응용통계연구 (2010)
23(5), 967–975
DOI: 10.5351/KJAS.2010.23.5.967

주성분회귀분석에서 주성분선정을 위한 새로운 방법

김부용[1] · 신명희[2]

[1]숙명여자대학교 통계학과, [2]웅진코웨이(주) 고객전략팀

(2010년 7월 접수, 2010년 8월 채택)

요 약

데이터마이닝 분야에서의 회귀모형에는 연관성이 높은 설명변수들이 포함되어 다중공선성을 유발하는 경우가 많은데, 다중공선성이 야기하는 문제를 해결하기 위하여 주성분회귀분석을 적용할 수 있다. 이 분석에서는 적절한 주성분을 선정하는 과정이 핵심인데, 기존의 선정방법들은 다중공선성을 잘 해결하지 못하거나 모형의 적합성을 저하시킨다는 지적을 받고 있다. 따라서 본 논문에서는 다중공선성 문제와 적합성 저하 현상을 동시에 해결할 수 있는 새로운 선정방법을 제안하였다. 다중공선성에 의해 최소제곱추정량의 분산이 팽창되는 문제를 주성분회귀에 의해 해결할 수 있지만, 주성분의 일부를 선정함에 따라 발생하는 편의도 동시에 통제해야 한다. 따라서 주성분회귀추정량의 평균제곱오차를 최소가 되게 하는 상태지수를 측정하고, 이 값에 영향을 미치는 주요 요인들을 컨조인트분석에 의해 파악하여 주성분 선정기준 모형을 구축하였다. 선정기준의 상한과 하한을 설정하고, 상태지수가 상한을 초과하면 해당 주성분을 제외시키고, 하한에 미달하면 해당 주성분을 포함시킨다. 그리고 상한과 하한 사이의 상태지수에 대응하는 주성분들에 대해서는 일반화선형검정을 순차적으로 적용하여 주성분을 선정하는 방법이다.

주요용어: 데이터마이닝, 다중공선성, 주성분회귀, 상태지수, 주성분선정.

1. 서론

통계학의 다양한 응용분야에서 널리 사용되는 회귀분석이 최근에는 데이터마이닝 분야에서 많이 활용되고 있다. 데이터마이닝 분야에서의 자료는 그 규모가 방대하다는 것뿐만 아니라, 수집 계획에 따라 엄격하게 통제되지 않는 상황에서 자료가 얻어지는 경우가 많다는 특징을 가지고 있다. 그래서 회귀분석 과정에서 적절한 선정방법에 의해 설명변수가 선정되더라도 그 설명변수들 사이에 높은 연관성이 존재하는 경우가 많다. 특히 기업의 재무 상태나 영업능력 등을 나타내는 변수들이나 고객의 재산수준이나 경제능력 등을 나타내는 변수들은 태생적으로 높은 연관성을 가질 수밖에 없다. 이와 같이 연관성이 상당히 높은 설명변수들이 회귀모형에 포함되면 심각한 다중공선성의 문제가 발생하게 된다. 즉, 다중공선성이 존재하는 자료에 회귀분석에서 일반적으로 사용되는 최소제곱추정법을 적용하는 경우 그 추정량의 분산이 지나치게 팽창되는 현상이 유발되기 때문에 그 추정량에 바탕을 둔 통계적 추론은 심각하게 왜곡된다는 것이다. 이러한 다중공선성의 문제를 해결하기 위한 하나의 방안으로서 주성분회귀분석(principal components regression)을 채택할 수 있는데, 이 분석 과정에서는 적절한 주성분을 선정하는 것이 매우 핵심적인 역할을 한다.

주성분의 선정을 위해 제시된 기존의 방법들은 주로 고유치들의 상대적 크기를 기준으로 주성분을 선정하는 것들인데, 이러한 방법들은 객관적인 경계치에 바탕을 둔 것이 아니기 때문에 분석자가 적절한 경

본 연구는 숙명여자대학교 2009년도 교비연구비 지원에 의해 수행되었음.
[1]교신저자: (140-742) 서울특별시 용산구 청파동, 숙명여자대학교 통계학과, 교수.
E-mail: buykim@sookmyung.ac.kr

계치를 주관적으로 결정해야 한다는 어려움이 있으며, 결정된 경계치의 크기에 따라 주성분회귀분석의 결과가 상이하게 얻어진다는 한계를 가지고 있다. 더욱이 이 방법들은 주성분들과 반응변수의 관계를 전혀 고려하지 않는 방법이기 때문에 회귀모형의 적합성을 매우 낮게 할 수 있다는 지적을 받고 있다. 반면에 회귀분석에서의 변수선정 방법을 주성분 선정에 적용함으로써 모형의 적합성을 떨어뜨리지 않도록 하려는 선정방법들이 제시되었는데, 이러한 방법들은 다중공선성의 문제를 근원적으로 해결하지 못할 수 있다는 결함을 가지고 있다. 따라서 본 연구에서는 다중공선성의 문제를 적절히 해결하면서도 동시에 모형의 적합성을 유지시킬 수 있는 새로운 선정방법을 제안한다. 한편, 제안된 선정방법을 비교 평가하기 위하여 Monte Carlo 모의실험을 실행하고자 한다.

2. 주성분회귀분석

회귀모형에 채택된 설명변수들 사이에 상당히 높은 연관성 혹은 선형의존성이 존재하는 경우가 있는데, 특히 데이터마이닝 분야에서 다수의 설명변수가 회귀모형에 도입되면 그럴 가능성은 매우 높아진다. 설명변수 간에 상당히 높은 수준의 선형의존성이 존재하는 현상을 다중공선성이라 하는데, 최소제곱법에 의한 회귀계수 추정량의 분산을 매우 크게 팽창시키는 문제를 야기한다. 그러므로 회귀분석을 실행하기에 앞서 자료에 다중공선성이 존재하는지 진단하고 (진단하기 위한 척도로서는 분산팽창인자, 상태수 및 상태지수, 그리고 분산분해비율 등이 활용됨), 다중공선성이 존재한다는 사실이 확인되면 문제를 극복할 수 있는 적절한 방안을 강구해야 한다. 다중공선성에 관련된 설명변수를 모형에서 제외시키지 않으면서 다중공선성의 문제를 해결하기 위해서는 능형회귀분석이나 주성분회귀분석을 채택할 수 있는데, 본 논문에서는 주성분회귀분석에 관하여 연구하고자 한다.

주성분회귀분석은 추정량에 어느 정도의 편의를 허용하는 대신에 분산의 지나친 팽창을 막으려는 시도인데, 선형회귀모형 $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ ($\boldsymbol{y}$는 반응변수 n-벡터, X는 $n \times p$ 연속형 설명변수 행렬, $\boldsymbol{\beta}$는 회귀계수 p-벡터, $\boldsymbol{\epsilon}$는 오차항 n-벡터임)의 변수들을 중심화 및 척도화(centering and scaling)한 모형,

$$\tilde{\boldsymbol{y}} = \widetilde{X}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\epsilon}} \tag{2.1}$$

에서 출발한다. 모형 (2.1)에서 $\widetilde{X}^T\widetilde{X}$는 설명변수의 상관행렬이라고 할 수 있는데, $\widetilde{X}^T\widetilde{X}$의 주성분을 구하기 위한 방법 중의 하나가 $\widetilde{X}$의 비정칙치분해(singular value decomposition), 즉 $\widetilde{X} = UDV^T$이다 (비정칙치분해는 Gram-Schmidt 방법 등에 의해 구할 수 있음). 여기서 $U_{n\times k}$는 직교행렬이며, $V_{k\times k} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k]$는 $\widetilde{X}^T\widetilde{X}$의 고유치 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 0$에 대응하는 고유벡터들로 구성된 직교행렬이다. 그리고 $D_{k\times k}$는 비정칙치 $\mu_1, \mu_2, \ldots, \mu_k$로 구성된 대각행렬인데, $\mu_j^2 = \lambda_j$의 관계가 성립한다. 한편, $\widetilde{X}^T\widetilde{X}$의 고유치들을 원소로 갖는 대각행렬을 Λ라 하면 $\widetilde{X}^T\widetilde{X} = V\Lambda V^T$이므로 모형 (2.1)의 회귀계수 최소제곱추정량의 분산은 $\sigma^2(V\Lambda V^T)^{-1}$와 같이 표현된다. 따라서 매우 작은 고유치들이 최소제곱추정량의 분산을 크게 팽창시키는 역할을 한다는 사실을 알 수 있으며, 그래서 고유치의 크기를 바탕으로 한 진단척도들이 다중공선성을 진단하는데 사용되는 것이다.

직교행렬 V에 의해 회전된 행렬 $Z = \widetilde{X}V$을 구성할 수 있는데, Z를 주성분행렬이라 하고 Z의 각 열, $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_k$을 주성분이라 한다. 주성분들에 의해 표현된 회귀모형은

$$\tilde{\boldsymbol{y}} = Z\boldsymbol{\gamma} + \tilde{\boldsymbol{\epsilon}}, \quad \boldsymbol{\gamma} = V^T\tilde{\boldsymbol{\beta}} \tag{2.2}$$

인데, 모형 (2.2)에서의 최소제곱추정량 $\hat{\boldsymbol{\gamma}}$의 분산은 $\sigma^2\Lambda^{-1}$이므로 매우 작은 고유치들은 모형 (2.1)에서는 물론 모형 (2.2)에서도 최소제곱추정량의 분산이 크게 팽창되는 현상을 유발하게 된다. 따라서 매우 작은 고유치들이 최소제곱추정량의 분산의 팽창에 미치는 영향을 제거하기 위하여, $\widetilde{X}^T\widetilde{X}$의 고유치

$\lambda_1, \lambda_2, \ldots, \lambda_k$ 중에서 상대적으로 작은 고유치들에 대응하는 s개의 주성분들을 적절한 방법에 의해 선정하여 모형에서 제외시키고 $g(=k-s)$개의 주성분들만 포함된 주성분모형을 다음과 같이 설정한다.

$$\tilde{\boldsymbol{y}} = Z_g \boldsymbol{\gamma}_{PC} + \tilde{\boldsymbol{\epsilon}}, \tag{2.3}$$

여기서 $Z_g = [Z_{n\times g} | 0_{n\times s}]$이다. 모형 (2.3)에서의 추정량 $\hat{\boldsymbol{\gamma}}_{PC} = (Z_g^T Z_g)^{-1} Z_g^T \tilde{\boldsymbol{y}}$은 최종적으로 모형 (2.1)에서의 추정량으로 변환되어야 하는데, 모형 (2.2)에서의 관계식 $\boldsymbol{\gamma} = V^T \tilde{\boldsymbol{\beta}}$와 최소제곱추정법의 불변성(invariance) 원리에 의해 모형 (2.1)의 주성분회귀추정량은

$$\hat{\tilde{\boldsymbol{\beta}}}_{PC} = V \left(Z_g^T Z_g \right)^{-1} Z_g^T \tilde{\boldsymbol{y}} \tag{2.4}$$

와 같이 구할 수 있다. 한편, 원래 회귀모형에서의 주성분회귀추정량 $\hat{\boldsymbol{\beta}}_{PC}$은 중심화 및 척도화 과정을 역으로 적용하면 구할 수 있다.

다음은 주성분회귀추정량 $\hat{\tilde{\boldsymbol{\beta}}}_{PC}$이 갖는 중요한 통계적 특성들을 살펴보기로 한다. 우선 대각행렬 $Q = \mathrm{diag}[\boldsymbol{\ell}_g, \mathbf{0}_s]$, $\boldsymbol{\ell}_g = [1, \ldots, 1]^T$, $\mathbf{0}_s = [0, \ldots, 0]^T$을 정의하면 $\hat{\boldsymbol{\gamma}}_{PC} = Q\hat{\gamma}$의 관계가 성립하므로, 추정량 (2.4)를 $\hat{\tilde{\boldsymbol{\beta}}}_{PC} = VQV^T \hat{\tilde{\boldsymbol{\beta}}}$와 같이 표현할 수 있다. 그리고 직교행렬 V를 $V = \left[V_g \,\vdots\, V_s \right]$와 같이 분할하면 $\hat{\tilde{\boldsymbol{\beta}}}_{PC} = (I - V_s V_s^T) \hat{\tilde{\boldsymbol{\beta}}}$로 표현된다. 따라서 $E(\hat{\tilde{\boldsymbol{\beta}}}_{PC}) = \tilde{\boldsymbol{\beta}} - V_s V_s^T \tilde{\boldsymbol{\beta}}$이기 때문에 V_s가 영행렬(null matrix)이 아니라면 $\hat{\tilde{\boldsymbol{\beta}}}_{PC}$은 편의추정량임을 알 수 있다. 한편, 대각행렬 Λ를

$$\Lambda = \begin{bmatrix} \Lambda_g | & 0 \\ 0 & | \, \Lambda_s \end{bmatrix}$$

와 같이 분할하면 주성분회귀추정량 $\hat{\tilde{\boldsymbol{\beta}}}_{PC}$의 분산-공분산행렬은 $\sigma^2 V_g \Lambda_g^{-1} V_g^T$가 된다. 그런데 모형 (2.1)에서의 최소제곱추정량인 $\hat{\tilde{\boldsymbol{\beta}}}$의 분산-공분산행렬은 $\sigma^2 (V_g \Lambda_g^{-1} V_g^T + V_s \Lambda_s^{-1} V_s^T)$이므로, 역시 V_s가 영행렬이 아니라면 $\hat{\tilde{\boldsymbol{\beta}}}_{PC}$의 분산이 $\hat{\tilde{\boldsymbol{\beta}}}$의 분산보다 작게 된다. 이와 같이 주성분회귀를 적용함으로써 다중공선성에 관련된 설명변수들의 일부를 모형에서 제외시키지 않고서도 추정량의 분산이 팽창하는 것을 막을 수 있다는 것 알 수 있다.

3. 주성분선정을 위한 방법

주성분회귀분석에서 가장 중요한 과정은 어느 주성분들을 모형에 포함시킬 것이지를 결정하는 것이다. 이는 최적의 주성분선정이 이루어져야 설명변수가 갖는 정보의 손실이 최소화될 수 있기 때문이다. 주성분선정을 위한 기존의 방법들은 주로 분산-공분산행렬이나 상관행렬의 고유치들의 상대적 크기를 기준으로 하는 방법들인데, Pidot (1969)은 1.0보다 작은 고유치들에 대응하는 주성분을 모형에서 제외시키는 기준을 적용하였으며, Jolliffe (1972)는 0.71보다 작은 고유치들에 대응하는 주성분을 제외시키는 방법을 제시하였다. 그리고 Marquardt (1970)는 작은 고유치들의 누적 비율이 일정한 수준($10^{-1} \sim 10^{-7}$)을 초과하지 않는 범위내의 고유치들에 해당하는 주성분을 모형에서 제외시키는 방법을 제안하였다. 그런데 이러한 방법들은 객관적인 기준에 의한 것이 아니라 모의실험자나 연구자의 주관적인 판단에 바탕을 둔 경계치를 적용한다는 단점을 가지고 있다. 한편, Mansfield 등 (1977)은 기존의 선정방법들이 주성분들과 반응변수와의 관계를 전혀 고려하지 않는다는 한계를 지적하고, 각 주성분에 대한 통계적 유의성 검정에 바탕을 둔 후진제거법에 의해 주성분을 선정하는 방법을 제안하였다. 특히 Jolliffe (1982)와 Hadi와 Ling (1998)은 매우 작은 고유치들에 대응하는 주성분이 반응변수와 높은 연관성을 가지고 있는 경우에는 큰 편의를 유발한다는 사실을 밝히고, 반응변수를 고려하지 않는 방법을 적용하면

모형의 적합성이 크게 낮아질 수 있다는 문제점을 제기하였다. 이러한 방법에서는 유의수준만 결정하면 되기 때문에 선정기준이 객관적이라는 장점은 있지만, 적합성에만 의존하므로 다중공선성의 문제를 근본적으로 해결하지 못하는 경우가 있다는 단점을 가지고 있다. 특히 Mason과 Gunst (1985)는 Jolliffe (1982)가 주장한 내용을 반박하는 연구결과를 제시하기도 하였다. 따라서 본 논문에서는 기존의 선정방법들이 갖는 단점들을 극복하기 위한 새로운 방법을 제안한다.

3.1. 주성분회귀추정량의 평균제곱오차

자료에 다중공선성이 존재하는 경우 주성분회귀추정량은 최소제곱추정량보다 작은 분산을 갖지만, 동시에 주성분의 일부만을 선정함에 따른 편의의 증가를 피할 수 없다는 사실을 제 2장에서 확인하였다. 그러므로 주성분회귀추정량의 분산이 축소되는 것과 편의가 증가되는 것을 적절한 수준에서 조정할 수 있는 주성분 선정방법이 요구된다. 따라서 주성분회귀추정량의 평균제곱오차인 MSE($\hat{\tilde{\boldsymbol{\beta}}}_{PC}$)을 최소가 되게 하는 선정방법에 초점을 맞추기로 한다. 그런데 $\hat{\tilde{\boldsymbol{\beta}}}_{PC}$는 벡터이므로 각 회귀계수에 대한 주성분회귀추정량의 평균제곱오차인 MSE($\hat{\tilde{\beta}}_{PCj}$)을 합계한 총평균제곱오차($\Omega$: total mean square error)를 최적화 대상으로 삼는다.

우선, 주성분회귀추정량의 분산-공분산행렬은 $\mathrm{Var}(\hat{\tilde{\boldsymbol{\beta}}}_{PC}) = \sigma^2 V_g \Lambda_g^{-1} V_g^T$이므로, 각 회귀계수에 대한 주성분회귀추정량의 분산인 $\mathrm{Var}(\hat{\tilde{\beta}}_{PCj})$을 합한 총분산($\Gamma$: total variance)은 다음과 같이 정의할 수 있다.

$$\Gamma = \sigma^2 \mathrm{trace}\left(V_g \Lambda_g^{-1} V_g\right) = \sigma^2 \sum_{j=1}^{g} \frac{1}{\lambda_j}, \tag{3.1}$$

그런데, $\lambda_j \geq 0$이므로 Γ는 g에 대해 단조증가함수임을 알 수 있다. 한편 주성분회귀추정량의 편의벡터는 $\mathrm{bias}(\hat{\tilde{\boldsymbol{\beta}}}_{PC}) = V_s V_s^T \tilde{\boldsymbol{\beta}}$이고 모형 (2.2)에서 $\boldsymbol{\gamma} = V^T \tilde{\boldsymbol{\beta}}$이므로, 각 회귀계수에 대한 주성분회귀추정량의 편의인 $\mathrm{bias}(\hat{\tilde{\beta}}_{PCj})$을 제곱하여 합한 총제곱편의($\Psi$: total squared bias)는 다음과 같다.

$$\Psi = \left(\tilde{\boldsymbol{\beta}}^T V_s\right)\left(V_s^T \tilde{\boldsymbol{\beta}}\right) = \sum_{j=g+1}^{k} \gamma_j^2.$$

이 결과로부터 Ψ는 g에 대해 단조감소함수임을 알 수 있다.

이와 같은 특성들로부터 모형에 포함되는 주성분의 개수인 g의 크기에 따라 주성분회귀추정량의 분산이 축소됨과 동시에 편의는 증가된다는 사실을 확인할 수 있다. 따라서 주성분의 적절한 선정이 매우 중요한 과정이라 할 수 있다. 본 연구에서는 주성분회귀추정량의 정확도(accuracy)와 정도(precision)를 함께 측정할 수 있는 척도인 $\Omega = \Gamma + \Psi$을 최소가 되게 하는 주성분 선정방법을 제안하고자 한다.

3.2. 주성분 선정기준의 설정

고유치 대신에 상태지수(CI; condition index: $c_j = \lambda_{\max}/\lambda_j$)의 크기를 주성분 선정의 기준으로 삼기로 한다. 즉, Ω을 최소화하는 CI의 상한(c_U)과 하한(c_L)을 설정하고 c_j가 c_L보다 작으면 이에 대응하는 주성분들을 모형에 포함시키고 c_j가 c_U보다 크면 대응하는 주성분들을 모형에서 제외시킨다. 그러나 c_j가 c_L과 c_U 사이에 속하면 해당 주성분들에 대해 유의성검정을 순차적으로 적용하여 주성분을 선정하는 방법이다. 이 방법을 적용하면 다중공선성의 문제와 동시에 모형의 적합성이 낮아지는 문제를 어느 정도 해결할 수 있을 것으로 기대된다.

CI의 선정기준인 c_U와 c_L은 Monte Carlo 모의실험을 통하여 결정하였다. 모의실험을 위한 자료의 설명변수 값은 다양한 규모와 특성을 갖도록 생성하였는데, 설명변수의 분포(D: 정규분포, 균일분포,

Cauchy분포), 설명변수의 수(K: 3, 4, 5, 6, 7, 8, 9), 설명변수의 수에 비례한 관찰치의 수(T: 30, 40, 50, 60, 70, 80, 90, 100배), 다중공선성에 관련된 변수의 수(M: 2, 3, 4, 5)의 모든 수준들의 조합에 대하여 자료를 생성하였다. $m(\leq k)$개의 변수 간에 다중공선성이 존재하는 자료를 생성하기 위하여, 특징이 상이한 3가지 분포 중의 하나로부터 난수를 생성하여 행렬 $G_{n\times m}$을 구성하고, 균일분포로부터 별도의 난수를 생성하여 행렬 $H_{m\times m}$을 구성한 후, G의 각 열의 선형조합으로 이루어진 행렬 $X_{n\times m}=GH$을 생성하였다. $X_{n\times m}$에 절편을 위한 벡터 $\boldsymbol{\ell}_n=[1,\ldots,1]^T$와 다중공선성이 존재하지 않는 행렬 $X_{n\times(k-m)}$을 생성하여 삽입함으로써 설명변수 행렬 $X_{n\times p}$을 완성하였다. 한편, 오차항은 정규분포로부터 생성하였으며, 사전에 지정한 회귀계수 값 $\boldsymbol{\beta}^0=[1,\ldots,1]^T$과 행렬 $X_{n\times p}$을 적용하여 반응변수 벡터 $\boldsymbol{y}$를 생성하였다. 각 설명변수에 상이한 seed를 지정하였으며 각 반복(1,000회)에도 상이한 seed를 지정하여 자료를 생성하였다.

각 조합($D-K-T-M$)별로 생성된 자료마다 다양한 CI 경계치를 적용하여 각 경계치에 대응하는 Γ와 Ψ의 추정치를 각각 $\widehat{\Gamma}=\sum_{j=1}^{p}\sum_{m=1}^{1000}(\hat{\beta}_{PCj}^{(m)}-\hat{\bar{\beta}}_{PCj})^2/999$, $\widehat{\Psi}=\sum_{j=1}^{p}\sum_{m=1}^{1000}(\hat{\beta}_{PCj}^{(m)}-\beta_j^0)^2/1000$에 의해 얻었으며, Ω의 추정치 $\widehat{\Omega}=\widehat{\Gamma}+\widehat{\Psi}$을 구하였다. 각 조합별 자료에서 $\widehat{\Omega}$이 최소가 되게 한 CI 경계치(c^*)를 파악하였으며, c^*의 크기에 영향을 미치는 주요 속성들의 상대적인 중요도를 측정하기 위하여 컨조인트분석(conjoint analysis)을 활용하였다. 컨조인트분석 결과, 설명변수 수의 중요도가 가장 높으며, 다음은 관찰치 수, 다중공선성에 관련된 변수 수, 설명변수 분포의 순서로 중요한 것으로 나타났다.

이와 같은 분석결과를 바탕으로 주성분 선정기준인 c_L과 c_U을 결정하기 위한 모형을 구축하였다. 상대적으로 중요도가 높은 K, T, M을 설명변수로 채택하고(D는 중요도가 낮을 뿐만 아니라 회귀분석에 앞서 파악할 수 없는 속성이기 때문에 제외됨), 자료조합별로 파악된 c^*을 반응변수로 삼아 다음과 같은 선형모형,

$$\boldsymbol{c}^*=R\boldsymbol{\psi}+\boldsymbol{\eta},\quad R=[\mathbf{1}\,|\,K\,|\,T\,|\,M] \tag{3.2}$$

을 설정하였다. 생성된 모의실험자료(528개)에 모형 (3.2)을 적합하여 새로운 관찰치 $\boldsymbol{r}_0$의 95% 예측구간,

$$\widehat{c_0^*}\pm t_{\frac{\alpha}{2}}(524)\left[\left(\frac{\boldsymbol{c}^{*T}\boldsymbol{c}^*-\hat{\boldsymbol{\psi}}R^T\boldsymbol{c}^*}{524}\right)\left(1+\boldsymbol{r}_0^T\left(R^TR\right)^{-1}\boldsymbol{r}_0\right)\right]^{\frac{1}{2}}$$

을 구하였으며, 예측구간의 하한과 상한을 각각 주성분 선정을 위한 기준인 c_L과 c_U로 채택하였다. 즉,

$$\begin{aligned}c_L&=\boldsymbol{r}_0^T\hat{\boldsymbol{\psi}}-1.96\left[745.84\left(1+\boldsymbol{r}_0^TA\boldsymbol{r}_0\right)\right]^{\frac{1}{2}},\\ c_U&=\boldsymbol{r}_0^T\hat{\boldsymbol{\psi}}+1.96\left[745.84\left(1+\boldsymbol{r}_0^TA\boldsymbol{r}_0\right)\right]^{\frac{1}{2}},\end{aligned} \tag{3.3}$$

여기서

$$\hat{\boldsymbol{\psi}}^T=[-24.071,\ 16.272,\ 0.410,\ 4.986],$$

$$A=\begin{bmatrix}0.05128 & -0.00344 & -0.00023 & -0.00345\\ -0.00344 & 0.00069 & -8.61\times10^{-21} & -0.00034\\ -0.00023 & -8.61\times10^{-21} & 3.61\times10^{-6} & 4.38\times10^{-20}\\ -0.00345 & -0.00034 & 4.38\times10^{-20} & 0.00175\end{bmatrix}$$

이며, $\boldsymbol{r}_o$는 주성분회귀분석의 대상이 되는 자료의 설명변수 수, 관찰치 수 그리고 다중공선성에 관련된 변수의 수로 구성된 벡터다. 그런데 식 (3.3)에서 c_L과 c_U을 구하기 위해서는 다중공선성에 관련된 변

수의 수가 몇 개인지 사전에 파악해야 한다. 이를 위해서는 Belsley 등 (1980)이 제시한 분산분해비율,

$$\pi_{ij} = \frac{v_{ji}^2/\mu_i^2}{Q_j}, \quad Q_j = \sum_{i=1}^{k} \frac{v_{ji}^2}{\mu_i^2},\ i, j = 1, \ldots, k$$

(단, μ_i는 $\tilde{X}$의 비정칙치이고 v_{ji}는 직교행렬 V의 원소임)을 활용할 수 있는데, CI가 30보다 크면서 동시에 π_{ij}가 0.5보다 큰 경우의 수를 다중공선성에 관련된 설명변수의 수로 파악할 수 있다.

3.3. 주성분 선정방법

제안된 선정방법은 두 단계로 구성되었다. 첫째 단계에서는 식 (3.3)에 의해 구한 c_U보다 c_j가 크면, 그 c_j에 대응되는 주성분이 강력한 다중공선성에 관련된 것으로 판단하여 모형에서 제외시키고, c_L보다 작은 c_j에 대응되는 주성분은 다중공선성과의 관련이 매우 약한 것으로 판단하여 모형에 포함시킨다. 둘째 단계에서는 c_L과 c_U 사이에 속하는 c_j에 대응되는 주성분은 강력한 수준의 다중공선성은 아니더라도 무시할 수 없는 수준의 다중공선성과 관련이 있는 것으로 판단하여 유의성 검정을 실행하는데, 일반화선형검정 (Montgomery 등, 2006)에 바탕을 둔다. 즉, 첫째 단계에서 c_L보다 작은 CI에 대응하는 a개의 주성분을 모형에 포함시킨 상황에서의 잔차제곱합을 $\mathrm{SSE}_a (= \tilde{\boldsymbol{y}}^T\tilde{\boldsymbol{y}} - \hat{\boldsymbol{\gamma}}_a^T Z_a^T \tilde{\boldsymbol{y}})$라 하고, c_L과 c_U 사이에 있는 b개의 CI에 대응하는 주성분들 중에서 가장 유의성이 높을 것으로 판단되는 j-번째 주성분이 추가된 모형에서의 잔차제곱합을 $\mathrm{SSE}_{a+1} (= \tilde{\boldsymbol{y}}^T\tilde{\boldsymbol{y}} - \hat{\boldsymbol{\gamma}}_{a+1}^T Z_{a+1}^T \tilde{\boldsymbol{y}})$라 하면, j-번째 주성분의 유의성에 대한 가설 $H_0 : \gamma_j = 0$을 검정하기 위하여 검정통계량,

$$F_0 = \frac{\mathrm{SSE}_a - \mathrm{SSE}_{a+1}}{\mathrm{SSE}_{a+1}/(n-a-1)} \sim F(1,\, n-a-1) \tag{3.4}$$

을 적용할 수 있다. 식 (3.4)에 바탕을 둔 유의성검정을 전진도입법과 같은 방식에 따라 b개의 CI에 대응하는 주성분에 순차적으로 적용하여 모형에 포함될 주성분을 결정한다. 이와 같은 선정방법을 적용하면 다중공선성과 강력하게 관련이 있는 주성분은 안전하게 제거할 수 있으며, 다중공선성과 상당한 정도의 관련이 있는 주성분은 반응변수에 대한 기여도를 유의성검정을 통하여 판단함으로써 최적의 주성분 g개를 선정할 수 있게 된다. 특히 이 단계에서는 유의수준만 결정하면 된다는 점에서 기존의 방법들과 달리 객관적인 선정기준을 적용한다는 특징을 갖는다.

3.4. 적용 사례

본 논문에서 제시한 선정방법을 Marquardt와 Snee (1975)의 자료(Acetylene data)에 적용하되, 3개의 변수(X_1, X_2, X_3)와 교호작용(X_1X_2, X_1X_3, X_2X_3)이 설명변수로 포함된 회귀모형을 채택하였다. 우선 다중공선성의 존재 여부를 진단하기 위하여, 중심화 및 척도화된 자료로부터 $\tilde{X}^T\tilde{X}$의 고유치($\lambda_j = 3.75, 2.09, 0.096, 0.053, 0.0003, 0.00006$), 상태지수($\kappa_j = 1.0, 1.79, 38.9, 70.5, 12630.5, 57010.9$), 분산분해비율($\pi_{6j} = 0.861, 0.999, 0.101, 0.998, 0.001, 0.954$)를 계산하였다. 이러한 진단척도들을 통하여 이 자료에는 강력한 다중공선성이 존재하며 4개의 변수가 다중공선성에 관련이 있음을 알 수 있었다. 따라서 주성분회귀분석을 실행하기로 하고 적절한 주성분을 선정하고자 하였다. 식 (3.3)에서 $\boldsymbol{r}_0^T = [1, 6, 16/6, 4]$이므로 $c_L = 40.598$, $c_U = 148.598$을 얻었으며, c_U보다 큰 상태지수(κ_5, κ_6)에 대응하는 주성분을 제외시키고 c_L보다 작은 상태지수($\kappa_1, \kappa_2, \kappa_3$)에 대응하는 주성분을 일단 선정하였다. 그리고 c_L과 c_U 사이의 상태지수(κ_4)에 대응하는 주성분(z_4)은 유의성 검정을 통하여 선정여부를 결정하였는데, 식 (3.4)에 의해 검정통계량 계산치($F_0 = 15.391$)와 p-값(0.002)를 얻었다. 이를 바탕으로 주성분 z_4을 추가로 선정하였으며, 식 (2.4)에 의해 주성분회귀추정치($\hat{\beta}_{PC,j} = -145.284, 0.153, -0.228, -86.115, -0.0002, -0.097, 20.746$)를 구하였다.

표 **4.1.** 선정방법별 주성분회귀추정량의 평균제곱오차 추정치

K	T	M	MAR1	MAR2	PIDO	JOLL	PCR-C
3	30	2	0.9142	0.9274	0.7781	0.8241	0.5866
	60	2	1.1316	1.1521	0.7041	0.7393	0.5586
	90	2	1.0735	1.0764	0.2958	0.3316	0.2523
5	30	2	0.7659	0.9437	0.3343	0.3946	0.2522
		3	0.4988	0.5565	0.1995	0.2343	0.1374
		4	0.2734	0.3566	0.1373	0.1616	0.0795
	60	2	0.7067	0.8000	0.2211	0.2451	0.1537
		3	0.4426	0.5065	0.1269	0.1458	0.0832
		4	0.2525	0.3501	0.1015	0.1200	0.0580
	90	2	0.8044	0.9443	0.1985	0.2007	0.1675
		3	0.4526	0.5259	0.0981	0.1243	0.0770
		4	0.2542	0.3477	0.0853	0.1010	0.0478
7	30	2	0.7262	0.9112	0.1574	0.1666	0.1166
		3	0.4166	0.5746	0.1128	0.1331	0.0743
		4	0.2835	0.4185	0.0834	0.0964	0.0526
		5	0.1854	0.3250	0.0829	0.0955	0.0423
	60	2	0.5468	0.7190	0.1801	0.1872	0.0913
		3	0.3864	0.5463	0.0680	0.0770	0.0549
		4	0.2719	0.3989	0.0698	0.0863	0.0376
		5	0.1655	0.3121	0.0526	0.0604	0.0292
	90	2	0.6159	0.8069	0.1117	0.1148	0.0775
		3	0.3951	0.5461	0.0658	0.0807	0.0496
		4	0.2853	0.4005	0.0547	0.0630	0.0307
		5	0.1679	0.3136	0.0500	0.0633	0.0240
9	30	2	0.5375	0.7107	0.1290	0.1412	0.0835
		3	0.3530	0.5150	0.0744	0.0812	0.0531
		4	0.2825	0.4429	0.0613	0.0820	0.0384
		5	0.2093	0.3767	0.0558	0.0627	0.0303
	60	2	0.5174	0.7114	0.0977	0.1186	0.0795
		3	0.3421	0.5061	0.0573	0.0636	0.0436
		4	0.2712	0.4329	0.0620	0.0679	0.0250
		5	0.1910	0.3497	0.0407	0.0465	0.0210
	90	2	0.5189	0.6732	0.0851	0.1085	0.0607
		3	0.3264	0.5029	0.0423	0.0542	0.0387
		4	0.2612	0.4254	0.0410	0.0524	0.0199
		5	0.2011	0.3542	0.0328	0.0389	0.0172

4. 주성분 선정방법의 평가

새로운 주성분 선정방법을 평가하기 위하여 모의실험을 실행하였다. 모의실험을 위한 자료는 3.2절과 동일한 방법으로 생성하였는데, 모형을 개발하기 위한 추정용 자료와 별개의 평가용 자료를 생성하기 위하여 난수생성용 seed를 다르게 지정하였다. 다중공선성이 존재하는 설명변수는 정규분포를 바탕으로 생성하였으며, 설명변수의 수(3, 5, 7, 9), 설명변수의 수에 비례한 관찰치의 수(30, 60, 90배), 다중공선성에 관련된 변수의 수(2, 3, 4, 5)의 조합에 따라 1,000개씩의 회귀자료를 생성하였다. 각 선정방법을 적용한 주성분회귀추정량의 성능을 평균제곱오차의 관점에서 비교 평가하였다. 즉, 다양한 규모의

36개 자료세트에 Pidot (1969)의 방법(PIDO), Marquardt (1970)의 방법(MAR1: 0.1을 적용한 경우, MAR2: 0.2를 적용한 경우), Jolliffe (1972)의 방법(JOLL)과 본 연구에서 제안한 선정방법(PCR-C)을 적용하여 주성분회귀 추정치를 구하였으며 평균제곱오차는 $(\widehat{\Gamma}+\widehat{\Psi})/p$에 의해 추정하였다. 선정방법별 주성분회귀추정량의 평균제곱오차 추정치는 표 4.1에 수록되었는데, 모든 자료에서 기존의 방법들보다 PCR-C에 의한 추정량의 평균제곱오차가 월등히 작게 추정되었으므로 제안된 방법이 상대적으로 우수하다고 할 수 있다. 한편, PCR-C 다음으로 Pidot (1969) 방법, Jolliffe (1972) 방법, Marquardt (1970) 방법 순서로 우수한 것으로 나타났다.

5. 결론

데이터마이닝분야에서 회귀분석이 많이 활용되는데, 자료수집과정이나 모형에 포함된 설명변수들의 특성상 다중공선성의 문제가 야기될 가능성이 높다. 본 논문에서는 다중공선성 문제를 해결할 수 있는 주성분회귀분석에 관하여 연구하였는데, 특히 주성분을 선정하는 방법을 연구하였다. 기존의 선정방법들이 갖는 한계점을 극복하기 위한 새로운 방법을 제안하였는데, 주성분회귀추정량의 평균제곱오차를 최소화하는 상태지수의 상한과 하한을 설정하고, 자료에서 계산된 상태지수가 상한보다 크면 해당 주성분을 모형에서 제외시키고, 하한보다 작으면 해당 주성분을 모형에 우선 포함시키며, 상한과 하한 사이의 상태지수에 대응하는 주성분들에 대해서는 유의성검정을 바탕으로 한 전진도입 방식에 의해 선정 여부를 결정하는 방법이다. 모의실험을 통해 기존의 선정방법보다 제안된 방법이 상대적으로 우수한 것으로 평가되었다. 그러나 본 논문에서 제시한 방법은 모의실험에 바탕을 둔 것이기 때문에 자료의 규모가 모의실험의 범위를 벗어나는 경우에는 일반화에 한계가 있을 수 있다.

참고문헌

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression Diagnostics*, John Wiley.

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

Jolliffe, I. T. (1972). Discarding variables in a principal component analysis. I: artificial data, *Applied Statistics*, **21**, 160–1733.

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

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.

Marquardt, D. W. (1970). Generalized inverse, ridge regression, biased linear estimation, and nonlinear estimation, *Technometrics*, **12**, 591–612.

Marquardt, D. W. and Snee, R. D. (1975). Ridge regression in practice, *The American Statistician*, **29**, 3–20.

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

Montgomery, D. C., Peck, E. A. and Vining, G. G. (2006). *Introduction to Linear Regression Analysis*, John Wiley & Sons, Inc.

Pidot, Jr., G. B. (1969). A principal components of the determinants of local government fiscal patterns, *The Review of Economics and Statistics*, **51**, 176–188.

Procedure for the Selection of Principal Components in Principal Components Regression

Bu-Yong Kim[1] · Myung-Hee Shin[2]

[1]Department of Statistics, Sookmyung Women's University
[2]Team of Customer Strategy, Woongjin Coway Co. Ltd

(Received July 2010; accepted August 2010)

Abstract

Since the least squares estimation is not appropriate when multicollinearity exists among the regressors of the linear regression model, the principal components regression is used to deal with the multicollinearity problem. This article suggests a new procedure for the selection of suitable principal components. The procedure is based on the condition index instead of the eigenvalue. The principal components corresponding to the indices are removed from the model if any condition indices are larger than the upper limit of the cutoff value. On the other hand, the corresponding principal components are included if any condition indices are smaller than the lower limit. The forward inclusion method is employed to select proper principal components if any condition indices are between the upper limit and the lower limit. The limits are obtained from the linear model which is constructed on the basis of the conjoint analysis. The procedure is evaluated by Monte Carlo simulation in terms of the mean square error of estimator. The simulation results indicate that the proposed procedure is superior to the existing methods.

Keywords: Data mining, multicollinearity, principal components regression, condition index, selection of principal components.

This research was supported by Sookmyung Women's University Research Grants (2009).
[1]Corresponding author: Professor, Department of Statistics, Sookmyung Women's University, Seoul 140-742, Korea. E-mail: buykim@sookmyung.ac.kr