한국체육학회지 자연과학편, 제45권 제4호 547-556 The Korean Journal of Physical Education, 2006, Vol. 45(4)

다중공선성 상태의 주성분회귀와 능형회귀

김진욱 * 고려대학교

Principal components regression and ridge regression under the multicollinearity

Kim, Jin-Uk *Korea Univ.*

요약

회귀분석에서 설명변수들 사이에 선형적이거나 또는 거의 선형적인 종속관계에 있는 경우를 다중공선성이라고 하는데 이 같은 결과를 왜곡시키는 공선성이 존재하는 경우에는 일반적인 최소제곱 추정량의 대안을 고려해야만 한다. 대개 연구자들은 그 대안으로 주성분회귀나 능형회귀를 많이 사용한다. 본 연구의 목적은 다중공선성이 존재하는 경우의 두 회귀방법을 제시하고자 시뮬레이션 자료와 실제 자료를 사용하였다. 생성된 자료는 세 개의 설명변수와 하나의 반응변수를 가지는 다중회귀분석 모형을 설정하였고 설명변수 두 개는 서로 관련이 있도록 하였고 실제 자료는 유산소성 체력 자료를 이용하였다. 이 두 가지 모형으로 일반최소제곱, 주성분회귀, 능형회귀의 각 계수들을 추정하였다. 이론적인 측면을 같이 고려하면서 두 자료의 타당한 추정치를 얻고자 분석한 결과 첫 번째 자료는 능형회귀분석이, 두 번째 자료는 주성분회귀분석이 만족스러운 결과를 나타냈다. 다중회귀분석시 다중공선성이 존재하는 자료에서 회귀계수의 타당한 추정치를 얻고자 한다면 주성분회귀와 능형회귀가 권장된다.

Abstract

In regression analysis the alternatives of OLS (ordinary least squares) method for estimating regression coefficients should be considered if there is a severe multicollinearity among explanatory variables. It means that the multicollinearity distorts the results of study. Researcher, therefore, usually use prevailing two alternative methods, which are principal components regression (PCR) and ridge regression (RR). The purpose of this study is to propose the two alternative regression methods in a case of the multicollinearity problem and to give some discussions of the methods. For this study, two regression model were considered. One model contains three explanatory variables, two of three variables of highly correlated, and one response variable, by explanatory variables, which were created under the assumptions of regression analysis. The other model was oxygen consumption analysed by SAS OnlineDoc. From these variables, OLS, PCR and RR estimators were estimated and their variations were examined by means of ridge trace. PCR and RR showed results with satisfaction in a case of multicollinearity of a multiple regression.

Key words: regression, multicollinearity, principal components regression, ridge regression

* windoft@hanmail.net

서론

회귀분석(regression analysis)이란 광의적으로 정의를 하자면 변수들 간의 관련성을 함수로써 설명하는 것이라 할 수 있으며, 여기서 변수는 연속형변수 뿐만 아니라 이산형변수도 포함된다. 회귀분석은 자연과학과 사회과학의 여러 분야에서 널리 사용되는 통계적 분석 방법으로, 하나 또는 그 이상의 설명변수(explanatory variable)의 변화에 따른 반응변수(response variable)의 변화를 알아봄으로써 설명변수와 반응변수 간의 인과성에 대한 여러 정보를 얻기 위한 통계적 분석방법이다. 변수들이 표준화 되었다는 가정 하에서 추정된 함수식에 나타난 계수는 설명변수의 변화에 따른 반응변수의 변화를 알아보기 위한 수치로써 또는 설명변수들 간의 상대적인 효과 추정을 위해서도 사용된다.

회귀분석을 위한 회귀모형의 계수는 일반적인 최소제곱(ordinary least squares : OLS)에 의해서 추정하게 되는데, 특히 설명변수가 두 개 이상인 경우의 다중회귀분석의 최소제곱법을 살펴보면 반응변수에 대한 하나의 설명변수의 변화율은 나머지 설명변수를 상수로 두고 추정하기 때문에(Thomas & Finney, 1996) 변화율을 의미하는 그 변수의 추정 계수는 나머지 설명변수의 영향이 없다고 가정한 상태에서 구한 것이 된다. 실제 현상에서 연구자가 설정한 여러 개의 설명 변수들이 완전하게 서로 독립이라면 현상의 설명은 쉽게 이루어질 것이다. 하지만 설명변수들은 대체로 관련을 이루고 있으며 이런 경우 최소제곱에 의한 여러 설명변수의 추정계수는 잘못된 결과를 가져오게 된다. 또한 수치해석적 측면에서 보면 모형에서 사용되는 설명변수 상관계수행렬의 rank deficiency 의 문제가 되어 이로 인한 해는 수치적으로 보았을 때 불안정한 것이 된다(Golub & Van Loan, 1996; Strang, 1988). 이것은 선형방정식계(system of linear equations)를 다루는 모든 영역에 공통적으로 고민해야할 문제일 것이다. 이처럼 설명변수들 간의 관련성 문제를 회귀분석에서는 다중공선성(multicollinearity)이라 불러 다각적인 해결을 시도하고 있다.

잔차분석을 거친 후 설명변수간의 관련을 나타내는 다중공선성이 존재하는 경우에는 설명변수 변화에 대한 반응변수의 회귀계수가 안정적이지 못하고 변화가 심하게 되어 이 회귀계수는 해석에 사용되어서는 안된다. 이것은 OLS 에 의한 계수 추정치는 자료분석에 있어 잘못된 해석을 이끌게 된다는 것을 의미하며, 따라서 OLS 의 대안으로서 대부분의 통계학 교재에서는 주성분분석을 이용한 회귀분석(principal components regression : PCR)과 능형회귀(ridge regression : RR)를 권장하고 있다(Kutner, Nachtsheim, Neter, & Li, 2005; Montgomery, Peck, & Vining, 2001; Chatterjee, Hadi, & Price, 2000; Draper & Smith, 1998).

주성분회귀분석은 원래의 변수보다 적은 개수의 잠재적인 주성분을 선택하는 차원의 축소를 통해 해석의 용이성을 목적으로 하는 주성분분석(Morison, 2005; Johnson & Wichern, 1998)에 회귀분석을 결합한 것으로, 설명변수의 주성분분석을 통해 얻어진 주성분 점수로써 회귀분석을 실시하며 동시에 중요하지 않은 성분은 제외시키고, 이것을 토대로 원래의 회귀계수로 변환시켜 분석하는 방법이다(Lin & Wei, 2002). 여기서 원래의 설명변수 개수는 유지된다.

능형회귀분석은 역시 불안정한 ill-conditioned 의 OLS 회귀계수보다 안정적인 해를 얻기 위한 방법으로 회귀분석의 틀을 유지하면서 회귀계수의 추정량을 얻기 위한 알고리즘이라 할 수 있다(Chatterjee et al., 2000). 이 방법은 Hoerl & Kennard(1970)에 의해서 제안된 방법으로 주성분회귀와 마찬가지로 원래의 설명변수의 개수는 유지된다.

OLS 에 대한 대안으로써 주성분분석 그리고 능형회귀의 연구로는 파라미터를 선택하는 여러 방법들의 비교, OLS 와 비교 그리고 효율성에 대한 지속적 연구가 이루어져 왔으며(Khalaf & Shukur, 2005; Kubokawa & Srivastava, 2004; Kaçıranlar & Sakallıoğlu, 2001) 각 연구자들에 의해 제안한 방법들이 우수한 결과를 나타낸다고 보고하고 있다.

다중공선성이 심한 경우에는 변수의 선택을 통한 제거도 하나의 방법이기는 하지만, 연구의 성격상 필요한 변수 또는 이론적으로 중요한 변수라면 모형에 꼭 포함시켜야 할 경우가 있으며, 이럴 경우 언급된 두 가지 방법을 사용한다. 더욱이 다중공선성이 존재하는 경우에 단순히 변수 선택법을 사용하면 잘못된 결과가 나타나기 때문에 주성분 또는 능형회귀 분석 후에 변수 선택을 사용해야 한다(Chatterjee et al., 2000).

체육학의 여러 분야에서 역시 다중회귀분석은 많이 사용되고 있다. 그러나 다중공선성이 존재할 경우의 계수가 그대로 해석 되어 이에 대한 결과의 왜곡이 예상된다. 결국 이것은 현상의 본질을 알아내고자 하는 연구의 목적에 부합되지 않기 때문에 반드시 타당한 효과를 추정해야 한다. 따라서 다중공선성이 존재하는 경우의 자료에서 일반 회귀분석에 대한 대안으로 주성분 회귀분석과 능형회귀분석의 분석방법을 제시하는 것이 본 연구의 목적이다.

연구방법

1. 이론적 고찰

1) 회귀모형

만일 p 개의 설명변수와 한 개의 반응변수 그리고 n 개의 관측치가 있고 모든 변수가 단위벡터 형태로 표준화(centered and scaled) 되었을 때 회귀모형은 아래와 같이 행렬 형태로 표현될 수 있다.

$$\mathbf{y}= \mathbf{Z}\beta+ \epsilon \qquad (1)$$

여기서 $\mathbf{y}$ 는 $n\times 1$ 반응변수 관측치 벡터, $\mathbf{Z}=(\mathbf{Z}_1,\cdots.\mathbf{Z}_p)$ 는 n 개의 성분(설명변수 관측치)을 가진 p 개 설명변수 열벡터로 이루어진 $n\times p$ 행렬, β 는 $p\times 1$ 의 회귀계수벡터, 그리고 ϵ 은 $p\times 1$ 의 랜덤오차벡터이다. 회귀분석에서는 $E(\epsilon)=0$, $E(\epsilon\epsilon^T)=\sigma^2\mathbf{I}$ 라고 가정하며, 여기서 $\mathbf{I}$ 는 identity matrix 이다. 또한 $\mathbf{Z}^T\mathbf{Z}$ 는 상관계수행렬이며, β 의 최소제곱 추정량은 $\hat{\beta}=(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y}$ 이다.

2) 주성분회귀

상관계수행렬 $\mathbf{Z}^T\mathbf{Z}$ 의 eigenvalue $\lambda_1\geq\cdots\geq\lambda_p$ 와 이에 해당되는 eigenvector $\mathbf{v}_1,\cdots.\mathbf{v}_p$ 는 각각 행렬 Λ, $\mathbf{V}$ 로 나타낼 수 있다. Λ 는 eigenvalue를 대각원소로 가지는 대각행렬이며, $\mathbf{V}$ 는 각각의 eigenvalue $\lambda_1,\cdots\lambda_p$ 에 해당되는 eigenvector 의 열로 이루어진 행렬이 된다(Johnson & Wichern, 1998; Morison, 2005). 여기서 행렬 Λ, $\mathbf{V}$ 는 정방행렬이며 아래의 조건을 만족한다.

$$\mathbf{V}^T(\mathbf{Z}^T\mathbf{Z})\mathbf{V}=\Lambda,$$
$$\mathbf{V}^T\mathbf{V}=\mathbf{V}\mathbf{V}^T=\mathbf{I} \qquad (2)$$

eigenvector 행렬은 직교행렬이기 때문에 $\mathbf{V}\mathbf{V}^T=\mathbf{I}$ 가 성립하며 이것을 식(1)에 대입하면,

$$\mathbf{y}=\mathbf{Z}\mathbf{V}\mathbf{V}^T\beta+\epsilon \qquad (3)$$

또는 다음과 같이 주성분에 의한 회귀모형으로 정리될 수 있다.

$$\mathbf{y}=\mathbf{C}\gamma+\epsilon \qquad (4)$$

여기서 $\mathbf{C}=\mathbf{Z}\mathbf{V}$ 이며, $\gamma=\mathbf{V}^T\beta$ 이다. 행렬 $\mathbf{C}$ 는 설명변수 $\mathbf{Z}_1,\cdots.\mathbf{Z}_p$ 와 $\mathbf{V}$ 의 선형조합으로서, 서로 관련된 $\mathbf{Z}_1,\cdots.\mathbf{Z}_p$ 가 서로 직교(독립)하도록 eigenvector 행렬을 곱하여 변환한 것으로 $\mathbf{C}_1,\cdots.\mathbf{C}_p$ 의 열로 이루어졌으며 이것을 주성분(principal components)이라 한다. 주성분은 서로 독립이기 때문에 $\mathbf{C}_j^T\mathbf{C}_j=\lambda_j$, $\mathbf{C}_i^T\mathbf{C}_j=0\,(i\neq j)$ 을 만족한다. 또한 벡터 γ 는 원래의 회귀계수벡터가 eigenvector 행렬에 의해 변환된 것으로 주성분으로 변환된 회귀모형의 계수벡터를 의미한다. 따라서 식(4)는 식(3)에 나타난 설명변수와 계수 모두를 서로 직교하도록 선형변환 시킨 후의 회귀모형이라 할 수 있다. 식(4)에 나타난 γ 의 최소제곱 추정량은 $\hat{\gamma}=(\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{y}$ 이며 주성분들은 서로 독립이기 때문에 다중공선성이 발생하지 않는다. 원래의 모형과 변환 후의 모형 간에는 식(5) 와 같은 관계가 성립되며, 한 모형의 추정값이 주어지게 되면 언제든지 다른 모형으로의 변환이 가능하게 된다.

$$\mathbf{Z}=\mathbf{C}\mathbf{V}^T,\quad \beta=\mathbf{V}\gamma \qquad (5)$$

변환된 주성분의 분산은 바로 eigenvalue 이며, 만일 $\lambda_i=0$ 이 되면 i 번째 주성분 값들은 모두 0 이 된다. 주성분은 $\mathbf{Z}_1,\cdots.\mathbf{Z}_p$ 의 선형함수이기 때문에 $\lambda_i=0$ 이라는 것은 설명변수들 간의 정확한 선형적

종속관계가 존재하는 것이라 할 수 있다. 즉, λ_i 가 거의 '0' 의 값을 가지면($\lambda_i \approx 0$) 설명변수들 간의 관련성이 아주 높은 것이라 할 수 있으며 이것이 다중공선성의 증거가 되어 중요하지 않은 주성분은 제거할 수 있다. 그러나 주성분을 제거하더라도 실제 해석에 있어서 어려움이 있기 때문에 원래의 추정치 β 로 변환하여야 한다. 이 때 중요하지 않은 주성분으로의 변환계수 eigenvector 는 '0'으로 놓고 나머지만 가지고 식(5)에 의해서 변환을 시키면 된다.

3) 능형회귀

일반적인 최소제곱에 의한 회귀계수 추정량은 전술한 바와 같이

$$\hat{\beta} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y} \qquad (6)$$

이다. 그러나 설명변수들 간에 강한 관련성이 존재하는 경우 Hoerl & Kennard(1970)는 계수를 추정하기 위해서 $\mathbf{Z}^T\mathbf{Z}$ 보다는 $\mathbf{Z}^T\mathbf{Z}+k\mathbf{I}, (k \geq 0)$ 를 사용할 것을 제안하였으며 이를 이용한 능형회귀 추정량은 다음과 같다.

$$\hat{\mathbf{q}}(k) = (\mathbf{Z}^T\mathbf{Z}+k\mathbf{I})^{-1}\mathbf{Z}^T\mathbf{y} \quad (7)$$

이 추정량은 주성분회귀와 마찬가지로 편의 되어(biased) 있지만 안정적인 해를 얻을 수 있는 장점을 가지고 있다(Akdeniz, 2002). 여기서는 능형매개변수(ridge parameter, bias parameter) k 를 어떻게 선택하느냐가 최대 관심사가 된다. 만일 k 가 커지면 추정량의 편의는 증가하게 되고 회귀계수 추정치는 '0'에 가까워진다. 이 매개변수를 0 으로부터 점차 증가시키면서 회귀계수 추정치의 변화를 그래프로 그려보면 특정 k 지점 이후로 추정치들이 안정된 값을 나타내게 되는데, 안정된 값을 가지기 시작하는 부근의 최소의 k 를 선택하는 것을 ridge trace 방법이라고 한다. 또한 수치적으로 k 를 선택하는 기준은 여러 가지가 있지만(Golub, Health, & Wahba, 1979) 많은 능형회귀 추정량 중에서도 Hoerl, Kennard, & Baldwin 이 1975년에 발표한 k 선택법이 많은 상황에 비교적 잘 맞으며 많은 비교 연구들도 다음의 식을 토대로 하고 있다(Khalaf & Shukur, 2005).

$$k = \frac{p\hat{\sigma}^2(0)}{\sum_{i=1}^{p}[\hat{\beta}_i(0)]^2} \qquad (8)$$

여기서 $\hat{\beta}_1(0), \cdots, \hat{\beta}_p(0)$ 은 $k=0$ 인 경우의 최소제곱 추정량이며, p 는 변수의 개수 그리고 $\hat{\sigma}^2(0)$ 은 이 때의 잔차평균제곱(residual mean square)이다.

주성분회귀나 능형회귀의 추정치 선택은 분산팽창인자(variance inflation factor : VIF) 나 상태지수(condition index)를 같이 검토하여 이루어진다.

2. 자료분석

본 연구에서 다중공선성이 존재하는 경우의 자료를 고려하기 위해서 두 종류의 자료를 분석하였는데 하나는 시뮬레이션 기법을 이용하여 생성된 자료와 다른 하나는 SAS의 홈페이지에서 제공하는 OnlineDoc^R (SAS Institute, 1999)의 유산소성 체력(aerobic fitness) 자료이다.

첫 번째, 생성한 자료에 의한 회귀모형은 세 개의 설명변수와 하나의 반응변수를 가진다. 회귀모형에서 안정적인 해를 얻기 위해서 관측치의 개수는 15개로 설정 하였으며 따라서 변수는 15개의 성분을 가진 벡터이다($\mathbf{y}$: 반응변수, $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$: 설명변수). 회귀분석의 가정들을 만족시키기 위해서 각 변수는 표준정규분포로부터 독립적으로 랜덤하게 얻어진 난수를 이용하였으며, 각 변수의 분산이 동일하다는 가정을 만족시켜야 하므로 같은 표준편차를 이용하여 자료를 구성하였다. 분석에 사용될 자료는 MATLAB의 난수발생기에 의해서 세 개의 설명변수가 생성되었다. 설명변수가 서로 독립적이기 위해서 Gram-Schmidt 의 직교화(orthogonalization)인 QR 분해(QR decomposition)를 이용하였다(Strang, 1988). 그리고 설명변수 $\mathbf{x}_1$ 과 $\mathbf{x}_3$ 가 거의 선형적인 관계를 갖기 위해서 공간상에서 $\mathbf{x}_3$ 를 변환행렬에 변환시켜 $\mathbf{x}_1$ 과 상관관계가 0.9986 이

되도록 하였다.

반응변수는 $y = 2x_1 + 0.75x_2 + 0.25x_3$ 의 식에 의해서 생성되었으며 관측오차를 고려하여 표준편차 0.1 의 노이즈를 추가 하였다. 모든 벡터는 길이가 '1' 이 되도록 normalization 하였다.

두 번째, 실제 자료를 분석하기 위해서 선택된 유산소성 체력은 산소흡수율(Oxygen)을 반응변수로 하여 연령(Age), 체중(Weight), 1.5 마일 달리기 기록(RunTime), 안정시 심박수(RestPulse), 주행중 심박수(RunPulse), 그리고 주행중 최대심박수(MaxPulse)의 여섯 개의 설명변수로 이루어진 국내에도 많이 소개된 다중회귀분석 모형이다. 이 모형의 분석결과는 SAS 웹사이트에 나타나 있으며 이 싸이트의 분석 결과는 표준화 시키지 않은 경우이며, 본 연구에서는 표준화 시킨 후 회귀분석을 실시하였다. 만약 원래 자료의 추정치를 구하고자 한다면 다음의 식을 이용한다 (Montgomery et al., 2001; Chatterjee et al., 2000).

$$\hat{b}_i = \hat{\beta}_i \sqrt{\frac{SST}{S_i}} \quad , \; i = 1 \cdots . p \qquad (9)$$

$$\hat{b}_0 = \bar{y} - \sum_{i=1}^{p} \hat{\beta}_i \bar{x}_i \qquad (10)$$

여기서 $\hat{b}_i$ 는 원래의 회귀계수 추정량이며 $\hat{b}_0$ 는 절편항이다. 두 자료의 분석결과 회귀분석의 가정에 위반되는 사항을 발견할 수 없었으며 추가적인 주성분분석과 능형회귀분석은 일반적인 통계 패키지에서는 독립적으로 수행할 수 없기 때문에 SPSS(ver. 12) 와 MATLAB(ver. 6.1) 을 이용하여 자료의 분석을 하였으며 주성분회귀와 능형회귀 분석에 있어 다중공선성 존재의 판단 기준으로 VIF 와 상태지수를 사용하였으며 능형회귀의 경우에는 ridge trace 와 식(8)의 k 를 이용하였다.

결과 및 논의

1. 생성자료

다중공선성이 존재하는 생성된 자료의 상관계수행렬과 회귀분석결과는 <표 1>, <표 2>와 같다.

표 1. 생성자료의 상관계수행렬(n=15)

	y	x1	x2	x3
y	1.000			
x1	.825**	1.000		
x2	.358	.000	1.000	
x3	.819**	.999**	.000	1.000

** p<.01

표 2. 생성자료에 대한 회귀분석결과

variable	estimate	t	p	collinearity		
				eigenvalue	condition index	VIF
x1	5.935	1.086	.299	1.999	1.000	365.461
x2	.828	2.894	.013	1.000	1.414	1.000
x3	-4.035	-.738	.475	.001	38.208	365.461
R^2=.816		F(3,12)=17.781(p=.000)			MSE=.082	

회귀분석결과를 보면 R^2 은 .816 이며 분산분석에서 F=17.781(p=.000)로 회귀식이 유의하다는 것을 알 수 있다. 그러나 계수의 추정치를 보면 x_2 를 제외하고는 회귀식에 기여하지 못하며 공선성 부분을 보더라도 x_3 의 상태지수가 38.208 이며 VIF 는 x_1, x_3 가 모두 365.461 로 이 두 변수가 거의 선형결합 형태를 이루는 아주 심각한 다중공선성을 나타내고 있다고 할 수 있다. 따라서 계수의 추정치는 믿을 수 없는 것이 되며 결과를 보더라도 원래의 계수 2, .75, .25 와는 상당한 차이가 있음을 알 수 있다. 이에 대한 대안으로 주성분회귀분석과 능형회귀분석을 실시하였다

주성분회귀분석은 원래의 변수를 주성분으로 변환하여 회귀계수를 추정하였으며 세 번째의 eigenvalue 가 .001 로 '0' 에 가까운 값을 가지기 때문에 여기에 해당되는 eigenvector 는 '0' 으로 놓고 식(5) 에 의해 원래의 계수로 변환하였다. 그 결과는 <표 3> 의 PCR 과 같으며 추정된 계수는 .9499, .8278, .9499 로 원래의 계수와 달라 별로 만족스러운 결과를 나타내지는 못했다.

능형회귀분석에서 k=.0047 로 나타나 계수의 추정치를 구한 결과는 <표 3>의 RR 과 같으며 2.0725, .8239, -.1771 으로 원래의 계수와 비슷한 값을 가져 어느 정도 만족할 만한 결과를 보여주고 있다.

표 3. 생성자료에 대한 PCR 과 RR

variable	OLS	PCR	RR
x1	5.9350	.9499	2.0725
x2	.8276	.8278	.8239
x3	-4.0352	.9499	-.1770
MSE	0.0818	0.0874	0.0871
R^2	.8163	.8035	.8086

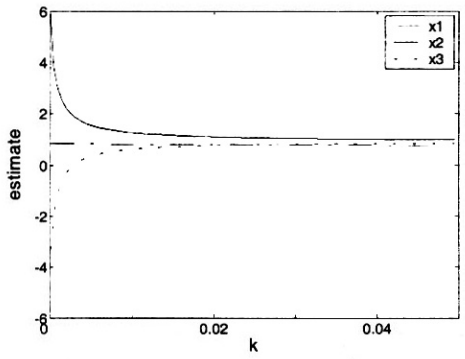

그림 1. 생성자료의 ridge trace

ridge trace 는 <그림 1>과 같으며 변수 x_1, x_3 의 추정치가 매우 변동이 심한 것을 알 수 있으며 k 가 약 .01 이후에 안정된 경향을 보여주고 있었다.

2. 유산소성 체력

유산소성 체력의 상관계수행렬과 회귀분석의 결과는 <표 4>, <표 5>와 같다.

표 4. 유산소성체력 자료의 상관계수행렬(n=31)

	Oxy	Age	Weight	Run Time	Rest Pulse	Run Pulse	Max Pulse
Oxy	1.000						
Age	-.304	1.000					
Weight	-.163	-.234	1.000				
Run Time	-.862**	.189	.144	1.000			
Rest Pulse	-.399*	-.164	.044	.450*	1.000		
Run Pulse	-.398*	-.338	.182	.314	.352	1.000	
Max Pulse	-.237	-.433*	.249	.226	.305	.930**	1.000

* p<.05, ** p<.01

표 5. 유산소성체력 자료에 대한 회귀분석결과

variable	estimate	t	p	collinearity		
				eigenvalue	condition index	VIF
Age	-.222	-2.317	.029	2.575	1.000	1.513
Weight	-.116	-1.388	.178	1.328	1.393	1.155
RunTime	-.685	-6.975	.000	.925	1.668	1.591
RestPulse	-.031	-.332	.743	.743	1.861	1.416
RunPulse	-.712	-3.151	.004	.369	2.643	8.437
MaxPulse	.522	2.270	.032	.060	6.533	8.744
R^2=.849		F(6,25)=23.361(p=.000)			MSE=.182	

유산소성 체력의 회귀분석 결과 R^2 은 .849 이며 분산분석은 F=23.361(p=.000)로 회귀식이 유의하다는 것을 말해주고 있다. 계수의 추정치를 보면 RestPulse 를 제외하고는 반응변수에 유의한 영향을 미친다는 것을 알 수 있다. 또한 각 변수의 추정치는 MaxPulse 를 제외하고는 모두 음수임을 알 수 있으며 변수의 값들이 증가 할수록 유산소성 체력은 낮다고 할 수 있다. 그러나 MaxPulse 는 의외의 결과를 나타내었는데(.522) RunPulse 나 MaxPulse 는 비슷한 성격의 변수이며, 두 변수의 상관계수를 보더라도 .930(p=.000) 로 양의 상관계수를 가지기 때문이다(표 4). 비록 .522 는 다른 변수들의 영향을 제거한 상태에서 추정된 값이지만, RunPulse만이 제거된 상태에서의 반응변수 Oxygen 과 MaxPulse 의 편회귀계수는 -.246(p=.261) 로 부호가 바뀌어 나타나므로 RunPulse 에 상당한 영향을 받는다는 것을 알 수 있다. 따라서 MaxPulse 는 추정치가 부호가 반대로 추정되었다는 것을 의심해 보아야 한다. 공선성 부분을 보면 MaxPulse 의 상태지수는 6.533 이며, RunPulse 와 MaxPulse 의 VIF 는 각각 8.437 과 8.744 로 그다지 크지는 않지만 어느 정도 두 변수는 관련성이 있으며 약한 다중공선성의 상태로 판단된다.

주성분회귀분석을 위하여 상관계수의 eigenvalue 와 eigenvector 를 구한 것이 <표 6>과 같다.

주성분분석(principal components analysis)의 입장에서 본다면 주성분의 개수를 선택하는데 있어 eigenvalue, 분산, 그리고 scree plot 등 여러 가지가 있으며 <표 6>을 본다면 4개 정도가 중요한 주성분이며

나머지 두 개가 중요하지 않은 주성분으로 요약될 수 있겠다.

표 6. 유산소성체력 자료 상관계수행렬의 eigenvalue 와 eigenvector

	eigenvector					
	v_1	v_2	v_3	v_4	v_5	v_6
Age	-.3156	-.5618	-.0760	.5340	-.5389	-.0597
Weight	.2370	.1939	-.9209	-.0193	-.2355	.0475
RunTime	.2792	-.6670	-.2370	.0732	.6430	-.0468
RestPulse	.3555	-.4192	.1349	-.6816	-.4634	-.0205
RunPulse	.5647	.0413	.2118	.3686	-.1156	.6967
MaxPulse	.5666	.1565	.1639	.3297	-.1126	-.7115
eigenvalue	2.5749	1.3278	.9251	.7432	.3687	.0603

특히 마지막 eigenvector 의 계수를 보면 RunPulse 와 MaxPulse 의 값이 .6967 과 -.7115 로 아주 높으며 다른 eigenvector 의 계수와는 달리 부호가 반대의 계수를 가져 잘못된 것임을 알 수 있다. 원래의 회귀계수로 변환시키기 위한 이 계수의 영향을 제거하기 위해서 v_6 은 '0'으로 하여 결과를 구한 것이 <표 7> 의 PCR 이다. 원래의 OLS 계수와 비교를 해 보면 거의 비슷한 결과를 가지며 특히 RunTime 의 계수가 영향력이 낮아지고 MaxPulse 의 부호가 바뀌었음을 볼 수 있다(-.0663). 이는 두 변수의 종속성을 고려해 보았을 때 적절한 계수라고 생각되며 이것으로 보았을 때 주성분회귀분석의 결과는 매우 만족스러운 결과를 보여주었다.

표 7. 유산소성체력 자료에 대한 PCR 과 RR

variable	OLS	PCR	RR
Age	-.2217	-.2711	-.2373
Weight	-.1160	-.0767	-.1024
RunTime	-.6846	-.7233	-.6709
RestPulse	-.0307	-.0477	-.0470
RunPulse	-.7121	-.1357	-.5047
MaxPulse	.5224	-.0663	.3109
MSE	.1892	.2408	.2332
R^2	.8486	.8073	.8427

능형회귀분석의 결과 ridge trace 를 보면(그림 2) 역시 MaxPulse 와 RunPulse 가 변화가 매우 크다는 것을 알 수 있으며, 특히 MaxPulse 는 점차 낮아지고 있으며 RunPulse 는 점점 높아지고 있어 두 변수의 추정치가 절대값으로 보았을 때 과대 추정되었다는 것을 알 수 있다. 두 변수간의 관련성에 기인한다고 하겠다.

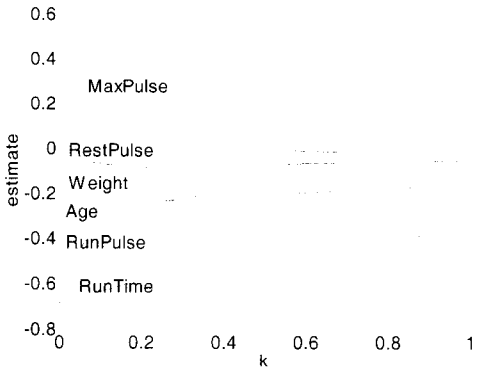

그림 2. 유산소성체력 자료의 ridge trace

표 8. k 에 대한 각 변수의 RR 추정치와 VIF 변화

k	q_1	q_2	q_3	q_4	q_5	q_6	VIF(q_1)	VIF(q_2)	VIF(q_3)	VIF(q_4)	VIF(q_5)	VIF(q_6)
0.00	-0.2217	-0.1160	-0.6846	-0.0307	-0.7121	0.5224	1.5128	1.1553	1.5909	1.4156	8.4372	8.7438
0.01	-0.2282	-0.1106	-0.6816	-0.0364	-0.6298	0.4384	1.4421	1.1175	1.5161	1.3641	6.3035	6.5198
0.02	-0.2329	-0.1066	-0.6775	-0.0413	-0.5680	0.3753	1.3804	1.0843	1.4494	1.3163	4.9131	5.0710
0.03	-0.2363	-0.1034	-0.6727	-0.0456	-0.5198	0.3262	1.3252	1.0543	1.3887	1.2717	3.9561	4.0742
0.04	-0.2388	-0.1008	-0.6676	-0.0494	-0.4811	0.2868	1.2750	1.0267	1.3331	1.2298	3.2687	3.3587
0.05	-0.2407	-0.0987	-0.6623	-0.0529	-0.4493	0.2546	1.2287	1.0008	1.2816	1.1903	2.7580	2.8273
0.06	-0.2420	-0.0969	-0.6568	-0.0561	-0.4228	0.2277	1.1858	0.9764	1.2339	1.1531	2.3677	2.4215
0.07	-0.2429	-0.0953	-0.6512	-0.0590	-0.4003	0.2049	1.1459	0.9533	1.1893	1.1179	2.0625	2.1044
0.08	-0.2435	-0.0939	-0.6457	-0.0617	-0.3809	0.1854	1.1085	0.9313	1.1477	1.0846	1.8191	1.8516
0.09	-0.2439	-0.0927	-0.6402	-0.0643	-0.3640	0.1684	1.0733	0.9103	1.1086	1.0530	1.6215	1.6467
0.10	-0.2440	-0.0916	-0.6347	-0.0666	-0.3492	0.1536	1.0402	0.8902	1.0719	1.0230	1.4589	1.4781
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.99	-0.1619	-0.0565	-0.3666	-0.1019	-0.1338	-0.0273	0.2284	0.2459	0.2239	0.2315	0.1392	0.1350

q_1:Age, q_2:Weight, q_3:RunTime, q_4:RestPulse, q_5:RunPulse, q_6:MaxPulse

또한 <표 8>에는 k 의 증가에 따른 추정치와 VIF 의 값을 나타내었으며 증가단위는 .01 의 값들만 나타내었다. k 가 증가함에 따라 마지막 두 변수의 VIF 가 점차 감소하는 것을 볼 수 있어 다중공선성의 영향이 점차 감소하고 있음을 보여주고 있다.

식(7)에 의한 k 는 .0336 이며 이것을 토대로 구한 계수는 <표 7>과 같다. 이것은 원래의 OLS 추정치와 유사한 결과를 보여주고 있으며 그다지 만족스러운 결과는 아니라 생각된다.

3. 논의

자료의 첫 번째 자료에서는 주성분회귀 보다는 능형회귀에서 타당한 결과를 나타냈다. 이것은 아마도 비교적 단순한 구조로 생성되었으며 eigenvector 계수의 대칭성으로 인해 주성분회귀분석에서는 좋은 결과를 얻지 못한 것 같다. 또한 능형회귀에서도 k 를 좀 더 크게 설정하면 마지막 추정치도 음수에서 양수로 변할 수도 있을 것이다.

두 번째 분석은 주성분회귀 분석에서 마지막 주성분 하나만 제거한 상태에서의 추정치는 어느 정도 타당한 결과를 보여주고 있다고 생각된다. 그러나 능형회귀분석은 원래의 OLS 분석과 결과가 유사하게 나타났으며 특히 MaxPulse 가 RunPulse 의 선형적 결합에 의해서 잘못 추정된 계수가 그대로 나타났다. 전체적으로 보았을 때 두 변수가 관련이 높기 때문에 주성분회귀와 능형회귀를 통해 RunPulse 와 MaxPulse 의 추정치의 절대값은 낮아졌으며 이것은 아주 바람직한 결과라 할 수 있다. 특히 주성분회귀에서의 MaxPulse 의 추정치가 낮아졌으며 음수로 바뀌어 이론적으로도 타당한 수치를 보여주었다고 할 수 있겠다.

능형회귀에 있어 식에 의해 계산된 k 로부터 얻어진 능형회귀 추정치는 원래의 회귀분석 결과와 유사한 결과를 나타내었지만 만일 k 를 잘 선택한다면 주성분회귀분석과 유사하게 나타날 수 있는 가능성도 배제할 수는 없다.

OLS 의 대안으로 주성분회귀분석의 타당성을 주장하는 사람들은 주성분의 점수들이 직교하기 때문에 다중공선성 문제가 해결되며, 몇 개의 주성분이 사용이 되든지 상관없이 원래의 설명변수의 개수는 유지 한다는 점과 서로 직교하는 주성분을 사용하기 때문에 회귀계수 추정치의 수치적 정확성이 증가 한다는 것이다(Hardi & Ling, 1998). 그러나 이들은 주성분회귀분석을 수행함에 있어서 조심해야 할 것들을 동시에 언급하고 있는데, 주성분회귀분석의 문제점은 변환된 주성분의 분산(λ)이 작다 하더라도 반응변수에 영향을 크게 미칠 수 있는 경우가 있기 때문이다. 또한 주성분 방법은 이상치(outlier)에 과도하게 영향을 받을 수 있다는 점이다. 물론 연구자는 본 분석 이전에 이상치에 대한 문제해결을 선행하겠지만 그렇지 않을 경우에는 문제가 될 수 있다. 즉, 주성분은 상관계수행렬을 토대로 구한 것이며 상관계수는 자료의 이상치에 영향을 받기 때문이다.

능형회귀의 장점은 추정계수의 안정성(stability)에 있다. 즉, 자료의 추가 또는 제거에 영향을 적게 받는다는 점이다. 단점은 주성분회귀와 같이 이상치에 대한 영향을 많이 받는다는 점이다. 어떠한 분석이건 이상치에 대한 대책은 마련해야 하겠다. Khalaf & Shukur (2005)는 추정치에 영향을 주는 요인으로는 표본의 크기, 설명변수들 간의 상관계수, 그리고 오차의 분산의 세 가지를 설정하여 이것들을 다양하게 조합하여 OLS, Hoerl & Kennard 의 능형회귀 그리고 그들의 방법으로 비교를 하여 새로운 방법에 의한 능형회귀의 우수성을 입증하기도 하였다.

주성분 분석과 능형회귀의 결과 중 어느 것이 더 우수한 것인가에 대한 연구는 지속적으로 이루어지고 있지만 일반적인 결론은 존재하지 않는다. 그렇지만 이들의 연구의 공통된 결론은 기존의 OLS 에 의한 계수의 추정량보다는 대안적인 방법이 더 나은 추정량이라는 점이다(Montgomery et al., 2001). 본 연구의 결과에는 두 방법 중 하나가 모두 나은 결과를 보여주지는 않은 것처럼 두 가지 방법 중 어느 것을 사용하는가에 대한 선택은 결국 연구자가 결정해야 할 것이다. 이처럼 상반된 결과를 보이는 것은 그만큼 현상이 복잡하며 하나의 도구로써는 부족하다는 것을 말해주고 있다. 그러나 회귀분석 서적에서는 주성분회귀분석 보다는 능형회귀를 더 선호하고 있는 듯 하다(Montgomery et al., 2001; Chatterjee et al., 2000; Kutner et al., 2005)

주성분회귀분석과 능형회귀분석에 의해서 구한 계수의 추정량들은 OLS 에 비해서 편의 되어 있기는 하지

만 정확한 편이라고 하였다(Chatterjee et al., 2000). 이것은 본 연구의 결과와도 어느 정도 일치하는 부분이라고 할 수 있겠다. 연구자는 다중공선성이 의심될 때는 이러한 방법들을 사용해야만 하며 변수의 선택 이전에 능형회귀나 주성분회귀 후 선택방법을 사용할 것을 권장하고 있다. 기계적으로 변수 선택을 사용하지 말고 자료의 구조를 탐색적인 자세로서 이론과 직관 그리고 상식에 의해서 접근을 해야 한다(Chatterjee et al., 2000). 본 연구도 이론적인 측면을 고려하여 적절한 추정치를 얻고자 노력하였으며 그 결과 어느 정도 만족스러운 해답을 얻을 수 있었다.

본 연구에서 다중공선성이 존재하는 두 개의 자료의 분석에서 주성분과 능형회귀분석의 결과 중 하나의 방법을 선택할 수 있었던 것은 선험적(a priori) 지식에 바탕을 둔 실증적으로 축적된 이론에 근거한 것이었다. 즉, 첫 번째 자료는 참값을 알고 있었으며, 두 번째 자료는 이론적인 고찰을 토대로 하였기 때문에 적절한 선택을 할 수 있었다. 따라서 어떤 방법을 선택하여야 하는가는 관련분야의 지식과 직관 그리고 경험을 종합적으로 고려해서 이루어져야 할 것이다.

결론 및 제언

본 연구는 다중회귀분석시 다중공선성이 존재하는 두 가지의 자료를 가지고 주성분회귀분석과 능형회귀분석을 실시한 것이다. 올바른 추정량을 구하고자 다중공선성이 존재하는 두 가지 종류의 자료를 분석한 결과, 기존의 OLS 에 의한 추정량보다는 대안적 방법들이 만족스러운 결과를 보여주었다. 그러나 주성분회귀나 능형회귀 두 방법 모두가 OLS 에 비해서 타당한 결과를 보여주지는 못하고 있으며, 두 개의 자료를 보았을 때 전통적인 최소제곱의 방법은 모두 추정량의 왜곡된 결과를 나타내고 있었다.

회귀분석의 목적은 여러 가지가 있을 수 있다. 특히 반응변수에 대한 설명변수의 영향력을 알고자 한다면 회귀계수의 추정값 자체가 중요한 수치이다. 회귀분석의 모형은 이론적 배경을 토대로 이루어지게 된다. 그러나 흔히 모형이 수립되더라도 탐색적인 방법을 통해서 모형의 수정이 지속적으로 이루어지게 되며, 모형에서 필요한 변수를 추가하거나 또는 제거하게 된다. 하지만 연구에 있어 매우 중요한 변수이거나 연구자가 꼭 필요하다고 판단되어지는 변수들의 종속변수에 대한 효과를 알고 싶은 경우에는 변수의 제거 방법을 사용하지 않고도 분석이 가능하다. 그러나 OLS 의 추정량이 이론적인 측면과 상반된 결과를 나타낸다거나 다른 설명변수를 추가 또는 제거 하였을 때 추정량의 변화가 심하거나, 또는 개별 관측·실험자료의 추가, 제거에 따라 추정량의 변화량이 크다면 다중공선성을 의심해 보아야 한다. 체육학 논문에서 보면 다중회귀분석에서 분산분석의 결과가 유의하지만 결정계수는 아주 낮은 경우 또는 반대의 경우가 많이 발견되며 이 또한 다중공선성의 징후에 해당되기 때문에 연구자는 이것을 해결하도록 노력해야 한다. 이럴 경우 주성분회귀분석이나 능형회귀분석을 사용하여야 하며, 어느 정도 타당한 추정치를 얻은 후에 변수 선택방법을 사용하여야 한다.

본 연구에서는 주성분회귀의 경우 하나의 주성분을 제거했으며, 능형회귀에서 k 의 선택 방법에 있어 하나의 식에 의존하여 추정량을 구하였지만 이것이 절대 기준이 될 수 없다. 따라서 연구자는 추가적인 인자분석(factor analysis)이나 ridge trace 를 살펴보면서 여러 가지 종합적인 방법에 의해서 타당한 추정량을 선택해야 할 것이다.

참고문헌

Akdeniz, F. (2002). More on the pre-test estimator in ridge regression. *Communications in Statistics - Theory and Methods, 31,* 987-994.

Chatterjee, S., Hadi, A. S., & Price, B. (2000). *Regression analysis by example* (3rd ed.). John Wiley & Sons.

Draper, N., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). John Wiley & Sons.

Golub, G. H., & Van Loan, C. F. (1996). *Matrix computation* (3rd ed.). The Johns Hopkins University Press.

Golub, G. H., Health, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics,*

21, 215-223.

Hardi, A. S., & Ling, R. F. (1998). Some cautionary notes on the use of principal components regression. *The American Statistician, 52,* 15-19.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: biased estimation for non-orthogonal problems. *Technometrics, 12,* 55-67.

Johnson, R. A., & Wichern, D. W. (1998). *Applied multivariate statistical analysis* (4th ed.). Prentice Hall.

Kaçıranlar, S., & Sakallıoğlu, S. (2001). Combining the LIU estimator and the .principal component regression estimator. *Communications in Statistics - Theory and Methods, 30,* 2699-2705.

Khalaf, G., & Shukur, G., (2005). Choosing ridge parameter for regression problems. *Communications in Statistics - Theory and Methods, 34,* 1177-1182.

Kubokawa, T., & Srivastava, M. S. (2004). Improved empirical Bayes ridge regression estimators under multicollinearity. *Communications in Statistics - Theory and Methods, 33,* 1943-1973.

Kutner, M., Nachtsheim, C., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). McGraw Hill.

Lin, M., & Wei, L. (2002). The small sample properties of the principal components estimator for regression coefficients. *Communications in Statistics - Theory and Methods, 31,* 271-283.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2001). *Introduction to linear regression analysis* (3rd ed.). John Wiley & Sons.

Morison, D. F. (2005). *Multivariate statistical methods* (4th ed.). Thomson Learning.

SAS Institute (1999). *SAS OnlineDoc* (ver. 8). from http://v8doc.sas.com/sashtml/

Strang, G. (1988). *Linear algebra and its applications* (3rd ed.). Thomson Learning.

Thomas, G. B., & Finney, R. L. (1996). *Calculus and analytic geometry* (9th ed.). Addison-Wesley.

논문투고일: 2006. 3. 31

심사완료일: 2006. 7. 2