제주교육대학교 논문집 (1998)
제27집 pp. 377~394

2단계 방법에서 기본집합 결정을 위한 초기 설정

김 종 우*

The Initial Setting of The Elementary Set on Two Stage Procedure

Kim, Chong-Woo

< 목 차 >

Ⅰ. 序 論
Ⅱ. 最少 제곱 推定量
Ⅲ. 초기 기본 집합
Ⅳ. 模擬實驗
Ⅴ. 結 論
* 참고문헌

Abstract

In this paper initial setting of elementary subset for two stage procedure is discussed. The outliers-testing procedure is conducted by elementary subset decision and detecting test, which is fluctuated by elementary subset. We compare the several methods.

* 제주교육대학교 컴퓨터교육과 부교수

Ⅰ. 序　論

1. 硏究의 背景

자료 분석의 방법으로 널리 사용되는 선형회귀분석(linear regression analysis)은 이상치(outlier)에 의한 영향이 매우 크게 나타나고 있다. 고전적인 최소제곱(OLS : ordinary least squares)에 의한 추정량은 붕괴점(breakdown point)가 0%에 이르며, 이상치 문제를 극복하기 위한 회귀 진단(regression diagnosis)분야는 직접접근방법, 로버스트(robust) 적합을 이용한 간접접근방법, 로버스트 적합 후에 직접접근방법을 사용하는 합성방법이 있다. 직접접근방법은 Prescott(1975), Tietjen, Moore와 Beck man(1973), Rosner(1975), Rosner(1983), Marasinghe(1985), Paul과 Fung(1991), Kianifard와 Swallow(1989), Hawkins(1991) 등에서 제시하고 있다. 이상치에 의한 영향을 적극적으로 배제시키기 위한 방법으로 로버스트 적합을 사용하여 이상치를 식별하려는 간접접근방법은 Huber(1964), Rousseeuw(1984), Rousseeuw와 van Zomeren(1990), Rousseeuw와 Leroy(1987), Atkinson(1994) 등에서 붕괴점 50%에 달하는 방법을 제시하고 있다. 직접접근방법과 간접접근방법의 절충점을 모색하는 합성방법은 자료를 두 개의 분리된 집합으로 나누어 자료를 식별하는 2단계 방법이다. 이러한 2단계 방법에 관한 연구물들로는 Atkinson(1986), Fung(1993), Woodruff와 Rocke(1994), Hadi(1992), Hadi와 Simonoff(1993), 김종우(1996) 등이 있다. 이러한 3가지 방법들 중에 앞의 두가지 방법의 문제점을 극복하고자 제안된 2단계 방법은 1단계에서 이상치를 배제 시키는 로버스트한 방법으로 기본집합을 구성하고, 이를 사용하여 2 단계에서 자료 중에 포함되어 있을 지도 모르는 이상치를 식별하는 방법이다. 그러나 이는 1단계에서 만들어지는 기본집합에 이상치가 포함되어 있지 않으리라는 기대를 갖고 행하는 방법인 것이다. 따라서 기본집합에 이상치를 가급적 포함시키지 않으려는 노력이 필요하다. 기본집합에 관한 연구물들은 Rosner(1983), Marasinghe (1985), Paul과 Fung(1991), Hawkins 외(1994), Rousseeuw(1984), Rousseeuw와 van Zom eren(1990), Rousseeuw

와 Leroy(1987), Atkinson(1986, 1994), Fung(1993), Woodruff와 Rocke(1994), Hadi(1992,1994), Hadi와 Simonoff(1993), 염준근외(1995), 김종우(1996) 등이 있다.

2. 硏究의 目的

본 연구에서 사용할 이상치 식별 방법은 2단계 방법으로 김종우(1996)에서 제안한 이상치 식별 방법을 개선하는 것이다. 이 방법에서 김종우(1996)은 1단계에서 기본집합을 결정하기 위하여 ELMS 방법을 사용하고 있다. ELMS 방법은 매우 높은 이상치 배제성을 지니고 있는 것으로 밝혀져 있다. 그러나 많은 계산양을 필요로 하고 있음에 따라 기본 집합의 크기가 inlier라 여겨지는 자료를 Hadi-Simonoff 방법과 같은 적절한 방법으로 취하고, 또한 기본집합을 구성 원소가 1개씩 늘 때마다 재구성하는 과정에서 필요 이상의 계산을 요구하는 문제점이 따른다. 따라서 ELMS 방법의 적절한 사용이 필요하다.

3. 硏究의 範圍

초기 기본집합의 이상치를 배제시키기 위한 방법의 모색을 위하여 기본집합의 초기 설정은 매우 의미가 크다. 이를 결정하는 방법으로 널리 사용되고 있는 OLS를 이용한 최소절대잔차를 이용하는 방법, 표준화된 잔차를 이용하는 방법, 최소중위수 잔차를 이용하는 방법을 비교하기 위하여, 기존에 연구된 OLS, Hadi-Simonoff, Hadi(1992), ELMS 방법을 사용한다.

Ⅱ. 最少 제곱 推定量

1. 最少 제곱 推定量의 性質

선형 회귀모형을 다음과 같이 설정하자.

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

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

$\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^T$는 $\boldsymbol{x}_i^T = (x_{i1}, x_{i2}, \ldots, x_{ip})$를 행으로 갖는 설명변수인 ($n \times p$) 행렬이다 (단, $p < n$).

$\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$는 모수인 ($p \times 1$) 벡터이다.

$\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I})$인 ($n \times 1$) 벡터이다.

이때, OLS에 의한 $\boldsymbol{\beta}$의 불편추정량은

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}\ , \tag{2}$$

이다.

식(1)의 모형에서 $\boldsymbol{Y}$는 $N(\boldsymbol{X\beta}, \boldsymbol{I}\sigma^2)$의 분포를 따르므로, 잔차 $\boldsymbol{e}$는

$$\begin{aligned} \boldsymbol{e} &= \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \\ &= (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} \\ &= (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}) \\ &= (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\varepsilon} \end{aligned} \tag{3}$$

이다. 여기서 $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$로서 「해트(hat)」 행렬이고, 잔차 $\boldsymbol{e}$의 평균과 분산은 다음과 같다.

$$E[\boldsymbol{e}] = E[\ (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\varepsilon}] \tag{4}$$

$$= (I - H)E[\varepsilon]$$
$$= 0$$
$$Var[e] = Var[(I-H)\varepsilon]$$
$$= (I - H)Var[\varepsilon](I - H)$$
$$= (I - H)\sigma^2$$

따라서, 잔차 e는 $N(0,(I-H)\sigma^2)$의 분포를 따르므로,

$$e_i/\sigma\sqrt{1-h_{ii}} \sim N(0,1) \quad (5)$$

이다. σ^2의 최소제곱추정량 s^2은

$$s^2 = e^T e/(n-p) \quad (6)$$

이다.

2. 標準化殘差의 性質

식(6)의 s^2을 사용한 i번째 관측치의 표준화잔차(studentized residual) r_i는 다음과 같다.

$$r_i = e_i/s\sqrt{1-h_{ii}} \quad (7)$$

여기서 h_{ii}는 헤트행렬의 대각선상의 원이며, 이러한 r_i를 내적으로 스튜던트화된 잔차(internally studentized residual)라고 부르기도 한다([23], p18). 이상치를 식별하는 방법으로 i번째 관측치가 회귀 적합에 사용되지 않는 잔차 e_i^*는 $e_i^* = y_i - x_i\hat{\beta}_{(i)}$, 여기서 $\hat{\beta}_{(i)} = (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T Y_{(i)}$이다. 아래 첨자($i$)의 사용은 i번째 관측치가 사용되지 않을 때이고, 윗 첨자 *의 표기는 이때 나타나는 잔차이다. 잔차 e_i^*를 사용한 제거잔차(deletion residual, externally studentized residual) r_i^*는

$$r_i^* = e_i^* / s_{(i)} \sqrt{1 + \boldsymbol{x}_i^T (\boldsymbol{X}_{(i)}^T \boldsymbol{X}_{(i)})^{-1} \boldsymbol{x}_i} \qquad (8)$$

이고, 여기서 $s_{(i)}^2 = \boldsymbol{e}_{(i)}^T \boldsymbol{e}_{(i)} / (n-p-1)$이다.

제거잔차 r_i^*의 최대 절대값 $\max |r_i^*|$가 이상치 인지의 판정은 Bonferroni t-검정을 사용한 기각치 $t_{(\alpha/2(n-1),\, n-p-1)}$값을 사용하여 할 수 있다([23],p.22). 제거잔차의 기각치를 표준화잔차에 적용하기 위하여 표준화잔차 r_i와 제거잔차 r_i^*의 관계를 조사하면,

$$r_i^* = r_i / [(n-p-r_i^2)/(n-p-1)]^{1/2}$$

이므로,

$$r_i = r_i^* \times [(n-p)/(n-p-1+r_i^{*2})]^{1/2} \qquad (9)$$

이다. 근사 기각치 c는

$$c = t_{(\alpha/2(n-1),\, n-p-1)} \times [(n-p)/(n-p-1+t^2_{(\alpha/2(n-1),\, n-p-1)})]^{1/2} \qquad (10)$$

이다([17]). 따라서 $\max |r_i|$의 기각치로서 c를 사용하여 이상치 판정을 할 수 있다.

이 결과는 Lund(1975)의 결과와 매우 근사하며, 표준화잔차의 기각치로 사용하는데 효율적이다. 집합의 크기가 k인 기본집합 J에 이상치가 포함되어 있을 것인가는 표준화잔차의 계산에 의하여 $\max |r_i|$를 구하고, 이를 식(10)에서 얻은 기각치 c를 사용하여 이상치 판정을 하여 결정할 수 있다.

$$\max |r_i| = \text{maximize} \left| \frac{e_i}{s_J \sqrt{1-h_{ii}}} \right| , \qquad (11)$$

여기서, $e_i = y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_J$ 단, $i \in J$,

$$s_J^2 = \boldsymbol{e}_J^T \boldsymbol{e}_J / (k-p) ,$$

$$h_{ii} = \boldsymbol{x}_i^T (\boldsymbol{X}_J^T \boldsymbol{X}_J)^{-1} \boldsymbol{x}_i .$$

〈표1〉 유의수준 α = 0.05에서 $\max|r_{ij}|$의 기각치

표본의 크기	독립변수의 수 p									
	1	2	3	4	5	6	8	10	15	20
5	1.92									
6	2.07	1.93								
7	2.18	2.08	1.93							
8	2.27	2.20	2.09	1.94						
9	2.35	2.29	2.21	2.10	1.94					
10	2.41	2.36	2.30	2.22	2.11	1.95				
12	2.52	2.48	2.44	2.39	2.32	2.23	1.95			
14	2.60	2.58	2.55	2.51	2.46	2.41	2.25	1.96		
16	2.67	2.65	2.63	2.60	2.57	2.53	2.43	2.26		
18	2.73	2.71	2.69	2.67	2.65	2.62	2.55	2.44		
20	2.78	2.77	2.75	2.73	2.71	2.69	2.64	2.56	2.14	
25	2.88	2.87	2.86	2.85	2.84	2.82	2.79	2.76	2.59	
30	2.96	2.95	2.94	2.94	2.93	2.92	2.90	2.88	2.79	2.16
35	3.02	3.02	3.01	3.01	3.00	2.99	2.98	2.96	2.91	2.64
40	3.07	3.07	3.07	3.06	3.06	3.05	3.04	3.03	2.99	2.84
45	3.12	3.12	3.11	3.11	3.11	3.10	3.09	3.08	3.06	2.96
50	3.16	3.16	3.15	3.15	3.15	3.15	3.14	3.13	3.11	3.04
60	3.23	3.22	3.22	3.22	3.22	3.22	3.21	3.21	3.19	3.15
70	3.28	3.28	3.28	3.28	3.27	3.27	3.27	3.27	3.26	3.23
80	3.33	3.33	3.32	3.32	3.32	3.32	3.32	3.32	3.31	3.29
90	3.37	3.37	3.36	3.36	3.36	3.36	3.36	3.36	3.35	3.34
100	3.40	3.40	3.40	3.40	3.40	3.40	3.40	3.39	3.39	3.38

〈표2〉 유의수준 $\alpha = 0.01$에서 $\max|r_{ij}|$의 기각치

표본의 크기	독립변수의 수 p									
	1	2	3	4	5	6	8	10	15	20
5	1.97									
6	2.16	1.98								
7	2.31	2.17	1.98							
8	2.43	2.32	2.17	1.98						
9	2.53	2.44	2.32	2.17	1.98					
10	2.62	2.54	2.45	2.33	2.18	1.98				
12	2.75	2.70	2.63	2.56	2.46	2.34	1.98			
14	2.86	2.82	2.77	2.72	2.65	2.57	2.34	1.99		
16	2.95	2.91	2.88	2.84	2.79	2.73	2.58	2.35		
18	3.02	2.99	2.96	2.93	2.89	2.85	2.74	2.59		
20	3.08	3.06	3.03	3.01	2.98	2.94	2.86	2.75	2.20	
25	3.20	3.19	3.17	3.15	3.14	3.11	3.07	3.01	2.77	
30	3.29	3.28	3.27	3.26	3.25	3.23	3.20	3.16	3.03	2.20
35	3.36	3.36	3.35	3.34	3.33	3.32	3.30	3.27	3.19	2.80
40	3.42	3.42	3.41	3.40	3.40	3.39	3.37	3.35	3.30	3.07
45	3.47	3.47	3.46	3.46	3.45	3.45	3.43	3.42	3.38	3.23
50	3.52	3.51	3.51	3.51	3.50	3.50	3.49	3.47	3.44	3.33
60	3.59	3.59	3.58	3.58	3.58	3.57	3.57	3.56	3.54	3.47
70	3.65	3.65	3.64	3.64	3.64	3.64	3.63	3.63	3.61	3.57
80	3.70	3.69	3.69	3.69	3.69	3.69	3.68	3.68	3.67	3.64
90	3.74	3.74	3.73	3.73	3.73	3.73	3.73	3.72	3.71	3.69
100	3.77	3.77	3.77	3.77	3.77	3.77	3.76	3.76	3.75	3.74

3. 잔차에 의한 이상치 식별

일반적으로 이상치 식별을 위하여 사용하는 OLS에서 잔차는 식(3)에 의하여 나타난다. OLS에 의한 잔차를 사용하여 이상치를 식별하는 방법은 계산이 매우 간편하다는 장점을 갖고 있으나, 각 관측치 잔차의 크기가 어느 정도를 이상치로 볼 것인가 하는 판단에 어려움을 갖게 한다. 이러한 문제를 극복한 표준화잔차 식(7)은 제거잔차 식(8),식(9)와의 관계에 의하여 기각치 식(10)을 구할 수 있고 이에 따른 이상치 판정은 매우 유용하다고 여겨진다([20],[21]). ELMS에서 사용하고 있는 이상치를 판정하는 잔차의 정도는 Rousseeuw(1984)에서 사용한 변형된 잔차의 크기를 사용하므로 2.5를 일정한 판단의 기준으로 삼는다.

Ⅲ. 초기 기본 집합

1. OLS를 사용한 초기 기본 집합

식(1)의 모형에서

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

여기서 Y는 $N(X\beta, I\sigma^2)$의 분포를 따르므로, 잔차 e는

$$e = Y - X\hat{\beta}$$

$$|e_{i_1}|_{1:n} \leq |e_{i_2}|_{2:n} \leq , \ldots , \leq |e_{i_n}|_{n:n} \qquad (12)$$

일 때, 초기 기본집합의 설정은 $J = \{ i_1, i_2, \ldots , i_k \}$ 이다.

2. Hadi와 Simonoff의 초기 기본집합

전체 자료를 대상으로 앞선 1의 과정인 OLS를 사용한 초기 기본 집합

$J=\{ i_1, i_2, \ldots, i_k \}$ 를 구한다. 각 잔차 e_i를 (7)에서 구한 표준화잔차로 변환하여 절대 표준화잔차의 크기순으로 초기 기본 집합 $J=\{ i_1, i_2, \ldots, i_k \}$ 를 구한다.

$$r_i = e_i / s\sqrt{1-h_{ii}}$$

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

$$e_i = y_i - \boldsymbol{x}_i \hat{\boldsymbol{\beta}} ,$$

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

$$h_{ii} = \boldsymbol{x}_i^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_i .$$

3. Hadi(1992)의 초기 기본 집합

이상치를 식별하기 위한 로버스트 방법을 사용하기 위하여 로버스트 위치 추정량 C_R과 스케일 추정량 S_R을 사용한다. 초기 기본 집합을 결정하기 위하여 각 관측치별로 $D_i(C_M, S_M)$을 계산한다. C_M은 $\boldsymbol{X}$의 열(column)별 중위수(median)로 구성된 벡터(vector)이고,

$$S_M = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - C_M)(x_i - C_M)^T$$

이다. 따라서 각 관측치를 $D_i(C_M, S_M)$에 따라 오름차순으로 정렬하고, 초기 기본 집합의 크기를 $(n+p+1)/2$로 하여, $J=\{ i_1, i_2, \ldots, i_k : k=(n+p+1)/2 \}$ 를 결정한다. 이상치 식별을 위한 위치 추정량 C_R과 스케일 추정량 S_R은 기본 집합 J 만을 사용한 C_J와 S_J에 의한 $D_i(C_J, S_J)$로 판정한다.

4. ELMS를 사용한 초기 기본집합

1 단계. 2 에서와 같이 전체 자료를 대상으로 OLS를 구하고, 이에 따른 잔차를 계산한다. 최소 절대 잔차 순으로 기본집합 J 의 크기 k만큼을 구한다.

$$r_i = e_i / s\sqrt{1-h_{ii}}$$

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

$$e_i = y_i - \boldsymbol{x}_i \hat{\boldsymbol{\beta}} .$$

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

$$h_{ii} = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i .$$

2 단계. 기본집합 J에서 $(k-1)$개의 원을 취하고 나머지집합 J^c에서 1개의 원을 취하여 새로운 기본집합 $J_{(i)j}$를 구성한다. 이들 중에서 각 기본집합들 중에 최소잔차중위수를 갖는 기본집합을 선택한다.

$$\underset{\hat{\boldsymbol{\beta}}_{J_{(i)}}}{\text{Minimize}} \ \underset{J_{i(j)}}{\text{med}} \ e_i^2 \qquad (13)$$

여기서 $J_{(i)j} = J_{(i)} \cup \{j\}$.

$$J_{(i)} = J - \{i\}, \quad i \in J \ (= \{i_1, i_2, \ldots, i_k\}),$$

$$j \in J^c \ (= \{i_{k+1}, i_{k+2}, \ldots, i_{n-k}\}),$$

$$e_i = y_i - \boldsymbol{x}_i \ \hat{\boldsymbol{\beta}}_{J_{(i)}} ,$$

$$\hat{\boldsymbol{\beta}}_{J_{(i)}} = (\boldsymbol{X}_{J_{(i)}}^T \boldsymbol{X}_{J_{(i)}})^{-1} \boldsymbol{X}_{J_{(i)}}^T \boldsymbol{Y}_{J_{(i)}} .$$

Ⅳ. 模擬實驗

1. 模擬實驗의 設計 및 節次

여러 가지 이상치 식별 방법들 중에 본 연구와 비교하기 위한 방법으로 OLS, Hadi와 Simonoff(1993), Hadi(1992), ELMS방법을 동일한 자료를 발생시켜 그 결과를 비교한다. 각 방법의 비교는 단순 선형회귀 모형과 다중 선형회귀 모형으로 나누어 실시한다. 모의실험을 위한 난수 발생과 프로그램의 실행은 SAS/IML을 사용하고, 이상치 식별을 위한 방법은 계획된 이상치의 식별 여부로 판정한다.

(1) 단순 선형회귀 모형에서 모의실험

단순 선형 모형일 때 $y_i = \beta_o + \beta_1 x_i + \varepsilon_i,\ i = 1, 2, \ldots, 25$ 이고, $\beta_o = 0$, $\beta_1 = 1$에서 정상적인 관측치(inliers)를 $x_i \sim U(0, 15)$, $\varepsilon_i \sim N(0, 1)$에서 구하고, 이상치 발생을 위하여 $y_i = x_i + 4, x_i = 20 - .05(i-1)$, $i = 26, 27, \ldots, l$ (단, l은 이상치 수)에서 1000회의 모의실험을 한다.

〈표4〉 단순 선형 회귀 모형에서 모의실험 비교표

이상치의 위치와 개수	OLS	Hadi-Simonoff	Hadi(1992)	ELMS
HL(1)	0	0	6	1
HL(3)	0	0	144	35
HL(5)	0	0	326	147
HL(10)	10	10	635	672
LL(1)	0	0	30	0
LL(3)	0	0	119	4
LL(5)	0	0	284	39
LL(10)	0	0	556	615

(주) 1. HL(n)은 High 지렛대 점인 이상치이고, n은 이상치 개수.
2. LL(n)은 Low 지렛대 점인 이상치이고, n은 이상치 개수.
3. 유의수준 $\alpha = 0.05$를 사용함.

(2) 다중 선형 회귀 모형에서 모의실험

다중 선형 모형일 때 $y_i = \beta_o + \beta_1 x_{i_1} + \beta_2 x_{i_2} + \varepsilon_i$, $i = 1, 2, \ldots, 25$. $\beta_o = 0$, $\beta_1 = 1$, $\beta_2 = 1$에서 정상적인 관측치(inliers)를 $x_{i_1} \sim U(0, 15)$, $x_{i_2} \sim U(0, 15)$, $\varepsilon_i \sim N(0, 1)$에서 구하고, 이상치 발생을 위하여 $y_i = x_{i_1} + x_{i_2} + 4$, $x_{i_1} = 20 - .05\,(i-1)$, $x_{i_2} = 20 - .05(i-1)$, $i = 26, 27, \ldots, l$(단, l은 이상치 수)에서 1000회의 모의실험을 한다.

〈표5〉 다중 선형 회귀 모형에서 모의실험 비교표

이상치의 위치와 개수	OLS	Hadi-Simonoff	Hadi(1992)	ELMS(1992)
HL(1)	0	0	25	17
HL(3)	0	0	292	79
HL(5)	3	0	461	246
HL(10)	280	278	794	791
LL(1)	0	0	45	0
LL(3)	0	0	246	24
LL(5)	0	0	436	126
LL(10)	32	34	620	756

(주) 1. HL(n)은 High 지렛대 점인 이상치이고, n은 이상치 개수.
2. LL(n)은 Low 지렛대 점인 이상치이고, n은 이상치 개수.
3. 유의수준 $\alpha = 0.05$를 사용함.

2. 模擬實驗의 結果 分析

모의실험을 위하여 발생시킨 자료는 정상치가 절대 잔차의 크기는 거의 2보다 크지 않은 값을 갖고 있으며, 계획된 이상치는 거의 대부분 4만큼 떨어진 곳에 위치하고 있다. 따라서 대부분의 이상치는 식별되어야만 정상적이라 할 수 있으나, 극단적으로 좌표 (0,-2)부근과 (25,27)근방에 정상치가 몰려 있는 경우는 계획된 회귀모형에서 이상치는 약 3 정도의 거리를 갖게 된다. 따라서 이와 같은 자료가 발생 할 경우에 이상치의 식별은 매우 어려운 것으로 알려지고 있다(김종우,1996). 모의실험의 결과 비교에서 High Leverage와 Low Leverage 모두에서 OLS, Hadi와 Simonoff(1993)는 매우 우수한 결과를 보이고 있으며, Hadi(1992), ELMS의 방법은 초기 기본 집합 구성에서 이상치를 다수 포함하는 결과를 나타내고 있다. 그러나 김종우(1996)에서 나타난 모의실험의 결과인 〈표4〉,〈표5〉에서 제시하고 있는 것과 같이 자료에 포함되어 있는 이상치의 식별에서는 OLS 방법은 매우 어려움을 보이고 있다.

V. 結　論

이상치를 파악하는 것은 자료의 분석에서 매우 중요한 의미를 갖는다. 2 단계 방법의 사용은 기본집합의 설정과 이를 사용하여 이상치를 식별하는 방법이다. 1단계에서 기본집합을 결정하기 위한 초기 기본집합을 구하는 방법은 앞서 제시된 것과 같이 잔차를 이용하는 방법을 사용한다. 이러한 이상치 파악 방법은 김종우(1996), Hadi와 Simonoff(1993) 등에서 다중이상치 식별을 위한 여러 가지 방법들 중에 우수함을 보이고 있다. 그러나 이 방법은 초기 기본집합과 표본에 따른 회귀추정량 $\hat{\beta}_J$를 결정할 때 사용하는 기본집합의 이상치에 대한 안정성에 크게 의존하고 있다. 따라서, 본 연구에서는 초기 기본집합을 결정하는 방법들의 비교를 위하여 OLS, Haidi(1992), Hadi-Simonoff(1993), ELMS(김종우,1996)에서 사용된 방법들의 비교를 하였다. 선형 회귀 모형에서 초기 기본 집합 구성을 위하

여 OLS 방법은 다른 방법들에 비하여 가장 빠르고 안정적으로 이상치를 배제시키는 것이 모의실험을 통하여 입증되었다. 그러나 자료 전체의 이상치 식별을 위한 방법에서 김종우(1996), Hadi-Simonoff(1993), Rousseeuw and Leroy(1987)은 이상치의 수가 증가함에 따라 급격하게 초기 기본집합에 이상치의 포함율은 높아지고 있으며, 특히 다변량의 경우에는 매우 비효율적임이 드러나고 있다. 따라서 Hadi-Simonoff(1993)에서 사용한 표준화 잔차에 의한 초기 기본집합 구성은 김종우(1996)에서 나타난 모의실험 결과와 비교에서와 같이 매우 효율적임을 알 수 있다. 그러나 이러한 초기 기본집합의 구성은 김종우(1996)에서 지적하고 있는 것과 같이 초기 기본 집합의 원을 하나씩 늘려감에 따라 이상치들을 배제시키는 효과를 갖음을 볼 수 있다. 그러나 이러한 방법에 의하여 얻어진 초기 기본집합에 이상치 포함 가능성은 Hadi와 Simonoff(1993)와의 모의실험에서 안정성이 입증되고 있다.

ELMS에 의한 관측치 반복 추출 방법은 가급적 이상치가 포함되지 않은 기본집합을 얻기 위하여 기본집합에 포함되어진 관측치들 중에 이상치가 포함되어 있을 경우에 최소 중위수를 갖는 기본집합을 재 선정하고, 자료상의 공선성을 피하기 위한 방법으로 기본집합의 크기를 늘리고 있다. 따라서 초기 기본 집합의 구성은 표준화 잔차에 의한 방법을 사용하고 이 집합에 포함되어 있을 가능성이 있는 이상치는 ELMS 방법에 의하여 배제시키는 것이 효과적으로 여겨진다. ELMS에 의한 방법은 기존의 기본집합에 이상치가 포함되어 있을 경우를 위하여 나머지집합의 관측치들을 하나씩 포함시킨 각 기본집합들 중에 최소 최대 절대 표준화잔차 minimize ($\max|r_i|$)$_j$를 갖는 기본집합을 선택한다. 이러한 방법에 의한 효과는 염준근외(1995)에서 효율적임이 보여지고 있다.

초기 기본 집합에 이상치를 가급적 포함시키지 않으려는 노력은 이상치 식별을 위한 2단계 방법에서 기본집합의 안정성에 큰 영향을 미친다. 그러나 Rousseeuw(1984)가 제시하고 있는 LMS 방법은 매우 비효율적이다. 따라서 연산양을 줄이고 안정적으로 이상치를 배제시키는 표준화잔차에 의한 초기 기본 집합의 구성은 매우 효과적이라 하겠다.

참 고 문 헌

o 參考 論文

〔1〕 김종우(1996), "다수 이상치 식별을 위한 2단계 방법에 관한 연구", 동국대, 박사학위논문.

〔2〕 염준근, 박종구,김종우(1995), "다변량 자료에서 다수 이상치 인식의 절차", 품질경영학회지, 제23권, 제4호, pp. 28-41.

〔3〕 Atkinson, A. C.(1986), "Masking Unmasked," *Biometrika*, Vol.73, No.3, pp.533-541.

〔4〕 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.

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

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

〔8〕 ----(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.

〔9〕 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.

〔10〕 Hawkins, D. M.(1991), "Diagnostics for Use With Regression Recursive Residuals," *Techometric*, Vol.33, No.2, pp. 221-234.

〔11〕 Huber, P. J.(1964), "Robust estimation of a location parameter," *Annal Statistics*,Vol.46, pp. 33-50.

[12] 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.

[13] Lund, R. E.(1975), "Tables for An Approximate Test for Outliers in Linear Models," *Techometric*, Vol.17, No.4, pp. 473-476.

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

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

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

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

[18] ----(1983), "Percentage Points for a Generalized ESD Many Outlier Procedure," *Techometric*, Vol.25, No.2, pp.165-172.

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

[20] 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.

[21] Tietjen, G. L., R. H. Moore, and R. J. Beckman(1973), "Testing for a Single Outlier in Simple Linear Regression," *Techometric*, Vol.15, pp.717-721.

[22] Woodruff, D. L., and Rocke, D. M. (1994), "Computable Robust

Estimation of Multivariate Location and Shape in High Dimension Using Compound Estimation,"*Journal of the American Statistical Association*, Vol.89, No 427, pp.888-896.

o 參考圖書

〔23〕 Atkinson, A. C.(1985), *Plots, Transformations, and Regression*, Clarendon Press, London.

〔24〕 Cook, R. D. and Weisberg, S.(1982), *Residuals and Influence in Regression*, Chapman and Hall.

〔25〕 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.

〔26〕 Rousseeuw, P. J. and Leroy, A. M.(1987), *Robust Regression and Outlier Detection*, John Wiley & Sons, New York.