Journal of the Korean Data Analysis Society (February 2013)
Vol. 15, No. 1 (B), pp. 177-186

선형회귀모형에서의 이상치 탐색방법들의 비교 연구*

박대인[1], 강현철[2], 한상태[3], 최호식[4]

요 약

통계적 자료 분석에서 이상치 검출은 기본적인 분석과정으로 모델 선택, 추론, 평가 등 거의 모든 분석과정에 상당한 영향을 끼친다. 경우에 따라서는 식별된 이상치 자체가 중요한 정보를 담기도 한다. 본 논문에서는 분석모형으로 선형회귀모형을 고려하고 이에 대한 여러 가지 이상치 탐색 방법들을 살펴보고자 한다. 이상치 탐색방법은 크게 지렛점 등의 전통적인 통계량 등을 활용하여 순차적으로 이상치를 식별하는 단일(single) 이상치 식별방법과 강건한 추정을 기반으로 여러 개의 이상치를 동시에 검출할 수 있는 다중(multilple) 이상치 식별방법(least median of square estimation, robust distance with minimum volume ellipsoid estimator)으로 구분할 수 있다. 본 논문에서는 다중 이상치를 포함하고 있다고 알려진 실제 자료를 통해, 최근 제안된 She, Owen (2011)의 축소추정법을 통한 다중 이상치 검출방법과 기존의 여러 방법들의 특징을 살펴보았다. 특히 수렁효과나 가면효과를 가진 자료에 대해서 She, Owen(2011)의 방법이 기존의 방법들보다 이러한 이상치들 간의 관계를 효과적으로 탐색할 수 있음을 확인하였다.

주요용어 : 이상치, 강건성, 선형회귀분석, 축소추정, 비볼록 벌점화.

1. 서론

통계적 자료 분석에서 이상치 검출(outlier detection)은 중요한 분석과정이다. 이상치는 대부분의 다른 관측치들에 비해 지나치게 크거나 작은 개체들을 통칭하는 용어로 모델 선택, 추론, 평가의 모든 분석과정에 상당한 영향을 미친다. 경우에 따라서는 이상치가 전체 자료에 대한 중요한 정보를 담을 때도 있는데, 예를 들어 데이터마이닝 분야의 사기탐지(fraud detection), 침입탐지(intrusion detection) 등과 같은 분야에서 이상치는 제거되어야 할 대상과 동시에 중요한 정보를 포함하고 있는 관측치로 다루어진다.

본 논문에서는 선형회귀모형에서 이상치 탐색과정을 살펴보기로 한다. 먼저 크기가 n개의 자료 $\{(\boldsymbol{x}_i, y_i) : \boldsymbol{x}_i \in R^p, y_i \in R\}_{i=1}^n$가 주어져 있을 때, $n \times p$인 자료 행렬을 $X = (\boldsymbol{x}_1^T, \boldsymbol{x}_2^T, \cdots, \boldsymbol{x}_n^T)^T$로 종속변수 벡터를 $\boldsymbol{y} = (y_1, \cdots, y_n)^T$라 놓자. 오차 ϵ_i가 평균 0, 분산 σ^2을 가진다고 할 때, 다음의 선형모형

$$y_i = \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i, \epsilon_i \sim (0, \sigma^2), i = 1, \cdots, n \tag{1}$$

---

*이 논문은 제1저자 박대인의 석사학위논문의 축약본입니다.
*2011년도 호서대학교의 재원으로 학술연구비 지원을 받아 수행된 연구임(2011-0016).
[1]135-513 충남 아산시 배방읍 세출리 산29-1, 호서대학교 정보통계학과 대학원 석사과정졸업.
E-mail : pcapcoms@naver.com
[2]135-513 충남 아산시 배방읍 세출리 산29-1, 호서대학교 정보통계학과 부교수. E-mail : hychkang@hoseo.edu
[3]135-513 충남 아산시 배방읍 세출리 산29-1, 호서대학교 정보통계학과 교수. E-mail : sthan@hoseo.edu
[4](교신저자) 135-513 충남 아산시 배방읍 세출리 산29-1, 호서대학교 정보통계학과 조교수.
E-mail : choi.hosik@gmail.com
[접수 2013년 1월 18일; 수정 2013년 2월 16일; 게재확정 2013년 2월 19일]

에 대한 최소제곱법에 의한 추정량을 $\hat{\boldsymbol{\beta}}^{LS}=(X^TX)^{-1}X^T\boldsymbol{y}$으로 놓자. 논의의 편의상 $\beta_0=0$으로 둔다. 그러면 예측값은 $\hat{\boldsymbol{y}}=(\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_n)^T=X\hat{\boldsymbol{\beta}}^{LS}=X(X^TX)^{-1}X^T\boldsymbol{y}=H\boldsymbol{y}$으로 주어진다. 여기서 $n\times n$인 행렬 H의 대각원소를 h_{ii}라고 놓으면 이 값은 기본적으로 이상치 탐색에 활용되는 통계량인 지레값(leverage; Hoaglin, Welsch, 1978)이 된다. 즉, 예측값 $\hat{y_i}$를 y_i로 편미분하면 $\partial\hat{y_i}/\partial y_i=h_{ii}$으로써 이 값은 i번째 관측치가 설명 변수들의 중심점으로부터 얼마나 떨어져 있는가를 나타내고 통상적으로 이 값이 크면 높은 지렛점(high leverage point)으로 분류한다. 이를 기초로 각 관측치 또는 변수가 가지는 영향력 및 특이성 등에 대한 지표 통계량으로, 스튜던트 잔차(studentized residual), 외적 스튜던트 잔차 R-student(externally studentized residual; Belsley, Kuh, Welsch, 1980), Cook의 거리(Cook, 1977), DFFITS(difference in fit standardized; Belsley, Kuh, Welsch, 1980) 통계량, Hadi 통계량(Hadi, Simonoff, 1993) 등 다양한 통계량들이 개발되어 왔다.

일반적으로 최소제곱법(ordinary least square; LS)을 통한 추정량은 이상치에 매우 민감하기 때문에 안정된 모형을 제공하지 못하는 것으로 널리 알려져 있다(Hoaglin, Welsch, 1978). 따라서 이를 활용한 지수화된 여러 통계량들을 통해서 이상치로 의심 가는 자료를 충분히 추적하고, 식별하는 과정은 일정정도의 한계를 가진다. 자료가 다중 이상치를 가질 경우 이러한 어려움은 보다 심화되는데, 이와 같은 경우에서는 일반적으로 적합단계에서 LS추정량 보다 강건한(robust)한 추정량을 구하는 방법들을 통해 이상치를 탐색한다. 만약 자료가 특이한 분포나 구조를 가질 때는 꼬리가 두터운 분포를 가정한 선형회귀분석(Jung, 2011)을 이용하거나 모호성을 반영하여 강건한 퍼지선형함수 추정하는 방법(Son, 2007)을 활용할 수 있다.

한편 자료의 크기가 커짐에 따라 이상치가 존재할 가능성은 커지게 되는데 이상치가 두 개 이상 함께 존재하게 되면 가면효과(masking effect)와 수렁효과(swamping effect)가 발생하는 것으로 알려져 있다. 가면효과는 보통 몇 개의 이상점들이 근접해 분포되어있는 경우 그중의 몇몇의 이상점이 식별이 안 되는 것을 의미하며, 수렁효과는 어떤 이상점의 큰 영향력에 의해 이상점이 아닌 다른 관찰치가 이상점으로 잘못 식별되는 경우를 의미한다. 이와 같이 다수의 이상점이 자료에 포함되어 있는 경우에는 그 이상점들을 식별하는 것이 용이하지 않다. 이상치를 탐색하는 방법론은 영향력 등을 통한 판별방법, 강건한 회귀추정량을 이용하는 방법 등 다양한 방법들이 개발되어 있다. 본 논문에서는 최근 제안된 She, Owen(2011)의 벌점화 방법론을 통해 다중 이상치를 검출하는 방법 및 그래프적으로 이들 간의 관계를 파악하는 방법을 소개하고, 기존의 여러 방법들과 비교하고 그 특징을 살펴보고자 한다.

본 논문의 구성은 다음과 같다. 2절에서는 이상치 탐색을 위한 전통적인 방법들과 최근 새롭게 제안된 비볼록 벌점화 회귀(nonconvex penalized regression)를 통한 축소추정(shrinkage estimation)방법론을 이용한 이상치 탐색 방법을 간단히 소개한다. 3절에서는 두 개의 실제 자료 분석을 통해 여러 추정방법들을 비교한다. 끝으로 결론 및 토의는 4절에 기술한다.

2. 이상치 탐색을 위한 방법

이 절에서는 최근 제안된 She, Owen(2011)의 방법을 중심으로 다중이상치 탐색 및 검출에 널리 쓰이는 방법들을 간단히 살펴보기로 한다.

2.1. 강건한 회귀방법(robust regression)

이 방법은 강건한(robust) 회귀 추정량을 통한 식별방법으로써 이상점 식별의 간접적인 접근방법

으로 볼 수 있다. 강건한 추정량 $\hat{\boldsymbol{\beta}}^R$은 이상점의 영향을 적게 받도록 한 추정량이기 때문에, 이를 통해 구해진 잔차 $\boldsymbol{e}=\boldsymbol{y}-X\hat{\boldsymbol{\beta}}^R$ 또는 표준화된 잔차는 이상치 식별에 유용할 수 있다. 가장 기본적인 방법은 Huber(1973)가 제안한 $\sum_{i=1}^{n}\rho(e_i/\sigma_i)$를 최소화하는 M-추정량을 들 수 있겠다. 여기서 $\rho(\cdot)$는 제곱보다 이상점에 덜 민감한 함수이며 대칭함수이다. σ의 추정량으로는 보통 mad(median absolute deviation)인 $\hat{\sigma}=1.483\,median\,|e_i-median(e_i)|$로 계산한다. M-추정량은 반응변수의 조건부 평균과 멀리 떨어져 있는 수직 이상점에 대하여 매우 강건하나, 지렛점에 대해서는 강건하지 않은 것으로 알려져 있다.

이상치 식별에 보다 직접적인 강건한 회귀추정방법으로는 Rousseeuw(1984)가 제안한 LMS(least median of squares)추정량과 LTS(least trimmed squares)추정량을 대표적으로 들 수 있다. LMS추정량은 잔차의 제곱합 대신에 잔차 제곱의 중앙값을 최소화시키는 해로 정의된다. 이 추정량은 지렛점뿐만 아니라 수직 이상점에 대해서도 강건하나 점근효율성(asymptotic efficiency)면에서는 다소 낮다는 것이 단점으로 지적되었다. 이점을 감안하여 LTS추정량은 순서화된 잔차 제곱의 합을 최소화한다. LMS에 비해 비교적 계산이 간단한 장점을 갖고 있다.

이외에도 특수한 분포를 가지는 자료에 대한 강건한 추정량으로는 독립변수의 경계에 오염된 관찰개체가 존재하는 경우에 선형회귀추정의 효율성을 개선한 대안적 추정량(alternative estimator; Park, 2009), 꼬리가 두터운 분포에 대한 강건한 회귀추정량(Jung, 2011)을 들 수 있다. 한편 Son (2011)은 강건한 퍼지 선형회귀추정에서의 이상치 검출방법을 제안하였다.

2.2. MVE추정량에 의한 이상점 식별방법

p차원의 확률변수 벡터가 다변량 정규분포를 따른다고 할 때, n개의 관측치들 간의 거리구조를 통해 이상치를 판별할 수 있겠다. 표본평균 $\bar{\boldsymbol{x}}=\sum_{i=1}^{n}\boldsymbol{x}_i/n$과 표본공분산행렬 $S=\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^T/(n-1)$로 놓으면 각 관측치($i=1,\cdots,n$)마다 다음으로 정의되는 마할라노비스 거리(Mahalanobis distance) $MD_i=\sqrt{(\boldsymbol{x}_i-\bar{\boldsymbol{x}})S^{-1}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^T}, i=1,\cdots,n$를 계산할 수 있다. MD_i는 표본평균과 표본공분산을 활용하기 때문에 강건한 거리가 아님을 쉽게 알 수 있다. 따라서 이와 같은 단점을 극복하기 위해 Rousseeuw, Zomeren(1990)은 위치모수(location parameter)와 공분산을 강건한 MVE(minimum volume ellipsoid estimation)추정량으로 대체하여 구한 거리 RD_i(robust distances)를 사용하는 방법을 제안하였다. 그러한 추정치들로 각각 표본평균과 표본공분산행렬을 대체하면, 마할라노비스 거리 MD_i와 동일한 형식으로 표현할 수 있다. 통상적으로 $RD_i>\sqrt{\chi^2_{1-\alpha}(p-1)}$이면 지렛점으로 식별할 수 있는데, $RD_i<\sqrt{\chi^2_{1-\alpha}(p-1)}$이고, $|e_i/\hat{\sigma}|<2.5$($\hat{\sigma}$는 추정된 표준편차)이면 해당되는 관찰치들을 정상점으로 판정하지만, $RD_i<\sqrt{\chi^2_{1-\alpha}(p-1)}$이고 $|e_i/\hat{\sigma}|>2.5$이면 수직 이상점으로 식별하고, $RD_i>\sqrt{\chi^2_{1-\alpha}(p-1)}$이고 $|e_i/\hat{\sigma}|>2.5$이면 나쁜 지렛점(bad high leverage point)으로 식별한다. 한편, $RD_i>\sqrt{\chi^2_{1-\alpha}(p-1)}$이고 $|e_i/\hat{\sigma}|<2.5$이면 좋은 지렛점(good high leverage point)으로 식별한다. 만약 2차원으로만 국한한다면 pC_2개의 조합의 개수만큼 이러한 그래프들을 추적 조사해야 하므로 시각적으로 식별하는 것에 한계를 가진다.

아울러, 2.1절에서 소개한 방법들을 통해서 얻은 표준화잔차와 2.2절의 강건한 거리를 동시에 적용하여 수직 이상점과 지렛점을 식별할 수도 있다. 즉, RD_i를 횡축으로, 표준화잔차를 종축으로

한 산점도를 그려서 확인하는 방법이다. 이와 같이 시각적으로 탐색하기 위한 여러 가지 방법들이 개발되어 있는데, 대표적인 방법으로는 추가변수그림(added variable plot), 편잔차 그림(partial residual plot), 편잔차 그림을 확장시킨 덧편차 그림(augmented partial residual plot)을 활용하여 이상치를 탐색할 수도 있다(Seo, Yun, 2010).

2.3. She, Owen이 제안한 이상점 식별방법

2.1절이나 2.2절의 방법들은 식 (1)의 모형에 기초하여 강건한 추정을 한다고 볼 수 있다. 반면에 She, Owen(2011)은 관측치 개별로 모수적 절편을 가지는 선형모형을 고려하고 이를 축소추정함으로써 보다 직접적으로 이상치를 식별하는 방법을 제안하였다. 이를 간단히 소개하자. 식 (1)의 모형에서 추가적으로 관측치마다 절편을 개별로 가지는 다음의 모형을 고려해 보자.

$$y_i = \beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta} + \gamma_i + \epsilon_i,\ i = 1,\cdots,n, \tag{2}$$

여기서 γ_i인 모수 $\boldsymbol{\gamma} = (\gamma_1,\cdots,\gamma_n)^T \in R^n$이다. 식 (2)의 모형에서 모든 관측치에 대해서 $\gamma_i = 0$이면 기존의 식 (1)의 모형과 동일함을 알 수 있다. 만약 특정 i에 대한 γ_i값이 0이 아니라면 이는 $\boldsymbol{x}_i$에서의 조건부 평균값과 γ_i만큼 차이가 난다는 것을 의미하고, 이는 회귀분석에서 이상치의 일반적 정의와 동일하게 된다. 따라서 식 (2)의 모형은 관측치마다의 이상치 여부를 탐색하고 또한 추정할 수 있는 모형이다. 또한 복수개의 γ_i값이 0이 아니라면 이는 여러 개의 이상치를 동시에 검출할 수 있다는 것을 의미하기도 한다. 이상치 검출 및 효과추정 면에서 수월성을 추구하는 반면에 추정될 모수의 개수는 $n+p$로 자료의 수 n보다 항상 크므로 축소추정치 방법을 활용할 필요가 있다. 특히 식 (2)의 모형에서 최소제곱추정치를 활용하면 $\hat{\gamma}_i$는 잔차 $y_i - \hat{\beta}_0^{LS} - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}^{LS}$로 추정됨을 알 수 있다. 이러한 단순 추정치를 피하기 위해 희소성(sparsity)를 가진 γ_i를 축소추정할 필요성이 있다. 가장 간단한 방법은 γ_i들에 대해서 l_1노음 $\sum_{i=1}^{n}|\gamma_i|$의 크기를 제한하여 추정하는 방법을 고려할 수 있다. l_1노음을 통한 축소추정치는 큰 크기의 참모수일수록 추정량의 편의성이 크게 되는 문제점이 있는 것으로 알려져 있다. 이를 개선하는 방법으로는 l_1노음 대신 SCAD(smoothly clipped absolute deviation; Fan, Li, 2001)나 MCP(minimax concave penalty; Zhang, 2010)와 같은 비볼록 벌점함수를 고려하면 개선되는 것으로 알려져 있다. 이러한 장점은 소위 신탁성질(oracle property)로 불리는 성질에서 기인된다고 설명된다. 여기서 어떤 추정량이 신탁성질을 가지고 있다고 함은 참값이 0인 γ_i를 미리 알고 있어서 나머지 0이 아닌 γ_i들로 구한 이상적인 추정량과 점근적으로 일치함을 의미한다.

She, Owen(2011)은 $\boldsymbol{\gamma}$에 대한 비볼록(nonconvex) 노음 벌점함수를 통해 축소추정하는 Θ-IPOD (iterative procedure for outlier detection)을 제안하였다. 논의의 편의상 $\beta_0 = 0$으로 두면, Θ-IPOD추정량은 다음으로 정의되는 목적 함수를 최소화하는 해 $(\boldsymbol{\beta},\boldsymbol{\gamma})$로 정의된다.

$$L(\boldsymbol{\beta},\boldsymbol{\gamma}) = \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta} - \gamma_i)^2 + \sum_{i=1}^{n}P(\gamma_i;\lambda), \tag{3}$$

여기서 $P(\gamma_i;\lambda)$는 γ_i와 조율모수(tuning parameter) λ에 특징지어지는 비볼록 벌점함수이다. 예를 들어 가장 비볼록성이 가장 큰 벌점함수는 다음의 식으로 정의되는 강분계작용소(hard thresholding

operator) $\theta_{hard}(z;\lambda) = 0I(|z| \le \lambda) + zI(|z| \le \lambda)$를 들 수 있다.

Table 1. The Stackloss data: x_1(air flow), x_2(water temperature), x_3(acid concentrations), y(stack loss)

id	x_1	x_2	x_3	y	id	x_1	x_2	x_3	y
1	80	27	89	42	11	58	18	89	14
2	80	27	88	37	12	58	17	88	13
3	75	25	90	37	13	58	18	82	11
4	62	24	87	28	14	58	19	93	12
5	62	22	87	18	15	50	18	89	8
6	62	23	87	18	16	50	18	86	7
7	62	24	93	19	17	50	19	72	8
8	62	24	93	20	18	50	19	79	8
9	58	23	87	15	19	50	20	80	9
10	58	18	80	14	20	56	20	82	15
					21	70	20	91	15

강분계작용소는 γ_i의 절대값이 λ_i보다 작은 범위에서는 $\gamma_i = 0$으로 되게 하며, 그 이상의 범위에서는 γ_i값을 축소하지 않는 특징을 가지고 있다. 비볼록 벌점함수를 가진 식 (3)의 목적식은 모수 $(\boldsymbol{\beta},\boldsymbol{\gamma})$에 대해서 비볼록함수이므로 식 (3)의 최소해가 전역최소해(global minimum)임을 보장하지 못한다. 따라서 Θ-IPOD는 대안적으로 순차적으로 $\boldsymbol{\beta}^{(t)}$와 $\boldsymbol{\gamma}^{(t)}$를 추정한다. 여기서 t는 알고리즘의 반복과정에서의 횟수를 나타낸다. Θ-IPOD 알고리즘은

$$L(\boldsymbol{\beta}^{(t)},\boldsymbol{\gamma}^{(t)}) \ge L(\boldsymbol{\beta}^{(t)},\boldsymbol{\gamma}^{(t+1)}) \ge L(\boldsymbol{\beta}^{(t+1)},\boldsymbol{\gamma}^{(t+1)})$$

을 항상 만족한다. 즉, 추정치 쌍 $(\boldsymbol{\beta}^{(t-1)},\boldsymbol{\gamma}^{(t)})$은 목적함수 $L(\boldsymbol{\beta},\boldsymbol{\gamma})$를 단조 감소시키고, Θ-IPOD 알고리즘을 통해 수렴된 해는 국소해(local minimum)임을 의미한다.

3. 이상치 식별방법의 비교

이 절에서는 지금까지 정리한 이상치 식별방법들을 실제자료 Stackloss 자료(Brownlee, 1965)에 적용했을 때 이들에 따른 결과를 살펴보도록 한다.

Table 1의 Stackloss 자료는 세 개의 독립변수 x_1(공기흐름), x_2(냉각수 주입 온도), x_3(산성 농도)가 반응변수 y(잃어버린 암모니아 천분률)에 미치는 영향을 살펴보고자 조사된 자료이다. 먼저 y의 상자그림을 통해 1, 2, 3번 개체가 이상치로 식별되었다. Table 2와 3은 각각 전체자료와 네 개의 개체들(21, 4, 3, 1)를 제거하고 구한 LS추정량에 대한 영향력 측도들이다. 21번 개체는 DFFITS 통계량(D_i), Cook의 거리 통계량(C_i), Hadi 통계량(H_i) 모두 가장 큰 값을 가지고 또한 각각의 기준치를 상회하여 영향을 크게 주는 측정값으로 판단되었다. 또한 이 개체의 R-student(R_i)의 절대값이 2보다 커서 이상치로 판단이 된다.

Table 4는 추정된 회귀계수와 이의 t값과 p-value를 나타낸다. 공기흐름변수가 가장 유의한 변수임을 알 수 있다. 21번 개체는 공기흐름변수 값이 70임에도 불구하고 잃어버린 암모니아 천분률 값이 15로 다른 값들과 비교해 볼 때 낮은 수치를 가지고 있다. 이러한 영향력 측도들을 기준으로 이상치로 의심 가는 관측치를 하나씩 제거하면 21, 4, 3, 1 순서로 제거되었다. 4번 개체는 공기흐름 변수 값이 62이고, 냉각수 입구 온도 24로 잃어버린 암모니아 천분률이 28이다. 이는 다른 관측개체 보다 큰 값을 가지므로 이상치로 의심된다. 그리고 3번, 1번 개체는 매우 높은 지렛점을 가지

고 있었으며, 또한 잃어버린 암모니아 천분률도 매우 높은 값을 가지고 있다.

Table 2. The influence measures

id	C_i	D_i	H_i	R_i	L_i
1	0.15	0.79	0.61	1.21	0.30
2	0.06	-0.48	0.53	-0.71	0.32
3	0.13	0.74	0.53	1.62	0.17
4	0.13	0.79	0.66	2.05	0.13
5	0.00	-0.12	0.09	-0.53	0.05
6	0.02	-0.28	0.20	-0.96	0.08
7	0.05	-0.44	0.37	-0.83	0.22
8	0.02	-0.25	0.31	-0.47	0.22
9	0.04	-0.42	0.30	-1.05	0.14
10	0.01	0.21	0.27	0.43	0.20
11	0.04	0.38	0.28	0.88	0.16
12	0.07	0.51	0.39	0.97	0.22
13	0.01	-0.20	0.21	-0.47	0.16
14	0.00	-0.01	0.26	-0.02	0.21
15	0.04	0.39	0.31	0.80	0.19
16	0.00	0.11	0.16	0.29	0.13
17	0.07	-0.50	0.75	-0.60	0.41
18	0.00	-0.07	0.19	-0.15	0.16
19	0.00	-0.09	0.22	-0.20	0.17
20	0.00	0.13	0.11	0.44	0.08
21	0.69	-2.10	1.56	-3.33	0.28

Table 3. The influence measures(removed 21,4,3,1)

id	C_i	D_i	H_i	R_i	L_i
1					
2	1.61	2.75	2.76	1.81	0.70
3					
4					
5	0.01	-0.16	0.14	-0.54	0.08
6	0.03	-0.38	0.31	-1.06	0.11
7	0.01	-0.22	0.36	-0.38	0.25
8	0.02	0.30	0.38	0.52	0.25
9	0.04	-0.42	0.35	-0.92	0.17
10	0.01	0.18	0.33	0.32	0.24
11	0.05	0.42	0.37	0.85	0.20
12	0.02	0.28	0.44	0.43	0.29
13	0.30	-1.34	1.35	-2.72	0.19
14	0.11	-0.69	0.56	-1.25	0.23
15	0.10	0.66	0.53	1.24	0.22
16	0.00	0.05	0.18	0.12	0.15
17	0.03	-0.33	0.78	-0.38	0.43
18	0.00	0.04	0.22	0.08	0.18
19	0.02	0.27	0.31	0.51	0.21
20	0.06	0.53	0.58	1.73	0.08
21					

Table 4. Estimated coefficients, t value, p-value and R^2

removed observations	$\hat{\beta}_1$	t	p-value	$\hat{\beta}_2$	t	p-value	$\hat{\beta}_3$	t	p-value	R^2
ϕ	0.716	5.312	0.000	1.295	3.521	0.003	-0.152	-0.971	0.344	0.914
21	0.889	7.483	0.000	0.817	2.514	0.023	-0.107	-0.861	0.402	0.949
21,4	0.957	10.133	0.000	0.556	2.104	0.053	-0.109	-1.124	0.279	0.969
21,4,3	0.904	10.442	0.000	0.586	2.513	0.025	-0.107	-1.254	0.232	0.972
21,4,3,1	0.798	11.832	0.000	0.577	3.484	0.004	-0.067	-1.094	0.296	0.975
21,4,3,1,2	0.686	7.834	0.000	0.567	3.701	0.003	-0.017	-0.272	0.789	0.942

결과적으로 네 개의 이상치(21, 4, 3, 1)를 제거한 모형이 R^2값이 가장 크고, 2번 개체를 추가적으로 제거한 모형은 R^2값이 네 개의 이상치를 제거한 모형 보다 감소하였다. Table 5에서 LMS의 $\hat{\sigma}=1.208$이며, LTS에서는 $\hat{\sigma}=1.036$으로 추정되었다. 임계치 ±2.5를 고려할 때, LMS에서는 1, 3, 4, 21번 개체가 LTS에서는 1, 2, 3, 4, 21이 임계치를 상회하여 이상치로 검출되었다. MVE추정방법에 따른 RD 기준으로는 $\sqrt{\chi^2_{0.975}(3)}=3.057$보다 큰 값을 가지는 1, 2, 3, 21번 개체가 지렛점이 높은 이상치로 검출되었다.

추가적으로 LTS 또는 LMS의 표준화잔차와 RD의 2차원 그래프를 탐색한 결과, LMS표준화잔차 대 RD의 그래프로부터 1, 3, 21은 나쁜 지렛점을, 2번 개체는 좋은 지렛점을 갖고 있음을 알 수 있었다. 그러나 2번 개체는 임계치 2.5의 경계선상에 놓여 있어 이를 이상치로 명확히 판정하기에는 힘든 면이 있다. LTS표준화잔차 대 RD 그래프를 통해서는 5개의 이상치로 검출된 것으로 판단

되는데, 4번 개체는 이상치이고, 이 중 1, 2, 3, 21번 개체들은 나쁜 지렛점을 가지고 있다.

Table 5. Estimated residuals via LMS, LTS, Mahalanobis and robust distance

id	LMS	LTS	Mahalanobis distance	Robust distance
	$e/\hat{\sigma}$	$e/\hat{\sigma}$		
1	6.42	8.19	2.254	5.528
2	2.28	3.38	2.325	5.637
3	6.21	7.69	1.594	4.197
4	7.25	8.71	1.272	1.589
5	-0.21	-0.19	0.303	1.189
6	-0.62	-0.57	0.773	1.308
7	-0.21	-0.04	1.853	1.716
8	0.62	0.92	1.853	1.716
9	-0.62	-0.60	1.361	1.227
10	0.62	0.40	1.746	1.936
11	0.62	0.30	1.466	1.494
12	0.21	-0.28	1.842	1.913
13	-1.86	-2.52	1.483	1.660
14	-1.45	-2.05	1.779	1.689
15	0.62	0.23	1.690	2.230
16	-0.21	-0.70	1.292	1.768
17	0.21	0.04	2.700	2.431
18	0.21	-0.04	1.503	1.523
19	0.62	0.54	1.593	1.710
20	1.86	2.02	0.807	0.675
21	-6.83	-8.09	2.177	3.657

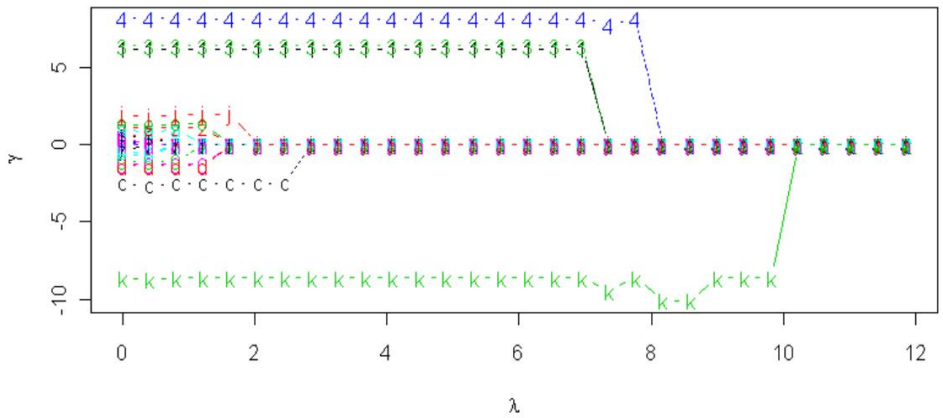

Figure 1. Solution path of γ_i via Θ-IPOD in Stackloss data

Figure 1은 Θ-IPOD 방법의 조율모수 λ에 따른 추정된 관측치마다의 γ_i값들의 변화를 나타낸 그래프이다. 큰 값의 λ에서는 모든 관측치들의 γ값은 0이다. λ가 감소하면서 관측치들의 모수적 절편효과 γ_i들의 패턴을 볼 수 있다. 첫 번째로 0이 아닌 값을 가지는 관측치는 21(k)번으로써, 가장 먼저 탐색이 되고, λ가 감소하면서 순차적으로 4번 개체, 그리고 1, 3번 개체가 0이 아닌 γ_i값을 가지는 관측치로 탐색이 된다. 이 관측치들은 영향력, LMS, LTS, MVE추정방법 등에서도 명백한

이상치로 판별되었다.

4. 결론 및 토의

본 연구에서는 표준화잔차, 영향력의 측도들과 강건한 회귀추정방법인 LMS추정량, LTS추정량, MVE 추정량 그리고 최근에 제안된 She, Owen(2011)의 Θ-IPOD 방법을 Stackloss 자료에 적용하여 이상치 식별 분석을 하였다.

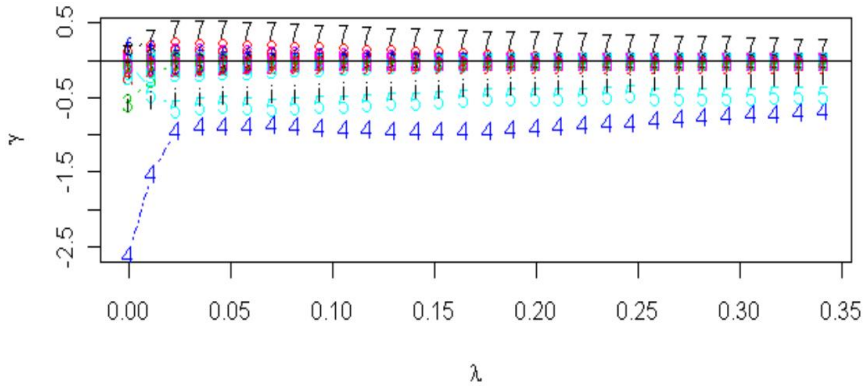

Figure 2. Solution path of γ_i via Θ-IPOD in Newyork River data

영향력 측도들을 이용한 검출 방법은 주로 한 개씩의 관찰치를 제외시킨 상태에서 그 관찰치의 영향정도를 파악하기 때문에 leave-one-out 방식으로 이상치를 단계적으로 처리한다. 이때 여러 영향력 측도들과 함께 표준화잔차를 판단하여 분석함으로써 비교적 이해하기 쉽고 적용하기 편리한 장점들을 가지고 있으나, 다수의 이상점이 존재하는 경우에는 많은 시간이 소비되며 이상치의 판단도 어렵기 때문에 효율성이 크게 떨어지는 단점이 있다.

한편, 다수의 이상점이 존재하는 경우에 적용할 수 있는 방법 중 대표적인 LTS, LMS추정량 및 MVE추정량에 바탕을 둔 이상점 식별방법을 비교하였는데, 영향력 측도들을 이용한 검출 방법과 거의 동일한 이상점을 식별하였다. 그런데, LTS, LMS추정량과 MVE추정량 개별로 이상점을 식별할 경우 다소 많은 관측치들을 이상치로 판정하였는데, 이는 이러한 추정량이 매우 강건하기 때문에 기인하는 것으로써, 실제 자료 분석에서는 임계값을 다소 크게 정해야 하는 임의성을 내포하고 있다. 반면에 LMS 또는 LTS의 표준화잔차 및 RD를 동시에 고려했을 때, 다수의 이상치가 존재할 때 구분이 명확하였으며, RD 거리 축을 통하여 지렛점이 높은 개체도 쉽게 구분이 가능한 장점이 있음을 확인하였다.

She, Owen(2011)이 제안한 Θ-IPOD의 그래프는 관측치들의 영향력을 조율모수 λ의 크기별로 살펴 볼 수 있기 때문에 추가적으로 관측치들 간의 관련성을 확인할 수 있겠다. 이는 가면효과와 수렁효과의 발생여부의 탐색을 직접적으로 확인하는 것이 가능함을 의미한다. 본 연구에서 분석한 Stackloss 자료의 결과에서는, Θ-IPOD 방법과 강건한 회귀추정방법 및 MVE추정방법에 근거한 방법의 결론이 비슷하게 도출되었다. 그러나 지면의 관계상 Stackloss 자료와 같이 여러 지수들을 제시하지는 않으나 Newyork River 자료(Brownlee, 1965)에 Θ-IPOD 방법을 적용했을 때, λ에 따른 추정된 γ_i값이 서로 교차함을 확인하였다. Figure 2에 제시된 것과 같이 λ의 값이 대략 0.05와 0.35사이에서는 4, 5, 7, 19(i)가 이상치임을 알 수 있다. 특히 λ값이 0일 때는 γ_i는 식 (1)의 모형에 대한 LS의 잔차로 볼 수 있는데, λ가 0에 가까워질수록 5번과 7번은 0에 가까워지는데 반해 4번 개체

의 값이 매우 커지는 것을 확인할 수 있다. 따라서 4, 5, 7번 관측치들 간의 상호작용에 의해서 5, 7번은 가면효과를 3번은 수렁효과가 발생하는 것을 알 수 있다. 이러한 구조는 영향력 측도들이나 LMS나 LTS의 표준화잔차와 RD 그래프상에서는 파악하기 힘들다. 따라서 가면효과나 수렁효과가 있는 경우에 기존의 방법들보다 Θ-IPOD 방법이 비교우위에 있다고 할 수 있다. 그러나 본 논문에서는 소 표본의 예제 자료를 이용하여 이상치 탐색 방법을 비교하였기 때문에 연구 결과를 일반화하기에는 다소 제한이 따른다. 보다 큰 표본이나 차원이 큰 자료에 대해서, 본 논문에서 고려한 여러 이상치 탐색방법들의 장단점을 비교하는 연구는 차후 과제로 남기고자 한다.

References

Brownlee, K. A. (1965). *Statistical theory and methodology in science and engineering,* 2nd edition, New York, John Wiley & Sons.

Belsley, D. A., Kuh, E., Welsch, R. E. (1980). *Regression diagnostics,* New York, John Wiley & Sons.

Cook, R. D. (1977). Detection of influential observations in linear regression, *Technometrics*, 19(1), 15-18.

Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association,* 96(456), 1348-1360.

Hadi, A., Simonoff, J. S. (1993). Procedures for the identifying of multiple outliers in linear models, *Journal of the American Statistical Association,* 88(424), 1264-1272.

Hoaglin, D. C., Welsch, R. E. (1978). The hat matrix in regression and anova, *The American Statistician*, 32(1), 17-22.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and monte carlo, *The Annals of Statistics,* 1(5), 799-821.

Jung, K. (2011). A Combined Robust Estimator Between the Least Squares Estimator and a t-type Regression Estimator, *Journal of the Korean Data Analysis Society,* 13(5A), 2235-2242.

Rousseeuw, P. J. (1984). Least median of squares regression, *Journal of the American Statistical Association,* 79(388), 871-880.

She, Y., Owen, A. B. (2011). Outlier detection using nonconvex penalized regression, *Journal of the American Statistical Association,* 106(494), 626-639.

Seo, H. S., Yun, M. (2010). Outlier detection methods using augmented partial residual plots in a partially linear model, *Journal of the Korean Data Analysis Society,* 12(2B), 1125-1134. (in Korean).

Son, B. Y. (2007). Robust Fuzzy Least-square Regression, *Journal of the Korean Data Analysis Society,* 9(1), 395-406. (in Korean).

Son, B. Y. (2011). Detection of outliers and influential observations in fuzzy linear regression, *Journal of the Korean Data Analysis Society*, 13(1B), 353-364. (in Korean).

Park, M. S. (2009). An alternative robust estimator in linear regression models, *Journal of the Korean Data Analysis Society,* 11(4B), 2265-2276. (in Korean).

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty, *The Annals of Statistics,* 38(2), 894-942.

Comparison Study of Outlier Detection Methods in a Regression Model*

Dae-In Park[1], *Hyuncheol Kang*[2], *Sang-Tae Han*[3], *Hosik Choi*[4]

Abstract

This paper is concerned with the outlier detection methods in regression model. Various influence measures for detecting outliers are illustrated and compared via real data sets. Including single outlier detection method and three multiple outliers detection methods are considered : procedure based on the least median of squares estimation, the robust distance with the minimum volume ellipsoid estimator, and She, Owen (2011) procedure. Comparison studies are conducted using two data sets which are known to contain multiple outliers. It appears, in general, that all of these procedures are effective in identifying the outliers. However, procedures based on the least median of squares estimation and the robust distance with the MVE estimator are difficult to identification of outliers when masking or swamping effects exist in the data. On the other hand, the procedure proposed by She, Owen (2011) is effective to detect such outliers.

Keywords : outlier, robustness, linear regression, shrinkage estimation, nonconvex penalization.

*This research was supported by the Academic Research fund of Hoseo University in 2011″(2011-0016).
[1]Graduate student, Department of Informational Statistics, Hoseo University, Asan 336-795, Korea.
E-mail : pcapcoms@naver.com
[2]Associate Professor, Department of Informational Statistics, Hoseo University, Asan 336-795, Korea.
E-mail : hychkang@hoseo.edu
[3]Professor, Department of Informational Statistics, Hoseo University, Asan 336-795, Korea.
E-mail : sthan@hoseo.edu
[4](Corresponding Author) Assistant Professor, Department of Informational Statistics, Hoseo University, Asan 336-795, Korea. E-mail : hosikchoi@hoseo.edu
[Received 18 January 2013; Revised 16 February 2013; Accepted 19 February 2013]