The Korean Communications
in Statistics Vol. 12 No. 3, 2005
pp. 625-634

V-mask Type Criterion for Identification of Outliers in Logistic Regression[1)]

Bu-Yong Kim[2)]

Abstract

A procedure is proposed to identify multiple outliers in the logistic regression. It detects the leverage points by means of hierarchical clustering of the robust distances based on the minimum covariance determinant estimator, and then it employs a V-mask type criterion on the scatter plot of robust residuals against robust distances to classify the observations into vertical outliers, bad leverage points, good leverage points, and regular points. Effectiveness of the proposed procedure is evaluated on the basis of the classic and artificial data sets, and it is shown that the procedure deals very well with the masking and swamping effects.

Keywords: logistic model, outlier, robust distance, clustering, V-mask

1. 서 론

데이터마이닝 분야에서 고객유치나 고객유지를 위한 고객성향 분석 및 고객세분화, 광고효과의 예측, 기업이나 개인의 신용평가, 리스크분석, 이탈예상 고객 판단 등에 로지스틱회귀분석이 많이 활용되고 있다. 그런데 데이터마이닝 분야에서의 자료 수집과정은 엄격하게 통제되지 않는 경우가 많기 때문에 로지스틱회귀분석 자료에 이상점이 다수 포함될 가능성이 높다. 자료에 이상점들이 포함되면 로지스틱회귀분석의 결과는 이상점에 의해 크게 왜곡될 수밖에 없으므로, 자료정제 과정에서 이상점을 식별하고 적절한 대책을 강구해야 한다. 예로서, SAS/E-miner는 이상점으로 판단되는 관찰치를 식별하여 제거하거나 대체한 후 모형구축을 할 수 있도록 하는데, 각 변수별로 극단적으로 크거나 작은 값을 이상점으로 식별하는 방법들을 채택하고 있다. 그러나 설명변수 각각에 대하여 식별한 이상점이 로지스틱회귀모형을 전제로 한 이상점과 일치하지 않을 수 있다. 더구나 제시된 식별방법들은 주관적인 판단에 전적으로 의존하고 있다. 따라서 로지스틱회귀에서의 이상점 식별에 효과적으로 적용할 수 있는 객관적인 방법의 개발이 요구된다.

로지스틱회귀에서의 이상점 식별에 관한 연구는 Pregibon(1981)이 선형회귀 진단방법을 확장하여 로지스틱회귀에 적용하려는 시도를 한 후 그다지 활발하게 진행되지 않았다. 그 이유는 지렛점이 포함될 가능성이 낮은 실험자료의 분석에 로지스틱회귀분석이 주로 활용되어 왔고, 잔차는 -1

1) This research was supported by Sookmyung Women's University Research Grants (2004).
2) Professor, Department of Statistics, Sookmyung Women's University, Seoul 140-742, Korea.
E-mail: buykim@sookmyung.ac.kr

과 +1 사이의 값으로 한정되기 때문에 분석결과에 미치는 이상점의 영향이 대수롭지 않다고 여긴 때문으로 분석된다. 그러나 데이터마이닝 분야에서의 관찰자료에는 이상점이 포함되는 경우가 많고, 로지스틱회귀에서 일반적으로 사용되는 최대우도추정법은 이상점에 의해 상당히 많은 영향을 받는다는 사실이 실제 자료들로부터 확인되었다. 더욱이 Jennings(1986)는 선형회귀에서의 회귀진단법을 로지스틱회귀에 확장하여 적용하는 것은 적절치 않다고 하였다. 따라서 본 연구에서는 로지스틱회귀에서 이상점 식별을 효과적으로 수행할 수 있는 새로운 방법을 제시하고자 한다.

2. 로지스틱회귀에서의 이상점 식별

로지스틱회귀에서는 주로 최대우도추정법을 사용하는데 추정치를 구하기 위하여 최적화기법인 반복재가중최소제곱추정(iterative reweighted least squares: IRLS) 알고리즘을 적용한다. 최대우도추정치가 이상점에 의해 얼마나 많은 영향을 받는지 확인하기 위하여 <표 1>에 수록된 자료들에 IRLS-알고리즘을 적용하였는데, 이상점들의 영향력이 막대함을 알 수 있었으며 이상점에 의해 영향을 받은 추정치를 바탕으로 한 통계적 추론들은 심각하게 왜곡될 우려가 있었다. 따라서 Pregibon (1982)과 Jennings(1986)가 강조했듯이 로지스틱회귀분석에서 이상점을 식별하는 과정은 매우 중요한 의미를 갖는다.

회귀이상점은 나쁜 지렛점과 수직이상점으로 구분될 수 있는데, 나쁜 지렛점을 식별하기 위해서는 Hadi(1994)와 같은 다변량자료에서의 이상점 식별법을 활용할 수 있다. 그런데 기존의 방법들은 마할라노비스 거리(Mahalanobis distance: MD)를 바탕으로 삼기 때문에 지렛점의 악영향으로 인하여 지렛점을 정확히 식별할 수 없다. 그러므로 MCD(minimum covariance determinant)-추정량과 같은 로버스트 추정량을 도입한 로버스트 제곱거리(robust squared distance: RSD)를 지렛점 식별에 적용하는 방법을 고려할 수 있다. 그런데 RSD의 분포에 바탕을 둔 식별방법들을 적용하면 과도한 식별이 이루어지는 문제점이 있다. 따라서 본 연구에서는 RSD의 계층적 군집화에 의한 방법을 강구하며 최적의 군집나무 경계선 높이를 결정하고자 한다. 한편, 수직이상점을 식별하기 위해서 로버스트 잔차(robust residual: RR)를 도입한 RSD-RR 산점도를 활용할 수 있는데, 이상점을 식별함에 있어서 주관적인 판단에 의존하지 않기 위하여 RSD의 중위수를 출발점으로 한 V-마스크 형태의 경계구역을 제안하고자 한다.

2.1 MCD-추정에 의한 RSD

로지스틱 회귀모형 $Y_i = E(Y_i) + \epsilon_i$, $E(Y_i) = exp(\boldsymbol{x}_i^T\beta)/\{1+exp(\boldsymbol{x}_i^T\beta)\}$ $(i=1,...,n,\ \beta$는 p-벡터, 확률변수 Y_i는 모수 $E(Y_i)=\pi_i$의 베르누이분포를 따름)에서 지렛점을 식별하기 위하여 마할라노비스 제곱거리 $MSD_i = (\boldsymbol{x}_i - \boldsymbol{m})^T S^{-1}(\boldsymbol{x}_i - \boldsymbol{m})$ (여기서 $\boldsymbol{m}$은 위치모수인 평균벡터이고 S는 형태모수인 공분산행렬임)을 활용할 수 있다. Mardia, Kent, and Bibby(1979)는 몇 가지 엄격한 가정 하에서 MSD는 $\chi^2(p)$-분포를 따른다는 것을 증명하였다. 한편 위치모수와 형태모수가 일반적인 방식, $\hat{\boldsymbol{m}} = n^{-1}\sum_{i=1}^{n} \boldsymbol{x}_i$, $\hat{S} = (n-1)^{-1}\sum_{i=1}^{n} (\boldsymbol{x}_i - \hat{\boldsymbol{m}})(\boldsymbol{x}_i - \hat{\boldsymbol{m}})^T$으로 추정되는 경우에 MSD는 베타분포를 따른다는 사실을 Gnanadesikan and Kettenring(1972)이 밝혔다. 즉,

$$\frac{(n-1)^2}{n} MSD \sim B\left(\frac{p}{2}, \frac{n-p-1}{2}\right) \tag{2.1}$$

MSD의 분포 (2.1)를 바탕으로 지렛점을 식별할 수 있지만, $\hat{m}$과 $\hat{S}$은 로버스트 추정량이 아니기 때문에 가림현상이나 붙음현상의 발생으로 인하여 정확한 지렛점 식별을 기대할 수 없다. 따라서 m과 S의 로버스트 추정량인 MCD-추정량이나 MVE(minimum volume ellipsoid)-추정량을 적용함으로써 이러한 현상들을 방지할 수 있다. 조합추정량인 MCD-추정량과 MVE-추정량은 붕괴점이 0.5로서 가장 높은 값을 갖는데, MCD-추정량이 점근적 정규성이라는 통계적 특성을 가지며 실용적 측면에서도 MVE-추정량보다 우수하다는 사실이 Rocke and Woodruff(1997)에 의해 제시되었으며 Hardin and Rocke(2004)도 MCD-추정량을 사용할 것을 추천하였다. 따라서 본 연구에서는 RSD를 정의하기 위하여 MCD-추정량을 채택하였다.

Rousseeuw(1985)가 제시한 MCD-추정량은 사전에 결정된 크기의 관찰치 부분집합들 중에서 공분산행렬의 행렬식이 최소가 되는 부분집합을 찾아서 그 부분집합에서의 평균과 공분산을 추정함으로써 얻을 수 있다. MCD-추정량 $(\hat{m}^*_J, \hat{S}^*_J)$을 MSD에 적용한 RSD는 다음과 같이 정의된다.

$$RSD_i = (x_i - \hat{m}^*_J)^T \hat{S}^{*-1}_J (x_i - \hat{m}^*_J), \tag{2.2}$$

$$\hat{m}^*_J = \frac{1}{h}\sum_{i\in J} x_i,\ \hat{S}^*_J = \frac{1}{h}\sum_{i\in J} (x_i - \hat{m}^*_J)(x_i - \hat{m}^*_J)^T,$$

J={h개의 원소로 구성된 집합: 원소의 수가 h인 모든 집합 K에 대해 $|\hat{S}_J{}^*| \leq |\hat{S}_K{}^*|$}.

여기서 h는 이상점이 포함되지 않은 관찰치들로 구성된 half-sample의 최소 크기를 의미한다. 그런데 (2.2)에서 정확한 MCD-추정량을 구하기 위해서는 모든 half-sample에 대해 공분산행렬의 행렬식을 계산해야 하기 때문에 막대한 계산이 요구된다. 따라서 계산효율성이 향상된 MCD-추정 알고리즘을 Rousseeuw and Driessen(1999)과 Hardin and Rocke(2004)이 제안하였으며 Woodruff and Rocke(1994)는 다차원 자료에서도 계산효율성이 높은 알고리즘을 제시하였다. 한편, MCD-추정량은 $h = [(n+p+1)/2]$ ($[\cdot]$는 최대정수 함수임)일 때 최대의 붕괴점을 갖는다는 사실이 Rousseeuw and Leroy(2003)에 의해 밝혀졌으므로 본 연구에서는 이 값을 h의 크기로 채택하였다.

2.2 RSD의 군집화에 의한 지렛점 식별

로지스틱회귀에서 지렛점을 식별하는 문제는 설명변수들로 구성된 다변량자료에서 이상점을 식별하는 문제와 동일하다고 할 수 있는데, 다변량자료에서 이상점을 식별하는 방법으로는 Rocke and Woodruff(1996), Kosinski(1999), Becker and Gather(1999), Viljeon and Venter(2002), Hardin and Rocke(2004) 등이 있다. (2.2)에 정의된 RSD를 지렛점 식별에 적용하기 위해서는 적절한 경계치를 결정해야 하는데, Hardin and Rocke (2004)는 RSD의 근사적 분포를 바탕으로 한 경계치를 제시하였다. 즉, MCD-추정량에 의해 정의된 RSD의 분포,

$$\frac{c(a-p+1)}{pa} RSD \sim F(p, a-p+1) \tag{2.3}$$

로부터 얻은 경계치 $\{pa/c(a-p+1)\}F_{1-\alpha}(p, a-p+1)$을 적용하여 지렛점을 식별할 것을 제안하였다. 그러나 이 경우 지렛점이 자연스럽게 구분되지 않는 상황에서도 경계치보다 큰 점들을 지렛점으로 과도하게 식별하는 문제가 발생한다. 더욱이 지렛점이 존재하지 않는 경우에도 RSD가 상대적으로 큰 점들을 지렛점으로 판단하는 오류를 범하게 된다. 따라서 정상점들로부터 자연스럽게 구분되는 점들만을 지렛점으로 식별하도록 RSD의 군집화에 의한 식별방법을 제안한다.

사전에 군집의 수를 지정해야 하는 군집화 방법은 사용할 수 없기 때문에, RSD를 군집화하기 위하여 단일결합군집법에 의한 계층적 군집화를 채택하였다. 군집화 결과를 바탕으로 정상점과 지렛점을 분류하기 위해서는 적절한 경계선 높이가 결정되어야 하는데, 본 연구에서의 경계선 높이는 Kim and Oh(2004)가 제시한 방법: $u^* = \bar{u} + \eta s_u$ (u^*는 경계선 높이, $\bar{u}$는 $n-1$개의 모든 군집에 대응하는 군집높이의 평균, s_u는 군집높이의 표준편차 추정량임)에 의해 결정하였다. 한편, <표 1>에 수록된 자료들의 RSD에 대해 계층적 군집나무를 구성하였으며, 경계선높이 u^*를 적절한 수준에서 결정하기 위한 η값의 최적치는 2.8이라는 사실을 밝혀냈다. 이 방법에 의한 지렛점 식별결과는 <표 1>에 수록되었는데, 대부분의 경우 지렛점을 정확히 식별하는 것으로 나타났다.

2.3 로버스트추정 알고리즘

IRLS-알고리즘에 의한 추정량은 지렛점에 의해 많은 영향을 받는다는 사실이 밝혀졌으므로, 지렛점의 영향을 적게 받는 로버스트 추정량을 구하기 위해서 IRLS-알고리즘을 수정해야 할 필요가 있다. 따라서 군집화에 의해 지렛점으로 식별된 관찰치에 RSD 크기에 역비례하는 가중치를 부여하는 방법을 도입하여 IRLS-알고리즘을 수정하였다.

Algorithm: RIRLS

<단계 1> RSD의 군집화를 바탕으로 지렛점을 식별하고, 가중치행렬 $W = diag[w_1, ..., w_n]$: $w_i = 1$ for $i \in A$, $w_i = w/RSD_i$ for $i \notin A$ (단, A는 정상점들의 지수집합을, w는 A에 속한 RSD 중에서 최대치를 의미함)을 구한 후, 설명변수 행렬을 $Z = WX$로 변환한다.

<단계 2> 반복수 $t = 0$을 지정하고, 회귀계수의 초기치 $\hat{\beta}^{(0)}$을 선정한다.

<단계 3> $\hat{\pi}_i^{*(t)} = z_i^T \hat{\beta}^{(t)}$, $\hat{\pi}_i^{(t)} = exp(\hat{\pi}_i^{*(t)})/[1 + exp(\hat{\pi}_i^{*(t)})]$을 계산한다.

<단계 4> 새로운 반응변수 값 $y_i^{*(t)} = \hat{\pi}_i^{*(t)} + (y_i - \hat{\pi}_i^{(t)})/v_i^{(t)}$, $v_i^{(t)} = \hat{\pi}_i^{(t)}(1 - \hat{\pi}_i^{(t)})$을 계산하고 가중치 행렬 $V^{(t)} = diag[v_1^{(t)}, ..., v_n^{(t)}]$을 구성한다.

<단계 5> 새로운 추정치 $\hat{\beta}^{(t+1)} = (Z^T V^{(t)} Z)^{-1} Z^T V^{(t)} y^{*(t)}$을 구한다.

<단계 6> 만약 $\| \hat{\beta}^{(t+1)} - \hat{\beta}^{(t)} \|_\infty \leq \delta$이면 ($\| \cdot \|_\infty$는 L_∞-norm, δ는 tolerance로서 아주 작은 양의 수임) 알고리즘을 종료하고, 그렇지 않으면 추정치를 최신화하고 <단계 3>으로 간다.

2.4 V-마스크 경계구역에 의한 이상점 식별

산점도를 바탕으로 이상점을 식별할 경우 흔히 주관적 판단에 의존하게 되는데, 본 연구에서는 객관적인 식별기준을 설정하고자 한다. 선형회귀와는 달리 로지스틱회귀의 RSD-RR 산점도는 RSD가 커짐에 따라 정상점에 대응하는 잔차는 0으로 수렴하고, 수직이상점이나 나쁜 지렛점에 대응하는 잔차는 ±1 근처에 산포되는 특징을 갖는다. 따라서 산점도의 특이한 형태를 반영한 V-마스크 형태의 경계구역을 이상점 식별기준으로 적용하는 것이 타당하다고 판단하였다.

RSD-RR 산점도상에서 V-마스크의 좌측 출발점을 (0, -1)과 (0, 1)로 선정하면 자료의 중앙에 위치한 정상점들이 수직이상점으로 식별되는 오류를 범하기 때문에 출발점을 적절한 위치로 이동시켜야 한다. 우선 적절한 출발점을 선정하기 위해 RSD의 분포를 활용할 수 있다. 즉, RSD의 분포 (2.3)에서 RSD의 기댓값은 $pa/\{c(a-p-1)\}$인데 이 값을 λ라 하면, (λ, -1)과 (λ, 1)을 V-마스크의 출발점으로 선정할 수 있다. 그러나 (2.3)은 근사분포이기 때문에 관찰치의 수가 크지 않은 경우에는 λ값이 부적절하게 얻어지는 경우가 있다. 따라서 본 연구에서는 V-마스크의 출발점을 결정하기 위하여 $\lambda = median(RSD_i)$을 적용하였다. 한편, 지렛점 중에서 좋은 지렛점의 잔차 절대값은 0.5보다 크지 않으며, 나쁜 지렛점의 잔차 절대값은 0.5보다 크기 때문에 대응하는 잔차의 크기가 ±0.5를 벗어나는지에 따라 좋은 지렛점과 나쁜 지렛점이 구분된다. 따라서 정상점으로 판정된 관찰치의 RSD중에서 최대치인 w와 λ를 기준으로 V-마스크를 형성할 수 있다. 즉, 출발점 (λ, -1)과 (w, -0.5)를 연결하는 직선 e_1을 하단 선으로, (λ, 1)과 (w, 0.5)를 연결하는 직선 e_2을 상단 선으로 하는 V-마스크를 다음과 같이 설정한다.

$$e_1 = \frac{1}{2(w-\lambda)} RSD - \frac{w-\lambda/2}{w-\lambda} \tag{2.4}$$

$$e_2 = \frac{-1}{2(w-\lambda)} RSD + \frac{w-\lambda/2}{w-\lambda} \tag{2.5}$$

두 직선 (2.4)와 (2.5)의 내부를 V-마스크 경계구역이라 부를 수 있는데, RSD가 0과 w사이에 있지만 V-마스크를 벗어나는 관찰치는 수직이상점으로 식별하고, RSD가 w보다 크면서 잔차가 ±0.5 밖에 위치하는 관찰치는 나쁜 지렛점으로 식별한다. 반면에 RSD가 w보다 크지만 잔차가 ±0.5 안에 위치하는 관찰치는 좋은 지렛점으로 식별한다. 이와 같이 관찰치들을 정상점, 수직이상점, 좋은 지렛점, 나쁜 지렛점으로 분류할 수 있다는 것이 제안된 식별방법의 우수한 특성이다.

3. 이상점 식별방법의 평가

제안된 식별방법을 <표 1>에 수록된 자료들에 적용하여 얻은 RSD-RR 산점도와 V-마스크가 <그림 1>~<그림 7>에 제시되었다. <표 1>에 설명된 [자료 A-1]의 산점도인 <그림 1>를 살펴보면 자료에 지렛점이 포함되지 않은 것으로 판단되며, {10, 41}-번 관찰치의 RR이 V-마스크를

벗어나기 때문에 2개의 관찰치가 수직이상점으로 정확히 식별된다. [자료 A-2]의 <그림 2>에서는 {1, 50}-번 관찰치의 RSD가 w와 일치하지만 RR이 ± 0.5를 벗어나므로 수직이상점으로 식별된다. [자료 B-1]에 대한 RSD-RR 산점도인 <그림 3>에서는 24-번 관찰치가 나쁜 지렛점으로, 23-번 관찰치는 수직이상점으로 식별되며, {9, 22, 27}-번 관찰치는 좋은 지렛점으로 판단된다. [자료 B-3]에 대한 <그림 4>에서는 {40, 41}-번 관찰치가 강력한 나쁜 지렛점으로 명확히 식별된다. [자료 C]에 대한 RSD-RR 산점도인 <그림 5>에서는 {91, 97}-번 관찰치가 수직이상점으로, 100번째 관찰치는 좋은 지렛점으로 식별된다. 그리고 [자료 D-1]에서는 {133, 147, 171 183}-번 관찰치가 수직이상점으로, {68, 76, 93, 106}-번 관찰치는 좋은 지렛점으로 식별된다. [자료 E]의 <그림 7>에서는 87번 관찰치가 수직이상점으로, 40번 관찰치는 좋은 지렛점으로 식별된다. 각각의 자료에 대하여 이상점을 식별한 결과는 <표 1>에 수록되었는데, 제안된 V-마스크 방법이 이상점을 정확히 식별하는 것으로 판단된다.

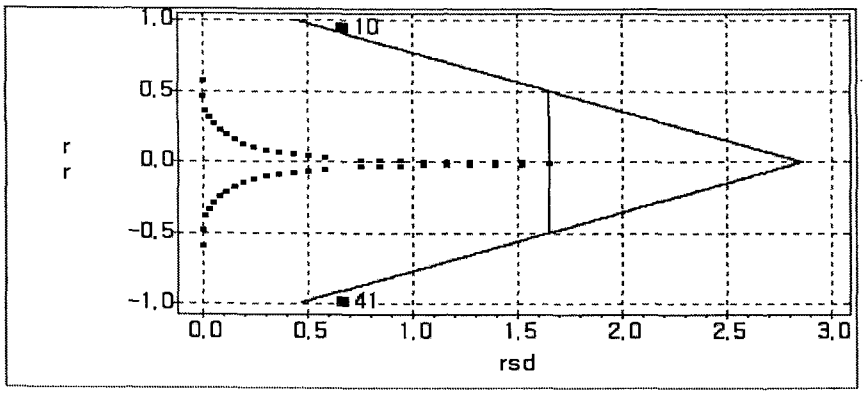
4. 결 론

로지스틱회귀분석에서 이상점은 제반 통계적 추론에 막대한 영향을 미치기 때문에 분석 결과가 심하게 왜곡된다. 따라서 데이터마이닝을 위한 로지스틱회귀분석에 앞서 자료에 이상점이 존재하는지를 확인하고 어느 관찰치가 이상점인지 식별하여 자료정제 과정에서 필요한 조치를 취하거나 로버스트 추정을 적용해야 한다. 본 연구는 로지스틱회귀에서의 이상점 식별을 위한 새로운 방법을 제시하였는데, RSD의 계층적 군집화에 의해 지렛점을 식별하고 이 식별결과를 바탕으로 로버스트 잔차를 구하고 RSD-RR 산점도에 V-마스크 형태의 경계구역을 적용하여 이상점을 식별하는 방법이다. 이 방법은 이상점식별에 매우 효과적인 것으로 평가되었는데, 특히 관찰치를 수직이상점과 나쁜 지렛점 그리고 좋은 지렛점과 정상점으로 분류할 수 있다는 장점을 가지고 있다.

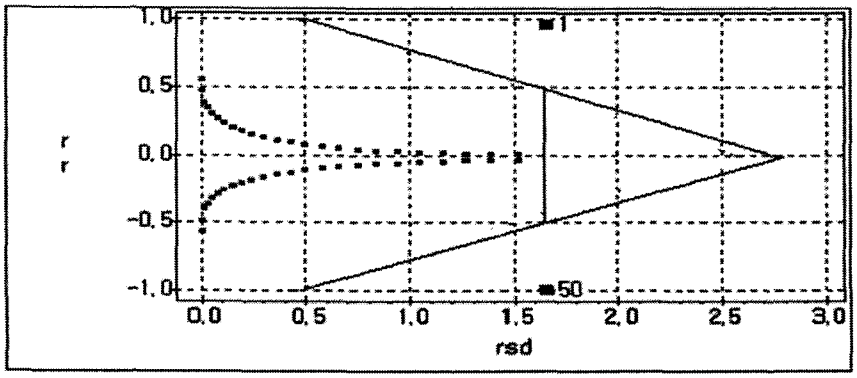
<표 1> 연구대상 자료의 특성 및 이상점 식별 결과

자료	자료의 특성	이상점 식별		
		좋은 지렛점	나쁜 지렛점	수직 이상점
A	인공자료: $x_i = i$ for $i = 1, 2, ..., 50$ $y_i = \begin{cases} 0 \text{ for } i = 1, ..., 23, 25, 27 \\ 1 \text{ for } i = 24, 26, 28, ..., 50 \end{cases}$	-	-	-
A-1	[자료 A]에 수직이상점 (x_{10}, y_{10})=(10, 1), (x_{41}, y_{41})=(41, 0)을 심은 자료	-	-	10, 41
A-2	[자료 A]에 수직이상점 (x_1, y_1)=(1, 1), (x_{50}, y_{50})=(50, 0)을 심은 자료	-	-	1, 50
A-3	[자료 A]에 좋은 지렛점 (x_1, y_1)=(−30, 0), (x_{50}, y_{50})=(80, 1)을 심은 자료	1, 50	-	-
A-4	[자료 A]에 나쁜 지렛점 (x_1, y_1)=(−30, 1), (x_{50}, y_{50})=(80, 0)을 심은 자료	-	1, 50	-
B	Pregibon(1981)에 수록된 자료	1, 3, 33	-	-

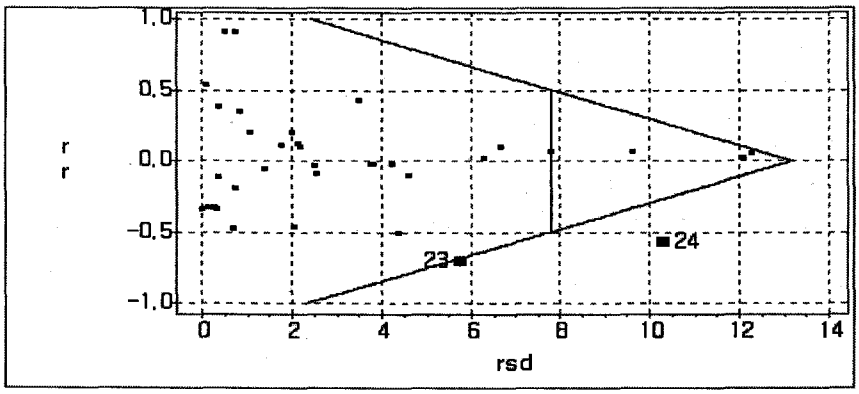
B-1	[자료 B]에서 좋은 지렛점 {1, 3, 33}-번을 제외한 자료	9, 22, 27	24	23
B-2	[자료 B]에 좋은 지렛점 (volume=3.9, rate=3.9, y=1), (volume=4.0, rate=4.0, y=1)을 추가한 자료	40, 41	-	-
B-3	[자료 B]에 나쁜 지렛점 (volume=3.9, rate=3.9, y=0), (volume=4.0, rate=4.0, y=0)을 추가한 자료	-	40, 41	-
C	Hosmer and Lemeshow(1989, p.3)에 수록된 자료	100	-	91, 97
D	Hosmer and Lemeshow(1989, p.247)의 자료에서 반응변수를 'low birth data'로, 설명변수를 'age', 'lwt', 'bwt'로 선정한 자료	23, 39, 68, 76, 93, 106, 130, 133, 147	-	1
D-1	[자료 D]에서 설명변수 'bwt'를 제외시킨 자료	68, 76, 93, 106	-	133, 147, 171, 183
E	김순귀 외(2003, p.214)에 수록된 자료	40	-	87



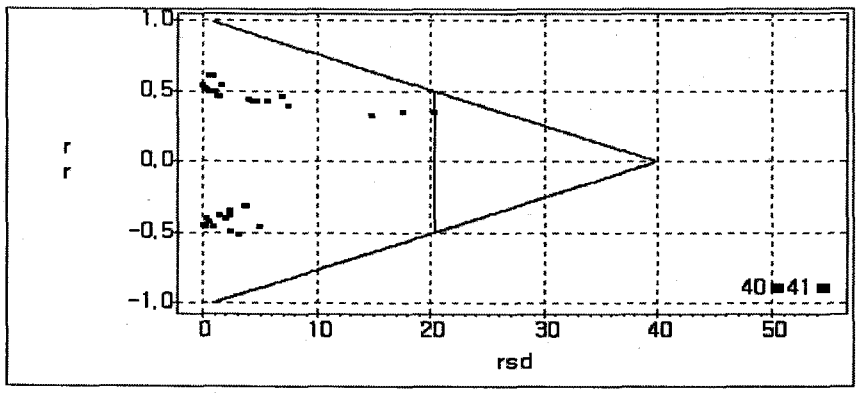
<그림 1> [자료 A-1]에 대한 RSD-RR 산점도 및 V-마스크



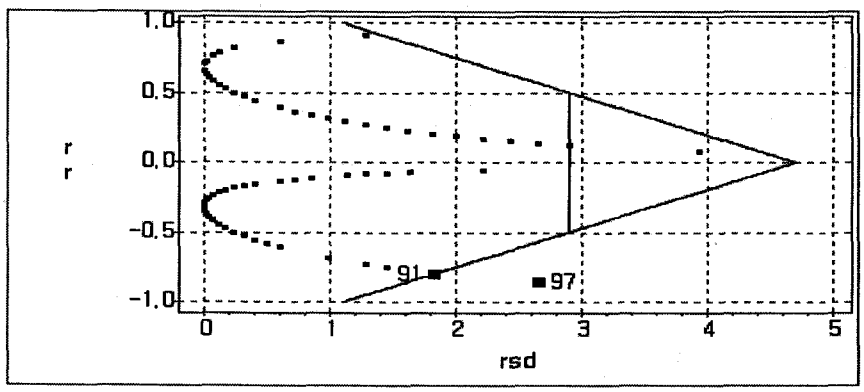
<그림 2> [자료 A-2]에 대한 RSD-RR 산점도 및 V-마스크



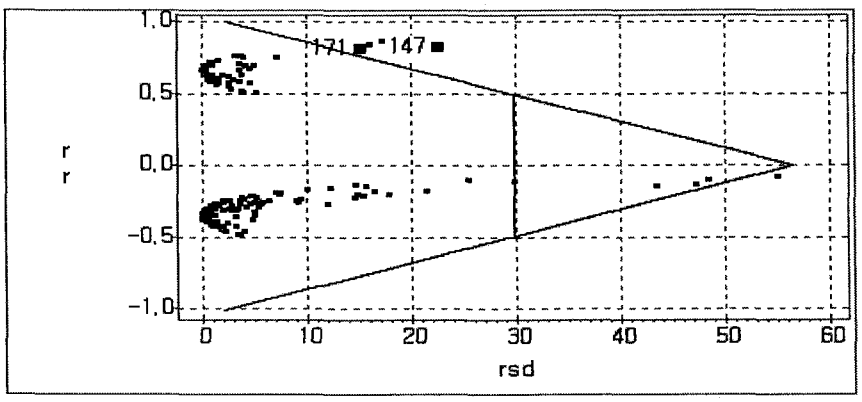
<그림 3> [자료 B-1]에 대한 RSD-RR 산점도 및 V-마스크



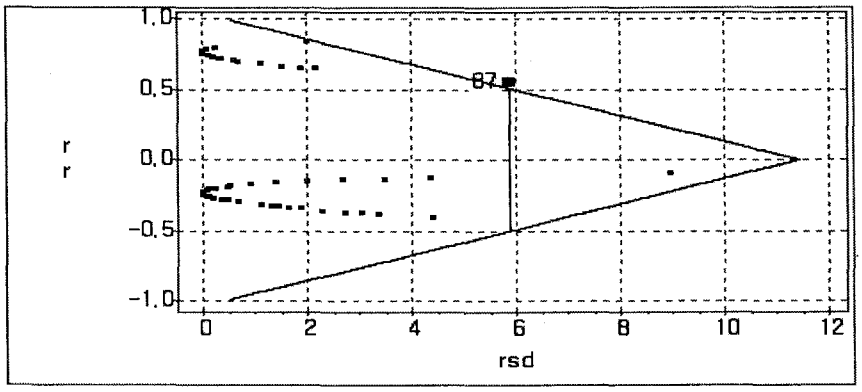
<그림 4> [자료 B-3]에 대한 RSD-RR 산점도 및 V-마스크



<그림 5> [자료 C]에 대한 RSD-RR 산점도 및 V-마스크



<그림 6> [자료 D-1]에 대한 RSD-RR 산점도 및 V-마스크



<그림 7> [자료 E]에 대한 RSD-RR 산점도 및 V-마스크

참고 문헌

[1] 김순귀, 정동빈, 박영술(2003). *SPSS를 활용한 로지스틱회귀모형의 이해와 응용*, 데이터솔루션.

[2] Becker, C. and Gather, U.(1999), The masking breakdown point of multivariate outlier identification rules, *Journal of the American Statistical Association*, Vol. 94, 947-955.

[3] Gnanadesikan, R. and Kettenring, J.(1972). Robust estimates, residuals, and outlier detection with multiresponse data, *Biometrics*, Vol. 28, 81-124.

[4] Hadi, A. S.(1994). A modification of a method for the detection of outliers in multivariate samples, *Journal of the Royal Statistical Society*, Vol. 56, 393-396.

[5] Hardin, J. and Rocke, D. M.(2004). Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator, *Computational Statistics & Data Analysis*, Vol. 44, 625-638.

[6] Hosmer, D. W. and Lemeshow, S.(2000). *Applied Logistic Regression*, John Wiley & Sons.

[7] Jennings, D. E.(1986). Outliers and residual distributions in logistic regression, *Journal of*

the American Statistical Association, Vol. 81, 987-990.

[8] Kim, B. Y. and Oh, M. H.(2004). Identification of regression outliers based on the clustering of LMS-residual plots, *The Korean Communications in Statistics*, Vol. 11, 485-494.

[9] Kosinski, A. S.(1999). A procedure for the detection of multivariate outliers, *Computational Statistics & Data Analysis*, Vol. 29, 145-161.

[10] Mardia, K., Kent, J. and Bibby, J.(1979). *Multivariate Analysis*, Academic Press.

[11] Pregibon, D.(1981). Logistic regression diagnostics, *The Annals of Statistics*, Vol. 9, 705-724.

[12] Pregibon, D.(1982). Resistant fits for some commonly used logistic models with medical applications, *Biometrics*, Vol. 38, 485-498.

[13] Rocke, D. M. and Woodruff, D. L.(1996). Identification of outliers in multivariate data, *Journal of the American Statistical Association*, Vol. 91, 1047-1061.

[14] Rocke, D. M. and Woodruff, D. L.(1997). Robust estimation of multivariate location and shape. *Journal of Statistical Planning and Inference*, Vol. 57, 245-255.

[15] Rousseeuw, P. J.(1985). Multivariate estimation with high breakdown point, *Mathematical Statistics and Applications*, Vol. B, eds. W. Grossmann, G. Pflug, I. Vincze, and W. Werz.

[16] Rousseeuw, P. J. and Driessen, K.(1999). A fast algorithm for the minimum covariance determinant estimator, *Technometrics*, Vol. 41, 212-223.

[17] Rousseeuw, P. J. and Leroy, A. M.(2003). *Robust Regression and Outlier Detection*, Wiley-Interscience.

[18] Viljoen, H. and Venter, J. H.(2002). Identifying multivariate discordant observations: a computer-intensive approach, *Computational Statistics & Data Analysis*, Vol. 40, 159-172.

[19] Woodruff, D. L. and Rocke, D. M.(1994). Computable robust estimation of multivariate location and shape in high dimension using compound estimators, *Journal of the American Statistical Association*, Vol. 89, 888-896.

[2005년 4월 접수, 2005년 9월 채택]