응용통계연구 (2008)
21(4), 571-580

로지스틱모형에서의 주성분회귀

김부용[1] · 강명욱[2]

[1]숙명여자대학교 통계학과; [2]숙명여자대학교 통계학과

(2008년 3월 접수, 2008년 6월 채택)

요 약

로지스틱회귀분석은 고객관계관리나 신용위험관리 등의 분야에서 많이 사용되는 기법인데, 이러한 분야에서의 로지스틱회귀모형에는 연관성이 높은 설명변수들이 다수 포함되어 다중공선성의 문제를 유발하는 경우가 있다. 다중공선성이 존재하는 상황에서 최우추정량은 심각한 결함을 갖는다는 사실은 잘 알려졌다. 이 문제를 해결하기 위하여 로지스틱주성분회귀를 연구하되, 분석상의 주요 과정인 주성분 선정을 위한 방법을 새롭게 제안하였다. 추정량의 분산을 최소가 되게 하는 상태지수 값을 측정하고, 이 값에 영향을 미치는 주요 요인들을 컨조인트분석에 의해 파악하여 주성분 선정기준을 결정하는 모형을 구축하였다. 제안된 방법은 다중공선성 문제를 적절히 해결하면서도 모형의 적합성을 향상시킨다는 사실이 모의실험을 통하여 확인되었다.

주요용어: 고객관계관리, 신용위험관리, 다중공선성, 로지스틱주성분회귀, 컨조인트분석.

1. 서론

의학이나 생명과학 분야의 실험자료 분석에 주로 사용되던 로지스틱회귀분석이 최근에는 고객관계관리(CRM)나 신용위험관리 및 신용평가 분야에서 예측이나 분류를 위하여 많이 활용되고 있다. 구체적으로, 기업이나 고객의 행동예측, 신용평점, 신용위험분석 등에 로지스틱회귀모형이 많이 사용되고 있는데, 이러한 분야에서의 로지스틱회귀모형에 포함된 설명변수들 중에서 기업과 개인의 재무상태나 경제능력 등을 나타내는 변수들이 상당히 높은 연관성을 가질 수밖에 없다. 예를 들면, 보험회사에서 계약성공확률 예측을 위한 로지스틱모형에는 설명변수로서 연령, 소득, 재산액, 보험가입금액, 대출가능금액, 직업 등이 채택될 수 있는데, 이와 같은 연속형 설명변수들 중에는 태생적으로 연관성이 높은 경우가 있다. 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.

따라서 로지스틱회귀모형에서의 다중공선성 문제를 적절히 해결할 수 있는 방안으로서, 주성분 분석을 로지스틱회귀모형에 접목시킨 로지스틱주성분회귀(LPCR: logistic principal components regression)를 제안하고자 한다. 특히, 로지스틱주성분회귀분석 과정에서 매우 중요한 단계 중의 하나인 주성분의 선정방법을 새롭게 개발하고자 한다. 주성분의 선정에 관한 기존의 연구들은 주로 고유치들의 상대적 크기를 기준으로 주성분을 선정하는 것들인데, 객관적인 경계치가 없기 때문에 주관적으로 결정해

본 연구는 숙명여자대학교 2007년도 교비연구비 지원에 의해 수행되었음.
[1]교신저자: (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

야 하는 어려움이 있고, 반응변수를 전혀 고려하지 않는 방법이기 때문에 모형의 적합성이 매우 낮아질 수 있다는 단점을 가지고 있다. 반면에 선형회귀분석에서의 변수 선정 방법을 원용하여 모형의 적합성을 높일 수 있는 주성분을 선정하는 방법들이 있는데, 다중공선성 문제를 해결하지 못하는 경우가 있다. 그러므로 본 연구에서는 다중공선성 문제를 적절히 해결하면서 동시에 모형의 적합성을 유지시킬 수 있는 방법을 제안하고, 제안된 방법의 효율성과 적합성을 평가하기 위하여 Monte Carlo 모의실험을 실행하고자 한다.

2. 로지스틱주성분회귀

CRM이나 신용위험관리 분야에서 계약의 성공확률을 예측하거나, 보험금의 부정청구를 탐지하거나, 고객의 신제품 구매여부를 판별하거나, 은행대출 신청고객의 신용을 평가하거나, 기업의 부도확률을 측정하려는 경우가 있다. 이런 경우에는 반응변수가 이항형으로 측정되기 때문에 다음과 같은 로지스틱회귀모형을 적용한다.

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

여기서 y_i는 이항형 반응변수이고, $\boldsymbol{x}_i^T$는 설명변수 행렬 $X_{n\times p}$의 i번째 행이며, $\pi(\boldsymbol{x}_i)$는 $\boldsymbol{x}_i$에서의 조건확률을 의미한다. 그리고 $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_k]^T$는 회귀계수인데, $\exp(\beta_j)\,(j = 1, 2, \ldots, k)$는 오즈비라고 불리며 해당 설명변수가 분류 결정에 미치는 영향의 정도를 파악하는데 활용된다.

2.1. 다중공선성

로지스틱회귀모형 (2.1)에 채택된 설명변수들은 높은 연관성을 갖는 경우가 있으며, 특히 CRM분야에서 다수의 설명변수가 모형에 도입되면 그럴 가능성은 매우 높다. 설명변수 간에 상당히 높은 수준의 선형의존성이 존재하는 현상을 다중공선성이라 하는데, 이는 추정량의 분산을 매우 크게 하는 등 심각한 문제를 야기하게 된다. Schaefer (1986)는 다중공선성이 최우추정량에 미치는 영향을 분석하였는데, 다중공선성이 존재함에도 불구하고 최우추정법을 적용하는 경우 예측이나 분류 결과가 왜곡될 수밖에 없다. 그러므로 자료에 다중공선성이 존재하는지 사전에 파악하고 적절한 조치를 취할 필요가 있다. 다중공선성을 진단하기 위한 척도로서는 분산팽창인자, 상태수, 상태지수(CI: condition index) 그리고 분산분해비율 등이 활용된다. 본 논문에서는 다중공선성 문제의 해결을 위한 방안으로서 로지스틱주성분회귀를 제안하고자 한다.

2.2. 로지스틱주성분회귀

주성분분석을 로지스틱회귀모형에 적용하여 다중공선성 문제를 해결하려는 시도가 로지스틱주성분회귀다. 연속형 설명변수들을 중심화 하여 비정칙치분해를 실행하면

$$\widetilde{X} = UDV^T, \quad \widetilde{X} = (\widetilde{x}_{ij})_{n\times k}, \quad \widetilde{x}_{ij} = x_{ij} - \bar{x}_j \tag{2.1}$$

인데, $U_{n\times k}$와 $V_{k\times k}$는 직교행렬이며 $D_{k\times k}$는 비정칙치 $\mu_1, \mu_2, \ldots, \mu_k$로 구성된 대각행렬이다. 행렬 $\widetilde{X}^T\widetilde{X}$의 고유치를 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 0$라 하고 고유치들을 원소로 갖는 대각행렬을 Λ라 하면

$$\widetilde{X}^T\widetilde{X} = V\Lambda V^T$$

인데, 행렬 $V = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k]$는 $\widetilde{X}^T\widetilde{X}$의 각 고유치에 대응하는 고유벡터 $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k$로 구성된 고유벡터행렬이다. 직교행렬 V에 의해 변환된 행렬

$$Z = \widetilde{X}V, \quad Z = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_k] \tag{2.2}$$

의 각 열을 구성하는 $\boldsymbol{z}_j$을 주성분이라 하며 주성분 (2.2)에 의해 변환된 로지스틱모형의 회귀계수를 추정하게 된다. 그런데 주성분선형회귀모형에서의 최소자승추정량의 분산은

$$\sigma^2(Z^TZ)^{-1} = \sigma^2\Lambda^{-1} \tag{2.3}$$

이므로 작은 고유치는 주성분모형에서는 물론 원래모형에서의 추정량 분산이 크게 팽창되는 현상을 유발하게 되는데, 고유치의 크기를 바탕으로 다중공선성을 진단하는 근거가 (2.3)에 있다. 한편, 로지스틱모형의 회귀계수 추정을 위한 반복재가중최소자승추정(IRLS: iterative reweighted least square estimation) 알고리즘의 각 반복과정에서는 기본적으로 최소자승추정법을 적용하므로 추정량의 분산이 팽창하는 것을 피할 수 없다. 따라서 어느 방식으로든 매우 작은 고유치에 대응하는 주성분을 제외시켜 변환한 로지스틱모형의 회귀계수를 추정하고 이를 다시 원래모형에서의 추정치로 역변환시킨다. 주성분 선정 과정을 거쳐서 모형에 포함될 주성분이 결정되면 최우추정법에 의해 회귀계수를 추정하게 되는데, 로그우도함수를 설정하고 그 함수를 최대가 되게 하는 값을 찾아 회귀계수 추정치로 결정한다. 변환된 로지스틱회귀모형에 따른 로그우도함수는

$$\begin{aligned} \ln L(y_i\,;\boldsymbol{\gamma}) &= \sum_{i=1}^{n}[y_i \ln \pi_i + (1-y_i)\ln(1-\pi_i)], \\ \pi_i = \pi(\boldsymbol{z}_i) &= \frac{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})}{1+\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})}, \quad i=1,\ldots,n \end{aligned} \tag{2.4}$$

이며, 함수 (2.4)가 최대가 되게 하는 $\boldsymbol{\gamma}$가 최우추정치 $\hat{\boldsymbol{\gamma}}$이 되는데 이를 구하기 위한 우도방정식은

$$Z^T(\boldsymbol{y}-\hat{\boldsymbol{\pi}}) = \mathbf{0}, \quad \hat{\boldsymbol{\pi}} = \frac{\exp(Z\hat{\boldsymbol{\gamma}})}{\mathbf{1}+\exp(Z\hat{\boldsymbol{\gamma}})} \tag{2.5}$$

이다. 방정식 (2.5)에서 $\hat{\boldsymbol{\gamma}}$을 구하기 위해 최적화기법인 IRLS-알고리즘을 사용하며 $(r+1)$번째 반복과정에서 $\hat{\boldsymbol{\gamma}}$은

$$\hat{\boldsymbol{\gamma}}^{(r+1)} = \hat{\boldsymbol{\gamma}}^{(r)} + \left(Z^TW^{(r)}Z\right)^{-1}Z^T\left(\boldsymbol{y}-\hat{\boldsymbol{\pi}}^{(r)}\right), \quad W^{(r)} = \mathrm{diag}\left[\hat{\pi}_i{}^{(r)}\left(1-\hat{\pi}_i{}^{(r)}\right)\right]$$

와 같이 얻어지는데, 적절한 반복완료기준에 의해 반복과정이 수렴할 때까지 진행하여 최종적인 추정치를 구한다 (IRLS-알고리즘에 관한 자세한 내용은 Kim (2005)을 참조할 수 있다). 추정알고리즘에 의해 주성분모형에서의 추정치 $\hat{\gamma}$을 구한 후, 로지스틱모형에서의 회귀계수 추정치 $\hat{\boldsymbol{\beta}}_{PC}$는 최우추정량의 불변성에 의해 다음과 같이 구할 수 있다.

$$\hat{\boldsymbol{\beta}}_{PC} = V\hat{\boldsymbol{\gamma}}. \tag{2.6}$$

그런데 로지스틱 회귀모형에 β_0을 포함시키기 위해서는 주성분모형에서 $\hat{\gamma}$을 구할 때 $\widetilde{Z} = [\mathbf{1}\,|\,Z]$을 적용하고, (2.6)에서 $\hat{\boldsymbol{\beta}}_{PC}$을 구할 때에는 V대신에 $\widetilde{V} = \begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & V \end{bmatrix}$, $\mathbf{0} = [0,\ldots,0]^T$을 적용한다.

3. 주성분의 선정방법

로지스틱주성분회귀에서의 주성분 선정방법에 관한 연구는 활발하게 이루어지지 않았다. 그러므로 선형회귀모형을 대상으로 하는 주성분회귀에서의 주성분 선정방법을 원용할 수 있는데, 선형회귀에서의 주성분 선정방법은 크게 두 가지로 나눌 수 있다. 첫째는 Mason과 Gunst (1985)와 같이 고유치들의 상대적 크기를 기준으로 주성분을 선정하는 것들인데, 고유치의 크기 순서로 합계를 구하여 일정

표 **3.1.** 속성 및 수준의 종류와 컨조인트분석 결과

속성	수준	부분가치 추정치	중요도
설명변수의 분포형태	정규분포 / 균일분포	−1.59873 / 1.59873	19.3
설명변수의 수(개)	3 / 4 / 5 / 6 7 / 8 / 9	−4.00228 / −1.49649 / −0.19960 / 0.02473 0.69955 / 2.24955 / 2.72454	40.6
관찰치의 수(배)	30 / 40 / 50 / 60 70 / 80 / 90	−1.45890 / −1.63185 / −0.64094 / 0.10906 −0.34549 / 2.40451 / 1.56361	24.4
다중공선성에 관련된 변수의 수(개)	2 / 3 / 4 / 5	0.54854 / −1.24875 / −0.65753 / 1.35774	15.7

한 수준(예를 들면, 90%)을 초과하는 고유치에 해당하는 주성분을 모형설정에서 제외시키는 방법이다. 그러나 작은 고유치에 대응하는 주성분이 오히려 적합성을 높여 주는 경우가 있다는 사실이 Jolliffe (1982)와 Hadi와 Ling (1998)에 의해 밝혀졌다. 그러므로 반응변수를 전혀 고려하지 않는 이 방법을 적용하면 모형의 적합성이 크게 낮아질 수 있다는 문제가 있고, 객관적인 방식이 아니라 모의실험이나 연구자의 주관에 근거하여 경계치가 결정된다는 단점도 가지고 있다. 둘째는 Mansfield 등 (1977)과 같이 앞의 선정방법들이 반응변수를 전혀 고려하지 않는다는 한계를 지적하고, 각 주성분의 유의성 검정에 바탕을 둔 변수선택법을 통하여 주성분을 선정하는 방법인데, 이 방식에서는 적합성에만 의존하므로 다중공선성 문제를 적절히 해결하지 못하는 경우가 발생할 수 있다. 본 논문에서는 선정방법들이 갖는 단점들을 극복하기 위하여 두 가지 방법을 절충한 새로운 방법, 즉 다중공선성 문제와 적합성이 낮아지는 문제를 동시에 해결할 수 있는 방법을 제안한다.

3.1. 컨조인트분석

다중공선성이 유발하는 심각한 문제는 추정량의 분산이 팽창한다는 것이므로 주성분 선정법의 초점을 추정량 분산의 최소화에 두기로 하였다. 추정량 분산의 크기에 영향을 미치는 주요 속성을 파악하고 속성들의 상대적인 중요도가 얼마인지 측정하기 위하여 컨조인트분석을 활용하였으며, 분석 결과를 바탕으로 CI의 경계치 상한과 하한을 구하는 모형을 구축하였다. 컨조인트분석을 위한 주요 속성으로는 설명변수의 수, 관찰치의 수(설명변수의 수에 비례한 크기), 다중공선성에 관련된 변수의 수, 설명변수의 분포형태 등을 선정하였으며 각 속성별 수준은 표 3.1과 같다.

컨조인트분석을 위한 자료는 속성 수준들의 모든 조합에 따라서 다양하게 생성하였다. 다중공선성이 존재하는 자료를 생성하기 위하여 Aguilera 등 (2006)이 사용한 방법을 적용하였으며, 자료생성과 분석에는 SAS/IML을 사용하였다. 분포의 특징이 상이한 정규분포와 균일분포로부터 각각 난수를 생성하여 행렬 $G_{n\times k}$을 구성하고, 균일분포로부터 별도의 난수를 생성하여 행렬 $H_{k\times k}$을 구성한 후, 설명변수 행렬은 $X_{n\times k} = GH$에 의해 생성하였다. 그리고 행렬 $X_{n\times k}$에 벡터 $\mathbf{1}_n = [1,\ldots,1]^T$을 삽입하여 설명변수 행렬 $X_{n\times p}$를 완성하였다. 한편, 사전에 지정한 회귀계수 값 $\boldsymbol{\beta} = [\beta_0 \,|\, \mathbf{1}_k^T]$, $\beta_0 = -\sum_{j=1}^{k}\sum_{i=1}^{n} x_{ij}/n$과 행렬 $X_{n\times p}$를 적용하여 (2.1)에서의 $\pi(\boldsymbol{x}_i)$을 계산하고, 자료의 완전분리 현상을 예방하기 위하여 균일분포로부터 난수 u를 생성하여 $\pi(\boldsymbol{x}_i) \geq u$이면 $y_i = 1$, $\pi(\boldsymbol{x}_i) < u$이면 $y_i = 0$으로 반응변수 값을 생성하였다.

컨조인트분석을 위하여 부적합한 프로파일을 제외하고 265개 조합별로 1,000개씩의 자료를 생성하였다. 각 자료에 로지스틱주성분회귀를 적용하여 CI의 경계치가 어느 크기일 때 회귀계수 추정치들의 분산 합이 최소가 되는지 조사하였다. 컨조인트분석을 위한 모형은 CI의 경계치를 반응변수로 하고, 모든

속성을 설명변수로 도입하되 다음과 같이 지시변수 형태로 표현하였다.

$$\boldsymbol{c} = \xi_0 + \xi_1 D + \sum_{j=2}^{7} \xi_j K_j + \sum_{j=8}^{13} \xi_j T_j + \sum_{j=14}^{16} \xi_j M_j + \epsilon, \tag{3.1}$$

여기서 $\boldsymbol{c}$는 추정치의 분산 합이 최소가 되는 CI의 경계치며, D는 설명변수의 분포형태, K는 설명변수의 수, T는 K에 비례한 크기의 관찰치 수, M은 다중공선성에 관련된 변수의 수를 나타내는 지시변수들이다. 모형 (3.1)을 바탕으로 한 컨조인트분석의 결과는 표 3.1에 수록되었는데, 설명변수 수의 중요도가 가장 높으며, 다음은 관찰치의 수, 설명변수의 분포형태, 다중공선성에 관련된 변수 수의 순서로 분석되었다.

3.2. 주성분 선정기준의 설정

컨조인트분석 결과 추정량 분산의 최소치가 얻어지는 CI의 크기에 가장 영향을 많이 미치는 요인은 설명변수의 수와 관찰치의 수로 나타났다. 다중공선성에 관련된 변수의 수와 설명변수의 분포형태도 CI의 크기에 영향을 미치기는 하지만 영향력은 상대적으로 작은 것으로 분석되었다. 그런데 회귀계수 추정치들의 분산 합이 최소가 되게 하는 CI의 크기는 모두 일정한 범위(5~25)에 있다는 사실을 발견하였다. 그러므로 설명변수의 수와 관찰치의 수를 바탕으로 CI의 경계치를 결정하는 모형을 설정하되, 경계치의 하한(C_L)과 상한(C_U)을 구하기로 하였다. 그러면 다중공선성에 관련된 변수의 수와 설명변수 분포형태의 영향력은 C_L과 C_U의 범위 내에 흡수되어 반영될 것으로 기대할 수 있다. 더욱이 다중공선성에 관련된 변수의 수와 설명변수의 분포형태는 로지스틱회귀분석에 앞서 사전에 알 수 있는 요인이 아니기 때문에 CI의 경계치를 결정하는 모형에는 포함시킬 수 없다. 따라서 설명변수의 수와 관찰치의 수를 설명변수로 하고 최소 추정량분산에 대응하는 CI를 반응변수로 채택한 선형모형

$$\boldsymbol{c} = R\boldsymbol{\psi} + \boldsymbol{\eta} \tag{3.2}$$

을 설정하였다. 여기서 $R = [\mathbf{1} \mid K \mid N]$인데, K는 설명변수의 수이고 N은 관찰치의 수이다. 컨조인트분석을 위한 265개 프로파일 자료에 (3.2)를 적용하여 $\boldsymbol{c}$의 95% 신뢰상한과 하한

$$\hat{c}_i \pm t_{\frac{\alpha}{2}}(262) \left[\left(\frac{\boldsymbol{c}^T\boldsymbol{c} - \hat{\boldsymbol{\psi}} R^T \boldsymbol{c}}{262} \right) \boldsymbol{r}_i^T (R^T R)^{-1} \boldsymbol{r}_i \right]^{\frac{1}{2}}$$

(단, $\boldsymbol{r}_i^T$는 행렬 R의 i번째 행임)을 각각 구하여, 다음과 같이 주성분 선정을 위한 경계치의 상한과 하한으로 채택하였다. 즉,

$$\begin{aligned} C_L &= \boldsymbol{r}_o^T \hat{\boldsymbol{\psi}} - (1.96) \left[(11.41465) \boldsymbol{r}_o^T A \boldsymbol{r}_o \right]^{\frac{1}{2}}, \\ C_U &= \boldsymbol{r}_o^T \hat{\boldsymbol{\psi}} + (1.96) \left[(11.41465) \boldsymbol{r}_o^T A \boldsymbol{r}_o \right]^{\frac{1}{2}}, \end{aligned} \tag{3.3}$$

여기서

$$\hat{\boldsymbol{\psi}}^T = \begin{bmatrix} 2.65004 & 0.96053 & 0.06571 \end{bmatrix},$$

$$A = \begin{bmatrix} 0.0864085173 & -0.0066986250 & -0.0006144280 \\ -0.0066986250 & 0.0012627787 & -0.0000223390 \\ -0.0006144280 & -0.0000223390 & 0.0000118001 \end{bmatrix}$$

이며, $\boldsymbol{r}_o$는 로지스틱회귀분석 대상 자료의 설명변수 수와 관찰치 수로 구성된 벡터다.

3.3. 주성분 선정방법

제안된 선정방법은 두 단계로 구성되었다. 첫째 단계에서는 (3.3)에 의해 구한 경계치 상한 C_U보다 CI가 크면, 강력한 다중공선성이 존재하는 것으로 판단하여 그 CI에 대응되는 주성분을 모형에서 제외시킴으로써 다중공선성의 문제를 해결한다. 둘째 단계에서는 CI가 C_U와 C_L 사이에 속하면, 강력한 수준의 다중공선성은 아니더라도 무시할 수 없는 수준의 다중공선성이 존재하는 것으로 판단하여 그 CI에 대응하는 주성분에 대하여 유의성 검정을 실행한다. 즉, CI의 크기 순서에 따라 각 주성분의 유의성에 대한 검정을 순차적으로 적용하여 해당 주성분을 모형에 포함시킬지 여부를 결정한다. 이와 같은 선정방법을 적용하면 강력한 다중공선성은 안전하게 제거할 수 있으며, 상당한 정도의 다중공선성이 존재하는 경우에는 반응변수에 대한 주성분의 기여도를 가설검정을 통하여 판단함으로써 최적의 주성분을 선정할 수 있다.

주성분 선정의 둘째 단계에서 실행하는 검정은 Hosmer과 Lemeshow (2000)에 기술된 우도비 검정통계량(DV: deviance)에 바탕을 둔다. 통계량 DV는 관심 있는 적합모형과 포화모형의 우도의 크기를 비교한 통계량인데, 식 (2.5)에서와 같이 $\hat{\pi}(\boldsymbol{z}_i)$를 $\hat{\pi}_i$로 간단히 표기하면 통계량은

$$\mathrm{DV} = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right] \tag{3.4}$$

가 된다. 그러므로 어느 주성분이 로지스틱모형에 포함되는 것이 유의적인지 아닌지를 검정하기 위해서는 그 주성분이 포함된 모형과 제외된 모형에서의 DV의 차이를 비교하게 된다. 즉, k개의 주성분을 포함하는 모형을 M_k이라 하고 M_k에서 q개의 주성분이 제거된 모형을 M_q이라 하면, 두 모형의 DV값의 차이

$$Q^2 = \mathrm{DV}(M_q) - \mathrm{DV}(M_k) \sim \chi^2_q \tag{3.5}$$

을 검정통계량으로 적용할 수 있다. 이 단계에서는 유의수준만 결정하면 된다는 점에서 기존의 방법들과 달리 객관적인 선정기준을 적용한다는 특징을 갖는다.

4. 로지스틱주성분회귀의 평가

새로운 주성분 선정방법을 적용한 로지스틱주성분회귀의 효율성과 적합성을 평가하기 위하여 Monte Carlo 모의실험을 실행하였다. 다중공선성이 존재하는 다양한 자료를 대량으로 생성한 후, 고유치 크기에 바탕을 둔 로지스틱주성분회귀 E-LPCR과 본 연구에서 제안한 C-LPCR을 각각의 자료에 적용하여 추정치의 분산과 정분류율을 측정하여 비교하였다.

4.1. 모의실험

각 설명변수는 정규분포와 균일분포로부터 생성하였는데, 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 조합에 대한 자료를 생성하였다. 다중공선성이 존재하는 자료를 생성하기 위하여 제 3장에서 사용한 방법을 적용하였으며, 다중공선성 진단을 위한 척도들을 적용하여 설명변수들 사이에 다중공선성이 어느 정도 존재하는지 확인하였다.

표 4.1. 효율성과 적합성의 비교

설명변수 수	관찰치 수	다중공선성 변수 수	설명변수 분포	AVE		ACCR	
				E-LPCR	C-LPCR	E-LPCR	C-LPCR
3	90	2	NR	0.2424	0.1887	0.7356	0.7099
			UF	0.1605	0.1837	0.7782	0.7692
	120	2	NR	0.1201	0.1380	0.7322	0.7262
			UF	0.1279	0.1475	0.7793	0.7719
	150	2	NR	0.1007	0.1164	0.7346	0.7302
			UF	0.1082	0.1213	0.7809	0.7752
4	160	2	NR	0.0989	0.1110	0.7363	0.7316
			UF	0.0996	0.1087	0.7725	0.7694
		3	NR	0.1746	0.1323	0.7794	0.7890
			UF	0.1711	0.1509	0.8171	0.8265
	200	2	NR	0.0845	0.0969	0.7356	0.7312
			UF	0.0859	0.0933	0.7709	0.7680
		3	NR	0.1342	0.1075	0.7822	0.7927
			UF	0.1341	0.1229	0.8216	0.8309
	240	2	NR	0.0767	0.0835	0.7349	0.7335
			UF	0.0767	0.0837	0.7741	0.7707
		3	NR	0.1249	0.0979	0.7818	0.7916
			UF	0.1112	0.1066	0.8240	0.8290
5	200	3	NR	0.1334	0.1154	0.7748	0.7829
			UF	0.1321	0.1236	0.8054	0.8118
		4	NR	0.2727	0.1349	0.7995	0.8309
			UF	0.2545	0.1684	0.8298	0.8569
	250	3	NR	0.1040	0.0910	0.7774	0.7831
			UF	0.1099	0.1043	0.8037	0.8045
		4	NR	0.1984	0.1050	0.8062	0.8314
			UF	0.1736	0.1341	0.8341	0.8528
	300	3	NR	0.0954	0.0826	0.7758	0.7803
			UF	0.0892	0.0919	0.8018	0.8033
		4	NR	0.1606	0.0935	0.8104	0.8330
			UF	0.1497	0.1119	0.8336	0.8531
6	300	3	NR	0.0888	0.0823	0.7723	0.7761
			UF	0.0877	0.0854	0.7989	0.8016
		4	NR	0.1427	0.0956	0.7979	0.8162
			UF	0.1382	0.1113	0.8177	0.8323
	360	3	NR	0.0768	0.0714	0.7774	0.7802
			UF	0.0692	0.0713	0.8002	0.8028
		4	NR	0.1261	0.0811	0.7989	0.8122
			UF	0.1086	0.0911	0.8193	0.8330
	420	3	NR	0.0711	0.0615	0.7780	0.7808
			UF	0.0626	0.0627	0.8040	0.8035
		4	NR	0.1089	0.0739	0.8023	0.8167
			UF	0.0970	0.0874	0.8222	0.8317

7	350	3	NR	0.0778	0.0725	0.7708	0.7739
			UF	0.0758	0.0775	0.7964	0.7993
		4	NR	0.1206	0.0917	0.7917	0.8056
			UF	0.1200	0.0937	0.8104	0.8182
	420	3	NR	0.0644	0.0612	0.7765	0.7780
			UF	0.0637	0.0654	0.7973	0.7977
		4	NR	0.1068	0.0741	0.7919	0.8070
			UF	0.0980	0.0832	0.8153	0.8217
	490	3	NR	0.0574	0.0548	0.7774	0.7793
			UF	0.0614	0.0558	0.7977	0.7985
		4	NR	0.0861	0.0673	0.7965	0.8061
			UF	0.0809	0.0757	0.8156	0.8200
8	480	3	NR	0.0557	0.0577	0.7747	0.7758
			UF	0.0557	0.0571	0.7945	0.7945
		5	NR	0.1222	0.0759	0.8087	0.8233
			UF	0.1309	0.0986	0.8165	0.8346
	560	3	NR	0.0503	0.0515	0.7748	0.7761
			UF	0.0498	0.0501	0.7960	0.7962
		5	NR	0.1134	0.0719	0.8101	0.8263
			UF	0.1078	0.0830	0.8209	0.8340
	640	3	NR	0.0444	0.0463	0.7763	0.7756
			UF	0.0438	0.0483	0.7959	0.7949
		5	NR	0.1033	0.0668	0.8074	0.8252
			UF	0.0950	0.0774	0.8214	0.8311
9	630	4	NR	0.0646	0.0546	0.7898	0.7960
			UF	0.0603	0.0580	0.8052	0.8090
		5	NR	0.0966	0.0651	0.8031	0.8191
			UF	0.0909	0.0743	0.8177	0.8271
	720	4	NR	0.0544	0.0479	0.7938	0.7964
			UF	0.0509	0.0532	0.8068	0.8098
		5	NR	0.0843	0.0600	0.8058	0.8172
			UF	0.0761	0.0695	0.8189	0.8244
	810	4	NR	0.0479	0.0449	0.7918	0.7967
			UF	0.0461	0.0497	0.8064	0.8072
		5	NR	0.0860	0.0560	0.8059	0.8196
			UF	0.0685	0.0637	0.8179	0.8260

제안된 방법의 효율성을 비교 평가하기 위하여 다양한 특성과 규모의 생성자료에 E-LPCR과 C-LPCR을 각각 적용하여 1,000개씩의 추정치를 구하고, 추정치의 분산의 평균(AVE: average of variances of the estimates, $\sum_{j=1}^{p} \widehat{\mathrm{Var}}(\hat{\beta}_j)/p$)을 계산하였다. 한편, E-LPCR과 C-LPCR의 적합성을 평가하기 위하여 각각 1,000개의 자료의 정분류율의 평균(ACCR: average of correct classification rate)을 측정하였다.

4.2. 평가 결과

두 방법을 적용하여 구한 AVE와 ACCR은 표 4.1에 수록되었다. 다중공선성에 관련된 변수의 수가 2인 경우를 제외하고 대부분의 경우에서 E-LPCR보다 C-LPCR의 AVE가 작으므로 제안된 방법의

효율성이 상대적으로 우수하다고 할 수 있다. 그런데 다중공선성에 관련된 설명변수의 수가 2인 경우는 다중공선성의 탐지와 문제해결이 용이하므로 실제로 별 문제는 없는 것으로 판단된다. 그리고 C-LPCR의 ACCR이 대부분의 경우에 크게 측정되었으므로 제안된 방법의 적합성 역시 E-LPCR보다 우수한 것으로 나타났다. 따라서 제안된 방법은 다중공선성 문제를 적절히 해결하면서도 모형의 적합성을 떨어뜨리지 않는다는 장점을 갖는 것으로 평가되었다.

5. 결론

고객관계관리나 신용위험관리 등의 분야에서 로지스틱회귀분석을 많이 활용하는데, 설명변수들의 특성상 흔히 다중공선성의 문제를 야기하게 된다. 본 논문에서는 로지스틱회귀분석 자료에 존재하는 다중공선성의 문제를 해결할 수 있는 주성분회귀분석에 관하여 연구하였다. 특히 주성분을 선정하는 방법을 새롭게 제안하였는데, 모의실험을 통해 기존의 주성분 선정방법을 적용한 경우보다 제안된 방법이 효율성과 적합성의 관점에서 상대적으로 우수한 것으로 평가되었다.

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.

Hadi, A. S. and Ling, R. F. (1998). Some cautionary notes on the use of principle components regression, *The American Statistician*, **52**, 15–19.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*, John Wiley & Sons, New York.

Jolliffe, I. T. (1982). A note on the use of principal components in regression, *Applied Statistics*, **31**, 300–303.

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

Mansfield, E. R., Webster, J. T. and Gunst, R. F. (1977). An analytic variable selection technique for principal component regression, *Applied Statistics*, **26**, 34–40.

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

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

Principal Components Regression in Logistic Model

Bu-Yong Kim[1] · Myung Wook Kahng[2]

[1]Dept. of Statistics, Sookmyung Women's University;
[2]Dept. of Statistics, Sookmyung Women's University

(Received March 2008; accepted June 2008)

Abstract

The logistic regression analysis is widely used in the area of customer relationship management and credit risk management. It is well known that the maximum likelihood estimation is not appropriate when multicollinearity exists among the regressors. Thus we propose the logistic principal components regression to deal with the multicollinearity problem. In particular, new method is suggested to select proper principal components. The selection method is based on the condition index instead of the eigenvalue. When a condition index is larger than the upper limit of cutoff value, principal component corresponding to the index is removed from the estimation. And hypothesis test is sequentially employed to eliminate the principal component when a condition index is between the upper limit and the lower limit. The limits are obtained by a linear model which is constructed on the basis of the conjoint analysis. The proposed method is evaluated by means of the variance of the estimates and the correct classification rate. The results indicate that the proposed method is superior to the existing method in terms of efficiency and goodness of fit.

Keywords: Customer relationship management, credit risk management, multicollinearity, logistic principal components regression, conjoint analysis.

This research was supported by the Sookmyung Women's University Research Grants 2007.
[1]Corresponding author: Professor, Dept. of Statistics, Sookmyung Women's University, Chungpa-dong, Yongsan-gu, Seoul 140-742, Korea. E-mail: buykim@sm.ac.kr
[2]Professor, Dept. of Statistics, Sookmyung Women's University, Chungpa-dong, Yongsan-gu, Seoul 140-742, Korea. E-mail: mwkahng@sm.ac.kr