응용통계연구 (2009)
22(3), 531–539

로버스트추정에 바탕을 둔 주성분로지스틱회귀

김부용[1] · 강명욱[2] · 장혜원[3]

[1]숙명여자대학교 통계학과, [2]숙명여자대학교 통계학과, [3](주)피스트글로벌 종합리스크사업팀

(2009년 1월 접수, 2009년 2월 채택)

요 약

로지스틱회귀분석은 고객관계관리를 위한 데이터마이닝 분야에서 많이 사용되는 기법인데, 이 분야의 모형설정 과정에서는 연관성이 매우 높은 설명변수들이 모형에 함께 포함되어 다중공선성의 문제를 유발하며, 더욱이 회귀자료에 이상점들이 포함되면 최우추정량은 심각한 결함을 갖게 된다. 두 가지 문제점을 동시에 해결하기 위하여 로버스트주성분로지스틱회귀를 적용할 수 있는데, 본 논문에서는 주성분의 선정기준을 결정하는 모형을 개발하고, 주성분모형에서의 추정치에 미치는 이상점의 영향을 축소하기 위한 로버스트추정법을 제안하였다. 제안된 추정법은 다중공선성과 이상점이 유발하는 문제들을 적절히 해결해 준다는 사실이 모의실험을 통하여 확인되었다.

주요용어: 다중공선성, 데이터마이닝, 이상점, 로버스트추정, 주성분로지스틱회귀.

1. 서론

로지스틱회귀분석은 고객관계관리(CRM) 분야에서 예측이나 분류를 위한 데이터마이닝 기법 중의 하나로 많이 활용되고 있다. 데이터마이닝을 위하여 로지스틱회귀분석을 사용하는 경우 수많은 변수들로부터 설명변수를 선정해야 하는데, 적절한 방법을 통하여 설명변수가 선정되더라도 다수의 설명변수가 모형에 포함되면 설명변수들의 일부가 높은 연관성을 가질 가능성이 매우 높다. 특히, 로지스틱회귀모형에 포함된 설명변수들 중에는 고객의 재산 상태나 경제능력 등을 나타내는 변수들이 많은데, 이러한 변수들은 태생적으로 높은 연관성을 가질 수밖에 없다. 이와 같이 연관성이 매우 높은 설명변수들이 로지스틱모형에 포함되면 다중공선성 문제가 발생하게 되는데, 로지스틱회귀분석에서 일반적으로 적용되는 최우추정량의 분산이 지나치게 팽창하기 때문에 이 추정량에 바탕을 둔 예측이나 분류는 심각하게 왜곡된다는 사실이 잘 알려져 있다. 로지스틱회귀모형에서의 다중공선성 문제를 해결하기 위해서는 주성분로지스틱회귀를 적용할 수 있는데, 주성분의 선정을 위한 적절한 기준의 설정이 요구된다. 따라서 본 연구에서는 상태지수에 바탕을 둔 주성분 선정기준 및 선정방법을 제안하고자 한다.

한편, CRM 분야에서는 엄격하게 통제되지 않는 상황에서 자료가 수집되는 경우가 많기 때문에 로지스틱회귀분석 자료에 이상점이 다수 포함될 가능성이 높다. 자료에 회귀이상점들이 포함되면 로지스틱회귀분석의 결과는 이상점의 영향에 의해 크게 왜곡될 수밖에 없으므로 이상점에 대한 적절한 대책을 강구해야 한다 (Croux와 Haesbroeck, 2003). 이상점들에 의해 유발되는 문제점을 해결하기 위한 방안중의 하나로서, SAS/E-miner와 같이 자료정제 과정에서 이상점을 식별하여 제거하는 방법을 고려할 수 있다. 그러나 이상점이 오염된 관찰치라는 사실이 확인된 경우에는 제거하는 방법이 효과적일

본 연구는 숙명여자대학교 2008년도 교비연구비 지원에 의해 수행되었음.
[1]교신저자: (140-742) 서울특별시 용산구 청파동 2가, 숙명여자대학교 통계학과, 교수.
E-mail: buykim@sm.ac.kr

수 있으나, 직접 이상점을 제거하는 방법은 문제 해결에 적절치 않기 때문에 (Kim, 2005), 본 연구에서는 주성분로지스틱모형의 추정치에 미치는 이상점의 영향을 적절히 축소할 수 있는 로버스트추정법을 제안하고자 한다. 특히, 지렛점들의 영향을 감소시키기 위하여 로버스트추정량인 MCD(minimum covariance determinant)-추정량이나 MVE(minimum volume ellipsoid)-추정량에 바탕을 둔 로버스트 제곱거리(RSD: robust squared distance)를 기준으로 지렛점을 식별하고 이들에게 적절한 가중치를 부여한다. 그리고 수직이상점에 대해서는 정상점들로부터 벗어난 정도에 비례하는 적절한 조정치를 부여하는 방식을 채택하여 그 영향력을 감소시킨다. 한편, 제안된 방법의 적합성을 평가하기 위하여 Monte Carlo 모의실험을 실행한다.

2. 주성분로지스틱회귀를 위한 주성분선정법

이항형 반응변수를 위한 로지스틱회귀모형은 일반적으로 다음과 같이 정의된다.

$$y_i = \pi(\boldsymbol{x}_i) + \epsilon_i, \quad \pi(\boldsymbol{x}_i) = \frac{\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)}{1+\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)}, \quad i = 1, \ldots, n, \tag{2.1}$$

여기서 y_i는 반응변수, $\boldsymbol{x}_i^T$는 설명변수 행렬 $X_{n\times(k+1)}$의 i번째 행, 그리고 $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_k]^T$는 로지스틱회귀계수를 의미한다. 모형 (2.1)에 도입된 설명변수들의 전부 혹은 일부 사이에 상당히 높은 연관성이 존재하는 현상인 다중공선성은 로지스틱회귀추정량의 분산을 매우 크게 팽창시키는 문제를 야기하게 된다. Schaefer (1986)는 다중공선성이 로지스틱회귀모형의 최우추정량에 미치는 악영향을 분석하였는데, 다중공선성이 존재함에도 불구하고 최우추정량을 적용하는 경우 각 회귀계수에 대한 통계적 추론은 물론 예측이나 분류 결과가 왜곡된다고 하였다. 그러므로 자료에 다중공선성이 존재하는지 사전에 진단하고 적절한 조치를 취할 필요가 있는데, 모형의 재설정이나 설명변수의 재정의에 따른 정보의 손실을 예방하면서 다중공선성 문제를 해결하기 위해서는 다음과 같은 주성분로지스틱회귀를 적용할 수 있다.

2.1. 주성분로지스틱회귀모형

주성분로지스틱회귀는 모형 (2.1)에 연속형 설명변수들이 포함된 상황에 적용되는데, 중심화 된 설명변수들의 행렬 X_c의 비정칙치분해(singular value decomposition), $X_c = \mathrm{UDV}^T$(단, $U_{n\times k}$와 $V_{k\times k}$는 직교행렬, $D_{k\times k}$는 비정칙치 $\mu_1, \mu_2, \ldots, \mu_k$로 구성된 대각행렬임)를 출발점으로 한다. 행렬 $X_c^TX_c$의 고유치를 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k (\geq 0)$라 할 때, 각 고유치에 대응하는 고유벡터 $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k$로 구성된 행렬이 V에 해당된다. X_c를 직교행렬 V에 의해 변환한 행렬, $Z = X_cV$의 각 열인 $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_k$를 주성분이라 한다. 주성분로지스틱모형에 포함될 주성분은 $X_c^TX_c$의 고유치들의 상대적인 크기를 기준으로 선정되는데, 매우 작은 고유치는 주성분로지스틱모형에서는 물론 모형 (2.1)에서의 추정량 분산을 크게 팽창시키는 역할을 하므로 작은 고유치에 대응하는 주성분을 적절한 기준에 의해 모형에서 제외시킨 후 주성분로지스틱회귀모형을 설정한다.

2.2. 주성분 선정법

Mason과 Gunst (1985) 등이 제안한 선형회귀모형에서의 주성분 선정법을 주성분로지스틱회귀모형에 포함될 주성분을 선정하는데 활용할 수 있지만, 주성분 선정기준이 객관적이지 못하다는 문제점을 가지고 있다. 그래서 Kim과 Kahng (2008)은 다중공선성 문제를 해결하면서도 로지스틱모형의 적합성이 유지되도록 하는 선정법을 제안하였다. 그런데 본 연구와 같이 로버스트추정량을 바탕으로 할 경우에는

유의성검정을 위한 통계량이 아직 파악되지 않고 있다. 따라서 상태지수($\lambda_{\max}/\lambda_j,\ j=1,\ldots,k$)의 크기에 따라 주성분을 선정할 수 있는 기준을 제안하고자 한다.

다중공선성이 유발하는 심각한 문제는 추정량의 분산이 과도하게 팽창한다는 것이므로, 주성분 선정법의 초점을 추정량 분산의 최소화에 두기로 하였다. 우선 모의실험을 실행하기 위하여 다중공선성이 존재하는 자료를 생성하였는데 Aguilera 등 (2006)이 사용한 방법을 적용하였다. 각 설명변수는 정규분포(0, 1)과 균일분포(0, 1)로부터 생성하고, 설명변수의 개수는 3, 4, 5, 6, 7, 8, 9개로 선정하고, 관찰치의 수는 설명변수 수의 30, 40, 50, 60, 70, 80, 90배로 선정하였다. 그리고 다양한 수준의 다중공선성이 발생하도록 다중공선성에 관련된 변수의 수를 2, 3, 4, 5개로 설정하여 각 조합별로 1,000개씩의 자료를 생성하였다. 각 자료에 주성분로지스틱회귀를 적용하여 상태지수의 경계치가 어느 크기일 때 회귀계수 추정치들의 분산 합이 최소가 되는지 조사하였다. 조사된 결과를 바탕으로 컨조인트분석을 실행하였는데, 추정치 분산의 최소치에 대응하는 상태지수의 크기에 가장 큰 영향을 미치는 요인은 설명변수의 수와 관찰치의 수인 것으로 나타났다. 설명변수의 분포형태와 다중공선성에 관련된 변수의 수도 영향을 미치는 것으로 분석되었으나 그 정도가 미미하고 분석에 앞서 알 수 있는 요인들이 아니기 때문에 다음과 같은 선형모형

$$\boldsymbol{c} = B\boldsymbol{\psi} + \boldsymbol{\eta}, \quad B = [\mathbf{1}\,|\,\boldsymbol{t}\,|\,\boldsymbol{g}] \tag{2.2}$$

을 설정하였다. 여기서 $\boldsymbol{c}$는 추정량 분산을 최소가 되게 하는 상태지수, $\boldsymbol{t}$는 설명변수 수, $\boldsymbol{g}$는 관찰치 수(설명변수의 배수)의 벡터이며 $\boldsymbol{\eta}$는 오차항이다. 모의실험을 통하여 측정된 자료에 모형 (2.2)를 적합한 결과 $\hat{\boldsymbol{\psi}}^T = [2.65004 \quad 0.96053 \quad 0.06571]$ (p-value: 0.0081, < 0.0001, < 0.0001)을 얻었으며 $\hat{\boldsymbol{c}} = B\hat{\boldsymbol{\psi}}$에 의해 상태지수의 경계치를 구하였다. 따라서 경계치 $\hat{c}_i$보다 작은 상태지수에 대응하는 주성분만을 모형에 포함시키면 다중공선성 문제를 해결할 수 있게 된다.

3. 로버스트주성분로지스틱회귀

주성분의 선정과정을 거쳐서 모형에 포함될 s개의 주성분이 결정되면, 주성분들로 구성된 행렬 Z_s에 절편을 삽입한 후 주성분로지스틱회귀계수 추정치 $\hat{\boldsymbol{\gamma}}$을 얻게 된다. $\hat{\boldsymbol{\gamma}}$은 반복재가중최소자승(IRLS: iterative reweighted least square)추정에 의해 구할 수 있는데, IRLS-알고리즘의 각 반복과정에서의 기본적인 계산은 최소자승추정과 같으므로 이상점들의 영향을 많이 받게 된다. 따라서 효과적인 방법에 의해 이상점들을 식별하고 각 이상점들의 영향을 적절한 수준에서 통제할 수 있는 로버스트추정이 요구된다.

3.1. 로지스틱회귀이상점의 식별

우선 행렬 Z_s에서의 지렛점을 식별하기 위하여 Hadi (1994)의 방법을 활용할 수 있으나, 이 방법은 Mahalanobis-거리를 바탕으로 하기 때문에 가림현상(masking effect)이나 붙음현상(swamping effect)이 유발되어 지렛점을 정확히 식별하는데 한계가 있다. 그러므로 MCD-추정량이나 MVE-추정량과 같은 로버스트추정량을 도입한 RSD를 지렛점 식별에 적용하는 방법을 고려할 수 있는데, MCD-추정량 $(\boldsymbol{\mu}_J,\ \Sigma_J)$은 다음과 같이 정의된다.

$$\Sigma_J = \frac{1}{h}\sum_{i\in J}(\boldsymbol{z}_{si}-\boldsymbol{\mu}_J)(\boldsymbol{z}_{si}-\boldsymbol{\mu}_J)^T, \quad \boldsymbol{\mu}_J = \frac{1}{h}\sum_{i\in J}\boldsymbol{z}_{si},$$

여기서 $h=[(n+k+1)/2]$([·]는 최대정수 함수임)이고 $\boldsymbol{z}_{si}$는 선정된 주성분들의 i번째 요소이다. 그리고 J는 h개의 원소로 구성된 지수집합인데, 원소의 수가 h인 모든 부분집합 H에 대해 $|\Sigma_J| \le |\Sigma_H|$을

만족시키는 집합이다. 한편, MVE-추정량 $(\boldsymbol{\mu}^*, \Sigma^*)$은 다음과 같이 정의된다. 즉, $\boldsymbol{\mu}^*$는 h개의 관찰치를 포함하는 타원체들 중에서 부피가 최소인 타원체의 중심이고, 공분산 추정량 Σ^*는 $\boldsymbol{\mu}^*$에 대응하는 타원체에 일치추정량을 위한 인자를 곱한 것으로 얻어진다. MCD-추정량과 MVE-추정량의 붕괴점은 $[(n-k+1)/2]/n$(단, $h=[(n+k+1)/2]$인 경우)인데 (Rousseeuw와 Leroy, 2003), 이는 로버스트 추정량들이 가질 수 있는 붕괴점의 최대치에 해당된다. 그런데 MCD-추정치나 MVE-추정치를 구하기 위해서는 많은 부분집합에 대해 공분산행렬의 행렬식이나 타원체의 부피를 계산해야 하기 때문에 많은 계산이 요구된다. 특히, 규모가 막대한 데이터마이닝 분야의 자료에서는 이러한 추정의 계산효율성이 심각하게 낮아지는 문제가 발생하는데, 계산효율성이 개선된 Woodruff와 Rocke (1994), Rousseeuw와 Driessen (1999), Hardin과 Rocke (2004) 등의 알고리즘을 적용할 수 있다.

MCD-추정량과 MVE-추정량을 편의상 $(\boldsymbol{\mu}_J^*, \Sigma_J^*)$로 표기하고, 이 추정량을 Mahalanobis-거리에 적용하면 RSD는 다음과 같이 정의된다.

$$\mathrm{RSD}_i = (\boldsymbol{z}_{si} - \boldsymbol{\mu}_J^*)^T \Sigma_J^{*-1} (\boldsymbol{z}_{si} - \boldsymbol{\mu}_J^*).$$

이제 RSD를 지렛점 식별에 활용하기 위해서는 경계치를 결정해야 한다. Kim (2005)은 RSD의 계층적 군집화에 의한 식별방법을 제안하였으나, 데이터마이닝 분야에서 방대한 자료에 계층적 군집화를 적용하면 계산효율성에 심각한 문제가 발생한다. 따라서 Hardin과 Rocke (2004)가 제시한 RSD의 근사적 분포를 바탕으로 지렛점을 식별하고자 한다. 즉, RSD의 분포,

$$\frac{d(a-k+1)}{ka}\mathrm{RSD} \sim F(k, a-k+1) \tag{3.1}$$

을 바탕으로 얻은 경계치, $\kappa = kaF_{1-\alpha}(k, a-k+1)/\{d(a-k+1)\}$을 적용하여 지렛점을 식별할 수 있다($d$와 a의 계산은 부록 참조).

한편, 수직이상점을 식별하기 위해서 로버스트잔차(RR: robust residual)를 도입한 RSD-RR 산점도를 활용할 수 있는데, 식별된 지렛점의 영향을 줄이기 위하여 Z_s에 다음과 같은 방식으로 가중치를 부여한다. 즉,

$$\begin{aligned} Z_a &= Q\left(Z_s - \mathbf{1}_n \boldsymbol{\mu}_J^{*T}\right) + \mathbf{1}_n \boldsymbol{\mu}_J^{*T}, \\ Q &= \mathrm{diag}[q_i], \quad q_i = \begin{cases} 1, & \text{for } i \in \mathrm{RP}, \\ \dfrac{\kappa}{\mathrm{RSD}_i}, & \text{for } i \notin \mathrm{RP}, \end{cases} \end{aligned} \tag{3.2}$$

여기서 $\mathbf{1}_n$은 단위벡터이고 $\mathrm{RP} = \{i | \mathrm{RSD}_i \le \kappa\}$는 정상점들의 지수집합이다. 가중치에 의해 변환된 행렬 Z_a에 절편을 추가한 후 지렛점의 영향을 적게 받는 추정치를 구하고, 이 추정치를 바탕으로 RR을 얻는다. RSD를 가로축으로 하고 세로축에 RR을 플롯한 RSD-RR 산점도에서 RSD의 중위수 τ를 출발점으로 삼고 식 (3.1)에서의 경계치 κ를 기준으로 경계구역을 설정하는데, $(\tau, -1)$과 $(\kappa, -0.5)$를 지나는 직선과 $(\tau, 1)$과 $(\kappa, 0.5)$를 지나는 직선이 구성하는 V-마스크 형태의 경계구역을 설정한다. 즉,

$$\begin{aligned} l_1 &= \frac{1}{2(\kappa-\tau)}\mathrm{RSD} - \frac{\kappa - \tau/2}{\kappa - \tau}, \\ l_2 &= \frac{-1}{2(\kappa-\tau)}\mathrm{RSD} + \frac{\kappa - \tau/2}{\kappa - \tau}. \end{aligned}$$

이러한 형태의 경계구역을 적용하면 수직이상점과 나쁜 지렛점은 물론 좋은 지렛점도 동시에 식별할 수

있다. 즉,

$$\begin{aligned} 0 \le \mathrm{RSD}_i \le \kappa,\ l_{2i} < |\mathrm{RR}_i| &\text{이면},\ i \in \mathrm{VO}, \\ \kappa < \mathrm{RSD}_i,\ 0.5 < |\mathrm{RR}_i| &\text{이면},\ i \in \mathrm{BL}, \\ \kappa < \mathrm{RSD}_i,\ 0.5 \ge |\mathrm{RR}_i| &\text{이면},\ i \in \mathrm{GL}, \end{aligned} \tag{3.3}$$

여기서 VO는 수직이상점, BL은 나쁜 지렛점, GL은 좋은 지렛점의 지수집합을 의미한다.

3.2. 로버스트주성분로지스틱회귀 추정

로지스틱회귀모형의 로버스트추정에 관한 연구로는 Copas (1988), Carroll과 Pederson (1993), Kordzakhia 등 (2001), Croux와 Haesbroeck (2003) 등이 있다. 이러한 추정법들은 대부분 영향력함수 접근법에 이론적 기초를 두고 있는데 실제로 영향력함수의 설정에는 어려움이 따른다. 따라서 본 연구에서는 Kim 등 (2007)이 제안한 로버스트추정법을 적용한다.

3.2.1. 가중치 및 조정치의 결정 IRLS-알고리즘에 의한 추정치는 이상점에 의해 많은 영향을 받으므로 로버스트추정치를 구하기 위해서는 이 알고리즘을 수정해야 할 필요가 있다. 식 (3.3)에 의해 식별된 지렛점들과 수직이상점들에 각각 적절한 가중치와 조정치를 부여하는 방식으로 IRLS-알고리즘을 수정함으로써 이상점들에 민감하지 않은 추정치를 얻을 수 있다. 우선, 행렬 Z_s에서 식별된 지렛점에 해당되는 관찰치에 식 (3.2)와 같은 가중치를 적용하여 지렛점이 미치는 영향을 적절히 감소시킬 수 있다. 그리고 수직이상점이나 나쁜 지렛점으로 식별된 관찰치들에 대해서는 반응변수 값을 적절히 조정해 주어야 하는데, 유계영향력함수(bounded influence function)와 유사한 방식으로 조정한다. 식별된 이상점 i에 대응하는 RR을 ζ_i라 하고, ζ_i와 같은 영역에서 수직으로 만나는 V-마스크의 경계선의 값을 w_i라 하면, 수직이상점과 나쁜 지렛점에 대해 다음과 같은 조정치를 각각 설정할 수 있다. 즉, 수직이상점의 경우에는 반응변수 값을

$$\tilde{y}_i = y_i - \frac{\zeta_i - \omega_i}{\eta(1 - |\omega_i|)} \tag{3.4}$$

와 같이 조정하고, 나쁜 지렛점의 경우에는 식 (3.4)에서 $\omega_i = 0.5$를 적용하여 반응변수 값을 조정한다. 여기서 $\eta(\ge 1)$는 조정치의 비율을 결정하는 인수인데, Kim 등 (2007)은 $\eta = 2.0$을 제안하였다.

3.2.2. 주성분로지스틱회귀계수의 로버스트추정 로버스트주성분로지스틱회귀 추정치는 가중치 (3.2)에 의해 변환된 주성분행렬 Z_a에 절편을 추가한 행렬 $\tilde{Z}_a$와 조정치 (3.4)에 의해 변환된 반응변수 벡터 $\tilde{\boldsymbol{y}}$에 IRLS-알고리즘을 적용하여 얻어진다. 즉, 초기치 $\tilde{\boldsymbol{\gamma}}^{(r)}$: $r = 0$을 지정한 후, $(r+1)$번째 반복과정에서의 추정치는

$$\begin{aligned} \tilde{\boldsymbol{\gamma}}^{(r+1)} &= \tilde{\boldsymbol{\gamma}}^{(r)} + \left(\tilde{Z}_a^T \tilde{W}^{(r)} \tilde{Z}_a\right)^{-1} \tilde{Z}_a^T \left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{\xi}}^{(r)}\right), \\ \tilde{W}^{(r)} &= \mathrm{diag}\left[\tilde{\xi}_i^{(r)}\left(1 - \tilde{\xi}_i^{(r)}\right)\right], \\ \tilde{\xi}_i^{(r)} &= \frac{\exp\left(\tilde{\boldsymbol{z}}_{ai}^T \tilde{\boldsymbol{\gamma}}^{(r)}\right)}{1 + \exp\left(\tilde{\boldsymbol{z}}_{ai}^T \tilde{\boldsymbol{\gamma}}^{(r)}\right)} \end{aligned} \tag{3.5}$$

와 같이 얻어지는데, 반복종료기준(예를 들면, $||\tilde{\boldsymbol{\gamma}}^{(r+1)} - \tilde{\boldsymbol{\gamma}}^{(r)}||_\infty \leq \delta$, $||\cdot||_\infty$는 L-infinity norm이며 δ는 tolerance)을 충족시킬 때까지 반복과정을 진행하여 최종 추정치를 구한다. 알고리즘에 의해 구한 주성분모형에서의 로버스트추정치 $\tilde{\boldsymbol{\gamma}}$은, 최우추정량의 불변성에 근거하여,

$$\hat{\boldsymbol{\beta}}_{PC} = V^+\tilde{\boldsymbol{\gamma}}, \quad V^+ = \begin{bmatrix} 1 & \mathbf{0}_k^T \\ \mathbf{0}_k & V \end{bmatrix}$$

(단, $\mathbf{0}_k$는 영벡터임)와 같이 로지스틱모형 (2.1)에서의 회귀계수 추정치로 변환된다.

4. 로버스트주성분로지스틱회귀의 평가

로버스트주성분로지스틱회귀 추정의 적합성을 평가하기 위하여 Monte Carlo 모의실험을 실행하였다. 다중공선성이 존재하는 다양한 자료를 대량으로 생성한 후, 모형 (2.2)에 의해 결정된 상태지수의 경계치를 적용하되 이상점을 고려하지 않는 주성분로지스틱회귀(PCL)와 로버스트추정을 도입한 로버스트주성분로지스틱회귀(RPCL)를 각각의 자료에 적용하여 각 방법의 평균정분류율(ACCR: average of the correct classification rate)을 비교하였다.

4.1. 모의실험

설명변수와 반응변수는 제 2장에서 사용한 방법과 동일하게 생성하였다. 설명변수의 분포형태(D, nr: 정규분포, un: 균일분포), 설명변수의 수(K), 관찰치의 수(N), 다중공선성에 관련된 변수의 수(M)를 다양하게 선정하여 자료를 생성하였는데, 지렛점과 수직이상점을 심기 위하여 자료의 일부를 임의로 오염시켰다. 이상점의 비중을 5%와 10%로 하되, 수직이상점(VO), 나쁜 지렛점(BL) 그리고 좋은 지렛점(GL)을 각 자료에 심었다. PCL과 RPCL의 적합성을 평가하기 위하여 정상점 자료와 오염된 자료들로부터 정분류율을 측정하고 전체의 평균을 구하였다. 모의실험 반복수는 1,000회이며 프로그램은 SAS/IML을 사용하였다.

4.2. 평가 결과

PCL과 RPCL의 ACCR를 평가하기 위한 모의실험 결과의 일부가 표 4.1에 수록되었다. RPCL의 경우 좋은 지렛점이 포함된 자료의 ACCR이 가장 높고 나쁜 지렛점이 포함된 자료의 ACCR이 가장 낮게 나타났는데, 좋은 지렛점의 비중이 클수록 ACCR이 높았고 수직이상점이나 나쁜 지렛점은 그 비중이 클수록 ACCR이 낮았다. 이상점의 형태나 비중이 상이한 모든 자료에서 PCL보다 RPCL의 ACCR이 크게 측정되었으므로 제안된 방법의 적합성이 우수하다고 할 수 있다. 따라서 제안된 방법은 다중공선성 문제를 적절히 해결하면서도 모형의 적합성이 높은 것으로 평가되었다.

5. 결론

고객관계관리를 위한 데이터마이닝 분야에서 로지스틱회귀분석을 많이 활용하는데, 설명변수들의 특성상 흔히 다중공선성의 문제를 야기하게 된다. 더욱이 이러한 분야의 자료에는 이상점이 다수 포함되는 경우가 많다. 본 논문에서는 로지스틱회귀분석 자료에 다중공선성과 이상점이 동시에 존재하는 경우, 이 문제를 해결할 수 있는 주성분로지스틱회귀분석에 관하여 연구하였다. 특히 주성분을 선정하는 방법을 제안하고 로버스트추정법을 도입하였는데, 모의실험을 통해 제안된 방법의 적합성이 상대적으로 우수한 것으로 평가되었다.

표 4.1. PCL과 RPCL의 평균정분류율 측정치

K	N	M	D	정상점		이상점 5%						이상점 10%					
						GL		VO		BL		GL		VO		BL	
				PCL	RPCL	PCL	RPCL	PCL	RPCL	PCL	RPCL	PCL	RPCL	PCL	RPCL	PCL	RPCL
3	90	2	nr	0.728	0.765	0.717	0.778	0.695	0.756	0.662	0.714	0.724	0.791	0.663	0.727	0.599	0.630
			un	0.776	0.826	0.762	0.835	0.747	0.813	0.722	0.778	0.763	0.845	0.713	0.786	0.658	0.712
	120	2	nr	0.729	0.764	0.716	0.776	0.690	0.755	0.659	0.716	0.724	0.790	0.659	0.728	0.591	0.625
			un	0.778	0.826	0.763	0.835	0.741	0.811	0.720	0.779	0.761	0.845	0.709	0.785	0.655	0.712
	150	2	nr	0.731	0.763	0.714	0.775	0.690	0.755	0.656	0.715	0.722	0.789	0.657	0.726	0.589	0.618
			un	0.779	0.825	0.758	0.834	0.743	0.813	0.720	0.781	0.759	0.844	0.709	0.787	0.653	0.711
4	160	3	nr	0.790	0.831	0.765	0.839	0.734	0.819	0.703	0.777	0.772	0.849	0.695	0.784	0.632	0.668
			un	0.828	0.880	0.809	0.886	0.812	0.865	0.753	0.829	0.809	0.893	0.787	0.831	0.674	0.737
	200	3	nr	0.793	0.831	0.766	0.840	0.738	0.820	0.707	0.779	0.771	0.850	0.697	0.786	0.638	0.665
			un	0.831	0.881	0.812	0.887	0.815	0.866	0.758	0.831	0.811	0.893	0.789	0.833	0.678	0.739
	240	3	nr	0.792	0.832	0.765	0.840	0.734	0.821	0.701	0.780	0.767	0.850	0.692	0.787	0.632	0.664
			un	0.829	0.881	0.812	0.887	0.812	0.867	0.756	0.833	0.812	0.893	0.786	0.835	0.677	0.739
5	200	3	nr	0.782	0.841	0.728	0.849	0.701	0.830	0.676	0.780	0.731	0.858	0.662	0.787	0.611	0.657
			un	0.812	0.888	0.777	0.894	0.757	0.874	0.734	0.833	0.773	0.900	0.715	0.834	0.657	0.720
	250	3	nr	0.782	0.841	0.729	0.849	0.705	0.831	0.675	0.784	0.731	0.857	0.665	0.791	0.611	0.657
			un	0.806	0.887	0.767	0.893	0.752	0.874	0.729	0.836	0.763	0.899	0.708	0.836	0.653	0.719
	300	3	nr	0.782	0.839	0.726	0.848	0.701	0.829	0.673	0.783	0.728	0.857	0.662	0.791	0.611	0.655
			un	0.804	0.887	0.765	0.893	0.747	0.874	0.726	0.837	0.762	0.899	0.707	0.837	0.649	0.719
6	300	4	nr	0.815	0.884	0.781	0.889	0.747	0.868	0.712	0.811	0.795	0.895	0.702	0.815	0.646	0.698
			un	0.832	0.919	0.807	0.923	0.778	0.902	0.747	0.856	0.817	0.926	0.734	0.854	0.666	0.725
	360	4	nr	0.812	0.882	0.773	0.888	0.735	0.867	0.701	0.812	0.789	0.894	0.696	0.818	0.636	0.697
			un	0.833	0.919	0.809	0.923	0.780	0.903	0.747	0.858	0.818	0.926	0.735	0.856	0.666	0.723
	420	4	nr	0.816	0.883	0.776	0.888	0.741	0.868	0.705	0.816	0.789	0.894	0.697	0.820	0.639	0.696
			un	0.831	0.919	0.805	0.923	0.779	0.903	0.746	0.860	0.811	0.927	0.735	0.858	0.666	0.723
7	350	4	nr	0.806	0.885	0.746	0.890	0.720	0.869	0.687	0.814	0.767	0.895	0.678	0.818	0.617	0.701
			un	0.818	0.921	0.779	0.925	0.758	0.905	0.730	0.851	0.786	0.928	0.712	0.848	0.648	0.708
	420	4	nr	0.807	0.885	0.744	0.891	0.716	0.871	0.685	0.820	0.766	0.896	0.674	0.823	0.614	0.700
			un	0.821	0.922	0.780	0.925	0.759	0.905	0.731	0.856	0.787	0.929	0.717	0.853	0.653	0.710
	490	4	nr	0.806	0.885	0.738	0.891	0.713	0.872	0.680	0.823	0.757	0.896	0.669	0.827	0.608	0.703
			un	0.820	0.921	0.778	0.925	0.757	0.905	0.729	0.856	0.782	0.929	0.716	0.853	0.649	0.709
8	480	5	nr	0.823	0.912	0.797	0.915	0.748	0.892	0.707	0.833	0.835	0.919	0.717	0.835	0.647	0.730
			un	0.834	0.940	0.815	0.942	0.776	0.920	0.735	0.860	0.859	0.941	0.737	0.856	0.662	0.716
	560	5	nr	0.825	0.913	0.795	0.917	0.744	0.894	0.704	0.836	0.835	0.920	0.714	0.838	0.643	0.732
			un	0.833	0.940	0.810	0.942	0.773	0.859	0.736	0.863	0.852	0.942	0.735	0.742	0.660	0.863
	640	5	nr	0.824	0.912	0.790	0.916	0.743	0.894	0.707	0.837	0.830	0.920	0.715	0.839	0.645	0.731
			un	0.830	0.939	0.805	0.942	0.771	0.919	0.734	0.863	0.845	0.942	0.731	0.859	0.658	0.714
9	630	5	nr	0.819	0.915	0.769	0.919	0.727	0.897	0.686	0.843	0.824	0.922	0.703	0.845	0.620	0.734
			un	0.827	0.941	0.788	0.944	0.763	0.921	0.728	0.855	0.833	0.945	0.723	0.851	0.646	0.709
	720	5	nr	0.817	0.914	0.763	0.917	0.721	0.896	0.679	0.843	0.821	0.921	0.700	0.845	0.615	0.733
			un	0.824	0.941	0.782	0.943	0.759	0.920	0.724	0.856	0.825	0.945	0.718	0.851	0.642	0.708
	810	5	nr	0.819	0.915	0.762	0.918	0.724	0.898	0.682	0.846	0.820	0.922	0.698	0.849	0.616	0.734
			un	0.825	0.941	0.783	0.944	0.761	0.921	0.723	0.856	0.821	0.946	0.718	0.852	0.641	0.707

부록: RSD분포에서 d와 a의 계산

$$d = P\left(\chi^2(k+2) < \chi^2_{h/n}(k)\right)/(h/n),$$

$$\alpha = (n-h)/n,$$

$$\chi^2_\alpha : 1-\alpha = P\left(\chi^2(k) \le \chi^2_\alpha\right),$$

$$d_\alpha = (1-\alpha)/P\left(\chi^2(k+2) \le \chi^2_\alpha\right),$$

$$d_2 = -P\left(\chi^2(k+2) \le \chi^2_\alpha\right)/2,$$
$$d_3 = -P\left(\chi^2(k+4) \le \chi^2_\alpha\right)/2,$$
$$b_1 = -2d_\alpha d_3/(1-\alpha),$$
$$b_2 = 1/2 + d_\alpha\left[d_3 - \chi^2_\alpha\left\{d_2 + (1-\alpha)/2\right\}/k\right]/(1-\alpha),$$
$$v_1 = (1-\alpha)b_1^2\left\{\alpha\left(d_\alpha\chi^2_\alpha/k - 1\right)^2 - 1\right\} - 2d_3d_\alpha^2\left\{3(b_1 - kb_2)^2 + (k+2)b_2(2b_1 - kb_2)\right\},$$
$$v_2 = n\left\{b_1(b_1 - kb_2)(1-\alpha)\right\}^2 d_\alpha^2,$$
$$a = 2v_2/(d_\alpha^2 v_1).$$

참고문헌

Aguilera, A. M., Escabias, M. and Valderrama, M. J. (2006). Using principal components for estimating logistic regression with high-dimensional multicollinear data, *Computational Statistics & Data Analysis*, **50**, 1905–1924.

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

Copas, J. B. (1988). Binary regression models for contaminated data, *Journal of the Royal Statistical Society, Series 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.

Hadi, A. S. (1994). A modification of a method for the detection of outliers in multivariate samples, *Journal of the Royal Statistical Society, Series B*, **56**, 393–396.

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.

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

Kim, B. Y. and Kahng, M. W. (2008). Principal components regression in logistic model, *The Korean Journal of Applied Statistics*, **21**, 571–580.

Kim, B. Y., Kahng, M. W. and Choi, M. A. (2007). Algorithm for the robust estimation in logistic regression, *The Korean Journal of Applied Statistics*, **20**, 551–559.

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.

Mason, R. L. and Gunst, R. F. (1985). Selecting principal components in regression, *Statistics & Probability Letters*, **3**, 299–301.

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*, John Wiley & Sons, New York.

Schaefer, R. L. (1986). Alternative estimators in logistic regression when the data are collinear, *Journal of Statistical Computation and Simulations*, **25**, 75–91.

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.

Principal Components Logistic Regression based on Robust Estimation

Bu-Yong Kim[1] · Myung Wook Kahng[2] · Hea-Won Jang[3]

[1]Department of Statistics, Sookmyung Women's University;
[2]Department of Statistics, Sookmyung Women's University;
[3]Enterprise Risk Team, FIST Global, Inc

(Received January 2009; accepted February 2009)

Abstract

Logistic regression is widely used as a datamining technique for the customer relationship management. The maximum likelihood estimator has highly inflated variance when multicollinearity exists among the regressors, and it is not robust against outliers. Thus we propose the robust principal components logistic regression to deal with both multicollinearity and outlier problem. A procedure is suggested for the selection of principal components, which is based on the condition index. When a condition index is larger than the cutoff value obtained from the model constructed on the basis of the conjoint analysis, the corresponding principal component is removed from the logistic model. In addition, we employ an algorithm for the robust estimation, which strives to dampen the effect of outliers by applying the appropriate weights and factors to the leverage points and vertical outliers identified by the V-mask type criterion. The Monte Carlo simulation results indicate that the proposed procedure yields higher rate of correct classification than the existing method.

Keywords: Datamining, multicollinearity, outlier, principal components logistic regression, robust estimation.

This research was supported by the Sookmyung Women's University Research Grants 2008.
[1]Corresponding author: Professor, Department of Statistics, Sookmyung Women's University, Chungpa-dong, Yongsan-ku, Seoul 140-742, Korea. E-mail: buykim@sm.ac.kr