Chapter 2. 주성분분석

기성복 바지를 살 때 우리 몸의 치수를 모두 알아야 하는가? 그렇지 않다. 허리 둘레와 기장만 알고 있으면 충분하다. 여기에 PCA(Principle Component Analysis: 주성분분석)가 숨어 있다면 믿으시겠습니까? 통계는 우리 일상 생활이다. 바지를 사려면 허리 둘레, 기장 이외 엉덩이 둘레, 허벅지 둘레, 무릎 높이 등 다른 하체에 대한 정보가 있어야 할 것 같지만 (허리, 기장) 두 측정치만 가지고 기성복을 사 입어도 잘 맞는다. 물론 그렇지 않은 사람은 맞춤 옷(혹은 Big & Tall) 집을 찾아야 한다. 이것이 가능한 것은 하체에 대한 많은 체형 측정 변수들이 2 개의 변수로 축약될 수 있고 그 변수가 다른 체형에 대한 정보의 대부분을 가지고 있기 때문이다. 변수 정보를 축약한 변수를 주성분 변수라 한다.

2.1 주성분분석 개념

2.1.1 고유치

행렬 A의 고유 방정식(characteristic equation) $|A_{n\times n} - \lambda I_n| = \underline{0}$를 만족하는 $\lambda_1, \lambda_2, ..., \lambda_n$들을 고유치(eigen value, characteristic value, latent value)라 하고 각 고유치에 대해 $A\underline{e}_i = \lambda_i \underline{e}_i$를 만족하는 벡터를 고유벡터(eigen vector)라 한다. 여기서는 고유치와 고유벡터 계산 방법에 대해서는 다루지 않겠지만 언급하고 싶은 것은 고유치와는 달리 고유벡터는 무수히 많이 존재한다는 것이다. 이 성질 때문에 주성분분석에서 요인 회전이 가능하다.

대칭 행렬에 대해서는 다음이 성립한다. 우리가 다변량분석에서 사용하게 될 공분산행렬이나 상관행렬은 모두 대칭 행렬이다.

(1)고유치는 실수이다

(2)대칭 행렬은 대각화가 가능하다(Diagnosable). $A = U^{-1}DU$ D는 대각원소가 A의 고유치인 대각 행렬이고 U는 직교 행렬이다.

(3)고유벡터는 orthogonal하다. 즉 $e_i' e_j = 0 \; for \; i \neq j$

(4)행렬의 계수와 0이 아닌 고유치의 수는 같다. 즉 0인 고유치가 존재하는 행렬은 full-rank가 아니며 역 행렬이 존재하지 않는다.

j번째 변수와 k번째 변수의 공분산은 $\sigma_{jk} = \mathrm{cov}(x_j \;\; x_k) = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_{ij} - \bar{x}_i)(x_{ik} - \bar{x}_k)$

$j = k$ 이면 공분산은 분산이므로 $\sigma_{kk} = \mathrm{var}(x_k) = \mathrm{cov}(x_k, x_k) = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_{ik} - \bar{x}_k)^2$ 그러므로 공분산행렬(covariance matrix)은 $\Sigma = \begin{bmatrix} \sigma_{11} \; \sigma_{12} \cdots \sigma_{1p} \\ \sigma_{21} \; \sigma_{22} \cdots \sigma_{2p} \\ \vdots \\ \sigma_{p1} \; \sigma_{p2} \cdots \sigma_{pp} \end{bmatrix}$ 이다.

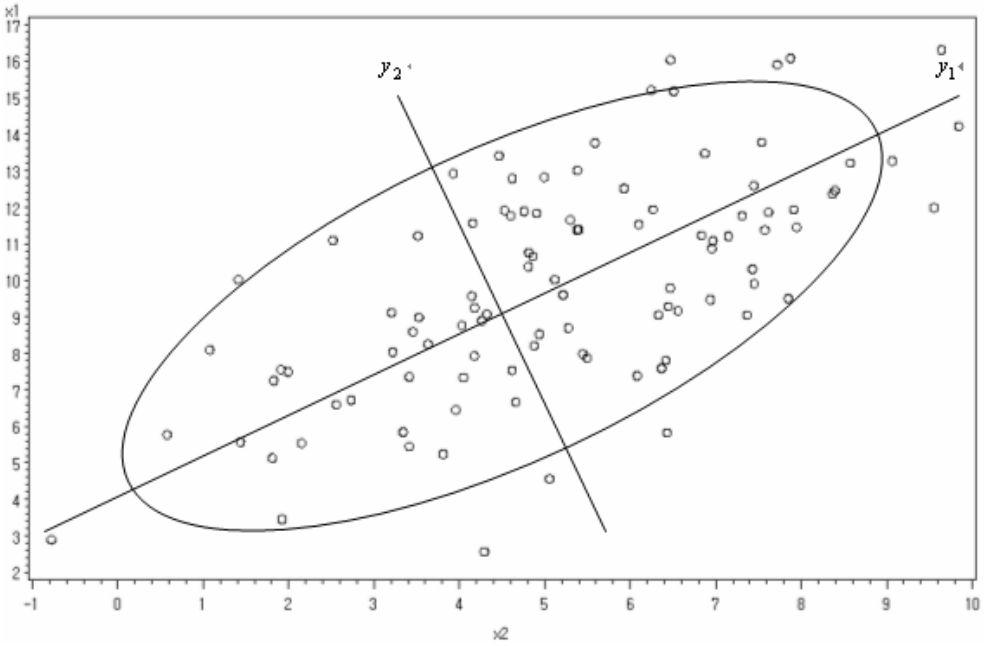
만약 표본 데이터인 경우에는 σ 대신 s을 사용하고 Σ 기호 대신 S 사용한다. j번째 변수와 k번째 변수의 상관계수는 $r_{jk} = \mathrm{cov}(x_j \;\; x_k) / \mathrm{var}(x_j)\,\mathrm{var}(x_k)$ 이므로 상관행렬(correlation matrix) $R = \begin{bmatrix} r_{11} \; r_{12} \cdots r_{1p} \\ r_{21} \; r_{22} \cdots r_{2p} \\ \vdots \\ r_{p1} \; r_{p2} \cdots r_{pp} \end{bmatrix}$ 이다.

공분산행렬 Σ의 고유치는 $|\Sigma - \lambda I| = 0$의 해이다. 공분산행렬은 대칭이므로 모든 고유치 값은 실수이고 행렬 차수만큼의 고유치가 존재한다. 행렬 Σ의 고유치를 $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ 하면 행렬 Σ의 고유치 λ_i 대해 $\Sigma e_i = \lambda_i \underline{e}_i$을 만족하는 u_i을 고유벡터라 한다. 고유벡터

는 수없이 많이 존재하는데 다변량에서는 고유치 λ_j에 대응하는 고유벡터를 e_j라 하면 $\underline{e}_i'\underline{e}_i = 1$이고 $\underline{e}_i'\underline{e}_j = 0 \text{ for } \lambda_i \neq \lambda_j$을 만족하는 고유벡터를 사용한다.

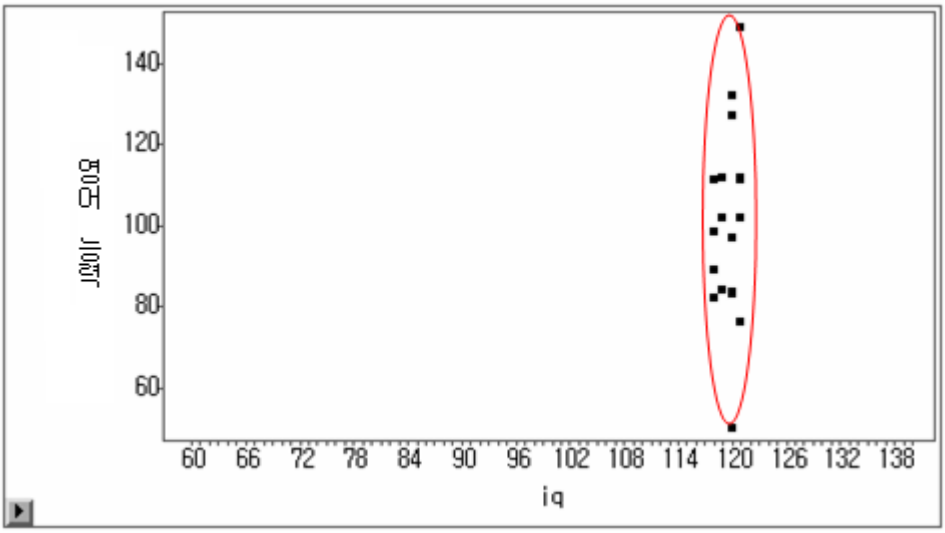
평균 $\underline{\mu} = \begin{bmatrix} 10 \\ 5 \end{bmatrix}$ 이고 공분산-분산 행렬 $\Sigma = \begin{bmatrix} 9 & 2 \\ 2 & 4 \end{bmatrix}$ 인 이변량 정규분포를 고려하자. $\Sigma = \begin{bmatrix} 9 & 2 \\ 2 & 4 \end{bmatrix}$ 의 고유치는 $\lambda_1 = 9.7$, $\lambda_2 = 3.2$ 이고 각 고유치에 대응하는 고유벡터는 $\underline{u}_1 = \begin{bmatrix} 0.94 \\ 0.33 \end{bmatrix}$, $\underline{u}_2 = \begin{bmatrix} -0.33 \\ 0.94 \end{bmatrix}$ 이다. 이로부터 이변량 자료를 추출하여 산점도를 그리면 아래 그림과 같다.

타원(ellipse)의 방정식은 $\frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} = 3(3 = p+1)$ 이고 타원 긴 쪽 길이는 $2\sqrt{3\lambda_1} = 10.8$, 타원 짧은 쪽 길이는 $2\sqrt{3\lambda_2} = 6.2$ 이다. 또한 y_1의 방향에서 분산은 $\lambda_1 = 9$ 이고 y_2의 방향에서 분산은 $\lambda_2 = 4$ 이다. 그러므로 만약 λ_2가 0 이 가까우면 자료는 직선 상에 모이게 된다. 0 이 되면 자료가 모두 직선 상에 모이므로 두 변수의 상관계수는 0 이다. $\lambda_1 = \lambda_2 = 0.5$ 이면 자료 산점도의 형태는 원이 되고 상관계수는 0 이다. 이처럼 고유치의 값은 두 변수 간의 상관 관계를 나타내는 지표가 된다. 변수 y_1과 y_2는 주성분이며 타원의 길이는 주성분의 원 변수 변동에 대한 설명력이다.



2.1.2 맛보기

19 명 학생의 영어 능력과 IQ 를 측정하여 산점도를 그렸다. 산점도를 보면 영어 능력과는 달리 IQ 의 변동(통계학에서는 이를 정보의 개념으로 해석)은 거의 없다. 다른 말로 표현하면 IQ 는 학생 개체들을 구별하는 역할이 미미하나 영어능력 변수는 학생들의 정보를 대부분 가지고 있다고 할 수 있다.



다음은 위의 자료에 대한 주성분분석 결과이다. 공분산행렬로부터 고유치를 구하고 그에 대응하는 고유벡터를 구했다.

공분산 행렬

	영어능력	IQ
영어능력	518.6520468	4.7309942
IQ	4.7309942	1.2280702

고유치

	Eigenvalue
1	518.695300
2	1.184817

고유벡터

Eigenvectors

	Prin1	Prin2
영어능력	0.999958	-.009142
IQ	0.009142	0.999958

영어능력 변수의 변동이 매우 크므로 제일 고유치 $\lambda_1 = 518.69$ 는 제이 고유치 $\lambda_2 = 1.18$ 에 비해 상대적으로 매우 크다. 이는 고유치의 크기가 변수의 분산(변동)을 나타내기 때문이다. 나중에 상세히 설명 하겠지만 주성분 변수 하나만으로 두 변수(영어능력, IQ)의 정보(변동)를 대부분(99.8%) 표현할 수 있다.

2.1.3 주성분분석 사용

원 변수의 축약으로 얻어진 주성분 사용 예제를 보면 주성분분석은 최종 분석이라기 보다는 다른 분석의 중간 단계임을 짐작할 수 있다.

(1)데이터 스크린

데이터가 수집되면 일변량인 경우에는 줄기-잎 그림, 상자-수염 그림을 그려 데이터의 분포 형태나 이상치 존재 여부를 파악한다. 이변량인 경우에는 산점도를 그려 두 변수간 함수 관계와 이상치 등을 판단한다. 그럼 변수가 3 개 이상인 경우 데이터를 어떻게 표현할 것인가? 3 차원 그래프를 산점도로 나타내는 Bubble 그림이 있지만 해석하는데 다소 어려움이 있고 3 개의 변수가 모두 고려된 관계를 설정하는데 충분하지 않다. 주성분분석은 원 변수들을 축약한 주성분을 이용하여 저차원 산점도를 그려 원 변수들의 함수 관계나 개체들의 특성이나 이상치의 존재 여부를 알아보는 다변량분석 방법이다.

(2)개체 순위

개체를 분류하는 경우 측정 항목(변수)이 1-2 개이면 평균 혹은 산점도에 의해 가능하다. 학생 데이터 예제 경우 IQ 는 개체(사람) 간 차이가 거의 없으므로 영어능력 높은 그룹, 낮은 그룹으로 나눌 수 있다. 이처럼 2 개 변수까지는 산점도나 평균만으로 개체 분류가 가능하나 $p \geq 3$ 개인 다변량 데이터 개체를 분류하려면 산점도만으로는 어려움이 있다. 원 변수를 축약한 주성분에 의해 개체를 분류하거나 군집 분석 결과에 대한 해석하는데 주성분을 이용하면 된다.

데이터 스크린과 같은 방법으로 3 개 이상의 변수가 1-2 개 주성분 변수로 변환되므로 평균이나 산점도에 개체 분류가 가능하다. $p \geq 3$ 다변량 데이터 개체 분류는 군집 분석을 실시하면 된다.

(3)회귀분석

다중 회귀분석에서 설명변수간의 상관 관계가 높으면 추정 회귀 계수의 분산이 커져 회귀계수의 부호까지 바뀌는 문제가 발생한다. 이를 다중공선성(multicollinearity) 문제라 한다. 최소자승 추정치(OLS: Ordinary Least Square)를 믿을 수 없게 된다. 이런 경우 ① 문제가 되는 설명변수를 제외하거나 ②능형 회귀분석(Ridge Regression: 추정치의 불편성을 희생하고 최소 분산을 갖는 추정치를 구하는 방법)을 이용하여 문제를 해결한다.

또 다른 방법은 ③설명변수로부터 주성분을 구하고 이 주성분을 설명변수로 이용하여 회귀분석을 실시하여 다중공선성 문제를 해결한다.

다음 절에 설명하겠지만 주성분 변수들은 서로 독립(상관 관계가 존재하지 않는다)이라는 성질이 있는데 이 성질을 이용하여 다중공선성 문제를 해결하게 된다. 그러나 주성분의 의미가 명확하지 않는다면(실제 이런 경우는 현실에서 빈번히 발생한다) 회귀분석 결과를 해석하는데 어려움이 있어 자주 사용되는 방법은 아니다.

(4)판별 분석

판별식을 이용하여 개체를 분류하는 분석을 판별 분석이라 한다. 판별 분석의 경우 분산-공분산행렬의 역 행렬을 구해야 하는데 측정 변수가 너무 많으면 계산이 오래 걸린다. 이런 경우 주성분분석 방법에 의해 변수의 수를 줄여 만든 새로운 변수(주성분)에 의해 판별 분석을 하게 된다. 그러나 요즈음은 컴퓨터 성능의 발달로 역 행렬을 구하는데 문제가 없어 판별분석 도구로 주성분분석을 사용하지는 않는다.

2.2 주성분분석하기

다음 원칙에 의해 p 개의 원 변수로부터 주성분을 얻는다. 주성분은 원 변수의 선형결합에 의해 구해지므로 새로운 변수이다. 그러므로 주성분과 주성분 변수는 동일한 개념이다.

(1)주성분 간에는 서로 상관 관계가 전혀 존재하지 않는다. (독립이다) 즉 주성분 변수간에는 정보의 겹치는 부분이 없다는 것을 의미한다.

(2)제일 주성분은 데이터의 변동(분산, 정보)을 가장 많이 설명하고 계속 구해지는 제이, 제삼, ... 주성분은 데이터의 나머지 정보들을 설명하고 크기는 점점 줄어든다.

변수가 p 개인 원 변수 벡터 $\underline{x}=\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$ 의 선형 결합의 형태로 주성분이 구해진다. 주성분 벡터를 $\underline{y}=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}$ 라 하면 $\underline{y}=L\underline{x}$ 에서 적절한 $L=\begin{pmatrix} l_{11}\ l_{12}\cdots l_{1p} \\ l_{21}\ l_{22}\cdots l_{2p} \\ \vdots \\ l_{p1}\ l_{p1}\cdots l_{pp} \end{pmatrix}$ (선형 계수 행렬)을 구하는 것이 주성분분석이다.

2.2.1 주성분 정의

주성분분석은 p 개의 원 변수의 선형 결합의 주성분을 이용하여 원 변수의 공분산 구조를 설명하는 방법이다. 공분산 구조를 설명한다는 것은 원 변수의 변동 합과 주성분 변수의 변동 합은 동일하다는 것을 의미한다. 그럼 선형 계수는 어떻게 구할 것인가?

변수 벡터 $\underline{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$ 가 공분산행렬 $\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots \sigma_{2p} \\ \cdots & & \\ \sigma_{p1} & \sigma_{p2} & \cdots \sigma_{pp} \end{bmatrix}$ 를 갖는다고 하자.

공분산행렬의 고유치를 $\lambda_1 \geq \lambda_2 . \geq .. \geq \lambda_p$ 라 하고 각 고유치(λ_i)에 대응하는 고유벡터를 e_i 라 하고 $y_i = \underline{e}'_i \underline{x}$ 라 하면 $Var(Y_i) = e_i' \Sigma e_i = \lambda_i$, $Cov(Y_i, Y_k) = e_i' \Sigma e_k = 0, \; for \; i \neq k$ 이 성립한다. 이런 경우 원 변수 X_i 들의 변동의 합은 고유치의 합과 동일하다. 다음의 Λ 은 대각 원소가 Σ 의 고유치인 대각 행렬이고 P 는 고유치에 대응하는 고유벡터로 구성된 직교 행렬이다.

$$\sum_{i=1}^{p} Var(x_i) = tr(\Sigma) = tr(P\Lambda P') = tr(\Lambda P'P) = tr(\Lambda) = \sum_{i=1}^{p} \lambda_i \; .$$

k 번째 주성분 $y_k = e_k' \underline{x} = e_{1k} x_1 + e_{2k} x_2 + ... + e_{pk} x_p$ 의 원 변수의 변동 설명 비율은 $\dfrac{\lambda_k}{\sum_{i=1}^{p} \lambda_i}$

2.2.2 주성분 구하기

(1)제일 주성분(first principal component): $\underline{a}_1' \underline{a}_1 = 1$ 을 만족하는 벡터 $\underline{a}_1$ 중 $\underline{a}_1' (\underline{x} - \underline{\mu})$ 의 분산 $V(\underline{a}'_1 (x - \mu))$ 을 최대화하는 $\underline{a}_1$ 을 찾은 후 $\underline{y}_1 = \underline{a}_1' (\underline{x} - \underline{\mu})$ 첫 번째 주성분이라 한다.

$\underline{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$ 는 평균 벡터이고 μ_i 는 변수 x_i 의 평균이다.

(2)제이 주성분: $\underline{a}_2' \underline{a}_2 = 1$, $\underline{a}_1' \underline{a}_2 = 0$ (첫 번째 주성분과 독립이다. 즉 첫 번째 주성분이 설

명하는 자료의 변동 부분과 겹치지 않는다)을 만족하고 $\underline{a}_2'(\underline{x}-\underline{\mu})$의 분산을 최대화하는 $\underline{a}_2$을 구하고 $y_2=\underline{a}_2'(\underline{x}-\underline{\mu})$을 두 번째 주성분이라 한다.

(3)제삼 주성분: $\underline{a}_3'\underline{a}_3=1$, $\underline{a}_1'\underline{a}_3=0$, $\underline{a}_2'\underline{a}_3=0$ (첫 번째, 두 번째 주성분과 독립)을 만족하고 $\underline{a}_3'(\underline{x}-\underline{\mu})$의 분산을 최대화 하는 $\underline{a}_3$을 구하고 $y_3=\underline{a}_3'(\underline{x}-\underline{\mu})$을 세 번째 주성분이라 한다.

위의 방법을 반복하여 변수의 개수만큼의 주성분들(y_1, y_2, …, y_p)을 구한다. 주성분은 원 변수의 개수만큼 존재하고 각 주성분은 서로 독립(겹치는 정보가 없다)이다. 각 주성분의 벡터가 아니라 하나의 변수임에 유의하기 바란다.

$$\underline{y}=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}=\begin{pmatrix} \underline{a}_1' \\ \underline{a}_2' \\ \vdots \\ \underline{a}_p' \end{pmatrix}\left(\underline{x}-\underline{\mu}\right)_{p\times 1}$$

계수 벡터 $\underline{a}_j$를 어떻게 구할 것인가?

변수 벡터의 분산-공분산행렬 Σ의 고유치 $\lambda_1\geq\lambda_2\geq..\cdots\geq\lambda_p$에 대응하는 고유벡터 $\underline{e}_1$, $\underline{e}_2$,…, $\underline{e}_p$은 앞에서 언급한 주성분 조건을 만족한다. 즉 $\underline{a}_1=\underline{e}_1$, $\underline{a}_2=\underline{e}_2$, …, $\underline{a}_p=\underline{e}_p$ 이라면 $\underline{e}_i\underline{e}_j=1,\ i=j$, $\underline{e}_i e_j=0,\ i\neq j$ 이다. 고유벡터를 주성분 계수로 사용하는 경우 다음이 성립한다.

①주성분 y_j의 분산은 고유치 λ_j와 같다.

②분산-공분산행렬의 대각 합 $tr(\Sigma)$은 원 변수 $x_1,x_2,,...,x_p$ 변동(분산)의 합이다.

③ $tr(\Sigma)=\sum_i V(x_i)=\sigma_{11}+\sigma_{22}+...+\sigma_{pp}$ 이므로 주성분 y_j의 변동 설명력은 $\lambda_j/\sum_{j=1}^{p}\lambda_j$ 이다.

(영어능력, IQ) 예제 데이터를 보면 공분산행렬의 대각합(518.65+1.23)과 공분산행렬로부터 구한 고유치의 합(518.70+1.18)은 같다. 제일 주성분의 원 변수 변동 설명력은 99.8%(=518.7/(518.7+1.18))이다.

2.2.3 주성분 점수

주성분의 값(관측치)을 주성분 점수라 한다. 주성분 점수는 위에서 구한 주성분 계수를 이용하여 구하면 된다. 다음은 r번째 개체의 j번째 주성분 점수 계산식이다.

$y_{rj} = \underline{e}_j' \underline{x}_r = e_{j1} * x_{r1} + e_{j2} * x_{r2} + ... + e_{jp} * x_{rp}$, $\underline{x}_r$은 r-번째 개체의 측정치($r = 1,2,...,n$)이다.

학생 데이터 예제에서 (영어능력, IQ) 변수에 대하여 j번째 학생(개체)의 영어능력은 110 이고 IQ=125 였다면 j번째 관측치의 주성분을 구하면 다음과 같다.

제일 주성분 점수: $Y_{j1} = 0.999958(110)_j + 0.009142(125)_j$

제이 주성분 점수: $Y_{j2} = -0.009142(110)_j + 0.999958(125)_j$

원 변수가 2 개인 경우 주성분도 2 개가 얻어진다. 주성분을 구할 때 서로 독립이 되도록 구했으므로 주성분 점수들(Y_1, Y_2) 간의 상관계수는 0 이다.

2.2.4 성분 부하벡터

공분산행렬의 고유벡터로부터 얻어지는 주성분 계수를 $\underline{e}_j$라 하면 $\underline{c}_j = \sqrt{\lambda_j}\underline{e}_j$는 성분부하벡터(component loading vector)이다. 성분 부하벡터는 주성분을 만들 때 사용되는 원 변수 앞의 선형 계수에 해당하므로 주성분의 이름을 부여하는데 사용한다. 성분 부하 값이 크다는 것은 그에 대응하는 원 변수의 영향이 크다는 것을 의미하므로 성분 부하 값이 큰 변수를 살펴 주성분의 이름을 부여하면 된다. 즉 부하 크기는 주성분 내의 각 변수의 중요성을 나타내므로 이를 이용하여 주성분 변수의 이름을 부여할 수 있다.

주성분은 원 변수의 선형 결합의 형태로 되어 있으므로 적절한 변수 이름을 부여하여 첫 번째 주성분, 두 번째 주성분이라 아니라 그 변수 명으로 부르는 것이 분석 결과 해석에 중요하다. 부하 벡터는 고유치의 제곱근을 곱해준 것 밖에 없으니 고유벡터의 값만으로 비교해도 충분하다.

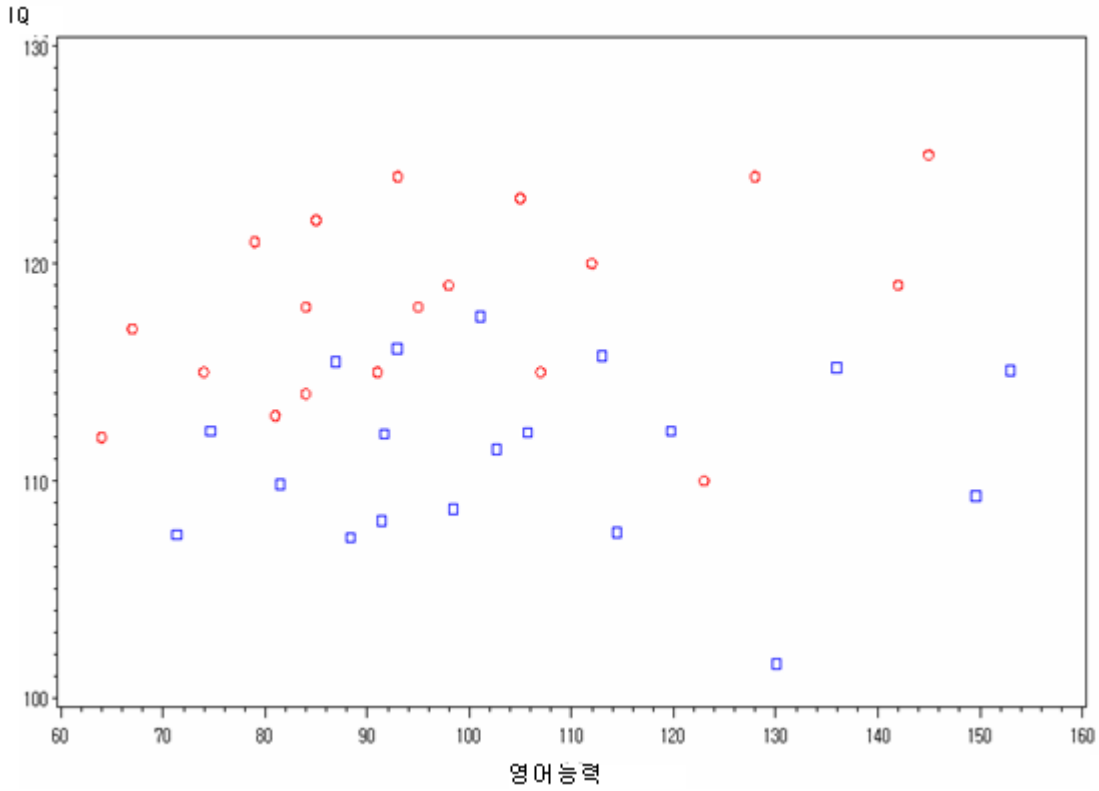
(1)성분 부하 값으로 주성분에 대한 원 변수의 영향력을 측정하므로 원 변수의 측정 단위는 유사해야 한다. 그러므로 원 변수의 측정 단위가 다른 경우에는 공분산행렬을 이용하기 보다는 상관행렬을 이용하여 고유치, 고유벡터를 구하는 것이 바람직하다.

(2)성분 부하의 크기 비교는 각 주성분 내에서만 가능하며 주성분간 성분 부하 값을 비교하는 것은 의미가 없다.

학생 데이터 예제에서 제일 주성분의 부하를 보면 영어능력 변수는 0.999958 이고, IQ 변수는 0.009 이다. 영어능력 변수의 영향력이 크므로 제일 주성분을 학습능력 주성분이라 이름 붙일 수 있다. 같은 이유로 두 번째 주성분은 IQ 변수의 부하 값이 크므로 지적 주성분이라 부를 수 있을 것이다. 이처럼 성분 부하 값은 주성분 이름을 부여하는데 사용된다.

2.2.5 주성분과 원 변수

학생 데이터 예제에서 공분산행렬로부터 고유치를 구하면 $\lambda_1 = 545.3, \lambda_2 = 16.3$ 이고 이로부터 얻은 고유벡터는 (0.0666, 0.998), (0.998, -0.0666)이다. 이 고유벡터는 주성분 점수를 구하는 선형 계수가 된다. 제일 주성분을 X-축, 제이 주성분을 Y-축으로 하여 산점도를 그려보자.



원은 원 변수, 네모는 주성분 값에 대한 산점도이다. 산점도의 형태는 크게 차이가 없어 보이나 주성분 산점도를 살펴보면(네모 관측치) 원에 비해 y 축(제이 주성분)의 변동은 줄어들었으며 x 축(제일 주성분)의 변동은 커진 것을 알 수 있다. 이는 주성분을 구할 때 제일 주성분이 변동을 가장 많이 설명하도록 설정하였기 때문이다.

2.2.6 주성분 개수 판단

주성분분석은 원 변수 차수를 줄이는데 목적이 있는데 실제 주성분의 개수는 원 변수의 수만큼 존재하게 된다. 즉 원 변수가 단지 서로 독립인 변수들로 변환된 것 뿐이다. 사실 변수의 개수(차수)를 줄여서는 원 변수들이 가진 정보(변동)를 100% 표현할 수 없다. (영어능력, IQ) 예제처럼 IQ 변수의 변동이 아무리 없어도 2 차원 공간에 표현된 정보(산점도)를 1 차원 공간으로 줄인다면 어느 정도 정보의 희생은 각오해야 한다. 산점도에서 오른쪽에서 불을 비춰 점들을 y 축에만 나타나게 한다면(2 차원 ► 1 차원) 영어능력이 115 점수 근처에 있는 3 명의 학생들은 거의 한 점에 표시된다. IQ 는 다소 차이가 있음에도 불구하고 3 명의 학생은 동일한 점으로 표현되므로 정보가 희생되는 결과를 초래한다. 그러므로 데이터 차수를 줄이기 위해서는 원 변수 변동에 대한 설명을 어느 정도 희생해야만 한다.

80% 규칙

k 번째 주성분의 설명력은 $\lambda_k / \sum_{j=1}^{p} \lambda_j$ 이다. 공분산행렬로부터 주성분을 구할 때 첫 번째 주성분의 설명력(λ_1)이 가장 크고, 두 번째 주성분의 설명력은 $\lambda_2, \cdots$ 이렇게 설명력의 크기 순으로 주성분을 구했다. 주성분 설명력의 합이 변수의 총 변동 중 약 80%가 되는 주성분까지만 사용하면 어떨까?

$$0.7 \le \frac{\hat{\lambda}_1}{tr(S)} + \frac{\hat{\lambda}_2}{tr(S)} + \ldots \le 0.9$$

실험실의 자료이면 주성분이 2-3 개 정도이면 90%, 사람의 의견을 점수화한 것은 주성분이 5-6 개 되어야 70% 정도가 된다. 특히 공분산행렬 대신 상관행렬을 사용하는 경우 고유치 값이 1 이상인 주성분만 사용하면 총 변동의 80%정도를 설명하므로 고유치가 1 이상인 고유벡터를 선형 계수로 사용하여 주성분 점수를 계산하면 된다.

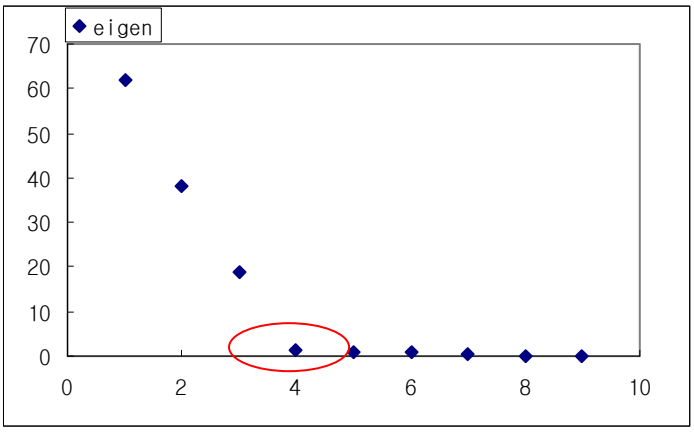
학생 예제 데이터에서 제일 고유치의 원 변수 변동 설명 비율이 99.8%로 대부분 차지하므로 2 개의 원 변수를 하나의 주성분으로 줄일 수 있을 것이라 판단할 수 있다.

공분산 행렬 고유치

	Eigenvalue	Difference	Proportion	Cumulative
1	518.695300	517.510484	0.9977	0.9977
2	1.184817		0.0023	1.0000

SCREE plot 사용

$(1, \hat{\lambda}_1), (2, \hat{\lambda}_2), \cdots$ SCREE 그림을 이용하여 갑자기 떨어지거나 0 에 가까워지는 것 이전 주성분만을 사용하면 된다. 다음은 원 변수가 9 개 있는 경우 SCREE plot 을 그린 것이다. 4 번째부터 뚝 떨어지고 0 에 가까우므로 주성분이 3 개이면 충분하다.



학생 데이터 예제에서 고유치가 518.69 에서 1.18 로 급격히 떨어지므로 주성분의 개수가 1 개이면 충분하다. 고유치 값의 변화를 통해 주성분 개수를 결정하기 위하여 SCREE plot 을 굳이 그릴 필요는 없다. 공분산행렬을 사용하는 경우 고유치의 누적 설명력이 80% 이상 되는 고유치, 상관행렬일 경우는 고유치가 1 이상인 고유치만 선택하면 된다.

2.2.7 상관행렬

측정 변수들의 측정 단위가 차이가 나는 경우 분산의 단위도 달라진다. 이로 인하여 측정 단위가 큰 변수의 분산이나 그 변수와 다른 변수들간 공분산이 커지므로 공분산행렬을 이용하여 구한 고유치나 고유벡터는 그 변수의 영향을 많이 받는다. 이런 문제점을 해결하기 위하여 공분산행렬에서 각 변수의 분산의 변동을 나누어 준 상관행렬로부터 고유치, 고유벡터를 구하고 이를 이용하여 주성분을 구하게 된다.

주성분분석은 변수의 변동(분산, 공분산)을 잘 설명하는 주성분을 찾는 것이 목적인데, 상관행렬을 사용하면 측정 단위 조정으로 인하여 변동에 대한 정보가 축소되는 경향이 있다. 그러므로 측정 단위의 차이가 크지 않다면 그대로 사용하거나 측정 단위를 조정하여 사용하면 된다. 예를 들어 다른 변수의 단위는 두 자리 정수인데 반해 소득(단위: 천원) 변수의 단위가 세자리 정수라면 소득 단위를 만원으로 하여 자릿수를 조정하고 공분산행렬을 이용하여 주성분분석을 실시하면 된다.

$$R=\begin{pmatrix} \frac{\sigma_{11}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{11}}} & \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} \cdots \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} \\ \frac{\sigma_{21}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{11}}} & \frac{\sigma_{22}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{22}}} \cdots \frac{\sigma_{2p}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{pp}}} \\ & \cdots \\ \frac{\sigma_{p1}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{11}}} & \frac{\sigma_{p2}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{22}}} \cdots \frac{\sigma_{pp}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{pp}}} \end{pmatrix} \rightarrow R=\begin{pmatrix} 1 & \rho_{12} \cdots \rho_{1p} \\ \rho_{21} & 1 \cdots \rho_{2p} \\ & \cdots \\ \rho_{p1} & \rho_{p2} \cdots 1 \end{pmatrix} \rightarrow \hat{R}=\begin{pmatrix} 1 & r_{12} \cdots r_{1p} \\ r_{21} & 1 \cdots r_{2p} \\ & \cdots \\ r_{p1} & r_{p2} \cdots 1 \end{pmatrix}$$

모집단 분산-공분산행렬 Σ의 추정치는 표본 분산-공분산행렬 $\hat{\Sigma}=S$이고 모집단 상관행렬 R의 추정치는 표본 상관행렬 $\hat{R}$이다. 상관행렬로부터 주성분을 구하는 것은 표본 분산-공분산행렬 S로부터 고유치를 구하고 그에 대응하는 고유벡터를 구하는 방법과 동일하다.

(1) S 대신 $\hat{R}$으로부터 고유치와 고유벡터를 구한다.

(2)주성분 개수의 선택은 고유치가 1 이상일 때만 선택하면 변동의 80%가 설명된다.

(3)주성분 변수의 변동 설명력은 다음과 같이 정리할 수 있다. p는 원 변수의 개수이다.

공분산행렬 Σ	상관행렬 R
$\lambda_j / \sqrt{tr(S)} = \lambda_j / \sum_{i=1}^{p} \lambda_k$	λ_j / p

2.2.8 주성분분석 절차

(1)원 변수 단위 점검: 원 변수의 단위를 살펴 단위가 유사하면 주성분분석을 위하여 공분산행렬을 이용한다. 단위가 다른 경우 맞출 수 있으면 (예: kg ► pound, 단위: 원 ► 단위: 천원) 맞춘 후 공분산행렬을 사용하고 불가능한 경우 상관행렬을 사용한다.

(2)고유치, 고유벡터(계수 벡터), 주성분의 원 변수 변동 설명력(%) 계산

(3)주성분 개수 정하기: 80% 규칙이나 고유치 1 이상, 실제는 주성분 2개 이하 선호,

(4)고유벡터를 이용하여 주성분 점수 구하고 적절한 이름 부여: 주성분 이름 부여는 다소 주관적이고 때로는 적절한 이름 부여가 불가능하다.

(5)주성분 점수 이용하여 2차 분석

①정규성 검정:원 변수가 정규 분포를 따르는지 선택된 주성분 점수에 의해 검정할 수 있다.

②이상치 진단: 선택된 주성분 산점도를 이용하여, 주성분이 하나인 경우에는 상자-수염 그림을 이용하여 이상치를 발견한다.

③개체 분류: 선택된 주성분을 이용하여 개체를 순서화(서열화)하거나 개체를 그룹화

하는데 사용할 수 있다.

④회귀분석: 설명변수간 다중공선성 문제가 발생하면 그 해결책으로 모든 주성분을 설명변수로 사용하여 회귀분석을 실시한다.

2.3 주성분분석 예제 1 (공분산행렬 사용)

A 회사에서는 48 명의 지원자에 대해 그들의 능력을 10 점 만점으로 15 개 영역(변수 이름 X1~X15)을에 대해 조사하여 가장 점수가 높은 지원자 6 명을 선발하려고 한다.

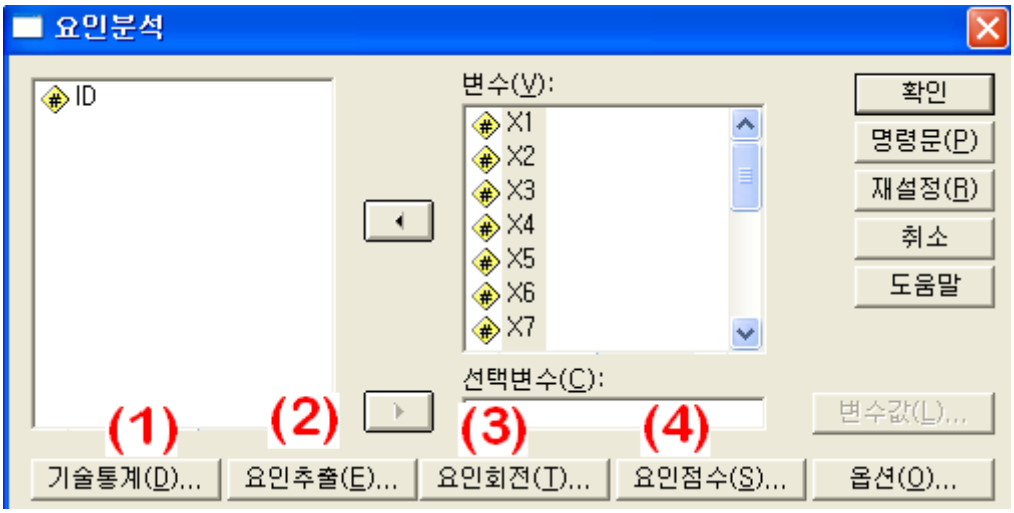
■🖫APPLICANT.SAV■ [Applied Multivariate Methods for Data Analysts, Dallas E. Johnson]

15 개 영역 점수를 1 개로 만들 수 없을까? 그러면 그 변수 값이 큰 6 명을 선발하면 된다. 아니 2 개로 만든다면? 산점도를 그려 점수가 높은 6 명을 선발하면 된다. 이처럼 15 개 변수가 1-2 개의 변수로 줄일 수 있다면 이것이 가능하다. 이를 주성분분석이라 한다.

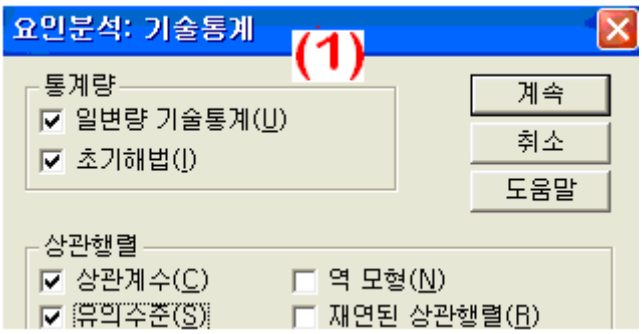
SPSS 에는 주성분분석이라는 메뉴는 없는데 이는 요인분석 메뉴에서 가능하기 때문이다. 요인분석에서 요인을 구하는 방법으로 가장 많이 사용되는 것이 주성분 방법인데 이것이 주성분 구하는 것과 동일하기 때문이다. 이에 대한 이론적 내용은 3 장에서 다루기로 한다.

15 개 변수가 모두 10 점 리커트 척도로 단위가 동일하므로 공분산행렬을 이용하여 주성분을 구하면 된다.

2.3.1 메뉴 설정



다음은 원 변수의 기술 통계나 상관계수를 출력하라는 설정이다. 상관계수가 높은 변수들은 어떤 주성분이 구해질 때 선형계수 값이 커짐을 발견할 수 있다. 주성분을 만드는데 원 변수의 선형 계수 값이 큰 변수를 살펴보면 서로는 상관 관계가 높음을 알 수 있다.

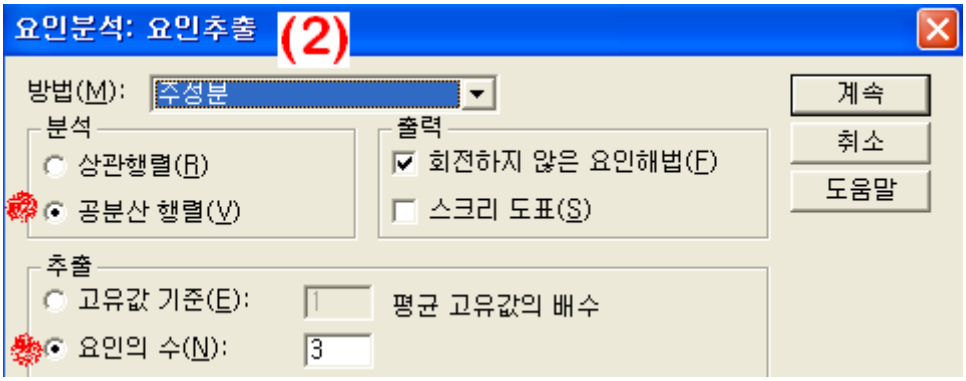


기술통계량

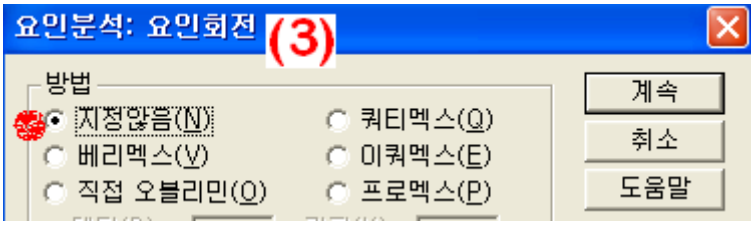
	평균	표준편차	분석수
X1	6.00	2.674	48
X2	7.08	1.966	48
X3	7.08	1.988	48
X4	6.15	2.896	48

		X1	X2
상관계수	X1	1.000	.239
	X2	.239	1.000
	X3	.044	.123
	X4	.306	.380
	X5	.092	.431

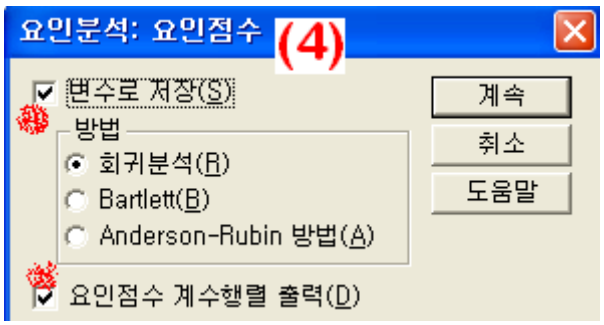
주성분분석을 위해서는 “요인분석”, “요인회전” 창 설정은 다음과 같이 하면 된다. 원 변수의 측정단위가 같으므로 공분산행렬을 사용하도록 설정하였다. SCREE 도표는 주성분의 개수를 결정하는 고유치 크기를 나타낸 그래프이다. 요인의 수를 3 개만 선택한 것은 주성분의 개수가 3 개 이상이면 축약의 의미가 없기 때문에 미리 설정한 것이다.



요인 회전은 요인분석에 사용되는 옵션이다. 주성분 분석에서는 "지정 않음"을 선택하면 된다.



주성분(Y_k) 점수를 데이터의 변수로 저장하도록 옵션을 설정하였다. 주성분 점수는 주성분 관측치이므로 데이터 스크린, 굽집/판별 분석, 회귀분석 등에 이용한다. 요인점수 계수행렬 출력 옵션을 선택하면 표준화된 주성분($Y_k / \sqrt{\lambda_k}$)을 구하는 계수가 출력된다. 이 계수를 이용하여 주성분 이름을 부여하게 된다. "회귀분석" 방법은 요인 점수를 구하는 방법이다. 이론적 내용은 3 장을 참고하기 바란다.

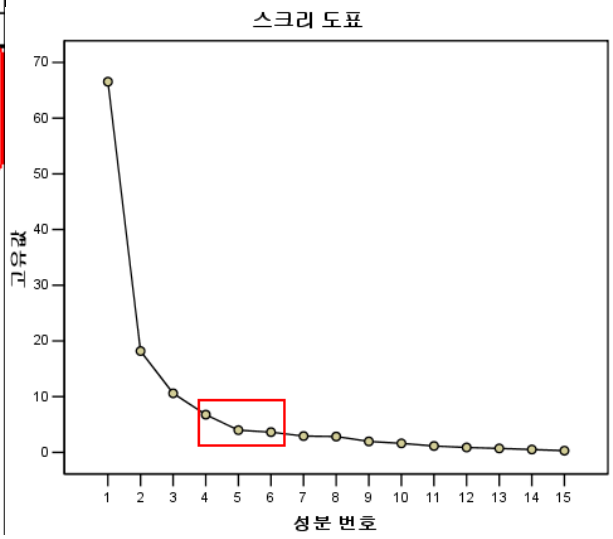


2.3.2 고유치와 주성분 개수

주성분은 원 변수의 개수(15)만큼 존재한다. “전체” 열은 고유치의 값이고, “%분산” 열은 주성분의 원 변수 변동 설명 비율이다. 마지막 “%누적” 열은 주성분 설명력의 누적이다. 출력 결과에 의하면 4 개의 주성분이면 누적 설명력이 83%이므로 15 개의 변수가 4 개의 주성분으로 축약될 수 있다. 그러나 주성분이 3 개인 경우 설명력은 78%으로 80%와 차이가 없고 주성분 개수가 적을수록 주성분분석 효율성이 높아지므로 3 으로 하는 것이 적절하다.

설명된 총분산

성분	초기 고유값[a] 전체	% 분산	% 누적
1	66,536	54,300	54,300
2	18,181	14,837	69,137
3	10,591	8,643	77,780
4	6,768	5,523	83,303
5	3,986	3,253	86,556
6	3,628	2,961	89,517
7	2,916	2,380	91,896
8	2,836	2,314	94,210
9	1,955	1,596	95,806
10	1,613	1,317	97,123
11	1,136	,927	98,050
12	,873	,712	98,762
13	,707	,577	99,339
14	,509	,415	99,754
15	,301	,246	100,000



상관행렬을 이용하는 경우와는 달리 공분산행렬을 이용하여 고유치를 구했으므로 설명력의 80%와 eigen-value 의 값이 1 이상인 주성분과는 일치되지 않는다. SCREE 도표는 고유치의 크기를 시각적으로 표시한 것 정도이다. 이것만으로는 적절한 주성분의 개수를 판단할 수 없다. 그러므로 자주 사용되는 그래프는 아니다. 고유치의 합은 원 변수의 분산의 합과 동일하다. 즉 원 변수 공분산행렬의 대각 원소의 합과 고유치의 합은 동일하다.

원 변수가 10 점 리커드 척도이므로 일반적으로 1~2 개 주성분만으로는 변동의 80%를 설명하는 경우는 흔하지 않다. 주성분분석이 이용되는 예제를 보면 회귀분석을 제외하고는 주성분 3 개 이상이면 주성분분석의 효과가 없다. 그래서 주성분의 설명력을 보지 않고 우리는 주성분을 3 개만 출력되도록 설정하였다. 미리 3 개로 설정하기 권한다. 만약 주성분 2 개만으로 원 변수 변동의 80% 이상을 설명한다면? 요인 개수를 2 로 하여 다시 실행해야 하나? 그럴 필요 없다. 앞의 2 개만 사용하면 된다.

2.3.3 주성분 구하기

6 명의 신입 사원을 어떻게 뽑을 것인가? 제일 주성분을 이용하여 첫 번째 주성분 점수가 가장 높은 6 명을 선발, 제이 주성분, 아니면 제삼 주성분만을 이용하거나 제일 주성분과 제이 주성분의 산점도를 이용하여 사원을 선발할 수 있다.

적절한 주성분은 4 개이지만 3 개만 출력되도록 하였다. 그 주성분들은 어떻게 만들어졌나? 주성분 점수는 수작업으로 계산할 필요가 없다. 이미 앞에서 "요인점수" 설정 창에서 "변수로 저장(S)" 옵션을 선택했으므로 데이터에 자동 저장되어 있다.

제일 주성분: $Y_1 = 1.216 * X1 + 1.079 * X2 + ... + 2.239 * X15$

제이 주성분: $Y_2 = 1.584 * X1 - 0.125 * X2 + ... + 2.008 * X15$

제삼 주성분: $Y_3 = 0.652 * X1 + 0.136 * X2 + ... + 0.055 * X15$

성분점수 계수행렬[a]

	성분		
	1	2	3
X1	,049	,233	,165
X2	,032	-,014	,025
X3	,007	,047	-,080
X4	,070	-,061	,534
X5	,069	-,134	-,141
X6	,131	-,146	-,121
X7	,037	-,179	,348
X8	,160	-,073	-,298
X9	,067	,494	,025
X10	,114	,009	-,103
X11	,112	-,084	-,221
X12	,126	-,053	-,047
X13	,139	-,019	,040
X14	,074	-,028	,315
X15	,111	,364	,017

2.3.4 주성분 이름 붙이기

주성분이 만들어질 때는 원 변수 모두 사용되었기에 각 주성분에는 모든 원 변수의 정보가 다 들어 있다. 원 변수의 정보가 모두 들어 있기는 하지만 영향의 정도는 다르므로 영향 정도를 이용하여 주성분 이름을 붙이게 된다. 주성분에 대한 원 변수의 영향 정도를 나타내는 값이 계수이다. 원 변수가 주성분 변수에 영향을 많이 준다는 의미는 계수 값의 크기가 크다는 것을 의미한다. 사용된 변수가 모두 10 점 리커트 척도로 값의 크기가 유사하므로 계수 크기를 이용하여 주성분 이름을 부여할 수 있다. 원 변수의 측정 단위가 다르면 공분산행렬 대신 상관행렬을 이용하면 된다.

각각의 주성분 내에서 계수가 큰 변수들을 묶은 후 변수들이 함께 나타내는 지표(index)를 이용하여 이름을 부여하면 된다. 상당히 주관적이고 어려운 작업(이로 인하여 회귀분석

의 설명변수로는 사용하지 않는다)이다. 음의 의미는 그 주성분 계산 시 다른 변수와 반대로 작용한다는 것으로 절대 크기가 양의 부호 계수 큰 것과 유사하다면 함께 고려되어야 한다.

- 제일 주성분 계수 크기에 의하면 X6(명석), X8(판매능력), X10(돌파력), X11(야망), X12(개념파악), X13(잠재력) 변수의 크기가 크므로 제 1 주성분은 정신적&지적 능력으로 할 수 있다.
- 제이 주성분에서는 X9(경험), X15(적합)이 큰 역할을 하므로 경험 주성분이라 할 수 있다.
- 제삼 주성분에서는 X4(호감), X7(진실), X14(사교)의 크기가 크므로 감성 변수로 이름하면 된다.

이처럼 각 주성분에 이름을 부여하는 것은 매우 주관적이고 오류를 범할 가능성이 높다. 기성복 바지처럼 2 개의 주성분 변수로 축약되고 그 주성분 변수에 적절한 이름(기장, 허리사이즈)을 쉽게 부여할 수 있다면 다행이지만 실제 응용에서는 쉽지는 않다.

2.3.5 주성분 점수 이용하기

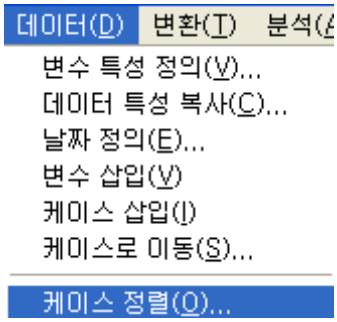
15 개의 원 변수를 축약하여 3 개의 주성분을 얻었다. 실제 4 개(설명력 83%)까지 필요하지만 주성분 3 개만으로 78% 정도 설명되고 4 개 이상(회귀분석 사용 제외)이면 효율성이 떨어지므로 3 개만 선택하였다. 주성분 점수(주성분의 관측치)는 “요인점수” 창에서 출력 옵션을 설정하였으므로 원 변수 옆 열에 3 개의 주성분 점수가 출력된다.

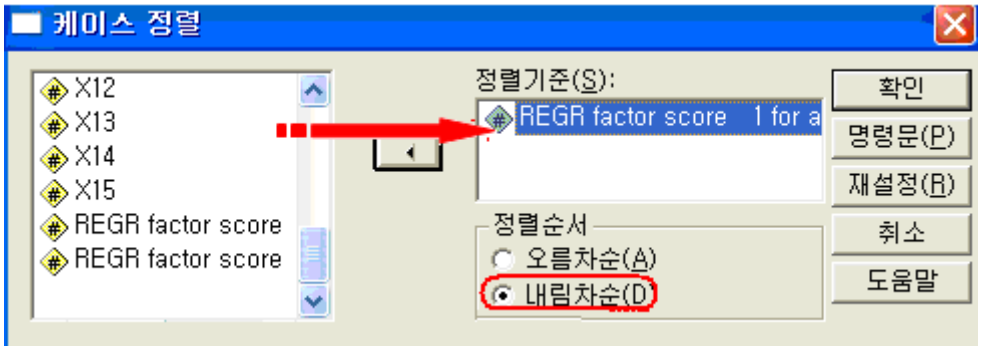
X10	X11	X12	X13	X14	X15	FAC1_1	FAC2_1	FAC3_1
8	9	7	5	7	10	.52765	-.08958	-.54000
9	9	8	8	8	10	1.24331	.09801	.02686
9	9	8	6	8	10	.80051	-.03955	-.09948

▣개체 순위

지적 능력 주성분, 경험 주성분, 감성 주성분을 이용하여 우수 지원자를 선발하면 된다. 주성분 3 개의 평균 점수가 가장 높은 지원자를 뽑으면 된다. 마케팅 부서라면 경험 주성분이 높은 지원자를 기획 부서는 지적 능력이 가장 높은 지원자를 선택하면 된다.

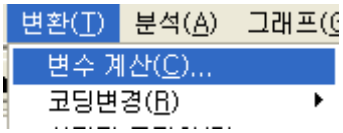
지적 능력이 가장 우수한 지원자를 선발하려면 제일 주성분 점수 크기 순으로 정렬하여 최고 점수 6 명을 선발하면 된다. 지원자 ID 40, 39, 8, 23, 22, 2 가 선발된다.

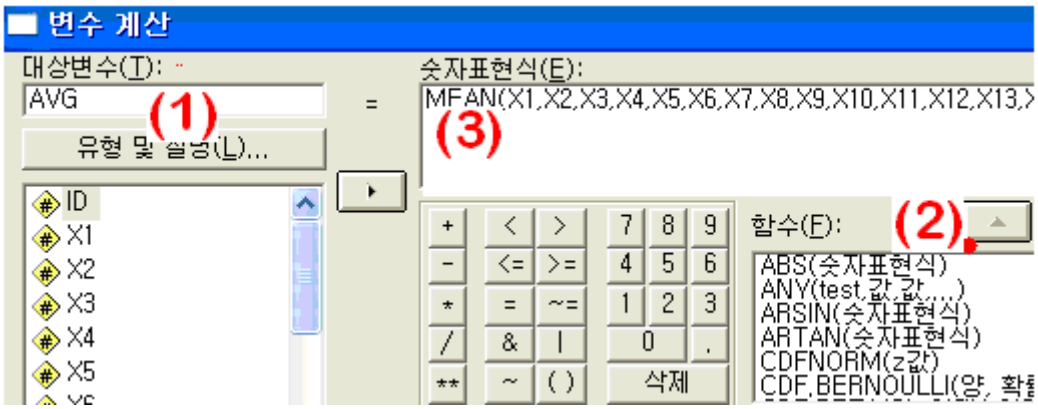




	ID	X1	X2	X3	X4	X5	X15	FAC1_1	FAC2_1	FAC3_1
1	40	10	6	9	10	9	10	1.69234	.85609	.57360
2	39	10	6	9	10	9	10	1.61581	.91339	.72387
3	8	9	9	9	8	9	10	1.35740	1.05203	.02561
4	23	7	10	7	9	9	8	1.30004	-.64651	.40179
5	22	9	8	7	8	9	8	1.27727	-.44525	.29810
6	2	9	10	5	8	10	10	1.24331	.09801	.02686
7	24	9	8	7	10	8	8	1.20384	-.54550	.86759
8	7	9	9	8	8	8	10	1.17611	1.15442	.09432

만약 원 변수 15 평균 점수에 의해 우수 지원자를 선발하면 다음과 같다. 여기서는 ID 7 번이 선발되었으나 제일 주성분(지적 능력)에 의해서는 ID 2 번이 대신 선발되었다.

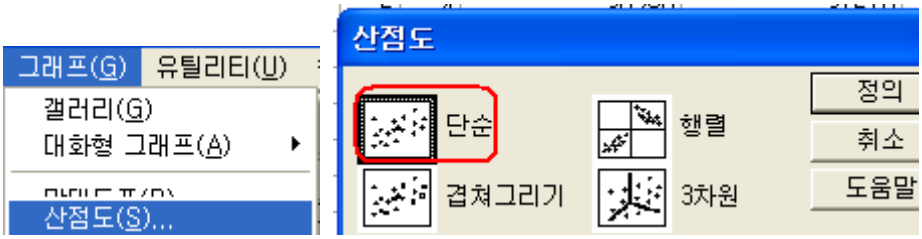




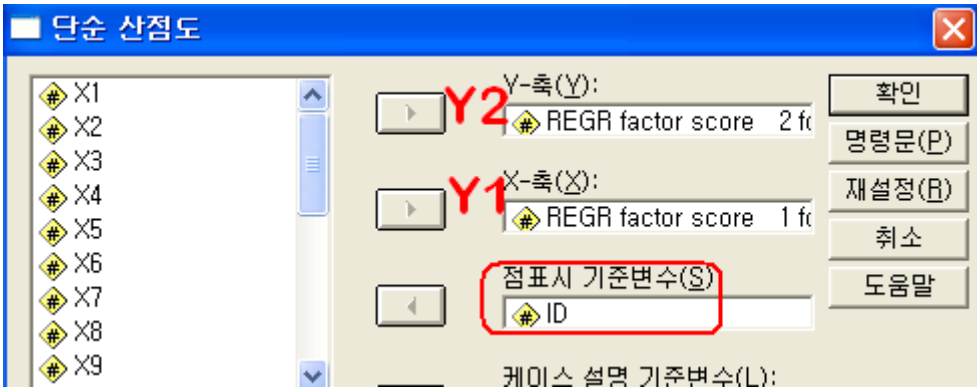
ID	X1	X2	X15	FAC1_1	FAC2_1	FAC3_1	AVG
40	10	6	10	1.69234	.85609	.57360	9.60
39	10	6	10	1.61581	.91339	.72387	9.47
8	9	9	10	1.35740	1.05203	.02561	9.00
23	7	10	8	1.30004	-.64651	.40179	8.60
7	9	9	10	1.17611	1.15442	.09432	8.60
22	9	8	8	1.27727	-.44525	.29810	8.53

▣이상치 발견

원 변수가 15 개나 되어 개체 중 이상치(산점도 행렬로 가끔 진단 가능하지만 거의 불가능)가 존재하는지 알 수 없었으나 이제 3 개의 주성분으로 축약하였으니 산점도를 그려 이상치 진단이 가능하다. 주성분이 3 개이므로 3 개의 산점도를 그려야 하지만, 여기서는 제일, 제이 주성분 산점도만 그리고 해석 방법을 설명하려고 한다.

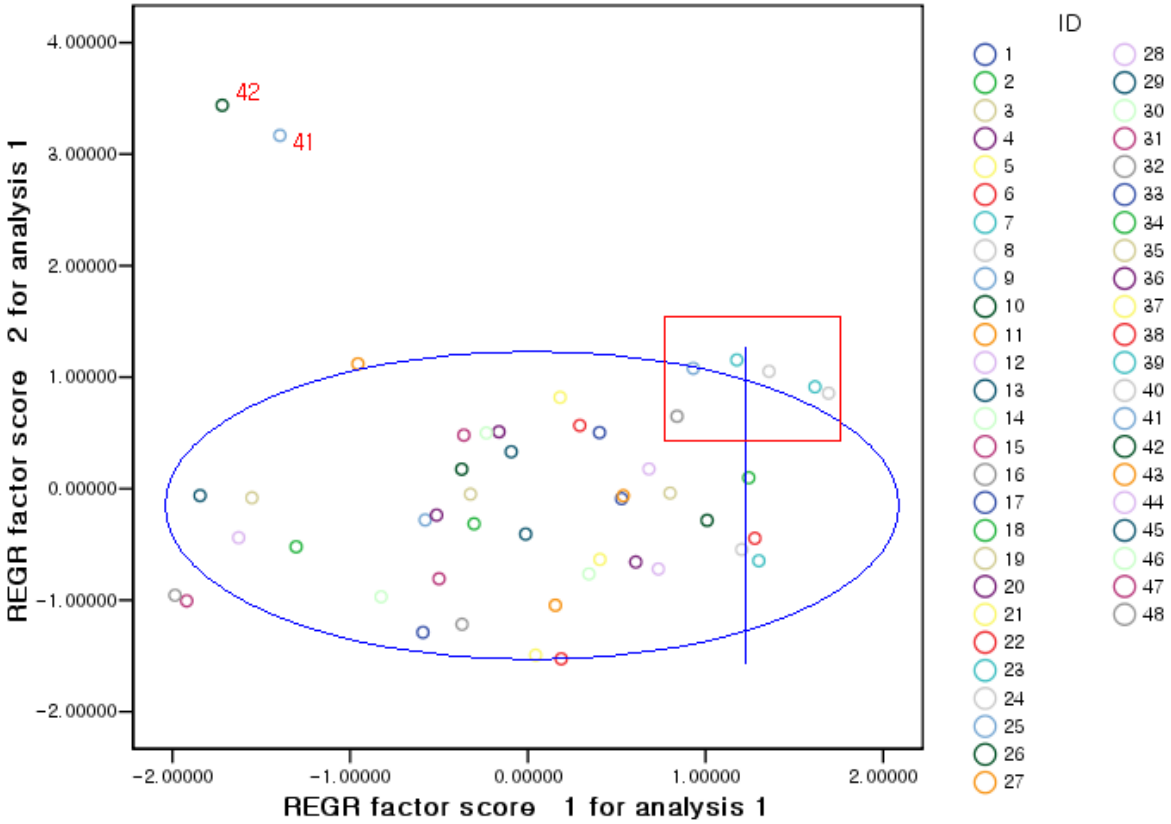


X-축을 제일 주성분, Y-축을 제이 주성분, 점들을 구별하는 변수로 ID 를 사용하였다.



산점도의 점들을 보면 제일 주성분의 설명력이 제이 주성분에 비해 크므로(고유치가 가장 크다) 타원의 현태가 나타난다. 제일 주성분(지적 능력)에 의해 지원자들 선발하면 수직선 오른편에 있는 6 명을 선발하면 된다. ID 41, 42 번 지원자는 제일 주성분은 낮고 제이 주성분은 매우 높아 이상치이다. 지적 능력은 낮으나 경험이 매우 많은 지원자이다.

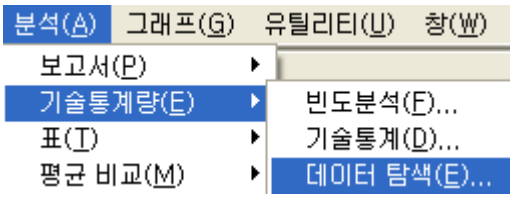
산점도는 개체를 분류하는데도 사용할 수 있으므로 다소 제한적이지만 주성분분석은 군집 분석에도 이용할 수 있다. 네모 상자 안의 개체들은 지적 능력과 경험 능력이 높은 그룹이다. 여기 속한 6 명을 뽑아도 될 것이다.

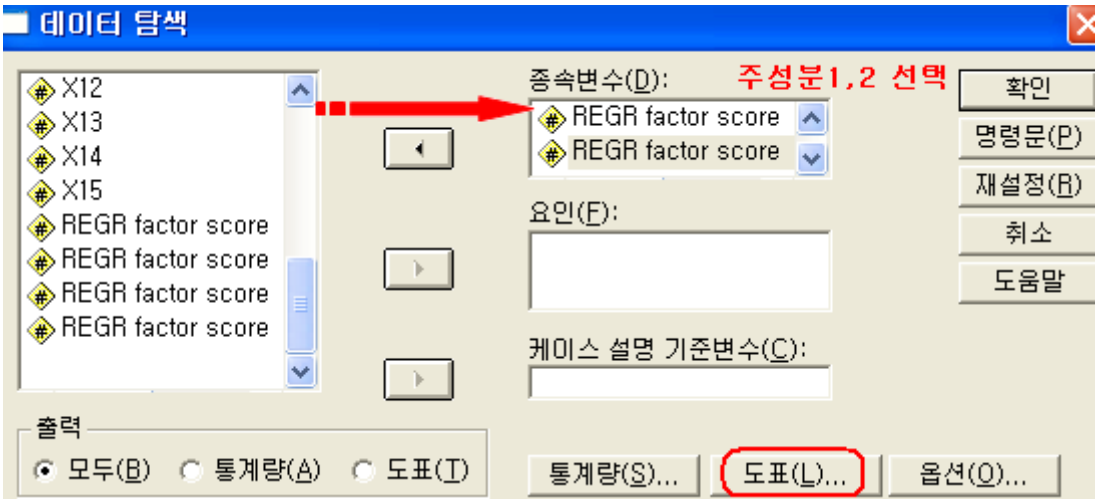


▣일변량 분석

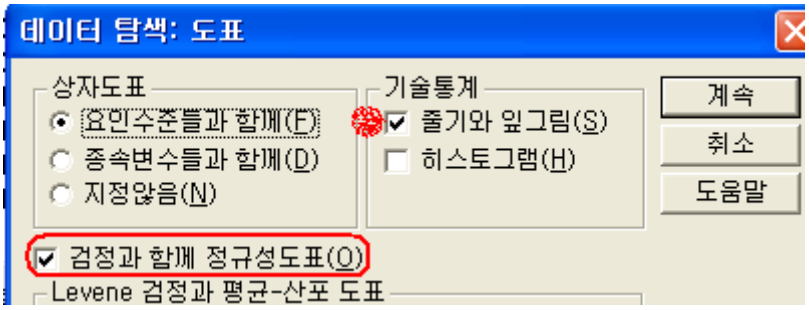
원 변수가 다변량 정규 분포(multivariate normal dist)를 따르면 주성분 변수는 일변량 정규분포를 따르므로 주성분 변수의 분포가 정규 분포이면 원 변수는 다변량 정규 분포임을 알 수 있다. (Why? 주성분은 원 변수의 선형 결합이므로 정규 분포를 따르는 변수의 합은 정규 분포이다) (정규분포 따르는지? Shapiro-Wilk 통계량) 또한 각 주성분 변수에 대한 상자-수염 그림(Box-whisker plot)을 그리면 이상치 존재 여부도 쉽게 파악 할 수 있다.

제일 주성분과 제이 주성분에 대한 일변량 분석을 실시해 보자.





줄기-잎 그림은 이상치 진단에 필요하고 “검정과 함께” 선택하면 정규성 검정 통계량이 출력된다.



정규성 검정은 K-S 나 S-W 어느 것을 사용해도 되나 S-W 을 주로 사용한다. 제일 주성분은 정규분포를 따르지만 제이 주성분은 정규분포를 따르지 않으므로 원 변수는 다변량 정규분포를 따른다고 할 수 없다. 근데 이 결과는 어디에 쓰나?

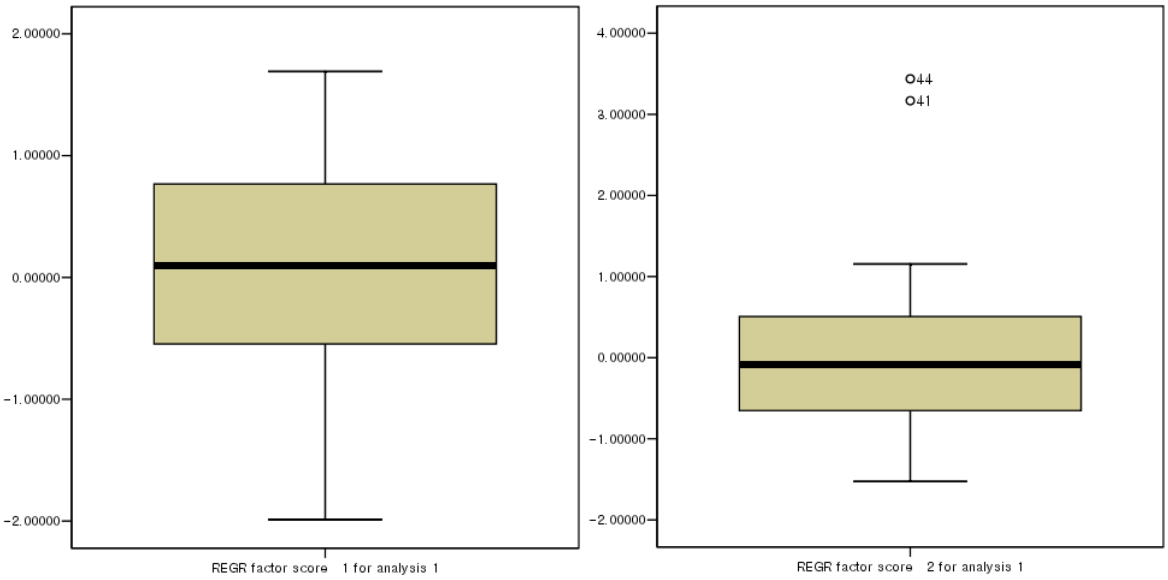
정규성 검정

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	통계량	자유도	유의확률	통계량	자유도	유의확률
REGR factor score 1 for analysis 1	.071	48	.200*	.962	48	.126
REGR factor score 2 for analysis 1	.120	48	.081	.897	48	.001

*. 이것은 참인 유의확률의 하한값입니다.

a. Lilliefors 유의확률 수정

예상대로 제일 주성분의 분산이 크다. 제이 주성분에서는 ID44, ID42 이상치가 있다. (제일, 제이 주성분 점수 산점도와 비교해 보라)



▣회귀분석에 이용

다중회귀($Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + e_i$) 모형에서 설명변수들간의 상관 관계가 크면 이로 인하여 추정된 회귀 계수의 분산이 커지고 부호가 바뀌는 문제가 발생한다. 다중공선성 발견 방법은 ①산점도 행렬이나 상관계수 크기 ②VIF(Variance Inflation Index, 분산팽창계수)와 Condition Index(상태지수)를 이용(책마다 차이는 있지만 대충 10 이상)하는 방법이 있다.

다중공선성의 해결 방법으로는 ①상관 관계가 높은 변수 제외 ②주성분분석 이용 ③능형 회귀(Ridge Regression) 등이 있다. 가장 선호되는 방법은 변수 제외 방법이다. 회귀분석에 주성분분석을 이용한다는 것은 주성분을 설명변수로 사용한다는 것이다. 원 설명변수

로부터 주성분을 구하고 이를 회귀 모형의 설명변수로 사용하는 것이다. 주성분은 서로 독립이므로 설명변수의 상관 관계로부터 발생하는 다중공선성 문제가 발생하지 않는다.

설명변수의 공분산행렬(설명변수들의 측정 단위가 다르면 상관행렬)로부터 주성분을 얻는다. 주성분이 새로운 설명변수로 이용되고 주성분 점수가 새로운 설명변수의 측정치가 된다. 새로운 회귀 모형은 다음과 같다.

$$Y_i = \beta_0 + \beta_1 P_{1i} + \beta_2 P_{2i} + \cdots + \beta_p P_{pi} + e_i$$

$P_1, P_2, ..., P_p$는 주성분 점수이다. 다음 두 조건이 만족할 때만 다중공선성 문제 해결로 주성분분석을 사용하기 바란다.

①다중공선성 문제를 발생하는 설명변수가 꼭 필요하다고 판단되거나 제외하면 설명변수 수가 너무 적은 경우

②주성분 이름을 부여하기 용이한 경우: 주성분에 대한 적절한 이름을 부여할 수 없으면 회귀분석 결과에 대한 해석이 어렵다.

③주성분을 회귀분석에 이용하는 경우 2~3개 주성분을 선택할 필요는 없다. 모든 주성분을 설명변수로 사용하고 변수 자동 선택 방법(stepwise, backward, forward)에 의해 선택하면 된다.

EXAMPLE 2-1 주성분 회귀분석에 사용

http://lib.stat.cmu.edu/DASL/Datafiles/carmpgdat.html 자동차 연비(MPG)에 영향을 미칠 것 같은 변수로 VOL(차 실내 공간), HP(마력), TS(최대 스피드), WEIGHT(차 중량)을 고려하여 데이터를 수집하였다. 표본의 크기는 30 이다. ■MPG.SAV■

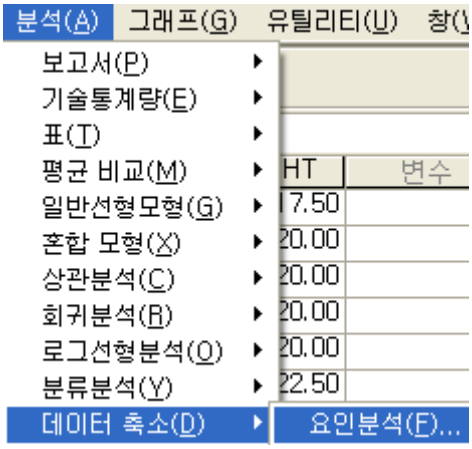
EXAMPLE1-17 의 회귀분석 결과를 이용하면 (HP, TS), (HP, WEIGHT), (TS, WEIGHT) 간의 상관 관계가 존재한다. 모든 설명변수를 고려한 회귀 모형 추정 결과는 다음과 같다. HP, TS 가 설명변수로 동시에 선택되었으므로 다중공선성 문제가 발생한다. 이로 인하여 HP 변수의 추정 회귀계수 부호가 양이다. 마력이 높을수록 연비가 높다?

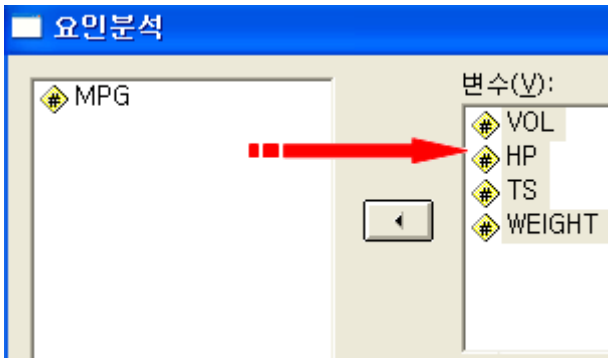
계수[a]

모형		비표준화 계수		표준화 계수	t	유의확률
		B	표준오차	베타		
1	(상수)	280,722	137,555		2,041	,052
	VOL	-,097	,045	-,230	-2,153	,041
	HP	,682	,706	1,391	,967	,343
	TS	-2,107	1,494	-1,538	-1,410	,171
	WEIGHT	-2,649	1,540	-1,088	-1,720	,098

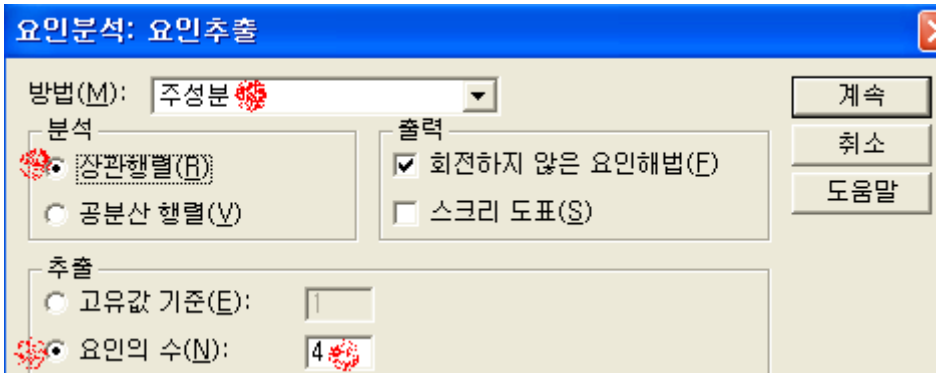
a. 종속변수: MPG

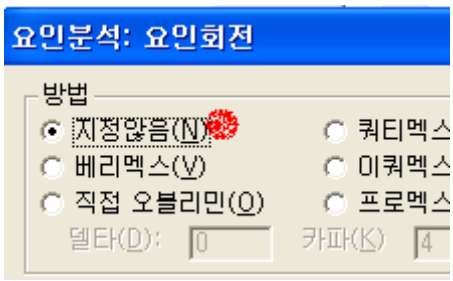
일반적으로 설명변수가 유의하고 다중공선성 문제가 발생하는 경우 주성분분석을 사용한다. 이 예제는 VOL, HP 2 개만 남고 다중공선성 문제는 발생하지 않으므로 주성분분석은 필요 없다. 그러나 TS, WEIGHT 도 연비에 영향을 미친다는 경험적 근거가 있다면 모든 변수를 고려하여 주성분을 얻어 이를 설명변수로 할 수 있다. 주성분은 원 변수의 선형 결합이기 때문이다.

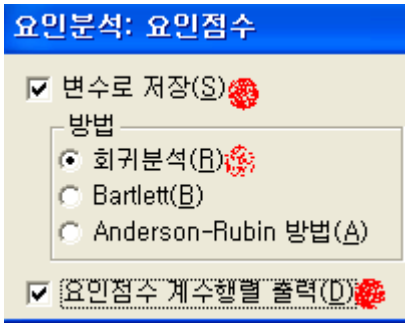




설명변수의 단위가 다르므로 상관행렬을 사용하였고 요인의 수는 설명변수의 개수만큼 설정해 주었다. 주성분 점수를 저장해 설명변수로 사용할 수 있도록 설정하였다.







주성분 점수를 설명변수로 하여 회귀분석을 실시해 보자. 처음에는 4 개 주성분 모두 설명변수 선택하고 유의한 변수 선택을 해야 한다. 제삼 주성분=>제사 주성분 순으로 유의하지 않아 제외하였다. (유의수준=0.15)

설명변수 제일 주성분, 제이 주성분은 어떤 의미가 있는가? 즉 주성분의 이름은? 성분점수 계수 행렬을 이용하면 된다. 주성분 1 은 이름 부여하기 다소 어렵다. 제이 주성분은 차 공간 변수라 할 수 있다. 이처럼 주성분 이름 부여는 다소 주관적이거나 불가능한 경우가 빈번히 발생한다. 그래서 주성분분석이 회귀분석의 다중공선성 문제 해결 방법으로 자주 사용되지 않는다.

성분점수 계수행렬

	성분			
	1	2	3	4
VOL	,109	,785	,754	,153
HP	,414	-,138	,066	-14,884
TS	,355	-,373	,679	11,257
WEIGHT	,333	,314	-1,052	6,446

회귀분석 결과를 보면 연비에 영향을 미치는 변수가 제일 주성분이다. 그래서 그게 뭔데? 마력, 차중량, 최대 속력이 섞인 변수?

계수[a]

모형		비표준화 계수		표준화 계수	t	유의확률
		B	표준오차	베타		
1	(상수)	41,090	,786		52,251	,000
	REGR factor score 1 for analysis 1	-7,135	,800	-,853	-8,920	,000
	REGR factor score 2 for analysis 1	-1,342	,800	-,160	-1,678	,105

a. 종속변수: MPG

2.4 주성분분석 예제 2 (상관행렬 사용)

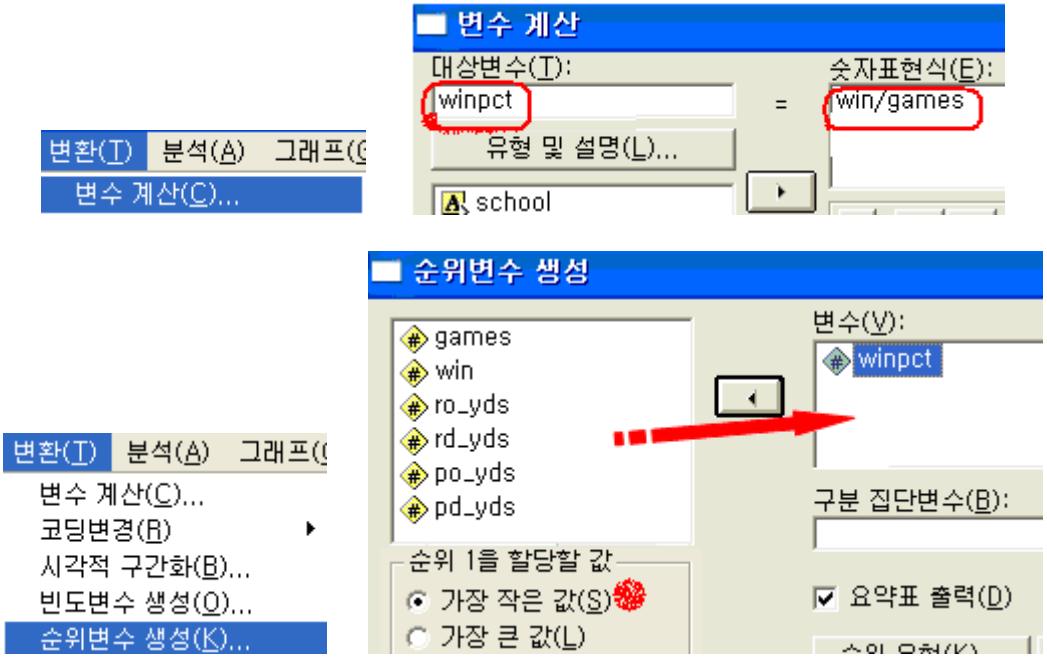
측정 변수들의 측정 단위가 차이가 나는 경우 원 변수의 분산 차이로 인해 발생하는 문제를 해결하기 위하여 공분산행렬에서 각 변수의 분산의 변동을 나누어 준 상관행렬로부터 고유치, 고유벡터를 구하게 된다.

EXAMPLE 2-2 상관행렬을 이용한 주성분분석

[http://lib.stat.cmu.edu/DASL]] . ■BIG8.SAV■ 학교명, 경기 수, 이긴 회수, Rushing 공격 야드(RO_YDS), Rushing 수비 야드, Passing 공격 야드, Passing 수비 야드,

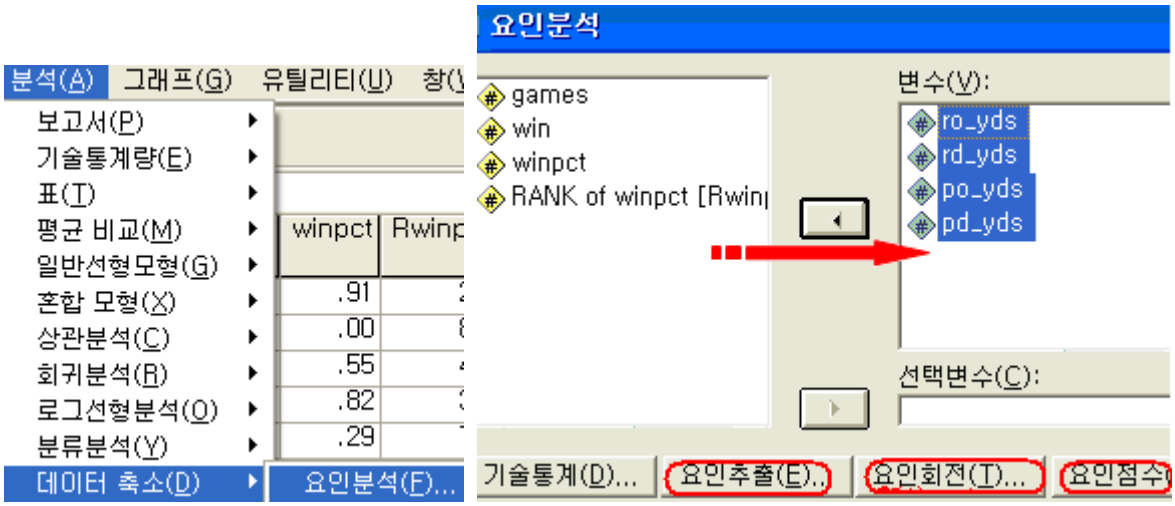
school	games	win	ro_yds	rd_yds	po_yds	pd_yds
Colorado	11	10.0	291.5	114.2	203.8	125.2
Iowa st.	11	.0	178.0	272.8	137.1	137.1
Kansas	11	6.0	247.1	171.2	140.9	135.8
Kansas st	11	9.0	125.6	167.5	237.6	94.3
Missouri	12	3.5	107.9	235.3	202.5	138.5
Nebraska	12	12.0	340.0	79.3	137.8	96.7
Oklahoma	11	6.0	182.2	148.5	173.9	107.5
Oklahoma	11	3.5	204.6	192.5	133.5	133.5

수비, 공격 야드 관련 4 개 변수에 의해 학교의 순위를 다시 매겨보고 승률에 의한 순위와 비교해 보자. 먼저 승률 변수를 계산하고 승률에 의해 순위를 부여하자. Nebraska => Colorado 순이다.

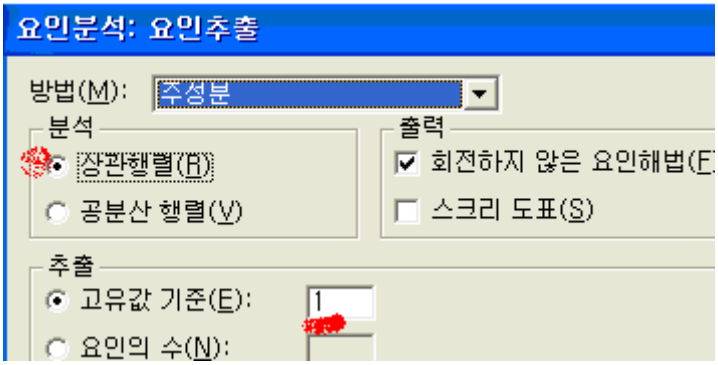


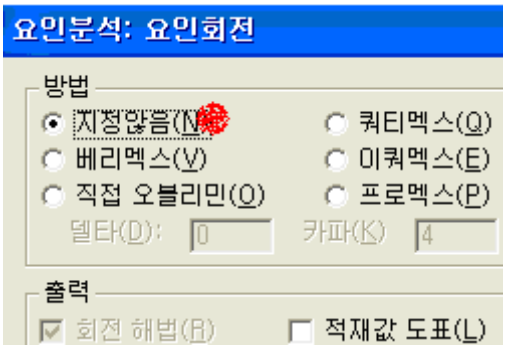
school	games	win	ro_yds	rd_yds	po_yds	pd_yds	winpct	Rwinpct
Colorado	11	10.0	292	114	204	125	.91	2.0
Iowa st.	11	.0	178	273	137	137	.00	8.0
Kansas	11	6.0	247	171	141	136	.55	4.5
Kansas st.	11	9.0	126	168	238	94.3	.82	3.0
Missouri	12	3.5	108	235	203	139	.29	7.0
Nebraska	12	12.0	340	79.3	138	96.7	1.00	1.0
Oklahoma	11	6.0	182	149	174	108	.55	4.5
Oklahoma st	11	3.5	205	193	134	134	.32	6.0

원 변수가 4 개이고 공격, 수비, 다른 공수 형태가 혼합되어 있어 이것의 평균으로 순위를 매기는 것은 문제가 있다. 4 개를 모두 고려한 순위는? 원 변수를 축약한 주성분을 이용하면 된다. 변수의 분산이 차이가 있으므로 상관행렬을 이용해야 한다.



상관행렬을 사용하였으므로 고유값이 1 이상인 고유치를 선택하면 적절한 주성분 개수가 선택된다.





상관행렬을 사용하였기 때문에 고유치의 합은 원 변수 개수인 4 이다. 주성분이 2 개이면 충분하다. 제일 주성분은 수비 능력으로 볼 수 있다. 제이 주성분은 패싱 공격 능력이라 할 수 있다. 허용한 야드가 적을수록 수비 능력이 높은 것이므로 제일 주성분이 작은 값일수록 순위 높은 팀이다. 제이 주성분에서 패싱 야드 계수가 음이므로 이것 역시 낮은 학교가 순위가 높을 것이다.

설명된 총분산

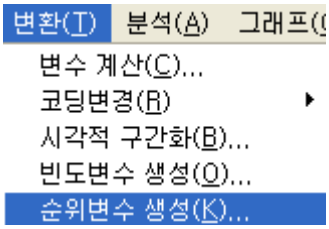
성분	초기 고유값			추출 제곱합 적재값		
	전체	% 분산	% 누적	전체	% 분산	% 누적
1	2.096	52.398	52.398	2.096	52.398	52.398
2	1.487	37.171	89.569	1.487	37.171	89.569
3	.362	9.062	98.631			
4	.055	1.369	100.000			

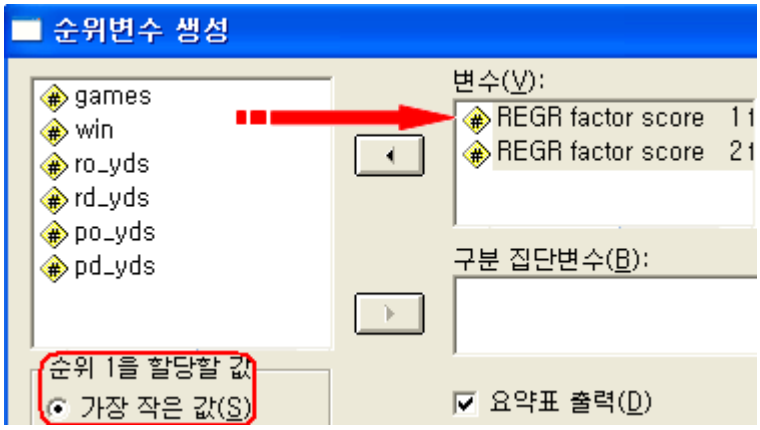
추출 방법: 주성분 분석.

성분점수 계수행렬

	성분	
	1	2
ro_yds	-.382	.367
rd_yds	.463	.058
po_yds	.024	-.627
pd_yds	.341	.376

football 성적을 수비능력과 패싱 공격 능력으로 순위를 매길 수 있다. 모두 낮을수록 순위가 높은 팀이다. 변환=>순위변수 생성을 이용하여 제일 주성분, 제이 주성분의 순위 변수를 만들자. "변수" 공간에 제일 주성분, 제이 주성분 모두 삽입하고 확인 버튼을 누른다.





제일 주성분의 순위와 승률은 어느 정도 일치하나 제이 주성분은 다소 차이가 있다. 제이 주성분은 패싱 공격 야드의 계수가 크지만 러싱 공격 야드, 패싱 수비 야드 계수도 패싱 공격 야드의 1/2 정도 되므로 여러 성분이 섞여 있어 승률과 다소 차이가 있다. 이처럼 주성분의 계수가 뚜렷이 구별되지 않거나 이름 부여가 애매한 경우에는 효율성이 감소된다. 이런 경우 원 변수의 축약으로 주성분을 만들기보다는 원 변수를 그룹화 하는 요인분석이 효과적이다.

school	gwa min	wi n	ro_	rd_	po_	pd_	win	Rwinpct	FAC1_1	FAC2_1	RFAC1_1	RFAC2_1
Colorado	*	*	*	*	*	*	1	2.0	-.733	-.113	2.0	4.0
Iowa st.	*	0	*	*	*	*	0	8.0	1.169	.803	7.0	6.0
Kansas	*	6	*	*	*	*	1	4.5	.060	.940	5.0	8.0
Kansas st.	*	9	*	*	*	*	1	3.0	-.084	-1.983	4.0	1.0
Missouri	*	4	*	*	*	*	0	7.0	1.292	-.559	8.0	2.0
Nebraska	*	*	*	*	*	*	1	1.0	-1.786	.546	1.0	5.0
Oklahoma	*	6	*	*	*	*	1	4.5	-.294	-.469	3.0	3.0
Oklahoma st	*	4	*	*	*	*	0	6.0	.376	.835	6.0	7.0

EXERCISE

경찰에 지원한 50 명의 신체적 특성 15 개를 측정하였다. ■🖫POLICE.SAV■

[Applied Multivariate Methods for Data Analysts, Dallas E. Johnson, 1998, p160]

ID: 지원자 번호/REACT: 시각적 자극에 대한 반응 시간/HEIGHT (cm) / WEIGHT (kg)

SHLDR: 어깨 넓이(cm) / PELVIC: 골반 넓이(cm) / CHEST: 가슴 넓이(cm)

THIGH: 허벅지 피부 두께 (mm) / PULSE: 맥박

DIAST: 심장 혈압 / CHNUP: 턱걸이 회수

BREATH: 폐활량 (liter) / RECVR: 런닝머신에서 제자리 달리고 5분 후 맥박

SPEED: 런닝머신에서 제자리 달리기 최대 속력

ENDUR: 런닝머신에서 달릴 수 있는 최대 시간(분) / FAT: 비만도

15 개 변수에 대한 주성분분석을 실시하여 변수를 축약하고 우수 지원자 10 명을 선발해 보자.