3,4,21
간접적인 방법	L1-추정	15,16,17,18,19,20	×	11	11,12,13,14	1,2,3,17	21
	M-추정	15,16,17,18,19,20	×	11	11,12,13,14	1,2,3,17	21
	LTS	15,16,17,18,19,20	11,20,30,34	4,6,8	1,2,3,4,5,6,7,8,9,10	1,2,3,6,11,17	1,3,4,21
	LMS	15,16,17,18,19,20	11,20,30,34	4,6,8	1,2,3,4,5,6,7,8,9,10	1,2,3,6,11,17	1,3,4,21
제안 방법 (MH&S)		14,15,16,17,18,19,20,21	11,20,30,34	×	1,2,3,4,5,6,7,8,9,10	1,2,3	1,3,4,21
이상점 [2]		14,15,16,17,18,19,20,21 혹은 15,16,17,18,19,20	11,20,30,34	4,6,8,19	1,2,3,4,5,6,7,8,9,10	1,2,3	1,3,4,21

(주) 1) ×는 이상점이 없다고 판정된 것임
2) 이상점은 여러 연구자에 의하여 이상점으로 알려진 점들임.

<표 1>은 위에서 소개한 여러 가지 자료들을 4장에서 소개된 분석방법과 Hadi와 Simonoff가 제시한 H&S, 본 논문에서 제시된 MH&S방법으로 이상점을 식별하여 정리한 것이다. <표 1>에서 볼 수 있듯이 간접적인 방법 중에서는 LMS와 LTS의 효율성이 매우 높은 것으로 나타났으며, 직접적인 방법으로는 우리가 제안한 MH&S방법의 효율성이 가장 높은 것으로 나타났다. 특히 Hadi와 Simonoff가 자기들의 방법의 효율성을 나타내기 위하여 만들었던 새로운 인공자료에서도 MH&S방법은 정확하게 이상점을 식별하였으며, 가장효과와 편승효과에 영향을 받도록 인위적으로 만들어진 Hawkins 등의 인공자료에서도 Hadi와 Simonoff 등의 방법에서는 이상점 식별에 실패하였으나, 우리가 제안한 MH&S방법은 정확하게 이상점을 식별하는 좋은 결과를 보여주고 있다.

6. 결론

그동안 연구되어온 다중이상점 식별방법들은 일반적으로 가장효과나 편승효과의 영향을 받는 것으로 알려져 왔으며, 많은 연구자들은 선형회귀에서 다중이상점을 찾는 문제를 이상점일 가능성이 있는 점들의 집합을 사용하는 방법과 이상점일 가능성이 매우 높은 점들의 집합을 사용하여 검정을 수행하는 두 방법으로 접근해 왔다. 그러나 이런 방법들이 처음 초기집합을 추출하는 과정에서 가장효과나 편승효과가 개입되어 이상점 식별에 실패하는 경우가 많은 것으로 분석된다.

한편 일반적으로 이상점 식별에 가장 이상적인 방법으로 인식되고 있는 로버스트 추정에 의한 간접적인 식별방법들은 계산이 매우 복잡한 데다가 식별기준도 분석자의 주관적인 기준에 따라 달라질 수 있는 단점을 지니고 있다.

따라서 본 연구에서는 직접적인 방법들의 장점인 검정방법이 단순하고 계산이 비교적 용이하다는 점과 로버스트 추정에 기초함으로써 이상점 식별이 정확하다는 간접적인 방법의 장점 사이에서 적절한 새 방법을 제시하고 있다. 우리의 방법은 Hadi와 Simonoff의 식별방법을 개선하는 형식을 취하고 있다. Hadi와 Simonoff의 식별방법은 맨 처음 이상점일 가능성이 없는 점들의 집합을 정하고서 시작하는데, 이 집합을 정할 때 그들이 제시한대로 표준화된 잔차와 같은 회귀진단값을 사용하면 가장효과나 편승효과가 개입될 수 있다. 우리는 이 첫단계에서 이상점일 가능성이 없는 점들을 정하기 위해 간접적인 방법의 하나로 간주할 수 있는 MAD가 최소인 모형을 추정하여 가장 적합이 잘되는 점들을 선택했다. 그러나 MAD가 최소인 모형을 추정하는 것은 매우 많은 계산이 필요하다. 우리는 Easton이 제한된 범위 내에서 흥미 있는 모형을 탐색했던 것처럼 제한된 범위 내에서 MAD가 최소인 모형을 탐색하는 방법을 제안하였는데, 이 방법은 최소제곱추정을 그대로 사용하기 때문에 계산이 매우 간단하고 기존의 회귀분석 패키지를 큰 수정 없이 사용할 수 있다. 특히 이 방법은 제한된 범위 내에서는 잔차제곱합이 최소가 되는 좋은 성질도 만족하고 있다.

우리는 새로 제안한 방법이 과연 실제 자료에 적합 시켰을 때 적절한 식별을 해줄 수 있는 지를 그 동안 많은 연구자들이 사용했던 자료들에 적용시켜 봄으로써 확인했다. 분석 결과, 우리의 방법은 비교해 본 직접적인 방법들 중에서 이상점을 가장 정확하게 찾아내는 것으로 확인되었으며, 간접적인 방법인 L1-추정이나 M-추정에 의한 것보다 더욱 정확하게 찾아내는 것을 확인할 수 있었다. 또 Hadi와 Simonoff가 제시한 인공자료에 대해서는 그들의 주장대로 LTS나 LMS에 의한 결과보다도 더 정확할 수 있음을 보여주고 있다..

참고문헌

[1] Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics : Identifying Influential Data and Sources of Collinearity*, New York: John Wiley & Sons.

[2] Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data,* John Wiley & Sons.

[3] Gentleman, J. F., and Wilk, M .B. (1975). Detecting Outliers II: Supplementing the Direct Analysis of Residuals, *Biometrics*, 31, 387-410.

[4] Easton, G.S. (1994). A Simple Dynamic Graphical Diagnostics Method for Almost Any Model, *Journal of American Statistical Association*, 89, 201-207.

[5] Hadi, A.S. (1992). Identifying Multiple Outliers in Multivariate Data, *Journal of the Royal Statistical Society*, Ser.B, 54,761-771.

[6] Hadi, A.S. and Simonoff, J.S. (1993). Procedures for the Identification of Multiple Outliers in Linear Models, *Journal of American Statistical Association*, 75, 1264 -1272.

[7] 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.

[8] Kianifard, F and Swallow, W.H. (1990). A Monte Carlo Comparison of Five Procedures for Identifying Outliers in Linear Regression, *Communications in Statistics, Part A-Theory and Methods,* 19, 1913-1938.

[9] Marasinghe, M.C. (1985). A Multistage Procedure for Detecting Several Outliers in Linear Regression, *Technometrics*, 27, 395-399.

[10] Paul, S.R and Fung, K.Y. (1991). A Generalized Extreme Studentized Residual Multiple Outlier Detection Procedure in Linear Regression, *Technometrics*, 33, 339-348.

[11] Prescott, P. (1975). An Approximation Test for Detecting Several Outliers in Linear Regression, *Technometrics,* 17, 129-132.

[12] Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*, John Wiley & Sons.