제주교육대학교 논문집(1996)
제25집 PP.193~204

Outward Testing을 위한 Clean Subset 결정 방법

김 종 우*

The Procedure of decision Clean Subset for the Outward Testing

Jong-Woo Kim
Dept. of Mathematics Education, Cheju National University of Education

Abstract

This article is concerned with clean subset procedures for outward testing in linear regression. The outward-testing procedure, which is controled by the initial subset and the minimum residuals, is suggested by two phases. The performance of this procedure is affected by the elemental subset, i.e. clean subset. The procedure, however, is fluctuated in detecting y outliers that are on high-leverage cases. Thus, we proposed ELMS algorithm for the clean subset and it is compared by the EDR-ESD procedure based on a suggestion of Davies and Gather.

1. 서 론

선형회귀분석(linear regression analysis)에서 인식하는 방법이다. 이상치와 영향력 관찰점을 인식하려는 회귀 진단(regression diagnosis)분야는 접근 방법에 따라 크게 두 가지 분야로 나눌 수 있다: 즉, 직접접근방법과 로버스트(robust) 적합을 이용한

* 제주교육대학교 수학교육과 부교수

간접접근방법이 있다.

직접접근방법은 이상치가 없으리라 여겨지는 자료들의 집합(clean subset)을 사용하여 잠재적인 이상치들의 집합(remained subset)을 인식해 가는 방법이다. 즉, clean subset에 포함된 관찰점의 색인(index)으로 구성된 집합을 M이라면, 자료에서 이 색인들에 해당하는 설명변수와 반응변수만을 추출하여 각각 Y_M, X_M을 구성하고, 이에 대응하는 β_M과 σ_M^2의 최소제곱 추정량은

$$\hat{\beta}_M = (X_M^T X_M)^{-1} X_M^T Y_M,$$
$$\hat{\sigma}_M^{\,2} = e_M^T e_M/(k-p), \quad \text{여기서} \quad e_M = Y - X\hat{\beta}_M \tag{1}$$

이다. 색인집합 M의 크기 k를 결정하는 한 가지 방법으로 Gentleman과 Wilk(1975)은 크기 k인 부분집합을 제거할 때에 잔차의 합이 가장 크게 축소하는 값으로 할 것을 제안했다. 즉, 색인집합 $M = \{i_1, \ldots, i_k\}$에 대응하는 관찰점들의 제거에 따르는 잔차제곱합은 다음과 같다.

$$\begin{aligned} Q_k &= SSE - SSE_{(M)} \\ &= e_M^T (I_M - H_M)^{-1} e_M, \end{aligned}$$

여기서 $SSE_{(M)}$는 색인집합 M에 대응하는 관찰점이 제거된 잔차제곱합이고,

e_M는 색인집합 M에 대응하는 자료의 잔차집합이며,

H_M는 H에서 색인집합 M에 대응되는 관찰점들로 구성된 부분행렬이다.

이것은 곧 clean subset을 최소잔차합을 갖는 크기 $(n-k)$인 부분집합을 사용하여 구하는 것과 같다. Prescott(1975)와 Rosner(1975,1983)는 "단일 관찰점 진단방법

(single case diagnostics)"을 사용하여 절대잔차가 가장 큰 관찰점부터 작은 관찰점순으로 이상치인지를 인식하고 검사하는 Inward testing을 제안하였고, Marasinghe(1985)는 "단일 관찰점 진단방법"의 연속적 이용에 의한 잠재적인 K개 이상치 집단을 선정한 후에 이들 중에서 이상치 존재를 확인하는 "다단계 접근법(multistage approach)"을 제안했다. Paul과 Fung(1991)은 잠재적인 K개 이상치 집단에 대한 진단으로 Cook의 거리(cook's distance)에 의한 이상치들과 GESR(generalized extreme studentized residuals)에 의한 이상치 집단을 잠재적인 이상치로 하여 GESR을 사용할 것을 제안했다. Kianifard와 Swallow(1989), Hawkins(1991)은 "단일 관찰점 진단방법"에 의해 자료를 정렬한 후에 최소 진단치를 갖는 K개 관찰점을 사용하여 이상치를 식별하는 순환잔차(recursive residual) 방법을 제안했다. Fung(1993)은 Rousseeuw(1984)의 LMS에 의하여 remained subset을 결정하고 수정된 Cook의 거리를 사용하여 이상치 여부를 확인하는 과정을 제안했다. 그러나 이들 진단통계량의 이상치 식별은 다중이상치가 존재할 때 발생하는 은폐효과와 수렁효과에 의해 크게 영향을 받거나 또는 자료상의 다공선성에 민감한 것으로 알려져 있으며, 잠재적인 이상치 집단의 크기 K에 대한 사전 지식을 요구하는 어려운 점이 있다[[12], [13], [21]]. Hadi와 Simonoff(1993)는 Hadi(1992)의 이상치 식별 알고리즘을 사용하여 이상치가 없으리라 예상되는 부분집합과 잠재적으로 이상치가 존재할지도 모르는 부분집합으로 나누어 이상치가 없으리라 예상되는 부분집합을 사용하여 관찰점들을 진단하고 이상치가 없으리라 예상되는 부분집합의 크기를 늘려 나가는 outward testing을 제시했다. Davies와 Gather(1993)은 최대 절대 잔차를 갖는 관찰점을 하나씩 배제시켜 가면서 clean subset의 원소 수가 $[(n+p-1)/2]$으로 구성시킨 후에 outward testing을 하였다. 이러한 outward testing은 붕괴점이 50%에 달하는 것으로 알려져 있다[Barnett와 Lewis,1984; Davies와 Gather,1993].

이상치 인식을 위한 간접방법으로 제시된 로버스트 추정량으로는 $(p+1)$개의 기저집합(elemental set)을 사용하여 잔차제곱의 중위수를 최소화하는 추정량을 선택하는 Rousseeuw(1984)의 LMS(least median of squares) 추정량과 Rousseeuw와 van Zomeren(1990)의 MD(mahalanobis 거리)의 $C(X)$와 $S(X)$에 로버스트 추정량을 사용하는 MVE(minimum volume ellipsoid)가 있다.

그러나 LMS는 이상치를 과도하게 지정하고 있으며[Atkinson,1986; Fung,1993],

MVE는 선정된 기저집합이 최소 볼륨을 갖는 타원을 형성하는 지에 의심을 갖게 한다[Cook와 Hawkins,1990].

본 연구에서는 선형회귀 구조를 갖는 모집단에서 outward testing을 위하여 이상치가 없으리라 예상되는 부분집합(clean subset)을 구성하기 위한 방법으로 ELMS와 EDR-ESD의 방법을 비교한다.

선형회귀모형을 다음과 같이 설정하자.

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

여기서 $Y = (y_1, y_2, \ldots, y_n)^T$는 반응변수인 $n \times 1$ 벡터이다.

$X = (x_1, x_2, \ldots, x_n)^T$는 $1 \times p$ 벡터인 $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$를 행으로 갖는 설명변수인 $n \times p$ 행렬이다 (단, $p < n$).

$\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$인 모수 $p \times 1$ 벡터이다.

$\varepsilon \sim N(o, \sigma^2 I_n)$인 $n \times 1$ 벡터이다(즉, $E(\varepsilon \mid X) = 0$ 이고 $\mathrm{var}(\varepsilon \mid X) = \sigma^2 I_n$, I_n은 계수 n인 단위 행렬).

이때, 최소제곱법(method of least squares: LS)에 의한 β와 σ^2의 최소제곱 추정량은

$$\hat{\beta} = (X^T X)^{-1} X^T Y,$$

$$\hat{\sigma}^2 = e^T e/(n-p), \quad \text{여기서} \quad e = Y - X\hat{\beta}$$

이다. 본 논문에서 사용할 잔차의 변형은

$$e_i / \hat{\sigma}\sqrt{1-h_{ii}}, \quad \text{여기서} \quad h_{ii} = x_i (X^T X)^{-1} x_i^T$$

으로, 여기서 h_{ii}는 Hat 행렬 H의 대각선상의 원소이다.

2. 제안된 방법

앞절에서 언급한 바와 같이 이상치를 찾는 직접접근 방법은 자료를 이상치가 없으리라 예상되는 부분집합(clean subset)과 자료 전체에서 clean subset을 제외하고 남아 있는 나머지 부분집합(remained subset)으로 분리한 다음에 clean subset을 사용하여 이상치를 찾는 것이다. 본 연구에서는 크기 $[(n+p-1)/2]$인 clean subset을 구하기 위하여 다음과 같은 두 가지 방법을 제시하고 이를 비교한다.

2.1 ELMS

Rousseeuw(1984)가 제안한 LMS는 $\hat{\beta}_M$를 결정하기 위해 행렬 X에서 크기 $(p+1)$인 부분집합을 무작위로 추출하여 다음과 같은 최소중위수잔차를 갖는 부분집합을 기저집합 J를 선택하는 것이다.

$$\underset{\hat{\beta}_J}{\text{Minimize}} \ \underset{J}{\text{med}} \ r_i^2 ,$$

여기서 $\hat{\beta}_J = (X_J^T X_J)^{-1} X_J^T Y_J$,

$$r_i = y_i - x_i \hat{\beta}_J .$$

이 추정량은 50%에 달하는 붕괴점을 갖고 있으나 오차의 정규성 가정 아래서 매우 낮은 효율성을 갖고 있다. 효율성을 높이기 위한 방법으로 Atkinson(1994)가 제시하고 있는 것처럼 적절하게 선택된 기저집합 M의 크기를 확장시키는 것은 보다 빠르게 최소잔차집합에 도달할 수 있으며, Hadi와 Simonoff(1993)가 지적하고 있는 X변량 사이에 높은 상관성을 갖는 자료에서도 안정성을 갖게 한다. 따라서 본

연구에서는 LMS의 기법을 확장한 방법으로 크기 $(p+1)$개인 초기 기저집합을 적절한 방법으로 설정하고, 이를 중심으로 $(p+1)$개의 최소잔차합을 갖는 기저집합을 재 설정한다. 이 기저집합을 사용한 선형회귀방정식에서 최소절대잔차를 갖는 관찰점을 선택하여 기저집합의 크기가 $(p+1)+1$개로 증가시키는 방법을 사용하여 $[(n+p-1)/2]$이 될 때까지 진행시킨다.

이 방법은 다음과 같은 3단계로 구성되어 있다.

ELMS 알고리즘

1 단계. 초기 색인집합 M의 설정.

전통적인 최소제곱법을 사용하여 절대잔차 $|r_i|$를 구하고 이를 오름차순으로 정렬하여 작은 순으로 크기 $k(=p+1)$인 초기 색인집합 M을 초기 기저집합으로 설정한다(단, p는 X의 계수).

$|r_{i_1}|_{1:n} \leq |r_{i_2}|_{2:n} \leq , \ldots , \leq |r_{i_n}|_{n:n}$일 때,

$M = \{ i_1, i_2, \ldots , i_k \}$ 이다.

2 단계. 최적 색인집합 M의 결정.

앞 단계에서 부분집합에 포함된 원들로 구성된 색인집합 M에서 $(k-1)$개의 원을 취하고, 나머지 부분집합에서 1개의 원을 취하여 새로운 색인집합 $M_{(i)j}$를 구성한다. 이 색인집합 $M_{(i)j}$은 총 $k\times(n-k)$개이며, 이들 중에서 최적 색인집합 M의 결정은 (1)에서 제시한 방법으로 $\hat{\beta}_{M_{(i)j}}$를 구하고, 이들 중에서 LMS에서 제시한 최소잔차중위수 $\min\ (\text{med}\ r_i^2)_{M_{(i)j}}$를 갖는 $M_{(i)j}$를 선택한다.

$M = \{1, 2, \ldots, k\}$ 와 $M^c = \{k+1, k+2, \ldots, k+n\}$ 에서

$M_{(i)j} = M_{(i)} \cup M_j^c$.

여기서 $M_{(i)} = M - \{i\}$, $i = 1, \ldots, k$,

$M_j^c = \{j\}$, $j = k+1, k+2, \ldots, n$.

$$\text{Minimize}_{\hat{\beta}} \ \text{med}_{M_{(i)j}} r_i^2 .$$

여기서 $\hat{\beta} = (X_{M_{(i)j}}{}^T X_{M_{(i)j}})^{-1} X_{M_{(i)j}}{}^T Y_{M_{(i)j}}$,

$r_i = y_i - x_i \hat{\beta}_M$.

3 단계. 색인집합 M의 크기 증가.

최적 색인집합 $M_{(i)j}$를 사용하여 구한 $\hat{\beta}_{M_{(i)j}}$에 의하여 결정된 색인집합의 원소 수가 $[(n+p-1)/2]$이 될 때까지 크기 k인 M에 $|r_i|_{(k+1):n}$에 해당하는 색인 i를 포함시켜 색인집합 $M_{(i)j}$의 크기를 하나 증가시키고, 2 단계를 반복 실행한다.

2.2 EDR-ESD

Davies와 Gather(1993)에서 제시된 EDR-ESD는 전체 자료에서 전통적인 최소제곱법을 사용하여 절대잔차를 구하고, 이들 중에서 최대 절대잔차를 갖는 관찰점을 배제하고 나머지들로 다시 위 방법을 사용하여 최대 절대잔차를 갖는 관찰점을 찾고, 또 이를 배제시키는 방법을 clean subset의 크기가 $[(n+p-1)/2]$에 도달할 때까

지 반복시킨다. 그리고 이렇게 구성한 clean subset을 사용하여 remained subset에서 이상치인 관찰점은 다시 clean subset에 포함시키고 그렇지 않은 관찰점은 remained subset에 남겨 두고 다시금 이 과정을 반복하는 방법이다. 이 방법은 Rosner(1975,1983) 등에서 제시된 "단일 관찰점 진단방법"을 확장한 개념이다.

여기서 ELMS와 clean subset에 대한 비교를 위하여 clean subset 구성에 관한 방법만을 고려하면, 다음과 같은 2 단계로 되어 있다.

EDR-ESD 알고리즘

1 단계. 색인집합 M의 구성.

크기 $k(=n)$인 초기 색인집합 M을 사용하여 전통적인 최소제곱법에서 절대잔차 $|r_i|$를 구하고 이를 오름차순으로 정렬한다.

$M = \{ i_1, i_2, \ldots, i_{k(=n)} \}$ 일 때,

$|r_{i_1}|_{1:n} \leq |r_{i_2}|_{2:n} \leq , \ldots , \leq |r_{i_n}|_{n:n}$이다.

2 단계. 최적 색인집합 M의 결정.

앞 단계에서 색인집합의 크기가 k일 때, $|r_i|_{k:n}$에 해당하는 색인 i를 제외시킨 색인집합 $M_{(i)}$를 새로운 색인집합으로 하여 색인집합의 원소 수가 $[(n+p-1)/2]$이 될 때까지 1, 2 단계를 반복 실행한다.

$M_{(i)} = M - \{ k \}$.

3. 모의실험

제안된 알고리즘의 안정성을 조사하기 위하여 단순 선형방정식일 때

$$y_i = \beta_o + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, 25.$$

여기서, $\beta_o = 0$, $\beta_1 = 1$,

$x_i \sim U(0, 15)$,

$\epsilon_i \sim N(0, 1)$

에서 정상적인 관찰점을 25개 구하고, 이상치 발생을 위하여

$$y_i = x_i + 4, \quad x_i = k - .05(i-1),$$

$$y_i = x_i - 4, \quad x_i = k - .05(i-1),$$

여기서 $k = -5, 7.5, 20$,

$i = 1, 2, 3, 4, 5$

를 사용하여 이상치 개수가 5개일 때 대하여 1000회의 모의실험을 하여 이상치의 인식율을 파악한다.

<표> ELMS와 EDR-ESD의 비교

계획된 이상치 위치(x축, y축)		이상치를 포함시킨 비율 (p_1)	$\hat{\beta}_0$의 범위 (p_2)	$\hat{\beta}_1$의 범위 (p_3)
(-5, -4)	ELMS	0.37	-2.94~ 1.58	0.79~1.41
	EDR	0	-3.18~-1.90	1.16~1.40
(-5, +4)	ELMS	0.4	-1.89~ 2.74	0.57~1.17
	EDR	1	1.75~ 3.00	0.59~0.81
(7.5, -4)	ELMS	0	-1.79~ 2.50	0.77~1.19
	EDR	0	-2.34~ 2.50	0.75~1.27
(7.5, +4)	ELMS	0	-1.62~ 1.89	0.77~1.22
	EDR	0	-1.53~ 2.16	0.73~1.16
(20, -4)	ELMS	0.48	-1.37~ 4.30	0.59~1.19
	EDR	1	-0.51~ 5.13	0.54~0.83
(20, +4)	ELMS	0.46	-5.66~ 2.00	0.77~1.48
	EDR	1	-5.46~-0.01	1.19~1.47

p_1은 clean subset에 이상치가 은폐효과를 갖게 하는 비율이고, p_2와 p_3는 이상치 진단을 위한 clean subset의 방향과 절편을 나타낸다. 따라서 p_1은 낮을수록, p_2와 p_3는 각각 0과 1근방의 범위를 나타내고 있을수록 clean subset을 잘 구성하고 있다고 할 수 있다. <표>에서 제시하고 있는 바와 같이 다중 이상치가 존재할 때에 EDR-ESD 방법은 거의 대부분의 경우에 masking 영향을 크게 받고 있음을 보여 주고 있다.

4. 결 론

이상치와 영향력 관찰점을 파악하는 것은 주어진 자료를 분석하는데 결정적인 의미를 지니고 있다. 널리 사용되고 있는 outward testing 방법은 일반적으로 주어진 자료를 이상치가 없으리라 여기는 clean subset과 이를 제외한 나머지들의 집합 즉, 이상치를 포함하고 있으리라 여겨지는 집합인 reminded subset으로 나누어 clean subset을 사용하여 주어진 자료에서 이상치를 식별하는 방법이다. 여기서 제안하고 있는 ELMS를 사용하여 clean subset을 구성하는 방법은 "단일 관찰점 진단방법"의 연속적 이용을 사용하는 EDR-ESD에 의한 clean subset을 구성하는 방법보다 <표>에서 제시하는 바와 같이 좋은 결과를 나타내고 있다. 즉, outward testing을 사용하여 이상치를 파악할 때 다수의 방법들이 초기 부분집합과 $\hat{\beta}$를 결정할 때에 사용하는 부분집합의 안정성에 크게 의존하고 있으나, ELMS방법은 이 초기 부분집합을 사용하여 반복적인 원의 재구성에 의한 방법과 최소중위수잔차를 계산하여 새롭게 부분집합을 구성하므로서 다중이상치에 의한 swamping, masking영향에서 기존의 방법에 비하여 잘 벗어날 수 있다.

그러나 나머지 집합에서 최소잔차합을 갖는 원을 취하여 clean subset에 포함시키는 방법은 초기 부분집합의 성향을 크게 개선하지 못하고 있으므로 clean subset을 설정하기 위한 향상된 방법이 필요하다고 여겨진다.

참 고 문 헌

[1] 염준근,박종구,김종우(1995), "다변량 자료에서 다수 이상치 인식의 절차", 품질경영학회지, 제23권, 제4호, pp. 28-41.

[2] Atkinson, A. C.(1986), "Masking Unmasked," Biometrika, Vol.73, No.3, pp.533-541.

[3] Atkinson, A. C.(1994), "Fast Very Robust Methods for the Detection of Multiple Outliers," Journal of the American Statistical Association, Vol.89, No.428, pp.1329-1339.

[4] Barnett, V., and Lewis, T.(1984), Outliers in Statistical Data (2nd ed.), John Wiley & Sons, New York.

[5] Belsey, D. A., Kuh, E., and Welsch, R. E.(1980), RegressionDiagnostics, Interscience, New York.

[6] Cook, R. D., and Hawkins, D. M.(1990), "Comment on Unmasking Multivariate Outliers and Leverage Points," Journal of the American Statistical Association, Vol.85, No.411, pp.640-644.

[7] Davies L. and Gather U.(1993), "The Identification Multiple Outliers(with discussion)",Journal of the American Statistical Association, Vol.88, No.423, pp.782-801.

[8] Fung, W-K.(1993), "Unmasking Outliers and Leverage Points : A Confirmation," Journal of the American Statistical Association, Vol.88, No.422, pp.515-519.

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

[10] Hadi, A.(1992), "Identifying Multiple Outliers in Multivariate Data," Journal of the Royal Statistical Society, Series-B, Vol.54, No.3, pp.761-771.

[11] Hadi, A.(1994), "A Modification of a Method for the Detection of Outliers in Multivariate Samples," Journal of the Royal Statistical Society,Series-B, Vol.56, No.2, pp.393-396.

[12] Hadi, A., and Simonoff, J. S.(1993), "Procedures for the Identifying of Multiple

Outliers in Linear Models," Journal of the American Statistical Association, Vol.88, No.424, pp.1264-1272.

[13] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, E.(1986), Robust Statistics: The Approach Based on Influence Functions, John Wiley & Sons, New York.

[14] Kianifard F., and Swallow, W. H.(1989), "Using Recursive Residuals, Calculated on Adaptively Ordered Observations, to Identify Outliers in Linear Regression," Biometrics, Vol.45, pp. 571-585.

[15] Marasinghe, M. G.(1985), "A Multistage Procedure for Detecting Several Outliers in Linear Regression," Tecnometrics, Vol.27, No.4, pp. 395-399.

[16] Paul, S. R., and Fung, K. Y.(1991), "A Generalized Extreme Studentized Residual Multiple-Outlier-Detection Procedure in Linear Regression, "Tecnometrics, Vol.33, No.3, pp. 339-348.

[17] Prescott, P.(1975), "An Approximation Test for Outliers in Linear Models," Tecnometrics, Vol.17, pp. 129-132.

[18] Rosner, B.(1975), "On the Detection of Many Outliers," Tecnometrics, Vol.17, No.2, pp. 221-227.

[19] Rosner, B.(1983), "Percentage Points for a Generalized ESD Many Outlier Procedure," Tecnometrics, Vol.25, No.2, pp. 165-172.

[20] Rousseeuw, P. J.(1984), "Least Median of Squares Regression," Journal of the American Statistical Association, Vol.79, No.388, pp.871-880.

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

[22] Rousseeuw, P. J., and van Zomeren, B. C.(1990), "Unmasking Multivariate Outliers and Leverage Points," Journal of the American Statistical Association, Vol.85, No.411, pp.633-639.