제주교육대학교 논문집(1994)
제23집 PP. 151~161

다중 이상치 식별을 위한 알고리듬 개발에 관한 연구

김 종 우*

〈목 차〉

요 약

회귀분석에서 널리 사용되고 있는 최소제곱에 의한 추정 방법은 다변량 자료에서 다수 이상치가 존재하는 경우에 masking효과와 swamping효과에 의하여, 이상치 파악과 영향력 관측점의 인식에 영향을 미친다. 이상치의 영향을 축소하는 여러가지 로보스트 추정량들 중 Rousseeuw가 제시한 Least Median of Squares Regression 방법은 효과적으로 이상치들을 인식하고 Influential Points를 파악할 수 있도록 하고 있다. 그러나 이 방법의 사용은 많은 계산을 요구하므로 수정된 Hadi의 방법을 사용하여 elementary set를 구하는 새로운 Algorithm을 제시한다.

1. 서 론

중회귀분석(Multiple Regression Analysis)에서 이상치와 영향력관찰점의 존재는 자료의 특성을 파악하는데 중요한 영향을 미치므로, 이들을 식별하기 위한 다양하고 효과적인 방법들이 많이 개발되어 왔다. 이러한 진단분야에서 과거 약 20년 동안 무수히 많은 진단 통계량들이 Belsely, et. al. (1980), Cook & Weiberg (1982) 및 Atkinson (1985) 등의 관련 서적과 연구물들을 통해 일반에게 잘 알려져 있으며, 또한 대부분의 「통계용 컴퓨터 소프트웨어」인 SAS, SPSS, BMDP, RATS, STATISTICA, … 에서 이러한 방법들은 이미 적용되어 사용하고 있다.

그러나 이들 대부분 진단통계량의 이상치 식별에는 한계가 있다. 이들이 사용하는 이상치 식별을 위한 잔차(Residual) e_i와 leverage point 식별을 위한 Hat Matrix에서 대각선상의 원 h_{ii}는 그 자체가 이상치에 크게 민감한 최소자승법(Least Square method :

* 제주교육대학교 수학교육과 조교수

LS)를 기초로 하고 있기 때문에, 다중 이상치의 존재시에는 이상치가 군집해 있는 방향으로 중심점을 끌어 당기는 영향에 의하여 이상치를 숨기려는 masking효과와 정상적인 점들이 중심점에서 멀리 떨어져 있는 점으로 인식되는 swamping효과에 의해 이상치와 영향력 관찰점을 인식하는데 문제점을 야기시키고 있다.

masking효과와 swamping효과를 극복하기 위한 방법으로 robust 개념을 사용한 이상치 추정량(Robust Estimator)이 70년대 중반부터 본격적으로 연구되기 시작했다. robust 추정량을 처음 소개한 사람은 Edgeworth로 알려져 있다. 그는 최소절대잔차(Least Absolute Residual) 추정량을

$$\min_{\theta} \sum_{i=1}^{n} | y_i - x_i\theta |$$

로 사용하였다. 이 추정량은 이상치 저항성(outlying resistency)에는 매우 강하지만 회귀 적합에 영향을 미치는 높은 leverage point에는 매우 취약함으로 밝혀져 있다.

다음에 Huber(1973)의 M 추정량은 0에서 유일한 최소값을 갖는 미분가능한 대칭함수 ρ에 대하여 잔차

$$r_i = y_i - x_i \hat{\theta}$$

의 함수 $\rho(r_i)$를 일반화함으로써 얻는다. 즉,

$$\min_{\theta} \sum_{i=1}^{n} \rho \frac{y_i - x_i\theta}{\sigma}$$

에 대응하는 추정치를 찾는 것이다. Huber의 M 추정량은 함수 ρ를 추정량이 Robust하도록 선택함으로써 얻는다. 여기서 Huber는 미니맥스 의사결정 원리에 의하여

$$\Psi(r) = \min(k, \max(r,-k))$$

단, k는 임의의 상수

를 제안했으나 높은 leverage point에서의 취약성 때문에 소위 일반화 M 추정량(Generalized M-estimator : GM)이 소개되게 되었다. GM은 가중치 함수 Ω_i를 이용하여 높은 leverage point의 영향을 제한하는 것을 목적으로 한다. 이러한 GM 계열의 추정량으로 Mallow(1975), Schweppe(1977), Hampel(1978), Krasker(1980), Krasker & Welsch(1982), Ronchetti & Rousseeuw(1985), Samarov(1985) 등은 Ψ와 Ω의 함수값을 조절하는 추정량들을 발표하였다.

그외에 대표적인 robust 추정량들로 L 계열 추정량들로 Bickel(1973), Rupert & Carroll(1980) 등이 있으며, 최근에 등장한 robust 추정량으로 Rousseeuw(1984)의 잔차제곱의 중위수를 최소화하는 Least Median of Squares(LMS) 추정량과 Least Trimmed Square(LTS) 추정량이 있다.

$$\text{LMS} : \min \sum_i r^2_{i:n}$$
$$\text{LTS} : \min \text{med}_i \; r_i^2$$

이 LMS 추정량을 사용하여 Atkinson(1986)은 LS에서 사용하고 있는 통계량인 잔차제곱의 합을 최소화하는 방법으로 수정된 Cook 통계량에 Rousseeuw의 LMS 추정량을 사용하는 robust 회귀추정 방법을 제안했다. 또한 Atkinson(1988)에는 LMS 추정량을 사용하여 transformation parameter의 값을 계산하는데 사용하는 방법을 제안하고 있으며, Rousseeuw(1990)는 Mahalanobis distance에 LMS 추정량과 MVE(Minimun Volume Ellipsoid)를 사용하여 다수의 이상치를 식별해 내고, leverage point를 찾아내는 방법을 제시하고 있다. Hadi(1992)는 LMS의 적용시에 발생하는 표본 추출의 처리시간이 길어짐을 줄이기 위한 방법으로 표본을 일정한 robust 추정량으로 중심에서 떨어져 있는 정도를 파악하여 이를 기준으로 ascending ordering 하여 표본을 추출함으로써 LMS에서 자료의 반복 추출 횟수를 줄이는 효과를 갖는 algorithm을 제시하고 있다.

본 연구에서는 Rousseeuw(1985)가 MVE에서 사용한 elementary set 추출의 문제점을 개선하기 위한 방법으로 ordering에 의하여 elementary set의 크기를 n에서 부터 시작하여 $[(n+p+1)/2]$까지 감소시켜 나가는 방법을 제시하고자 한다([ ]는 gauss기호). 본 연구에서는 이러한 descending ordering에 의한 방법의 문제점을 개선하는 방법을 제시하고자 한다.

2. Mahalanobis Distance

고전적인 방법에 의한 Mahalanobis Distance 계산식은

$$\text{MD}_i = \sqrt{(x_i - C(X))S(X)^{-1}(x_i - C(X))^t} \quad i=1,2,\ldots n \tag{1}$$

여기서 $C(X) = \sum_{i=1}^{n} x_i/n$

$$S(X) = \sum_{i=1}^{n} (x_i - C(X))^t (x_i - C(X))/(n-1)$$

거리 MD_i는 X상의 각 점 x_i가 그 집단의 중심으로 부터 얼마나 떨어져 있는가를 우리에게 알려준다. 그러나 이 값은 서론에서 지적한 바와같이 masking효과와 swamping효과에 의하여 다중 이상치들이 큰 MD_i값을 제시하지 못하고 있다. 이것은 $C(X)$와 $S(X)$

가 robust하지 못하기 때문이다; 즉, 조그만 이상치 집단도 이상치 집단 방향으로 $C(X)$를 끌어당기고, $S(X)$를 부풀리기 때문이다. 그러므로 (1)식에서 평균 $C(X)$와 공분산 $S(X)$를 robust 추정량으로 바꾸는 것은 당연하다고 여겨진다.

robust 추정량으로서 Campbell(1980), Stahel(1981), Donoho(1982), Hampel *et. al* (1986), Rousseeuw and van Zomeren(1990)등이 $C(X)$와 $S(X)$에 새로운 robust 추정량을 사용하였다. Rousseeuw(1985)가 제시한 MVE는 거의 50%에 달하는 breakdown point를 갖으므로서 이상치의 집단 식별을 매우 용이하게 해주고있다 (Lopuhaä and Rousseeuw, 1991). 그러나 이 방법의 적용에 따르는 크기 h인 MVE를 구하기 위하여는 $n \times p$행렬에서 $n!/h!(n-h)!$ 회에 달하는 계산을 필요로 하게 된다. 이러한 문제점을 극복하기 위하여 몇가지 algorithm이 Stahel(1981), Donoho(1982), Rousseeuw and Leroy(1987), Woodruff and Rocke(1993) 등에서 제시되고 있다.

Rousseeuw and Leroy(1987)가 제시하고 있는 resampling algorithm은 식(1)에 크기 $p+1$인 subsample에 대하여 j번째 subsample의 $C_j(X)$와 공분산 $S_j(X)$를 대입하여 거리를 구하고 있다.

$$D_i(C_j,S_j) = \sqrt{(x_i - C_j)^t S_j^{-1}(x_i - C_j)} \quad i = 1,2, \ldots, n \tag{2}$$

여기서 $C_j(X) = \sum_{i=1}^{n} w_i x_i / \sum_{i=1}^{n} w_i$

$$S_j(X) = \sum_{i=1}^{n} (x_i - C_j(X))^t (x_i - C_j(X)) / (\sum_{i=1}^{n} w_i - 1)$$

$w_i = w(RD_i)$는 n개 관찰점중에 $p+1$개를 택함.

MVE에 사용하기 위한 j번째 subsample의 결정 방법은 (2)에서 계산된 n개 값들중에 $100(h/n)$번째 percentile을 m_j라 하면, h개 관찰점을 갖는 C_j와 S_j의 크기는 $(m_j^p \det(S_j))^{1/2}$에 비례하므로 $m_j^p \det(S_j)$가 최소가 되는 j번째 subsample을 사용하게 된다. D_i가 Outlier인지를 판정하기 위해 Rousseeuw and van Zomeren(1990)은 h를 $[(n+p+1)/2]$로 사용한 다음 식을 제안하고 있다.

$$D_i(C_j,c_jS_j) = \sqrt{(x_i - C_j)^t (c_jS_j)^{-1}(x_i - C_j)} \quad i = 1,2,\ldots, n \tag{3}$$

여기서 $c_j = c_{np} m_j / \chi^2_{p,0.50}$

$$c_{np} = (1 + 15/(n-p))^2$$

c_j는 다중정규분포에서 자료가 추출될 수 있도록 하는 수정계수이다.

3. Hadi's Algorithm

Rousseeuw and van Zomeren (1990)가 제시하고 있는 robust 추정 방법은 Mahalanobis Distance (MD_i)에 MVE를 위하여 선별된 elementary set의 평균 C_j와 공분산행렬 S_j를 사용하여 D_i를 계산하는 것이다.

그러나 이 방법은 첫째, MVE를 위하여 선별된 j 번째 elemetary set에 이상치가 절대로 포함되어있지 않다는 가정을 필요로 한다. 만일 이상치를 포함한 elementary set을 구성할 때에 왜곡된 추정량 D_i를 중심으로 이상치를 판별하게 될 것이다. 둘째로, j 번째 elemetary set의 rank가 p가 안될 경우에 Rousseeuw and van Zomeren (1990)은 이 subsample을 포기하고 있다. 세째로, j 번째 elemetary set의 rank가 p일지라도 공분산 행렬 S_j는 singular matrix가 될 수 있는 문제점을 갖고 있다(Hadi, 1992).

이러한 문제점을 극복하기 위한 방법으로 Hadi (1992)는 3 단계에 걸쳐 subsample을 선별하고 robust 거리를 계산하는 방법을 제시하고 있다.

Step 0 : Initial rearrange the n observations in ascending order according to a suitably chosen robust distance.

$$D_i(C_R, S_R) = \sqrt{(x_i - C_R)^t S_R^{-1} (x_i - C_R)}$$

where C_R and S_R are robust location and covariance matrix estimators.

Step 1-1 : If the basic subset is of full rank, compute

$$D_i(C_R, S_R) = \sqrt{(x_i - C_b)^t S_b^{-1} (x_i - C_b)}$$

where C_b and S_b are the mean and covariance matrix of the basic subset.

Step 1-2 : If the basic subset is not of full rank, compute the eingenvalues of S_b: $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_p = 0$ and eigenvectors V_b.

$$D_i(C_R, S_R) = \sqrt{(x_i - C_b)^t V_b W_b V_b^t (x_i - C_b)}$$

where W_b is a diagonal matrix with jth diagonal element is

$W_i = \max(\lambda_j, \lambda_s)^{-1}$ with λ_s is the smallest non-zero eigenvalue of S_b.

Step 2 : Increase size of basic subset while repeat Step 0 and Step1.
Step 3 : Stopping Criterion.

Hadi가 제시하는 algorithm은 Rousseeuw가 MVE를 위하여 subsample을 추출하기 위한 resample algorithm을 개선한 방법으로 ascending ordering에 의하여 elementary set의 rank를 $p+1$에서 부터 시작하여 적어도 $[(n+p+1)/2]$까지 증가시켜 나가는 방법이다. 그러나 이러한 ascending ordering에 의한 방법은 이상치가 전체자료에서 50% 정도를 차지할 경우를 대상으로 하는 방법으로서, 실제 자료에서 나타나는 이상치가 전체 자료에서 차지하는 비율이 작은 경우에는 descending ordering 방식이 효과적이다.

4. Modified Hadi's Algorithm

Hadi가 제시한 algorithm에서 크기 h인 MVE를 구하기 위하여 $n \times p$ 행렬에서 robust 추정량을 사용하여 밝혀진 가장 큰 D_i부터 하나씩 제거해 나가면서 $[(n+p+1)/2]$가 될때까지 elementary subset을 구하는 algorithm은 다음과 같다.

Step 0 : Initial rearrange the n observations in descending order according to a suitably chosen robust distance.

$$D_i(C_R, S_R) = \sqrt{(x_i - C_R)^t S_R^{-1} (x_i - C_R)}$$

where C_R and S_R are robust location and covariance matrix estimators.

Step 1 : If the subset is of full rank, compute

$$D_i(C_R, S_R) = \sqrt{(x_i - C_s)^t S_s^{-1} (x_i - C_s)}$$

where C_s and S_s are the mean and covariance matrix of the subset which is removed the largest D_i.

If the subset is not of full rank, the next largest D_i is removed.

Step 2 : Repeat step 1 util the basic subset contains $[(n+p+1)/2]$ observations.

Step 3 : Measure the residuals using the basic subset.

Modified Hadi's algorithm은 먼저 Step 0에서 robust location과 scale을 사용하여 모든 X_i에 대하여 Mahalanobis distance를 구한다. Step1에서 구한 MD_i를 descending ordering으로 하여 ranking이 $[(n+p+1)/2]$보다 작은 것들을 순차적으로 배제시킨다. 일정한 기준에 도달한 자료만으로 구성된 elementary set를 사용하여 회귀계수 $\hat{\beta}_b$를 구하여 모든 X_i에 대하여 잔차 D_i를 계산한다. Hadi's algorithm의 최소 계산 횟수는 $[(n-p-1)/2]$를 필요로 하지만, 만일 이상치의 개수가 $[(n-p-1)/2]$보다 작을 경우, 즉 k일 때 Modified Hadi's algorithm은 $n-k$ 회의 연산 횟수를 요구하게 된다. 그러므로 $[(n+p+1)/2]$과 k를 비교하게 된다. 그러나 일반적인 자료의 특성상 이상치의 수는 최대 50%에 달할 수 있으나 그러한 경우에는 두 algorithm의 연산 횟수는 동일하게 나타나고 대부분의 경우에서처럼 이상치의 발생건수가 5% 미만일 경우에는 Modified Hadi's algorithm에서 사용하는 descending ordering 방식이 효과적이다.

5. 예

Brownlee (1965)의 Stackloss자료를 사용한 simulation에서 표준화잔차 (r_i)와 Mahalanobis 거리(MD_i), Modified Hadi's Algorithm을 사용한 D_i에 관한 비교 표이다.

〈표〉 Stackloss 자료

관찰값	X_1	X_2	X_3	Y	r_i	MD_i	LMS_i	D_i
1	80	27	89	42	1.193	2.25	6.85	8.2233785
2	80	27	88	37	-.726	2.32	2.64	4.8137877
3	75	25	90	37	1.546	1.59	6.62	7.0852096
4	62	24	87	28	1.882	1.27	7.83	6.0708259
5	62	22	87	18	-.542	0.30	0.10	0.6393562
6	62	23	87	18	-.965	0.77	-0.17	0.4753786
7	62	24	93	19	-.834	1.85	0.82	2.1659976
8	62	24	93	20	-.485	1.85	1.64	2.7666451
9	58	23	23	87	-1.046	1.36	0.00	0.245927
10	58	18	18	80	.437	1.75	0.00	0.8376744
11	58	18	18	89	.884	1.47	0.63	0.9416001
12	58	17	17	88	.969	1.84	0.00	0.3755046
13	58	18	18	82	-.480	1.48	-2.34	2.1367473
14	58	19	19	93	-.017	1.78	-1.00	0.427413
15	50	18	18	89	.809	1.69	0.97	0.467439

관찰값	X_1	X_2	X_3	Y	r_i	MD_i	LMS_i	D_i
16	50	18	18	86	.299	1.29	-0.07	0.6582029
17	50	19	19	72	-.611	2.70	-0.49	3.7029301
18	50	19	19	79	-.153	1.50	0.00	1.6799549
19	50	20	20	80	-.203	1.59	0.63	1.0995264
20	56	20	20	82	.454	0.81	1.76	0.5160211
21	70	20	20	91	-2.638	2.18	-6.87	3.1051625

cutoff value : $LMS_i = 2.5$, $\sqrt{\chi^2_{3,.975}} = 3.06$, $D_i = 3.06$

5. 인 용 문 헌

김성수, (1992), "Regression Diagnostics Using Dynamic Graphical Mathods", 서울대.

김승구, (1991), "높은 파손점과 효율성을 갖는 로보스트 회귀추정량에 관한 연구", 동국대.

박성현, (1991), 회귀분석, 박영사.

Akinson, A. C. (1986), "Masking unmasked", Biometrica 73, 533~541.

Akinson, A. C. (1988), "Transformed Unmasked", Technometrics 30, 311~318.

Andrews, D. F., and Pregibon, D. (1978), "Finding the Outliers that Matter", The Journal of the Royal Statistical Society Series-B 40, 85~91.

Atkinson, A. C. (1981), "Two graphical displays for outlying and influential observations in regression", Biomerica 68, 13-20.

Atkinson, A. C. (1985), Plots, Transformationsand Regression, Oxford. U. K. : Oxford University Press.

Beckman, R. J. and Cook, R. D. (1983), "Outlier……s", Tecnometrics 25, NO. 2, 119~149.

Belsey, D. A. and Kuh, E. and Welsch, R. E. (1980), "Regression Diagnostics", Wiley-Interscience.

Coakley, C. W. and Hettmansperger, T. P. (1993), "A Bounded Influence, High Breakdown, Efficient Regression Estimator", Journal of the American Statistical Association 88, NO 423, 872~880.

Cook, R. D. and Wnag, P. C. (1983), "Transformations and Influential Cases in

Regression", Technometrics 25, NO 4, 337~343.

Cook, R. D. (1981), "Detection of Influential Observations and Outliers in Regression", Technometrics 23, NO 1, 15~18.

Davies, L. and Gather, U. (1993), "The Identification of Multiple Outliers", The Jounal of American Statistical Association 88, No 423, 782~801.

Dhar, S. K. (1991), "Minimun distance estimation in an additive effects outliers model", The Annals of Statistics 19, NO 1, 205~228.

Draper, N. R. and John, J. A. (1981), "Influential Observations and Outliers in Regression", Tecnometrics 23, NO 1, 21~26.

Fung, W. K. (1993), "Unmasking Outliers and Leverage Point : A Confirmation", The Jounal of American Statistical Association 88, NO 422, 515~519.

Gray, J. B. and Ling, R. F. (1984), "K-Clustering as a Detection Tool for Influential Subsets in Regression", Technometrics 26, NO 4, 305~330.

Hadi, A. (1992), "Identifying Multiple Outliers in Multivariate Data", The Jounal of the Royal Statistical Society Series-B 54, NO 3, 761~771.

Hawkins, D. M. (1991), "Diagnostics for Use With Regression Recursive Residuals", Tecnometrics 33, NO 2, 221~234.

Hoaglin, D. C. and Welsch, R. E. (1978), "The Hat Matrix in Regression and ANOVA", The American Statistician 32, NO 1, 17~22.

Hocking, R. R. (1983), "Developments in Linear Regression Methodology : 1959-1982", Technometrics 25, NO 3, 219~230.

Kim, Soon-Kwi, (1993), "A Study on Applications of Regression Diagnostic Method to Technometrics, and the Statistical Quality Control", Journal of the KSQC 21, NO 1, 55~64.

Lopuhaä, H. P. and Rousseeuw, P. J. (1991), "Breakdown points of affine equivariant estimators of multivariate location and covariance matrices", The Annals of Statistics 19, NO 1, 229~248.

Mansfield, E. R. and Helms, B. P. (1982), "Detecting Multicollinearity", The American Statistician 36, NO 3, 158~160.

Marasinghe, M. G. (1985), "A Multistage Procedure for Detecting Several Outlier in Linear Regression", Technometrics 27, NO 4, 395~399.

Mason, R. L. and Gunst, R. F. (1985), "Outlier-Induced Collinearities", Technometrics 27, NO 4, 401~407.

Oman, S. D. (1983), "Regression Estimation for a bounded Response Over a Bounded Region", Technometrics 25, NO 3, 251~261.

Parker, I. (1988), "Transformations and Influential Observations in Minimum Sum of Absolute Errors Regression", Technometrics 30, NO 2, 215~220.

Paul, S. R. and Fung, K. Y. (1991), "A Generalized Extreme Studentized Residual Multiple-Outlier-Detection Procedure in Linear Regression", Tecnometrics 33, NO 3, 339~348.

Rosner, B. (1983), "Percentage Points for a Generalized ESD Many-Outlier Procedure", Technometrics 25, NO 2, 165~172.

Rousseeuw, P. J. and Zomeren, B. C. (1990), "Unmasking Multivariate Outliers and Leverage Points", The Jounal of American Statistical Association 85, NO 411, 633~639.

Rousseeuw, P. J. (1984), "Least Meadian of Squares Regression", The Jounal of American Statistical Association 79, NO 388, 871~880.

Schall, R. and Dunne, T. T. (1990), "Influential Variables in Linear Regression", Technometrics 32, NO 3, 323~330.

Solomon, H. (ed.), (1961). Studies in Item Analysis and Prediction, Stanford University Press, California.

Stewart, G. W. (1987), "Collinearity and Least Squares Regression", Statistical Science 2, NO 1, 68~100.

Walker, E. and Birch, J. B. (1988), "Influence Measures in Ridge Regression", Technometrics 30, NO 2, 221~227.

Webster, J. T. and Gunst, R. F. and Mason, R. L. (1974), "Latent Root Regression Anlysis", Tecnometrics 16, NO 4, 513~522.

Weisberg, S. (1983), "Some Principles for Regression Diagnostics and Influence Analysis", Tecnometrics 25, NO 3, 240~244.

Willan, A. R. and Watts, D. G. (1978), "Meaningful Multicollinearity Measures", Tecnometrics 20, NO 4, 407~411.

Woodruff, D. L. and Rocke, D. M. (1993), "Heuristic Search Algorithms for the Minimum Volume Eilipsoid", American Statistical Association 2, No 1, 69~95.

〈Abstract〉

An Algorithm for Indentifying of Multiple Outliers

Kim, Chong-Woo

Identification of multiple outliers is difficult because of the masking effect and the swamping effect. Recently, among the various robust estimator of reducing the effect of outliers, LMS (Least Meadian Square) estimator has been to be a suitable method proposed to expose outliers and leverage points.

However, as you know it, the data anlysis method with LMS estimator is to be taken the median of the squared residuals in the sample which is extracted the sample space. Then this model causes the trouble, for the number of the chosen sample is nCp, i.e. as the size of sample space n is increasing the number is increasing fastly. And the covariance matrix may be the singular matrix, so that matrix is approching collinearity. Thus we propose a procedure for the sampling in LMS method and study the algorthm for finding the effective elementary set.