응용통계연구 제 20권 3호, 2007년, pp. 551–559

로지스틱회귀모형의 로버스트 추정을 위한 알고리즘*

김부용[1] 강명욱[2] 최미애[3]

요 약

로지스틱회귀에서 일반적으로 사용되는 최대우도추정법은 이상점에 대해 로버스트하지 않다. 따라서 본 논문에서는 로지스틱회귀모형의 로버스트 추정을 위한 알고리즘을 제안하고자 한다. 이 알고리즘은 V-마스크 형태의 경계기준에 의해 나쁜 지렛점과 수직이상점을 식별하고, 식별결과를 바탕으로 이상점의 영향력을 감소시키기 위한 효과적인 방안을 모색한다. 이상점의 영향력 감소는 가중치와 조정치를 적절히 선정함으로 가능하며, 그 결과 붕괴점이 높은 추정치를 얻게 된다. 제안된 알고리즘을 다양한 자료에 적용하여 정분류율을 측정하여 비교하였는데, 새로운 알고리즘이 최대우도추정보다 정확한 분류를 해 주는 것으로 평가되었다.

주요용어: 로지스틱회귀, 이상점 식별, V-마스크 경계기준, 로버스트 추정.

1. 서론

실험자료 분석에 주로 활용되던 로지스틱 회귀분석이 최근 데이터마이닝 분야에서 많이 활용됨에 따라 이상점의 영향력에 관심을 기울이게 되었다. 데이터마이닝 분야에서의 자료 수집과정은 엄격하게 통제되지 않는 경우가 많기 때문에 자료에 이상점이 다수 포함될 가능성이 높은데, 자료에 이상점들이 포함되어 있음에도 불구하고 최대우도추정법을 적용하여 얻은 로지스틱회귀분석 결과는 이상점에 의해 크게 왜곡된 것일 수밖에 없다. 따라서 이상점을 식별하고 이상점의 영향에 지나치게 민감하지 않은 로버스트 추정법을 적용하여 통계적 분석을 실행해야 한다.

선형회귀 분야와 달리 로지스틱회귀에서의 이상점 식별 및 로버스트 추정법에 관한 연구는 활발하게 진행되지 않은 상황이다. 그 이유는 의학이나 생물학 등의 분야에서 이상점의 발생 가능성이 매우 낮은 잘 설계된 실험으로부터 수집된 자료의 분석에 로지스틱회귀분석이 주로 활용되어 왔기 때문이다. 로지스틱회귀에서의 이상점식별에 관한 연구로는 Pregibon (1981)과 Jennings (1986)가 있으며, Kim (2005)은 V-마스크 형태의 이상점

* 본 연구는 숙명여자대학교 2006년도 교내연구비 지원에 의해 수행되었음.
1) (140-742) 서울특별시 용산구 청파동 2가, 숙명여자대학교 통계학과, 교수
E-mail: buykim@sm.ac.kr
2) (140-742) 서울특별시 용산구 청파동 2가, 숙명여자대학교 통계학과, 교수
E-mail: mwkahng@sm.ac.kr
3) (443-742) 경기도 수원시 영통구 매탄3동 416, 삼성전자 컴퓨터사업부, 사원
E-mail: miae21.choi@samsung.com

식별 기준을 제시하였다. 한편, 로지스틱회귀에서의 로버스트 추정법에 관한 연구 중에서 Pregibon (1982)은 이탈도함수 대신에 Huber-type 손실함수와 유사한 미분가능한 단조 비감소함수를 대체하여 이상점에 덜 민감한 최대우도추정치를 얻는 방법을 제안하였다. 그러나 Copas (1988)는 Pregibon 추정치에 매우 큰 편의가 존재한다는 사실을 주장하고, 오분류모형에 바탕을 둔 최대우도추정치를 제안하였다. 그리고 Carroll과 Pederson (1993)은 Copas가 제안한 추정량이 불일치추정량에 해당되기 때문에 이를 수정하여 Mallows-type의 추정량을 제안하였으며, Bianco와 Yohai (1996)는 유계함수를 도입하여 Pregibon 추정량이 일치추정량이 되도록 개량하였다. 한편, Kordzakhia 등 (2001)은 이상점의 영향을 완화시키기 위해 M-추정에 바탕을 둔 로버스트 추정법을 제시하기도 하였다. 이와 같은 추정법들은 대부분 영향력함수 접근법에 이론적 기초를 두고 있는데 실제로 적절한 영향력함수의 선정에는 어려움이 따른다. 따라서 본 연구에서는 V-마스크 형태의 식별방법에 바탕을 둔 로버스트 추정알고리즘을 새롭게 제시하고자 한다.

2. V-마스크에 의한 이상점 식별

로지스틱회귀에서는 주로 최대우도추정법을 사용하는데, Kim (2005)은 최대우도추정치가 이상점에 의해 막대한 영향을 받는다는 사실을 확인하고, 수직이상점과 나쁜 지렛점을 식별할 수 있는 새로운 방법을 제안하였다. 로지스틱 회귀모형 $Y_i = E(Y_i) + \epsilon_i$, $E(Y_i) = \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})/(1+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))$ $(i = 1, \ldots, n$, $\boldsymbol{\beta}$는 p-벡터)에서 지렛점을 식별하기 위해서는 널리 알려진 마할라노비스 제곱거리 (Mahalanobis squared distance: MSD)를 활용할 수 있다. Gnanadesikan과 Kettenring (1972)은 위치모수벡터($\boldsymbol{m}$)와 형태모수행렬($\boldsymbol{S}$)을 일반적인 방법으로 추정하는 경우에 MSD는 베타분포를 따른다는 사실을 밝혔는데, 이러한 MSD의 분포를 바탕으로 지렛점을 식별할 수 있다. 그러나 위치모수와 형태모수의 일반적인 추정량 $\hat{\boldsymbol{m}}$과 $\hat{\boldsymbol{S}}$은 로버스트 추정량이 아니기 때문에 가림현상이나 붙음현상의 발생으로 인하여 정확한 지렛점 식별을 기대할 수가 없다. 따라서 Rousseeuw (1985)가 제시한 붕괴점이 0.5인 MCD (minimum covariance determinant)-추정량을 적용함으로써 이러한 현상들을 방지하여 정확한 식별을 할 수 있다. 사전에 결정된 크기의 관찰치 부분집합들 중에서 공분산행렬의 행렬식이 최소가 되는 부분집합을 찾아서 그 부분집합에서의 위치모수와 형태모수를 추정함으로써 MCD-추정량을 얻을 수 있는데, MCD-추정량 $(\hat{\boldsymbol{m}}_J^*, \hat{\boldsymbol{S}}_J^*)$을 MSD에 적용한 로버스트 제곱거리 (robust squared distance: RSD)는 다음과 같이 정의할 수 있다.

$$RSD_i = (\boldsymbol{x}_i - \hat{\boldsymbol{m}}_J^*)^T \hat{\boldsymbol{S}}_J^{*-1} (\boldsymbol{x}_i - \hat{\boldsymbol{m}}_J^*) \tag{2.1}$$

단,

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

$$J = \left\{ A \mid n(A) = h \text{ and } |\hat{\boldsymbol{S}}_A^*| \le |\hat{\boldsymbol{S}}_K^*| \text{ for all } K \text{ with } n(K) = h \right\}.$$

여기서 J는 h개의 원소로 구성된 집합으로 원소의 수가 h인 모든 집합 K에 대해 $|\hat{\boldsymbol{S}}_J^*| \le |\hat{\boldsymbol{S}}_K^*|$를 만족한다. 또한 h는 이상점이 포함되지 않은 관찰치들로 구성된 half-sample의 최

소 크기를 의미하는데, MCD-추정량은 $h = [(n+p+1)/2]$ ($[\cdot]$는 최대정수 함수임)일 때 최대의 붕괴점을 갖는다는 사실이 Rousseeuw와 Leroy (2003)에 의해 밝혀졌다. 한편, 정확한 MCD-추정치를 구하기 위해서는 막대한 계산이 요구된다는 문제점을 극복하기 위하여 Woodruff와 Rocke (1994)와 Rousseeuw와 Driessen (1999)은 계산효율성이 향상된 MCD-추정 알고리즘을 개발하였다.

MCD-추정량에 의한 RSD의 근사적 분포를 바탕으로 Hardin과 Rocke (2004)는 지렛점 식별을 위한 경계치를 제시하였다. 그러나 지렛점이 자연스럽게 구분되지 않는 상황에서도 경계치보다 큰 점들을 지렛점으로 과도하게 식별하는 문제가 발생한다. 그러므로 본 논문에서는 정상점들로부터 자연스럽게 구분되는 점들만을 지렛점으로 식별하도록 RSD의 계층적 군집화에 의한 식별방법을 제안한다. 즉, 로버스트 추정량인 MCD-추정량을 도입한 로버스트 제곱거리를 계층적으로 군집화 하여 지렛점을 식별하는 방법이다.

일단 식별된 지렛점에 대해서는 적절한 가중치를 부여함으로서 지렛점의 영향을 적게 받는 추정치를 구할 수 있으며, 이 추정치를 바탕으로 로버스트 잔차 (robust residual: RR)를 얻게 된다. RSD를 가로축으로 하고 세로축에 RR을 플롯한 산점도를 RSD-RR 산점도라고 할 수 있는데, 로지스틱회귀의 RSD-RR 산점도의 특징을 감안하여 RSD의 중위수를 출발점으로 지정한 V-마스크 형태의 경계구역을 다음과 같이 설정한다. 즉,

$$l_1 = \frac{1}{2(r-\lambda)} RSD - \frac{r-\lambda/2}{r-\lambda}, \tag{2.2}$$

$$l_2 = \frac{-1}{2(r-\lambda)} RSD + \frac{r-\lambda/2}{r-\lambda}. \tag{2.3}$$

여기서 $\lambda = \text{median}(RSD_i)$이고, r는 정상점으로 판정된 관찰치의 RSD중에서 최대치에 해당된다. 두 경계선 (2.2)와 (2.3)의 내부를 V-마스크 경계구역이라 할 수 있는데, 이와 같은 경계구역을 적용하면 수직이상점과 나쁜 지렛점은 물론 좋은 지렛점도 동시에 식별할 수 있다. 즉, RSD가 0과 r사이에 있지만 V-마스크를 벗어나는 관찰치는 수직이상점으로 식별하고, RSD가 r보다 크면서 잔차가 ± 0.5 밖에 위치하는 관찰치는 나쁜 지렛점으로 식별한다. 반면에 RSD가 r보다 크지만 잔차가 ± 0.5 안에 위치하는 관찰치는 좋은 지렛점으로 식별한다.

3. 로버스트 추정알고리즘

최대우도추정량은 가장 효율적인 추정량으로 알려졌지만 이상점이 존재하는 경우에는 매우 왜곡된 현상을 나타낸다 (Croux와 Haesbroeck, 2003). Kim (2005)은 로지스틱회귀 추정에서 이상점의 악영향이 크다는 사실을 자료의 분석을 통하여 확인하였으며, 새로운 식별법에 의하여 다양한 특성을 갖는 이상점들을 효과적으로 식별할 수 있음을 입증하였다. 따라서 식별된 이상점들에 적절한 가중치와 조정치를 부여하는 방식으로 최대우도추정법을 수정함으로써 이상점에 민감하지 않은 추정치를 얻을 수 있는 알고리즘을 개발하고자 한다.

로지스틱회귀에서의 우도방정식은 비선형이므로 추정치를 직접 구할 수 없기 때문에 최적화기법 중의 하나인 반복재가중최소제곱 (iteratively reweighted least square: IRLS) 추정법을 적용해서 추정치를 구하는 것이 일반적이다. 따라서 본 논문에서는 IRLS를 수정하여 로버스트 추정치를 얻을 수 있는 알고리즘을 제안하고자 한다. 우선 수직이상점의 영향을 줄이기 위해서는 IRLS-알고리즘의 각 단계에서 잔차의 함수로 표현되는 적절한 가중치를 적용하여 추정치를 구하는 방식을 채택할 수 있다. 그러나 IRLS-알고리즘에 의한 추정치는 지렛점에 의해서도 많은 영향을 받는다는 사실이 밝혀졌으므로, 특히 나쁜 지렛점의 영향을 적게 받는 로버스트 추정량을 구하기 위해서 IRLS-알고리즘을 적절히 수정해야 할 필요가 있다. 즉, 수직이상점뿐만 아니라 나쁜 지렛점에 대해서도 동시에 로버스트한 추정치를 얻기 위해서는 수직이상점과 나쁜지렛점의 영향력을 구분하여 파악한 후 각각을 조정해야 한다. 따라서 지렛점으로 식별된 관찰치에는 정상점보다 작은 가중치를 부여하되, 정상점들의 최대 RSD인 r을 초과하는 지렛점에 대해서는 RSD의 크기에 역비례하는 가중치를 부여하는 방법을 도입하여 IRLS-알고리즘을 수정하였다. 즉, 식별된 지렛점들에 대해서 각각 다른 크기의 가중치,

$$v_i = \begin{cases} 1, & \text{if } RSD_i \le r, \\ \dfrac{r}{RSD_i}, & \text{if } RSD_i > r \end{cases}$$

를 적용하여 지렛점의 영향을 줄이도록 하였다.

한편, 수직이상점이나 나쁜 지렛점으로 식별된 관측치의 잔차를 ψ_i라 하고, ψ_i와 수직으로 만나는 V-마스크의 경계선(l_1, l_2)에서의 잔차를 α_i라 하였을 때, 수직이상점과 나쁜 지렛점에 대해 다음과 같은 조정치를 각각 설정할 수 있다. 즉, y값이 0일 경우에는 조정치를 더해 주고, y값이 1일 경우에는 조정치를 빼는 방식으로 반응변수 값을 다음과 같이 조정하는 것이다.

$$\text{수직 이상점의 경우: } \tilde{y}_i = \begin{cases} y_i + \dfrac{|\psi_i| - |\alpha_i|}{\eta(1 - |\alpha_i|)}, & \text{if } y_i = 0, \\ y_i - \dfrac{|\psi_i| - |\alpha_i|}{\eta(1 - |\alpha_i|)}, & \text{if } y_i = 1, \end{cases} \tag{3.1}$$

$$\text{나쁜 지렛점의 경우: } \tilde{y}_i = \begin{cases} y_i + \dfrac{|\psi_i| - 1/2}{\eta(1 - 1/2)}, & \text{if } y_i = 0, \\ y_i - \dfrac{|\psi_i| - 1/2}{\eta(1 - 1/2)}, & \text{if } y_i = 1. \end{cases} \tag{3.2}$$

여기서 $\eta(\ge 1)$는 조정치의 크기의 비율을 설정해 주는 인수인데, 본 연구에서는 $\eta = 2.0$을 채택하였다. 위와 같은 로버스트 추정치를 얻는 과정을 자세히 기술하면 다음과 같다.

알고리즘 RIRLS:

단계 0 RSD-RR 산점도와 V-마스크에 의해 로지스틱회귀 이상점을 식별한다.

단계 1 가중치행렬 $\boldsymbol{V} = \text{diag}\{v_1, \ldots, v_n\}$; $v_i = 1$ for $i \in A$, $v_i = r/RSD_i$ for $i \notin A$ (단, A는 정상점들의 지수집합, $r = \max\{RSD_i,\ i \in A\}$임)을 구성한 후, 설명변수 행렬을 $\boldsymbol{Z} = \boldsymbol{VX}$로 변환한다. 그리고 잔차 ψ_i와 수직으로 만나는 l_1 혹은 l_2의 값을 α_i라 하고 반응변수 값을 다음과 같이 조정한다. 즉, $\tilde{y}_i = (|\psi_i| - |\alpha_i|)/(\eta(1 - |\alpha_i|))$ if $y_i = 0$, $i \in O$; $\tilde{y}_i = (\eta - |\psi_i| - (\eta - 1)|\alpha_i|)/(\eta(1 - |\alpha_i|))$ if $y_i = 1$, $i \in O$ (단, O는 수직 이상점들의 지수집합), $\tilde{y}_i = (2|\psi_i| - 1)/\eta$ if $y_i = 0$, $i \in L$; $\tilde{y}_i = ((\eta + 1) - 2|\psi_i|)/\eta$ if $y_i = 1$, $i \in L$ (단, L은 나쁜 지렛점들의 지수집합).

단계 2 반복수 $t = 0$을 지정하고, IRLS 추정치인 $\hat{\boldsymbol{\beta}}^{(0)}$를 초기값으로 지정한다.

단계 3 $\hat{\pi}_i^{(t)} = \exp(\hat{\pi}_i^{*(t)})/(1 + \exp(\hat{\pi}_i^{*(t)}))$, $\hat{\pi}_i^{*(t)} = \boldsymbol{z}_i^T \hat{\boldsymbol{\beta}}^{(t)}$을 계산하고, 가중치 행렬 $\boldsymbol{W}^{(t)} = \text{diag}\{w_1^{(t)}, \ldots, w_n^{(t)}\}, w_i^{(t)} = \hat{\pi}_i^{(t)}(1 - \hat{\pi}_i^{(t)})$을 구성한다.

단계 4 변환된 반응변수 값 $y_i^{*(t)} = \hat{\pi}_i^{*(t)} + (\tilde{y}_i - \hat{\pi}_i^{(t)})/w_i^{(t)}$을 생성한다.

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

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

4. 알고리즘의 평가

제안된 추정알고리즘 RIRLS을 평가하기 위하여 표 4.1에 수록된 자료들을 대상으로 최대우도추정치와 로버스트 추정치를 구하였으며 (단, 각 설명변수가 변환되지 않은 상태의 로지스틱모형을 전제로 하였음), 추정법별로 정분류율 (accurate classification rate: ACR)을 측정하였다.

분석대상 자료 중에는 이상점이 포함되지 않은 것과 다양한 특성의 나쁜 지렛점이나 수직 이상점이 포함되도록 몇 개의 관찰치를 인위적으로 교체한 자료가 있다. 알고리즘 RIRLS에 의한 추정치와 최대우도추정치는 표 4.2에 수록되었는데, 나쁜 지렛점이나 수직 이상점이 존재하지 않는 자료의 경우에는 추정치에 별 차이가 없지만 이상점이 포함된 자료의 경우에 추정치에 상당한 차이가 있음을 알 수 있다.

새로운 알고리즘의 평가를 위하여, 각 자료별로 알고리즘 RIRLS과 MLE을 적용하여 사후확률을 추정하고 경계치 0.5를 기준으로 관찰치를 분류한 후 정분류율을 구하였다. 정분류율의 측정 결과를 통하여 어느 추정법이 이상점의 영향력을 보다 더 효과적으로 줄여주는지를 확인할 수 있다. RIRLS와 MLE의 정분류율을 비교한 결과, 이상점이 포함된 자료의 경우 RIRLS에 의한 추정치에 바탕을 둔 분류가 더 정확한 것으로 나타났다.

표 4.1: 추정알고리즘의 평가를 위한 자료

자료명	자 료 내 용
A	인공자료 ($x_i = i$ for $i = 1, \ldots, 50$, $y = 0$ for $i = 1, \ldots, 25$, $y = 1$ for $i = 26, \ldots, 50$)의 관찰치 8개 교체: $(x_1, y_1) \leftarrow (-30, 1), (x_2, y_2) \leftarrow (29, 1), (x_3, y_3) \leftarrow (3, 1)$, $(x_{10}, y_{10}) \leftarrow (10, 1), (x_{41}, y_{41}) \leftarrow (41, 0), (x_{48}, y_{48}) \leftarrow (48, 0)$, $(x_{49}, y_{49}) \leftarrow (79, 0), (x_{50}, y_{50}) \leftarrow (80, 0)$.
B	자료 A에서의 인공자료의 관찰치 6개 교체: $(x_1, y_1) \leftarrow (-10, 1), (x_2, y_2) \leftarrow (2, 1), (x_3, y_3) \leftarrow (3, 1)$, $(x_{48}, y_{48}) \leftarrow (48, 0), (x_{49}, y_{49}) \leftarrow (49, 0), (x_{50}, y_{50}) \leftarrow (60, 0)$.
C	Vaso-Constriction of Skin 자료. (Finny, 1947; http://support.sas.com/ onlinedoc/913 /getDoc/en/statug.hlp /logistic_sect59.htm)
D	자료 C의 관찰치 5개 교체: $(x_{1,21}, x_{2,21}, y_{21}) \leftarrow (0.4, 0.5, 1), (x_{1,26}, x_{2,26}, y_{26}) \leftarrow (3.5, 3.0, 0)$, $(x_{1,29}, x_{2,29}, y_{29}) \leftarrow (3.0, 3.5, 0), (x_{1,34}, x_{2,34}, y_{34}) \leftarrow (0.5, 0.4, 1)$, $(x_{1,39}, x_{2,39}, y_{39}) \leftarrow (3.5, 3.0, 0)$.
E	Wais 점수와 치매증상과의 관계 자료. (http://ftp.sas.com/samples/A55201)
F	자료 E의 관찰치 5개 교체: $(x_{22}, y_{22}) \leftarrow (28, 1), (x_{36}, y_{36}) \leftarrow (29, 1), (x_{47}, y_{47}) \leftarrow (2, 0)$, $(x_{50}, y_{50}) \leftarrow (3, 0), (x_{54}, y_{54}) \leftarrow (25, 1)$.
G	Erythrocyte sedimentation rate 자료. (http://www.ed.uiuc.edu/courses/EdPsy490AT/lectures /5logreg2_02.pdf)
H	자료 G의 관찰치 5개 교체: $(x_{1,5}, x_{2,5}, y_5) \leftarrow (5.0, 45, 0), (x_{1,12}, x_{2,12}, y_{12}) \leftarrow (2.0, 12, 1)$, $(x_{1,22}, x_{2,22}, y_{22}) \leftarrow (6.0, 46, 0), (x_{1,27}, x_{2,27}, y_{27}) \leftarrow (6.2, 41, 0)$, $(x_{1,32}, x_{2,32}, y_{32}) \leftarrow (1.3, 10, 1)$.
I	다발성골수종 환자 자료. (Krall 등, 1975; http://ttest.co.kr/cgi-bin/tboard /read.cgi?board=raw&y_number=12)
J	자료 I의 관찰치 5개 교체: $(x_{1,5}, x_{2,5}, y_5) \leftarrow (5.1, 27, 0), (x_{1,6}, x_{2,6}, y_6) \leftarrow (4.7, 26, 0)$, $(x_{1,14}, x_{2,14}, y_{14}) \leftarrow (5.1, 31, 0), (x_{1,52}, x_{2,52}, y_{52}) \leftarrow (20.2, 81, 1)$, $(x_{1,61}, x_{2,61}, y_{61}) \leftarrow (17.3, 82, 1)$.

표 4.2: RIRLS와 MLE에 의한 추정치와 정분류율

추정법		자료 A	자료 B	자료 C	자료 D	자료 E
MLE	$\hat{\beta}_0$	−1.2288	−1.6348	−9.1866	−0.6276	2.4040
	$\hat{\beta}_1$	0.0482	0.0641	3.6604	0.0538	−0.3235
	$\hat{\beta}_2$			2.5927	0.3498	
	ACR	0.8400	0.8600	0.8474	0.6154	0.7071
RIRLS	$\hat{\beta}_0$	−2.8148	−1.9541	−9.1597	−2.4557	2.3839
	$\hat{\beta}_1$	0.1100	0.0801	3.6867	0.8215	−0.3170
	$\hat{\beta}_2$			2.6828	0.9885	
	ACR	0.8864	0.8800	0.8461	0.7179	0.7393

추정법		자료 F	자료 G	자료 H	자료 I	자료 J
MLE	$\hat{\beta}_0$	−1.0379	−12.7921	1.2988	6.0798	−1.1736
	$\hat{\beta}_1$	0.0219	1.9104	0.3360	−0.2214	−0.0098
	$\hat{\beta}_2$		0.1558	−0.1000	−0.0446	0.0387
	ACR	0.6852	0.7179	0.6875	0.6066	0.7231
RIRLS	$\hat{\beta}_0$	−0.0982	−12.0565	−3.7230	6.0798	1.1587
	$\hat{\beta}_1$	−0.0729	1.7057	0.7640	−0.3214	−0.0986
	$\hat{\beta}_2$		0.1518	0.0063	−0.0146	0.0163
	ACR	0.6852	0.8154	0.8125	0.8744	0.7692

5. 결론

이상점은 로지스틱회귀분석에서의 통계적 추론 과정에 막대한 악영향을 미치기 때문에 분석 결과가 심하게 왜곡될 수 있다. 따라서 데이터마이닝을 위한 로지스틱회귀분석에 앞서 자료에 이상점이 존재하는지 확인하고 어느 관찰치가 이상점인지 식별해야 하며, 이상점의 영향을 적절히 줄일 수 있는 로버스트 추정법을 적용해야 한다. 본 논문은 로지스틱회귀에서의 로버스트 추정을 위한 새로운 알고리즘을 제시하는데, RSD-RR 산점도에 V-마스크 형태의 경계구역을 적용하여 이상점을 식별하고, 적절한 가중치와 인수를 선정하여 이상점의 영향력을 조정하는 방식으로 로버스트 추정치를 얻는 알고리즘이다. 이 추정알고리즘을 몇 개의 자료에 적용해 본 결과, 분류의 정확도 관점에서 최대우도추정보다 우수한 것으로 평가되었다.

참고문헌

Bianco, A. M. and Yohai, V. J. (1996). Robust estimation in the logistic regression model, *Robust Statistics, Data Analysis, and Computer Intensive Methods* (Rieder, H. ed.), 17–34, Springer-Verlag, New York.

Carroll, R. J. and Pederson, S. (1993). On robustness in the logistic regression model, *Journal of the Royal Statistical Society,* Ser. B, **55**, 693–706.

Copas, J. B. (1988). Binary regression models for contaminated data, *Journal of the Royal Statistical Society,* Ser. B, **50**, 225–265.

Croux, C. and Haesbroeck, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression, *Computational Statistics & Data Analysis*, **44**, 273–295.

Finney, D. J. (1947). The estimation from individual records of the relationship between dose and quantal response, *Biometrika*, **34**, 320–334.

Gnanadesikan, R. and Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data, *Biometrics*, **28**, 81–124.

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

Jennings, D. E. (1986). Outliers and residual distributions in logistic regression, *Journal of the American Statistical Association*, **81**, 987–990.

Kim, B. Y. (2005). V-mask type criterion for identification of outliers in logistic regression, *The Korean Communications in Statistics*, **12**, 625–634.

Kordzakhia, N., Mishra, G. D. and Reiersolmoen, L. (2001). Robust estimation in the logistic regression model, *Journal of Statistical Planning and Inference*, **98**, 211–223.

Krall, J. M., Uthoff, V. A. and Harley, J. B. (1975). A step-up procedure for selecting variables associated with survival, *Biometrics*, **31**, 49–57.

Pregibon, D. (1981). Logistic regression diagnostics, *The Annals of Statistics*, **9**, 705–724.

Pregibon, D. (1982). Resistant fits for some commonly used logistic models with medical applications, *Biometrics*, **38**, 485–498.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point, *Mathematical Statistics and Applications* (Grossmann, W., Pflug, G., Vincze, I. and Wertz, W. eds.), 283–297, Reidel, Dordrecht.

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

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

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*, **89**, 888–896.

[2007년 5월 접수, 2007년 7월 채택]

Algorithm for the Robust Estimation in Logistic Regression*

Bu-Yong Kim[1] Myung Wook Kahng[2] Mi-Ae Choi[3]

ABSTRACT

The maximum likelihood estimation is not robust against outliers in the logistic regression. Thus we propose an algorithm for the robust estimation, which identifies the bad leverage points and vertical outliers by the V-mask type criterion, and then strives to dampen the effect of outliers. Our main finding is that, by an appropriate selection of weights and factors, we could obtain the logistic estimates with high breakdown point. The proposed algorithm is evaluated by means of the correct classification rate on the basis of real-life and artificial data sets. The results indicate that the proposed algorithm is superior to the maximum likelihood estimation in terms of the classification.

Keywords: Logistic regression, outlier identification, V-mask criterion, robust estimation.

* This Research was supported by the Sookmyung Women's University Research Grants 2006.

1) Professor, Department of Statistics, Sookmyung Women's University, Chungpa-dong 2-ga, Yongsan-gu, Seoul 140-742, Korea
E-mail: buykim@sm.ac.kr

2) Professor, Department of Statistics, Sookmyung Women's University, Chungpa-dong 2-ga, Yongsan-gu, Seoul 140-742, Korea
E-mail: mwkahng@sm.ac.kr

3) Computer Systems Division, Samsung Electronics Co., Gyeonggi-do 443-742, Korea
E-mail: miae21.choi@samsung.com