CHAPTER 4

주성분 분석

기성복 바지를 살 때 우리 몸의 치수를 모두 알아야 하는가? 그렇지 않다. 허리 둘레와 기장만 알고 있으면 충분하다. 여기에 PCA(Principle Component Analysis: 주성분 분석)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 하고 원 변수를 축약할 수 있는지 알아보는 것을 주성분 분석이라 한다.

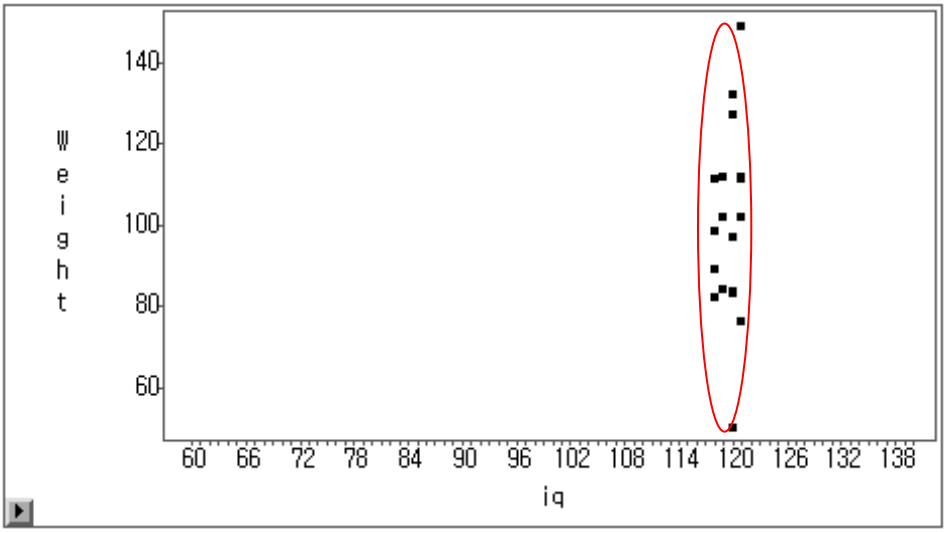
3 장 지원자 예제(지원자의 능력에 대해 15 개 항목 측정)에서 지원자의 능력을 나타내는 하나의 지표를 얻는 것은 쉬운 일이 아니다. 15 개 항목의 단순 평균이나 다양한 형태의 가중치를 이용한 가중 평균 등을 사용할 수 있지만 단점이 있음을 보았다. 기업의 재무 상황을 평가하기 위하여 유동 비율, 부채 비율, 자본 회전율, 자기 자본 비율 등 30 개 재무 관련 변수를 측정하였다고 하자. 30 개의 재무 변수를 이용하여 기업 평가하는 하나의 지표를 얻는 것은 쉬운 일이 아니다. 주성분 분석은 다변량 변수를 축약하여 1-2 개의 지표 변수를 만들 때 사용한다.

주성분 분석은 (1) $p \geq 3$ 인 변수를 1-2 개의 주성분 변수(많아도 3 개를 넘으면 의미가 없다.)로 줄이고 (2)새롭게 만들어진 주성분 변수에 대한 개념을 정의(이름 부여)하는데 주 목적이 있다. 주성분 변수는 p 개 원 변수들의 선형 결합에 의해 만들어지므로 p 개의 변수(차원) 정보가 1-2 개의 차원(주성분)만으로 표현되게 된다. 물론 모든 정보가 다 표현될 수 없어 어느 정도의 희생이 필요하다. 이를 80% 규칙이라 한다.

4.1. 맛보기

4.1.1. 예제

19 명 학생들의 몸무게(단위: pound)와 IQ 를 측정하였다고 하자. 아래 산점도를 보면 몸무게와는 달리 IQ 의 변동(통계학에서는 이를 정보의 개념으로 해석)은 거의 없다. 그러므로 IQ 는 학생 개체들을 구별하는 역할이 미미하나 몸무게 변수는 학생들의 정보를 대부분 가지고 있다.



다음은 위의 자료에 대한 주성분 분석 결과이다. 공분산 행렬로부터 고유치를 구하고 그에 대응하는 고유 벡터를 구했다.[2 장 SAS/IML 이용]

Simple Statistics

	Weight	iq
Mean	100.0263158	119.6842105
StD	22.7739335	1.1081833

Covariance Matrix

	Weight	iq
Weight	518.6520468	4.7309942
iq	4.7309942	1.2280702

Eigenvalues of the Covariance Matrix

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

Eigenvectors

	Prin1	Prin2
Weight	0.999958	-.009142
iq	0.009142	0.999958

몸무게 변수의 변동이 매우 크므로 고유치 $\lambda_1 = 518.69$ 은 다른 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(몸무게, IQ)의 정보(변동)를 대부분(99.77%) 표현할 수 있다.

4.1.2. 주성분 분석 사용

(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. 주성분 분석은 $p \geq 3$ 개 변수를 가진 데이터를 저 차원(1-2 차원) 그래프로 나타내어 개체들의 특성이나 이상치의 존재 여부를 알아보는 방법이다.

(2)군집

개체를 분류하는 경우 측정 항목(변수)이 1-2 개이면 평균 혹은 산점도에 의해 가능하다. 4.1.1 절 예제 경우 IQ 는 개체(사람) 간 차이가 거의 없으므로 몸무게 많은 그룹, 낮은 그룹으로 나눌 수 있다. 이처럼 2 개 변수까지는 산점도와 평균만으로 개체 분류가 가능하나 $p \geq 3$ 개인 다변량 데이터 개체를 산점도 만으로는 어려움이 있으므로 주성분에 의해 개체를 분류하거나 군집 분석 결과에 대한 해석으로 주성분 분석을 사용한다.

(3)판별 분석

판별식을 이용하여 개체를 분류하는 분석을 판별 분석이라 한다. 판별 분석의 경우 분산-공분산 행렬의 역 행렬을 구해야 하는데 측정 변수가 너무 많으면 계산이 오래 걸린다. 이런 경우 주성분 분석 방법에 의해 변수의 수를 줄여 만든 새로운 변수(주성분)에 의해 판별 분석을 하게 된다. 그러나 요즈음은 컴퓨터 성능의 발달로 역 행렬을 구하는데 문제가 없으므로 이 경우에는 주성분 분석을 거의 사용하지 않는다.

(4)회귀 분석

다중 회귀분석에서 설명 변수간의 상관 관계가 높으면 다중공선성(multicollinearity) 추정 회귀 계수의 분산이 커지므로 (변수들 간의 상관 관계가 높으면 $(X'X)^{-1}$ 계산 시 $|X'X| \approx 0$ 이 되므로 회귀 계수 추정치 분산 $s^2(\hat{b}) = MSE(X'X)^{-1}$ 이 매우 커진다) 최소 자승 추정치를 믿을 수 없게 된다. 이런 경우 ①문제가 되는 설명 변수를 제외하거나 ②능형 회귀분석(Ridge Regression: 추정치의 불편성을 희생하고 최소 분산을 갖는 추정치를 구하는 방법)을 이용하여 문제를 해결하거나 ③주성분 변수를 설명 변수로 이용하여 회귀 분석을 실시한다. 다음 절에 설명하겠지만 주성분 변수들은 서로 독립(상관 관계가 존재하지 않는다)이라는 성질이 있는데 이 성질을 이용하게 된다. 그러나 주성분 의미가 명확하게 해석되지 않는다면 회귀분석 결과 해석에 어려움이 있어 자주 사용되지는 않는다..

4.2. 주성분 구하기

다음 원칙에 의해 주성분(principal components)을 얻는다.

(1)주성분 변수 간에는 서로 상관 관계가 전혀 존재하지 않는다. (독립이다)

(2)첫 주성분은 데이터의 변동(분산, 정보)을 가장 많이 설명하고 계속 구해지는 2, 3, …번째 주성분은 자료의 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 주성분 분석이다.

참고

주성분은 원 변수의 선형 결합에 의해 구해지므로 새로운 변수이다. 그러므로 이 책에서는 주성분과 주성분 변수를 혼용해서 사용하겠다.

4.2.1. 주성분 정의

주성분 분석은 p 개의 원 변수의 선형 결합의 주성분 변수를 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 공분산 행렬 $\Sigma = \begin{bmatrix} \sigma_{11}\ \sigma_{12}\ \cdots \sigma_{1p} \\ \sigma_{21}\ \sigma_{22}\ \cdots \sigma_{2p} \\ \cdots \\ \sigma_{p1}\ \sigma_{p2}\ \cdots \sigma_{pp} \end{bmatrix}$ 를 갖는다고 하자.

공분산 행렬의 고유치를 $\lambda_1 \geq \lambda_2 . \geq .. \geq \lambda_p$ 라 하고 각 고유치(λ_i)에 대응하는 고유 벡터를 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.

변수 X_i 들의 변동의 합은 고유치의 합과 동일하다. 다음의 Λ 은 대각 원소가 Σ 의 고유치인 대각 행렬이고 P 는 고유치에 대응하는 고유 벡터로 구성된 직교 행렬이다.

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

그러므로 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

$\frac{\lambda_k}{\sum_{i=1}^{p}\lambda_i}$ 이다.

4.2.2. 주성분 구하기

(1)첫 번째 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)두 번째 주성분(second principal component)

$\underline{a}_2^{'}\underline{a}_2 = 1$, $\underline{a}_1^{'}\underline{a}_2 = 0$(첫 번째 주성분과 독립이다. 즉 첫 번째 주성분이 설명하는 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) 세 번째 주성분(third principal component)

$\underline{a}_3^{'}\underline{a}_3 = 1$, $\underline{a}_1^{'}\underline{a}_3 = 0$, $\underline{a}_2^{'}\underline{a}_3 = 0$ (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 변수의 개수만큼 존재하고 각 주성분은 서로 독립(겹치는 정보가 없다)이다. 각 주성분의 벡터가 아니라 하나의 변수임에 유의하기 바란다.

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

(4)주성분 계수 벡터 $\underline{a}_j$ 를 어떻게 구할 것인가?

변수 벡터의 분산-공분산 행렬 Σ 의 고유치 $\lambda_1 \geq \lambda_2 \geq .. \cdots \geq \lambda_p$ 에 대응하는 고유 벡터 $\underline{e}_1$, $\underline{e}_2$,..., $\underline{e}_p$ 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.

고유 벡터를 주성분 계수로 사용하는 경우 다음이 성립한다.

①주성분 y_j 의 분산은 고유치 λ_j 와 같다.

②분산-공분산 행렬 trace $tr(\Sigma)$ 은 원 변수 $x_1, x_2, ,..., x_p$ 변동(분산)의 합이다.

③즉 $tr(\Sigma) = \sum_i V(x_i) = \sigma_{11} + \sigma_{22} + ... + \sigma_{pp}$ 이므로 주성분 y_j 의 변수 변동의 설명력은

$\lambda_j / \sum_{j=1}^{p} \lambda_j$ 이다.

(5)주성분 추정

모집단의 분산-공분산 행렬 Σ 는 알 수 없으므로 이에 대한 추정치로 표본 분산-공분산 행렬 S 을 이용한다. 표본 분산-공분산 행렬 $S(=\hat{\Sigma})$ 로부터 고유치를 구하고 그에 대응하는 고유 벡터를 구하여 이를 주성분의 추정치로 사용하면 된다. (몸무게, IQ) 예제에서 공분산 행렬의 추정치 S 는 다음과 같다.

Covariance Matrix

	Weight	iq
Weight	518.6520468	4.7309942
iq	4.7309942	1.2280702

위의 공분산 행렬의 고유치, 고유 벡터는 다음과 같다.

	Eigenvectors	
Eigenvalue	Prin1	Prin2
518.695300	0.999958	-.009142
1.184817	0.009142	0.999958

원 변수들의 분산의 합(518.65+1.228)은 고유치의 합과 같다. 각 고유 벡터 원소의 제곱 합은 1 이고 $(-0.009142)^2+(0.999958)^2=1$ 각 고유 벡터 곱은 0 이다. 각 고유 벡터는 서로 독립임을 알 수 있다.

4.3. 주성분 점수 및 이름

4.3.1. 주성분 점수

주성분을 이용하여 각 개체에 대한 주성분 점수를 계산하고 이를 이용하여 개체를 좌표에 나타낼 수 있다. 이를 통하여 개체(데이터) 이상치 발견이 가능하고 주성분 변수간 관계를 파악할 수 있다. 다음은 r번째 개체의 j번째 주성분 점수이다.

$y_{rj}=\underline{e}_j'(\underline{x}_r-\mu)$, $\underline{x}_r$은 r-번째 개체의 측정치 ($r=1,2,...,n$)

Eigenvectors

	Prin1	Prin2
Weight	0.999958	-.009142
iq	0.009142	0.999958

학생 데이터 예제에서 (IQ, Weight) 변수에 대하여 모평균 벡터 $\underline{\mu}$의 추정치는 표본 평균 벡터이므로 $\hat{\underline{\mu}}=\bar{\underline{x}}=\begin{pmatrix}100.2\\119.68\end{pmatrix}$ 이다.

j 번째 학생(개체)의 몸무게가 110(pound)이고 IQ=125 였다면 j 번째 관측치의 첫 번째 주성분 점수는 $Y_{j1} = 0.999958(110 - 100.02)_j + 0.009142(125 - 119.68)_j$ 이고 두 번째 주성분 점수는 $Y_{j2} = -0.009142(110 - 100.02)_j + 0.999958(125 - 119.68)_j$ 이다.

주성분 점수는 수작업을 통해 계산하지 않아도 된다. SAS 에서는 OUT= 옵션을 이용하면 주성분 점수 데이터를 얻을 수 있다. (4.5 절 참고)

4.3.2. 주성분 부하 벡터

공분산 행렬의 고유 벡터 $\underline{e}_j$ 로부터 얻어지는 $\underline{c}_j = \sqrt{\lambda_j}\underline{e}_j$, $j = 1,2,\ldots, p$ 들을 성분 부하 벡터 (component loading vector)라 한다. 성분 부하 벡터는 주성분을 만들 때 사용되는 원 변수의 선형 계수에 해당하므로 주성분의 이름을 부여하는데 사용할 수 있다. 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 고유 벡터의 값만으로 비교해도 충분하다. 학생 데이터 예제에서의 정보 부하를 보면 첫 번째 주성분의 몸무게 변수는 0.99 이고, IQ 변수는 0.009 이므로 첫 번째 주성분은 신체 특성 변수로 이름을 붙일 수 있다. 같은 이유로 두 번째 주성분은 IQ 변수의 부하 값이 크므로 두 번째 주성분은 지적 특성 변수로 부를 수 있다. 이렇게 성분 부하 값으로 주성분 이름을 부여하는데 주의해야 할 것이 있다.

(1)성분 부하 값으로 주성분에 대한 원 변수의 영향력을 측정하므로 원 변수의 측정 단위는 유사해야 한다. 그러므로 원 변수의 측정 단위가 다른 경우에는 공분산 행렬을 이용하기 보다는 상관계수 행렬을 이용하여 고유치, 고유 벡터를 구하는 것이 바람직하다.(4.5 절 참고)

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

4.4. 주성분 개수

주성분 분석은 변수 차수를 줄이는데 목적이 있는데 실제 주성분의 개수는 원 변수의 수만큼 존재하게 된다. 즉 원 변수가 단지 서로 독립인 변수들로 변환된 것뿐이다. 사실 변수의 개수(차수)를 줄여서는 원 변수들이 가진 정보(변동)을 100% 표현할 수 없다. (몸무게, IQ) 예제처럼 IQ 변수의 변동이 아무리 없어도 2 차원 공간에 표현된 정보(산점도)를 1 차원 공간으로 줄인다면 어느 정도 정보의 희생은 각오해야 한다. 4.1 절의 산점도에서 오른쪽에서 불을 비춰 점들을 y 축에만 나타나게 한다면(2 차원 ► 1 차원) 몸무게가 115 파운드 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 변동의 설명을 희생해야만 한다.

주성분 개수가 원 변수의 개수와 동일하므로 원 변수의 변동을 어느 정도 설명(일반적으로 80%)하는 주성분만을 선택하여 2 차 분석을 실시하게 된다. 변동에 대한 설명 정도를 희생하여 차수를 줄이게 된다.

4.4.1. 총 변동 설명 비율

k 번째 주성분의 설명력은 $\lambda_k / \sum_{j=1}^{p} \lambda_j$ 이다. 공분산 행렬로부터 주성분을 구할 때 첫 번째 주성분의 설명력(λ_1)이 가장 크고, 두 번째 주성분의 설명력은 $\lambda_2, \cdots$ 이렇게 크기 순으로 주성분을 구하였다. 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 공분산 행렬 대신 상관 계수 행렬을 사용하는 경우 고유치 값이 1 이상인 주성분만 사용하면 총 변동의 80%정도를 설명하므로 고유치가 1 이상인 고유 벡터를 선형 계수로 사용하여 주성분 점수를 계산하면 된다.

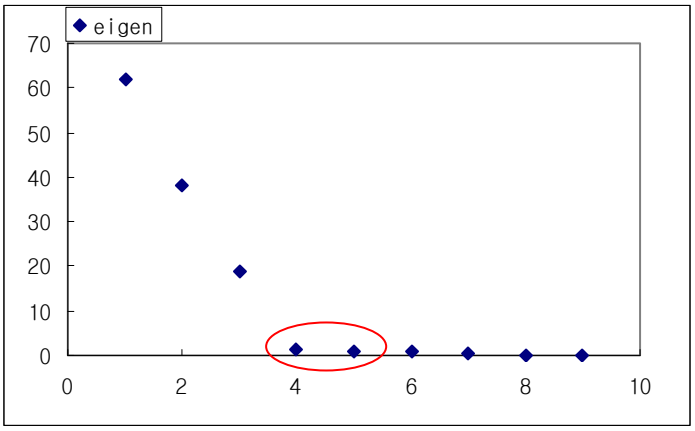
앞의 예제에서 첫 번째 고유치의 전체 설명 비율이 99.8%로 대부분 차지하므로 변수를 2 개를 하나의 주성분으로 줄일 수 있다. 첫 번째 주성분은 성분 부하 값에 의해 이름을 붙인 결과 신체 특성 변수였다. 주성분의 이름을 부여하는 것은 다소 주관적이다.

Eigenvalues of the Covariance Matrix

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

4.4.2. SCREE plot 사용

$(1, \hat{\lambda}_1)$, $(2, \hat{\lambda}_2)$, … 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. 4.5 절에서 설명하겠지만 누적 설명력이 80% 이상 되거나 고유치가 1 이상인 것만 선택하면 된다.

4.5. 상관 계수 행렬

측정 변수들의 측정 단위가 차이가 나는 경우 분산의 단위도 달라진다. 이로 인하여 측정 단위가 큰 변수에 대한 분산, 공분산이 커지므로 공분산 행렬을 이용하여 구한 고유치나 고유 벡터는 그 변수의 영향을 많이 받는다. 이런 문제점을 해결하기 위하여 공분산 행렬에서 각 변수의 분산의 변동을 나누어 준 상관 계수 행렬로부터 고유치, 고유벡터를 구하고 이를 이용하여 주성분을 구하면 된다.

주성분 분석은 변수의 변동(분산, 공분산)을 잘 설명하는 주성분을 찾는 것이 목적인데, 상관 계수 행렬을 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, 공분산 행렬을 이용하여 주성분 분석을 실시하면 된다.

다음은 모집단 상관 계수 행렬이다.

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

모집단 분산-공분산 행렬 Σ 의 추정치는 표본 분산-공분산 행렬 $\hat{\Sigma}=S$ 이고 모집단 상관 계수 행렬 R 의 추정치는 표본 상관계수 행렬 $\hat{R}$ 이다.

상관 계수 행렬로부터 주성분을 구하는 것은 표본 분산-공분산 행렬 S 로부터 고유치를 구하고 그에 대응하는 고유 벡터를 구하는 방법과 동일하다. 차이가 있다면

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

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

(3)주성분 변수의 변동 설명력은 다음과 같이 정리할 수 있다.

공분산 행렬 Σ	상관 계수 행렬 R
$\lambda_j / \sqrt{tr(S)} = \lambda_j / \sum_{i=1}^{p} \lambda_k$	λ_j / p

4.6. 예제

4.6.1. 변수가 2 개인 경우

지금까지 살펴보았던 개념들을 보다 쉽게 이해하기 위해 학생데이터를 예제로 하여 다음 작업을 실행하였다.

(1)IQ(y 축)와 몸무게(x 축)의 산점도를 그리시오.

(2)IQ 와 몸무게 공분산 행렬을 구하시오.

(3)공분산 행렬로부터 고유치를 구하시오. λ1=545.25 λ2=16.27

(4)첫 번째 주성분 계산을 위한 계수, 고유 벡터는 (0.0666, 0.998)이고 두 번째 주성분을 위한 고유 벡터는 (0.998, -0.0666)이다. 주성분 점수(변수) y1 을 y 축, y2 를 x 축으로 하여 산점도를 그리시오.

(5)(1)의 산점도와 (4)의 산점도를 비교 해석하시오.

◈SAS 자료 만들기 & 문제 (4)

```
data one;
input weight iq;
y1=0.0666*iq+0.998*weight;
y2=-0.0666*weight+0.998*iq;
cards;
123 110
```

고유 벡터를 이용하여 주성분 값(점수, score)을 구하였다. $(\underline{x}-\underline{\mu})$ 를 이용해야 하나 비교를 위해 측정 단위를 맞추려고 $\underline{x}$ 를 사용하였다.

weight	iq	y1	y2
123	110	130.080	101.588
74	115	81.511	109.842
145	125	153.035	115.093
64	112	71.331	107.514
84	118	91.691	112.170

◆문제 (2)

```
proc corr data=one cov;
    var weight iq;
    var y1 y2;
run;
```

주성분 y1 과 y2 는 상관 계수가 0 이다.

	weight	iq	y1	y2
weight	542.9052632	35.1578947	544.1609684	-1.0699116
iq	35.1578947	18.6184211	36.3275658	16.2396684
y1	544.1609684	36.3275658	545.4920624	0.0137902
y2	-1.0699116	16.2396684	0.0137902	16.2784452

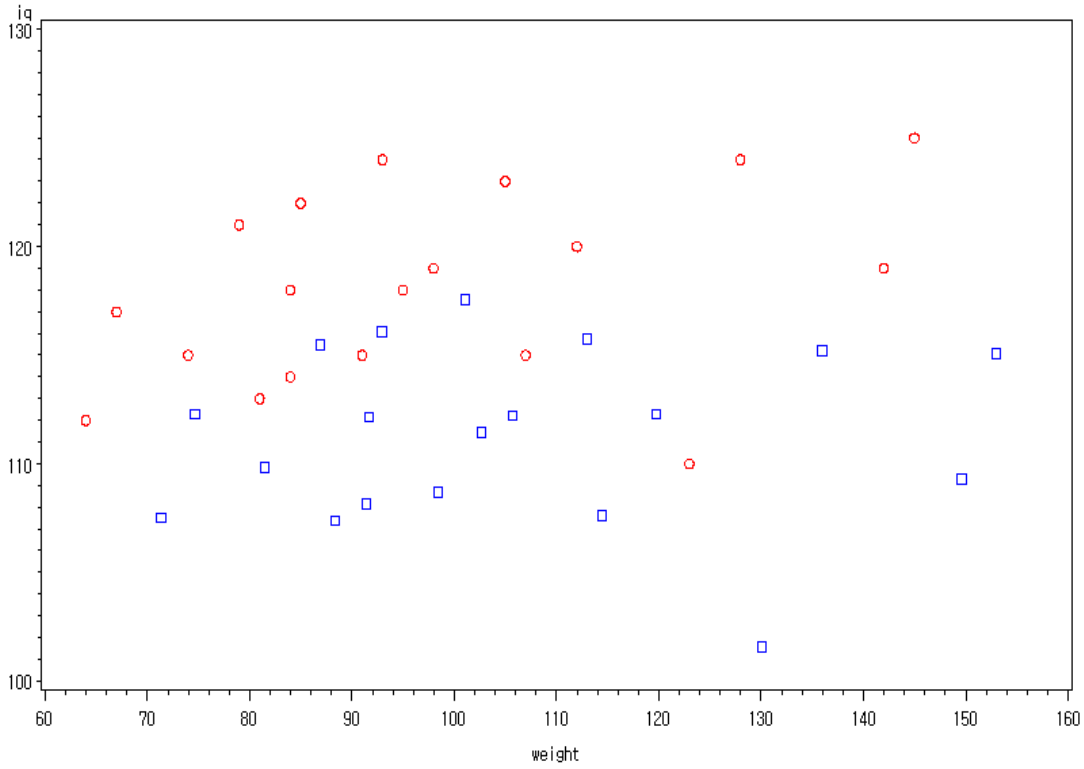
◆문제 (3)

```
proc iml;
    reset print all;
    x={542.9 35.16,35.16 18.62};
    call eigen(m,e,x);
run;
```

```
M       545.24744 E     0.9977787 -0.066616
        16.272561       0.0666162 0.9977787
```

◆문제 (1), (5)

```
proc gplot;
    symbol1 v=circle c=red;
    symbol2 v=square c=blue;
    plot iq*weight y2*y1/overlay;
run;
```



원은 원 변수의 자료, 네모는 주성분 값에 대한 산점도이다. 산점도의 형태는 크게 차이가 없어 보이나 첫 번째 주성분(신체 특성 변수)은 (몸무게, IQ)의 변동 중 545.25/(545.25+16.27)=97(%)를 설명하고 있다. 산점도에서 네모를 살펴보면 원에 비해 y 축(IQ)의 변동은 줄어들었으며 x 축(몸무게)의 변동은 커진 것을 알 수 있다.

4.6.2. 공분산 행렬 사용

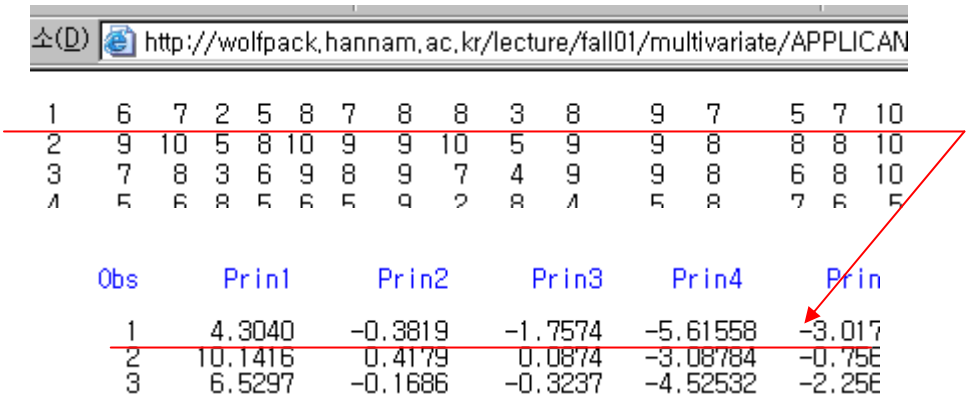
회사 지원자 48 명의 능력을 측정한 자료에 대해 주성분 분석을 시행해 보자. 이 예제에서 주성분 분석의 주목적은 측정 변수를 2~3 개의 주성분 변수로 줄여 그것을 이용하여 우수 지원자를 선택하는 것이다. [APPLICANT.txt]

15 개 항목이 서로 다른 능력을 측정하는 것이면 단순 평균을 사용하여 우수 지원자를 선발할 수 있다. 그러나 15 개 항목 중 6 개가 유사한 능력(보다 현실적 상황)을 측정한다면 그 능력이 뛰어난 지원자가 선발될 것이다. 주성분 분석은 15 개 항목을 1~2 개 항목(이를

주성분이라 한다)으로 줄여 지원자를 산점도에 나타냄으로써 우수 지원자를 선발할 수 있게 한다.

```
DATA APPLICANT;
      INFILE "C:\TEMP\APPLICANT.TXT";
      INPUT ID L AP AA LI SC LC HO SM EX DR AM GC PO KJ SU;
RUN;
PROC PRINCOMP DATA=APPLICANT OUT=SCORE COVARIANCE;
      VAR L--SU;
RUN;
PROC PRINT DATA=SCORE;
      VAR PRIN1-PRIN15;
RUN;
```

(1)OUT 옵션은 개체들의 주성분 점수를 SCORE 라는 SAS 자료에 저장한다. SAS 는 점수의 변수 명으로 PRIN1(첫 번째 주성분), PRIN2, …… 로 자동으로 설정한다. SCORE 라는 SAS 데이터에 주성분 점수가 저장되어 있다. 우선 SCORE 데이터를 출력해 보면 다음과 같다. 첫 번째 지원자의 15 개 항목 점수가 첫 번째 주성분 점수는 4.3040, 두 번째 점수 –0.3819, …… 이렇게 변환되었다. 측정 단위가 원 변수와 다른 것은 $(x-\mu)$가 사용되었기 때문이다.



(2)COVARIANCE(COV) 옵션은 분산-공분산 행렬을 이용하여 주성분 분석을 실시하라는 명령이다. Default 는 상관 계수 행렬(R)을 이용한 주성분 분석이다.

(3)VAR L--SU 는 모든 변수들 열거하는 대신 사용한다. 이것은 VAR L AP AA…… SU; 와 동일하다. –가 두 개 임에 유의하기 바란다.

◈출력 결과

우선 기초 통계량이 출력되고 그 아래 추정된 공분산 행렬이 출력된다. 공분산 행렬의 차수는 원 변수의 개수 p 개와 같다.

Simple Statistics

	l	ap	aa
Mean	6.000000000	7.083333333	7.083333333
StD	2.673749459	1.966023455	1.987549901

Covariance Matrix

	l	ap	aa	
l	7.14893617	1.25531915	0.23404255	2.2
ap	1.25531915	3.86524823	0.48226950	2.0
aa	0.23404255	0.48226950	3.95035461	0.0
l i	2.2978723A	2.09397163	0.00886525	7.8

$\hat{\sigma}_{11} = s_{11} = 7.15$, $\hat{\sigma}_{12} = \hat{\sigma}_{21} = s_{12} = s_{21} = 1.26$, $\ldots S = \hat{\Sigma}$

Total Variance 122.53501773

Eigenvalues of the Covariance Matrix

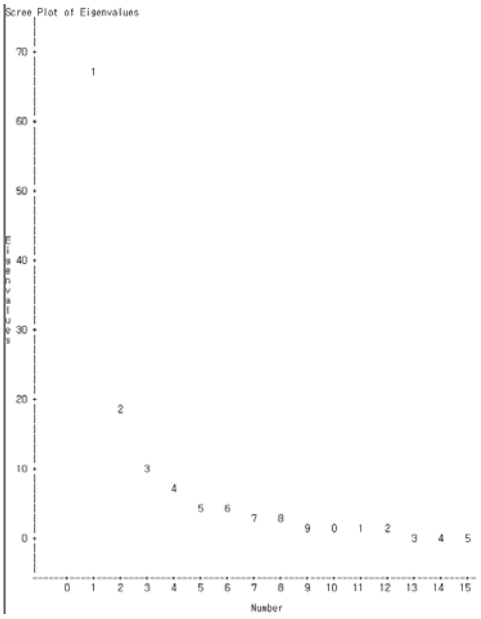
	Eigenvalue	Difference	Proportion	Cumulative
1	66.5364216	48.3558875	0.5430	0.5430
2	18.1805340	7.5895485	0.1484	0.6914
3	10.5909855	3.8230376	0.0864	0.7778
4	6.7679478	2.7823032	0.0552	0.8330
5	3.9856446	0.3579742	0.0325	0.8656
6	3.6276704	0.7119224	0.0296	0.8952
7	2.9157480	0.0802174	0.0238	0.9190
8	2.8355306	0.8804081	0.0231	0.9421
9	1.9551225	0.3416313	0.0160	0.9581
10	1.6134912	0.4770192	0.0132	0.9712
11	1.1364720	0.2636018	0.0093	0.9805
12	0.8728702	0.1661951	0.0071	0.9876
13	0.7066751	0.1981398	0.0058	0.9934
14	0.5085353	0.2071663	0.0042	0.9975
15	0.3013690		0.0025	1.0000

$\hat{\lambda}_1 = 66.5, \hat{\lambda}_2 = 18.2, \hat{\lambda}_3 = 10.6, \hat{\lambda}_4 = 6.8 \ldots$

(15x15) 공분산 행렬로부터 구한 고유치의 합(66.53+18.18+……+0.30)과 15 개 원 변수 총 변동 (122.53 = $\sum_{i=1}^{15} s_{ii} = 7.15 + 3.87 + 3.95 + ..$)와 합은 동일하다. 고유치와 주성분 변수의 설명력 비율($\lambda_i / \sum \lambda_i$)와 누적 설명력 비율이 출력된다. Difference 는 고유치와 다음 고유치 값의 차이다. 추정된 공분산 행렬(S)로부터 구한 첫번째 고유치는 66.53 이고 이 고유치는 전체 변동 중 54%(=66.54/122.54)를 설명하고 있다. 48.36 은 (66.54-18.18)에서 구해진다.

Cumulative = 주성분 설명력의 누적이다. 출력 결과에 의하면 4 개의 주성분이면 누적 설명력이 83%이므로 15 개의 변수가 4 개의 주성분으로 축약될 수 있다. 상관 계수 행렬을 이용하는 경우와는 달리 공분산 행렬을 이용하여 고유치를 구했으므로 설명력의 80%와 eigen-value 의 값이 1 이상인 주성분과는 일치되지 않는다. 원 변수가 10 점 리커드 척도이므로 1~2 개 주성분만으로는 변동의 80%를 설명하지는 못한다. 그러나 본 예제에서 주성분 분석의 목적은 우수 지원자 선발 이므로 주성분 두 개만(설명력 69%) 이용해도 될 것이다. 물론 첫 번째 주성분을 이용하여 첫 번째 주성분 점수가 가장 높은 6 명을 선발해도 된다.

```
PROC FACTOR DATA=APPLICANT SCREE COVARIANCE;
        VAR L--SU;
RUN;
```



SCREE plot 에 의하면 첫 번째 주성분 이후 두 번째 주성분부터 설명 비율은 현저히 떨어지고 첫 번째 주성분으로는 총 변동의 54% 밖에 설명하지 못한다. 이런 일이 발생하는 이유는 앞에서 언급했듯이 각 변수의 측정 단위가 1-10 점 점수인 리커드 척도 형태이기 때문이다. 실험실 자료나 사회 과학 데이터 변수인 경우에는 고유치 1~2 개 정도면 80% 정도를 설명한다.

다음은 각 고유치로부터 고유 벡터를 구하면 $\hat{a}_1 = e_1, \hat{a}_2 = e_2, ..., \hat{a}_{15} = e_{15}$ 는 다음과 같다.

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5
l	0.149129	0.371461	0.200481	-.277311	0.636939
ap	0.132250	-.029296	0.041918	0.134231	0.042210
aa	0.029611	0.101846	-.131030	0.603168	0.167474
li	0.203126	-.093042	0.619733	0.126399	0.053473
sc	0.231436	-.235740	-.189273	-.072088	-.025117
lc	0.336870	-.195978	-.124714	0.052788	0.231817
ho	0.120238	-.300549	0.447178	0.255587	-.334369
sm	0.379017	-.090010	-.281581	-.172303	-.177778
ex	0.164016	0.636212	0.025043	0.166245	-.191487
dr	0.316050	0.012486	-.113315	-.134844	-.338054
am	0.312106	-.122150	-.244517	-.147307	0.105416
gc	0.338764	-.074347	-.050497	0.206271	0.258316
po	0.357165	-.024920	0.041308	0.317232	0.108875
kj	0.226076	-.044837	0.385206	-.459715	-.026846
su	0.274483	0.470867	0.016815	-.015962	-.349972

r 번째 개체(학생)의 j 번째 주성분 점수의 추정치는 $y_{rj} = \underline{e}_j^{'}(\underline{x}_r - \mu)$, ($\underline{x}_r$ 은 r-번째 개체의 측정치)이다. 이 추정치 점수는 OUT=SCORE 옵션에 의해 SAS 데이터 score 에 저장되어 있으므로 PROC PRINT 하면 다음 출력을 얻는다. 첫 번째 지원자의 첫 번째 주성분 점수(4.304)를 구하는 식은

$$0.149*(L-\overline{L})+0.132*(AP-\overline{AP})+0.0296*(AA-\overline{AA})+...+0.2745*(SU-\overline{SU})$$

$$=0.149*(6-6)+0.132*(7-7.08)+0.0296*(2-7.08)+...+0.2745*(10-5.96)=0.4304$$

Obs	Prin1	Prin2	Prin3	Prin4	Prin5
1	4.3040	-0.3819	-1.7574	-5.61558	-3.01763
2	10.1416	0.4179	0.0874	-3.08784	-0.75680
3	6.5297	-0.1686	-0.3237	-4.52532	-2.25663
4	-1.3281	2.1759	1.0981	2.81460	-0.14716
5	1.4804	3.4916	3.2469	2.45018	-0.93626
6	2.3832	2.4176	1.0401	1.18088	-0.72581
7	9.5935	4.9223	0.3070	-0.09617	-0.19334

4.6.3. 상관 행렬 사용

측정 변수들의 측정 단위가 차이가 나는 경우 분산의 크기의 차이가 커지므로 이런 문제점을 해결하기 위하여 공분산 행렬에서 각 변수의 분산의 변동을 나누어 준 상관 계수 행렬로부터 고유치, 고유벡터를 구할 수 있다. 그러나 지원자 예제의 경우 변수의 측정 단위가 같으므로 상관 행렬을 이용할 필요가 없다. 앞에서 언급하였더라도 가능하면(측정 단위가 차이가 많이 나면 측정 단위를 바꾸어서라도) 공분산 행렬을 사용할 것을 권한다. 여기서는 사용 방법만을 살펴보기로 하자.

```
DATA APPLICANT;
      INFILE "D:\TEMP\APPLICANT.TXT";
      INPUT ID L AP AA LI SC LC HO SM EX DR AM GC PO KJ SU;
RUN;
PROC PRINCOMP DATA=APPLICANT OUT=SCORE1;
      VAR L--SU;
RUN;
PROC PRINT DATA=SCORE1;
      VAR PRIN1-PRIN15;
RUN;
```

Correlation Matrix

	l	ap	aa	li	s
l	1.0000	0.2388	0.0440	0.3063	0.092
ap	0.2388	1.0000	0.1234	0.3796	0.430
aa	0.0440	0.1234	1.0000	0.0016	0.001
li	0.3063	0.3796	0.0016	1.0000	0.302
sc	0.0921	0.4308	0.0011	0.3024	1.000

[추정된 상관 행렬]

상관 계수는 공분산 행렬에 의해서도 구해진다. (L, AP)의 상관 계수는 다음과 같다.

$$0.2388 = \frac{1.2553}{\sqrt{7.149}\sqrt{3.8652}}$$

상관 행렬로부터 구해진 고유치는 공분산 행렬(S)의 고유치와 다르다. 여기서는 고유치가 1 이상인 주성분(4 개)만 선택하면 80% 이상 설명한다는 사실과 일치한다. 상관 계수 행렬에서는 i 번째 주성분의 설명력 비율은 λ_i / p 이다. 이는 상관 계수 행렬의 대각 원소가 1 이기 때문이다.

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	7.51379418	5.45749301	0.5009	0.5009
2	2.05630117	0.60048169	0.1371	0.6380
3	1.45581948	0.25792178	0.0971	0.7351
4	1.19789771	0.45874509	0.0799	0.8149
5	0.73915262	0.24457355	0.0493	0.8642
6	0.49457907	0.14331724	0.0330	0.8972
7	0.35126183	0.04135981	0.0234	0.9206
8	0.30990202	0.05294047	0.0207	0.9412

고유치가 다르므로 고유 벡터도 다르다. 이 고유벡터는 성분 부하 값이므로 이를 이용하여 주성분의 이름을 부여하게 된다.

	Prin1	Prin2	Prin3
l	0.162440	0.428846	0.315375
ap	0.213108	-.035266	-.022878
aa	0.040184	0.236919	-.430470
li	0.225078	-.129796	0.465825
sc	0.290481	-.248896	-.241026
lc	0.314870	-.130990	-.150037
ho	0.158117	-.405450	0.283928
sm	0.324256	-.029492	-.185975
ex	0.134068	0.553139	0.082591

성분 부하 값이 다르므로 주성분 점수는 다르다. 다음은 상관 계수 행렬로부터 얻은 주성분 점수이다.

Obs	Prin1	Prin2	Prin3
1	1.29016	-0.48287	0.62912
2	3.49758	0.03013	0.59806
3	2.20620	-0.41185	0.81637
4	-0.49476	0.53592	0.07592
5	0.41889	0.98975	0.93321
6	0.80042	0.51814	0.35932
7	3.16806	1.63399	0.26072
8	3.72140	1.58604	0.00735
9	2.61145	1.38955	0.70842
10	2.43248	0.23884	-3.43444
11	1.13481	0.79704	-4.24543
12	1.82350	-0.36406	-2.92683

4.7. 주성분 해석하기

4.7.1. 주성분 이름 붙이기

앞 절에서 설명하였듯이 변수들의 측정 척도 단위가 유사하다면 (많이 다르지 않다면) 공분산 행렬로부터 고유치와 고유 벡터를 구한다. 고유 벡터를 계수로 하여 원 변수의 선형 결합인 주성분을 벡터로 표현하면 다음과 같다.

$$\text{주성분 변수 벡터 } \underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_p \end{bmatrix} = \begin{bmatrix} \underline{e}_1' \\ \underline{e}_2' \\ \cdots \\ \underline{e}_p' \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_p \end{bmatrix} = L\underline{x}$$

주성분 변수는 원 변수의 선형 결합 형태로 표시되고 주성분 분석의 경우 원 변수의 크기 단위는 비슷하므로 주성분 값의 크기는 선형 계수에 의해 결정된다. 즉 주성분의 크기는 원 변수에 곱해지는 계수(즉 고유 벡터)에 의해 결정되므로 주성분 계수(고유 벡터)를 이용해 주성분 변수의 이름을 부여할 수 있다. 지원자 예제에서 공분산 행렬로부터 구한 고유 벡터는 다음과 같다.

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5
l	0.149129	0.371461	0.200481	-.277311	0.636939
ap	0.132250	-.029296	0.041918	0.134231	0.042210
aa	0.029611	0.101846	-.131030	0.603168	0.167474
li	0.203126	-.093042	0.619733	0.126399	0.053473
sc	0.231436	-.235740	-.189273	-.072088	-.025117
lc	0.336870	-.195978	-.124714	0.052788	0.231817
ho	0.120238	-.300549	0.447178	0.255587	-.334369
sm	0.379017	-.090010	-.281581	-.172303	-.177778
ex	0.164016	0.636212	0.025043	0.166245	-.191487
dr	0.316050	0.012486	-.113315	-.134844	-.338054
am	0.312106	-.122150	-.244517	-.147307	0.105416
gc	0.338764	-.074347	-.050497	0.206271	0.258316
po	0.357165	-.024920	0.041308	0.317232	0.108875
kj	0.226076	-.044837	0.385206	-.459715	-.026846
su	0.274483	0.470867	0.016815	-.015962	-.349972

각각의 주성분 내에서 계수가 큰 변수들을 묶은 후 변수들이 함께 나타내는 지표(index)를 이용하여 이름을 부여하면 된다. 상당히 주관적이고 어려운 작업이다. 음의 의미는 그 주성분 계산 시 다른 변수와 반대로 작용한다는 것으로 절대 크기가 양의 부호 계수 큰 것과 유사하다면 함께 고려되어야 한다. 앞에서 살펴보았듯이 적절한 주성분 개수(전체 변동 중 약 80% 설명)는 4 개였다.

- 1 번째 주성분의 계수 크기에 의하면 LC(명석), SM(판매능력), DR(돌파력), AM(야망), GC(개념파악), PO(잠재력) 변수의 크기가 크므로 제 1 주성분은 정신적&지적 능력으로 할 수 있다.
- 2 번째 주성분에 의하면 EX(경험), SU(적합)이 큰 역할을 하므로 경험 주성분이라 할 수 있다.
- 3 번째 주성분에 의하면 LI(호감), HO(진실), KJ(사교)의 크기가 크므로 심성 변수로 이름하면 된다.
- 4 번째 주성분에 의하면 AA(성적), KJ(사교적)의 계수 크기가 크다. 성적 변수와 사교적 변수 간에는 반대 개념이 된 지표? 그래도 학교 성적 주성분이라 하는 것이 적당.

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.

4.7.2. 주성분 계수 나타내기

주성분을 구하는데 사용되는 계수를 주성분에 따라 산점도를 그리면 계수의 크기에 따라 원 변수를 분류하고 주성분에 적절한 이름을 붙이는데 편리할 것이다. 이 산점도는 주성분 이름 부여 시 편리하게 이용될 수 있다. 우선 첫 번째 주성분(Prin1)과 두 번째 주성분(Prin2)의 계수에 대한 산점도를 그려보자.

주성분 분석을 실시하여 통계량 값들을 SAS data OUT1 에 저장한다. [OUTSTAT=OUT1] 앞에서 설명하였듯이 OUT 옵션은 주성분 점수 값을 저장한 것이다.

```
PROC PRINCOMP DATA=APPLICANT OUT=SCORE OUTSTAT=OUT1 COVARIANCE;
    VAR L--SU;
RUN;
PROC PRINT DATA=OUT1;
RUN;
```

다음은 OUT1 의 SAS 데이터를 출력한 것이다.

Obs	_TYPE_	_NAME_	l	ap
1	MEAN		6.0000	7.0833
2	N		48.0000	48.0000
3	COV	l	7.1489	1.2553
18	EIGENVAL		66.5364	18.1805
19	SCORE	Prin1	0.1491	0.1323
20	SCORE	Prin2	0.3715	-0.0293

우리가 필요한 것은 고유 벡터(계수)이므로 _TYPE_ 변수가 SCORE 인 것만 남겨두면 된다. 그리고 산점도를 그리려면 데이터를 전치(transpose)해야 한다.

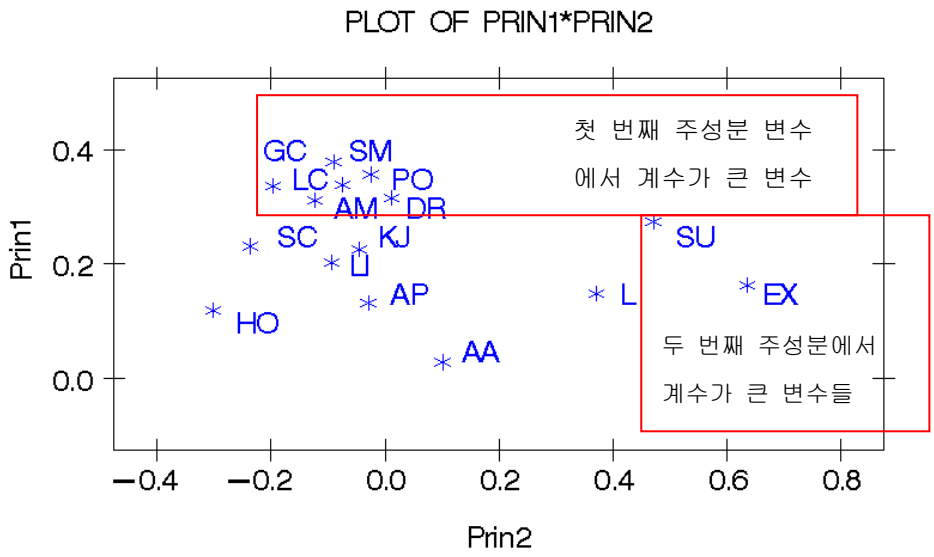
```
DATA OUT1;
    SET OUT1;
    IF (_TYPE_="SCORE");
RUN;

PROC TRANSPOSE DATA=OUT1 OUT=OUT2;
RUN;

PROC PRINT DATA=OUT2;
RUN;
```

Obs	_NAME_	Prin1	Prin2
1	l	0.14913	0.37146
2	ap	0.13225	-0.02930
3	aa	0.02961	0.10185
4	li	0.20313	-0.09304
5	sc	0.23144	-0.23574
6	lc	0.33687	-0.19598
7	ho	0.12024	-0.30055
8	sm	0.37902	-0.09001
9	ex	0.16402	0.63621
10	dr	0.31605	0.01249
11	am	0.31211	-0.12215
12	gc	0.33876	-0.07435
13	po	0.35716	-0.02492
14	kj	0.22608	-0.04484
15	su	0.27448	0.47087

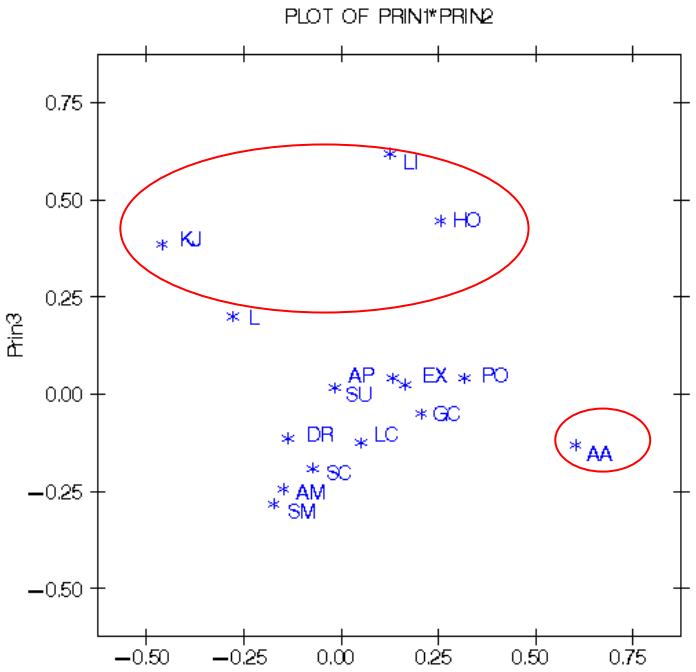
이는 주성분 부하 값과 동일하며 4.6.2 절의 PRINCOMP 출력결과와 동일하다.

이제 Prin1 과 Prin2 를 _NMAE_ 변수를 ID 로 하여 산점도를 그리면 된다. SAS 에는 산점도를 그리는 PLOTIT 이라는 MACRO 함수가 정의되어 있다.

```
TITLE "PLOT OF PRIN1*PRIN2";
%PLOTIT(DATA=OUT2, LABELVAR=_NAME_, PLOTVARS=PRIN1 PRIN2,
        COLOR=BLACK, COLORS=BLUE);
```



첫 번째 주성분, 두 번째 주성분 이름을 부여하는데 이 그림이 사용된다. 위 그래프는 페이지 72 고유 벡터의 계수 큰 것을 네모 친 결과와 일치한다. 이 산점도는 시각적 효과로 인하여 주성분의 적절한 이름을 부여하는데 도움을 준다. 3 번째 4 번째 주성분에 대해서도 같은 방법을 사용하면 된다. 즉 %PLOTIT 에서 PRIN3 PRIN4 로 바꾸어 주기만 하면 아래의 산점도를 얻을 것이다.



4.7.3. 주성분 점수 이용하기

(1)이상치 발견

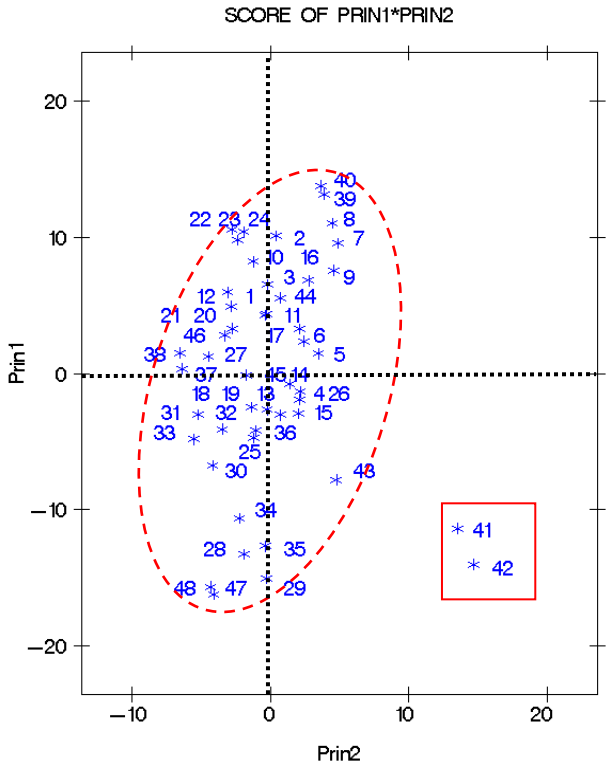
주성분 점수를 이용하여 산점도를 그리면 점들의 분포 형태는 타원의 형태를 띄며 타원의 길이는 고유치의 크기에 비례한다. (2.7.4 절 참고) 산점도에서 타원으로부터 떨어진 개체가 이상치가 된다. 주성분 점수는 OUT=SCORE 옵션에 의해 주성분 점수가 저장되어 있다.

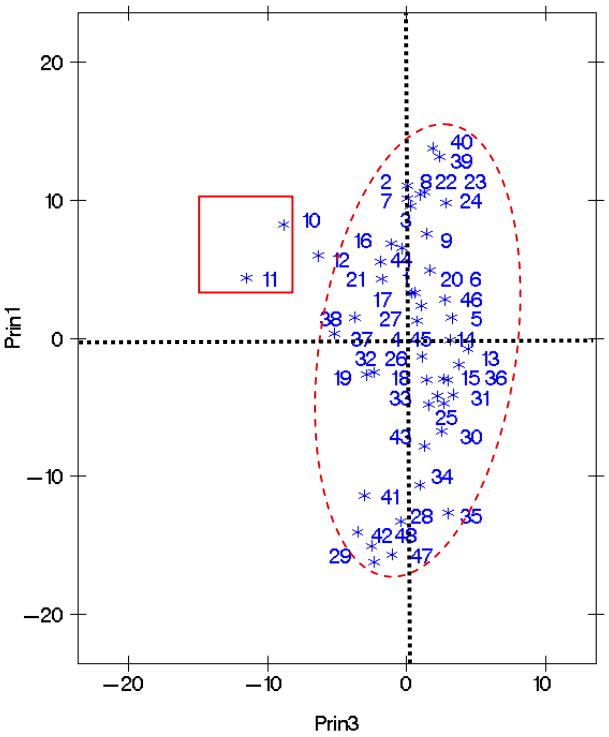
```
PROC PRINT DATA=SCORE;
    VAR ID PRIN1-PRIN15;
RUN;
```

Obs	id	Prin1	Prin2
1	1	4.3040	-0.3819
2	2	10.1416	0.4179
3	3	6.5297	-0.1686
4	4	-1.3281	2.1759
5	5	1.4804	3.4916
6	6	2.3832	2.4176
7	7	9.5935	4.9223

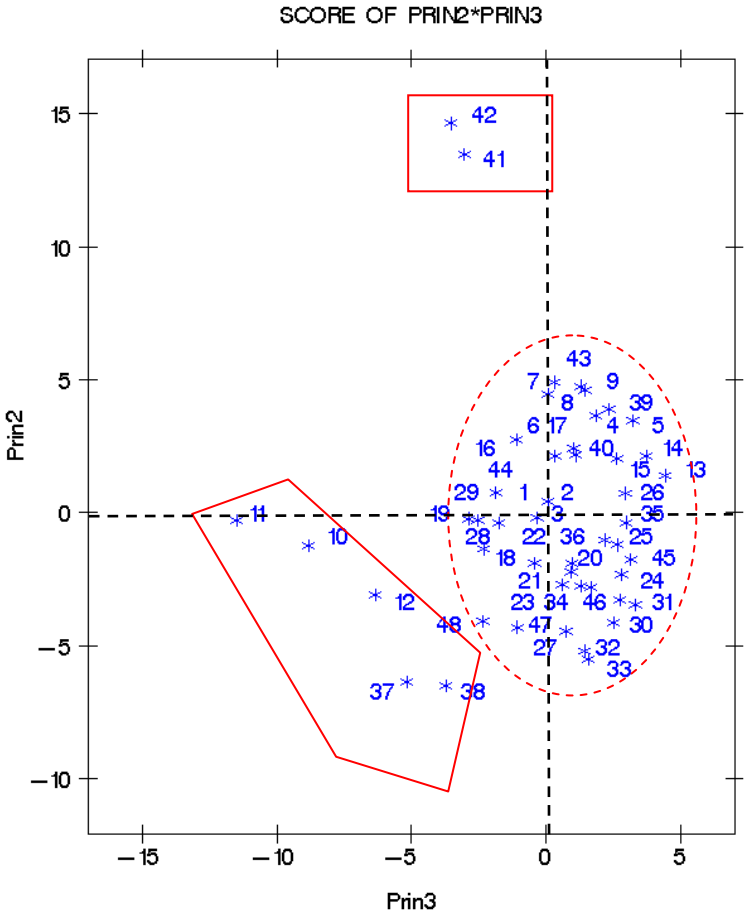
앞에서와 같이 PLOTIT macro 함수를 이용하여 산점도를 그리면 된다. DATA=, LABELID=, PLOTVARS= 의 이름을 적절히 바꾸어 주면 된다.

```
TITLE "SCORE OF PRIN1*PRIN2";
%PLOTIT(DATA=SCORE, LABELVAR=ID, PLOTVARS=PRIN1 PRIN2,
        COLOR=BLACK, COLORS=BLUE);
TITLE "SCORE OF PRIN1*PRIN3";
%PLOTIT(DATA=SCORE, LABELVAR=ID, PLOTVARS=PRIN1 PRIN3,
        COLOR=BLACK, COLORS=BLUE);
TITLE "SCORE OF PRIN2*PRIN3";
%PLOTIT(DATA=SCORE, LABELVAR=ID, PLOTVARS=PRIN2 PRIN3,
        COLOR=BLACK, COLORS=BLUE);
```


SCORE OF PRIN1*PRIN2
Prin1
Prin2
20
10
0
−10
−20
−10
0
10
20


Prin1
Prin3
20
10
0
−10
−20
−20
−10
0
10

까만 점선에 의해 개체를 4 분류로 나눌 수 있다. 1 사분면(요인 1 이 높고 요인 2 도 높은 그룹), 2 사분면(요인 1 이 높고 요인 2 도 낮은 그룹), 3 사분면, 4 사분면으로 개체를 나눌 수 있다. 주성분 점수를 구할 때 평균을 빼주므로 주성분 값을 0 을 중심으로 2 개로 나누면 된다. 만약 주성분이 하나로 충분하면(80% 정도) 요인 1 에 의해 높은 그룹(0 이상)과 낮은 그룹(0 미만)으로 나누면 된다.



3 개의 주성분 점수 산점도로부터 (10, 11, 12, 37, 38) 지원자는 순한(mild) 이상치이고 (41, 42) 개체는 극심한 (타원에서 많이 떨어짐)한 이상치이다. (41, 42) 지원자는 다른 지원자에 비해 두 번째 주성분(경험) 점수가 높다. 즉 경험이 많은 지원자이다. (10, 11, 12, 37, 38) 지원자들은 다른 지원자에 비해 심성 점수가 유난히 낮음을 알 수 있다.

(2)개체 분류

정신적&지적 능력(요인 1)에 의해 지원자 중 6 명을 뽑는다면 결과는 다음과 같다. 그렇다고 PRIN1 에 다른 항목의 점수가 반영되지 않은 것은 아니다. 주성분은 모든 원 변수의 선형 결합으로 만들어졌으므로 선형 계수 값이 작을지라도 영향은 존재한다.

```
PROC SORT DATA=SCORE;
     BY DESCENDING PRIN1;
RUN;
PROC PRINT DATA=SCORE;
     VAR ID PRIN1;
RUN;
```

Obs	id	Prin1
1	40	13.8044
2	39	13.1802
3	8	11.0723
4	23	10.6044
5	22	10.4187
6	2	10.1416

경험(요인 2) 우선하여 지원자 중 6 명을 뽑는다면 결과는 다음과 같다.

```
PROC SORT DATA=SCORE;
     BY DESCENDING PRIN2;
RUN;
PROC PRINT DATA=SCORE;
     VAR ID PRIN2;
RUN;
```

Obs	id	Prin2
1	42	14.6597
2	41	13.5076
3	7	4.9223
4	43	4.7752
5	9	4.6013
6	8	4.4857

심성이 고운 사람(요인 3)을 뽑는다면 결과는 다음과 같다.

```
PROC SORT DATA=SCORE;
    BY DESCENDING PRIN3;
RUN;
PROC PRINT DATA=SCORE;
    VAR ID PRIN3;
RUN;
```

Obs	id	Prin3
1	13	4.4261
2	14	3.7461
3	31	3.3333
4	5	3.2469
5	45	3.1581
6	35	2.9770

어느 점수로 사용하느냐에 따라 합격한 6 명이 다르다. 만약 주성분 3 개 점수의 합을 사용한다면 다음 결과를 얻는다.

```
DATA SCORE1;
    SET SCORE;
    PRIN=MEAN(PRIN1, PRIN2, PRIN3);
RUN;
PROC SORT DATA=SCORE1;
    BY DESCENDING PRIN;
RUN;
PROC PRINT DATA=SCORE1;
    VAR ID PRIN;
RUN;
```

Obs	id	PRIN
1	39	6.47683
2	40	6.44045
3	8	5.21379
4	7	4.94092
5	9	4.55193
6	2	3.54899

4.8. 주성분 분석 이용

주성분 분석은 원 변수를 축약한 새로운 변수(주성분)를 이용하여 변수의 개수를 줄이는 분석 방법이다. 주성분 변수 벡터를 $\underline{y}$, 원 변수 벡터를 $\underline{x}$라 하면 $\underline{y} = L\underline{x}$이다. 주성분 변수는 원 변수의 선형 결합으로 구해진다. 1)원 변수가 가진 정보는 어떻게 표현할 것인가? 정보는 변동으로 표현되므로 공분산 행렬을 이용하면 된다. 2)계수 행렬(L)은 어떻게 구할 것인가? 공분산 행렬로부터 고유치를 구하고 크기 순으로 정렬하고(고유치의 크기가 각 주성분 변수의 총 변동의 설명력) 각 고유치에 대해 $e_i'e_i = 1, e_i'e_j = 0$을 만족하는 고유 벡터를 구하여 이를 계수로 사용한다. 3)주성분 변수의 개수 결정? 원 변수 총 변동의 80% 정도를 설명하는(상관 행렬을 사용하는 경우는 고유치가 1 이상인 것만 택하면 된다) 고유치까지 택하면 된다. 고유치의 크기가 큰 주성분은 총 변동 설명력이 크다. (공분산 행렬 사용 $\lambda_i / \sum_{i=1}^{p} \lambda_i$, 상관 계수 행렬 사용 λ_i / p)

(1)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? Normal plot, Shapiro-Wilks W statistic) 또한 각 주성분 변수에 대한 상자-수염 그림(Box-whisker plot)을 그리면 이상치 존재 여부도 쉽게 파악 할 수 있다.

다음은 지원자 예제에서 얻는 4 개의 주성분 변수에 대한 일변량 분석 방법과 결과 해석이다. 앞에서 살펴본 것처럼 주성분 변수는 OUT 옵션에 의해 SAS data 에 저장할 수 있으므로 주성분 결과를 SCORE 라는 곳에 저장하였다. PROC UNIVARIATE 를 이용하여 정규 분포 가설 검정 통계량과 줄기 잎 그림(STEM-LEAF PLOT), 상자-수염 그림(BOX-WHISKER PLOT)을 그려보자.

```
PROC PRINCOMP DATA=APPLICANT OUT=SCORE COVARIANCE;
    VAR L--SU;
RUN;

PROC UNIVARIATE DATA=SCORE NORMAL PLOT;
    VAR PRIN1-PRIN4;
RUN;
```

먼저 정규 분포 가설을 검정하는 W-통계량 출력 결과를 살펴보자.

변수: Prin1
정규성 검정

검정		----통계량----		-------p값--------
Shapiro-Wilk	W	0.962353	Pr < W	0.1257
Kolmogorov-Smirnov	D	0.070579	Pr > D	>0.1500

변수: Prin2

검정		----통계량----		-------p값--------
Shapiro-Wilk	W	0.897215	Pr < W	0.0005
Kolmogorov-Smirnov	D	0.119941	Pr > D	0.0832

변수: Prin3

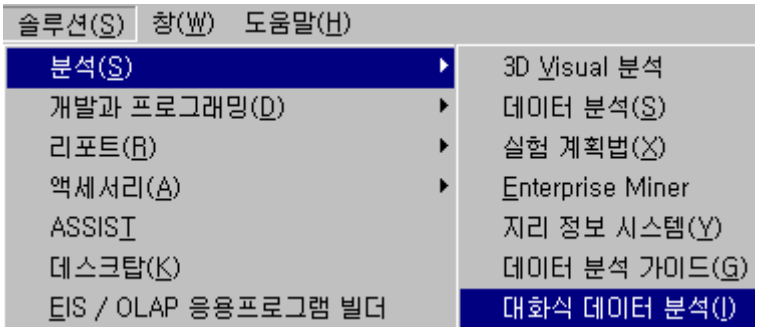
검정		----통계량----		-------p값--------
Shapiro-Wilk	W	0.87729	Pr < W	0.0001
Kolmogorov-Smirnov	D	0.156048	Pr > D	<0.0100

변수: Prin4

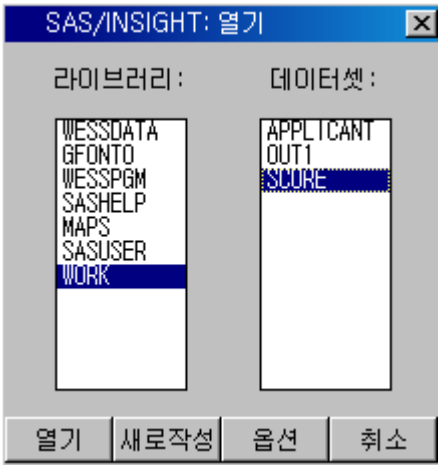
검정		----통계량----		-------p값--------
Shapiro-Wilk	W	0.972061	Pr < W	0.3041
Kolmogorov-Smirnov	D	0.070276	Pr > D	>0.1500

요인 1 과 요인 4 는 정규 분포를 따르지만 요인 2, 요인 3 는 정규 분포를 따르지 않는다. 그러므로 원 변수들은 다변량 정규 분포를 따르지는 않을 것이다.

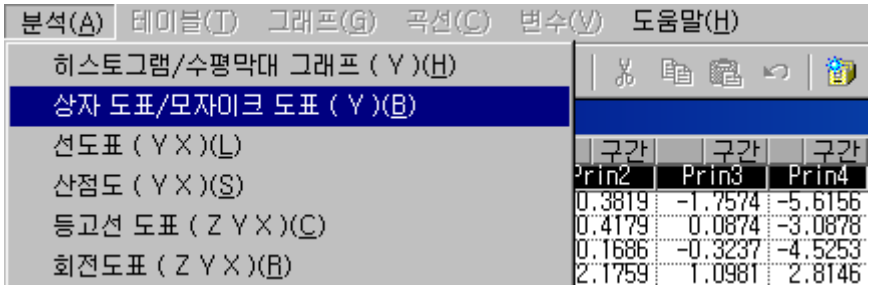
SAS/INSIGHT 를 이용해 Box-whisker plot 을 그려 보자. 일단 SCORE data 가 만들어지면 메뉴에서 다음과 같이 선택한다.

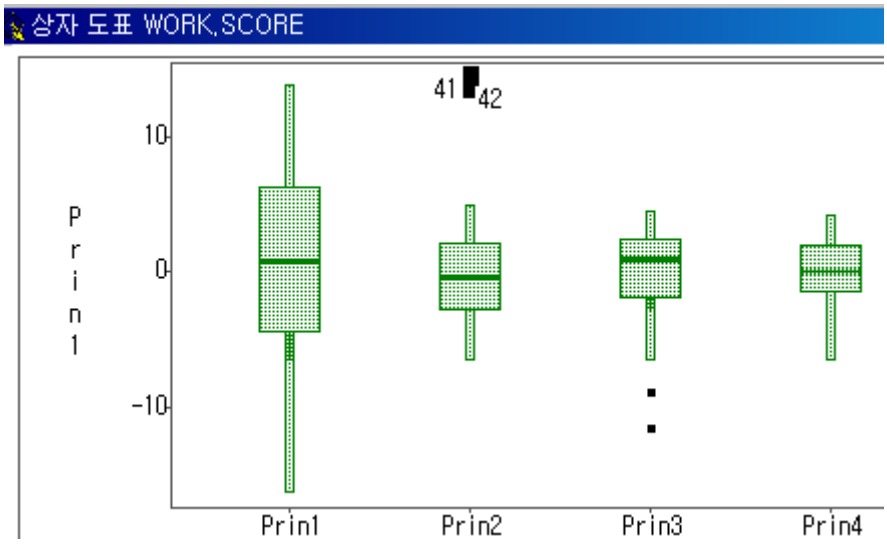


분석할 변수가 있는 SAS data 를 선택하는 마법사 화면이 나타나면 아래와 같이 선택한다.



분석할 변수를 선택하고(여러 변수를 선택하는 경우는 CTRL 을 누른 상태에서 변수를 마우스로 클릭해 준다) 메뉴에서 분석 메뉴를 선택하면 된다.





그려진 Box-whisker plot 을 해석하면 된다. 그림에서 알 수 있듯이 Prin2 는 오른쪽으로 Prin3 는 왼쪽으로 치우쳐 있어 정규 분포를 따르지 않을 가능성이 높음을 알 수 있다. bullet 점들은 이상치를 나타내는 것으로 Prin2 주성분 변수에 의하면 (41, 42)가 이상치임을 알 수 있다. Prin2 는 경험 주성분 변수로 (EX. SU)의 점수 반영이 높으므로 당연한 결과이다.

38	4	9	6	6	9	9	7	9	1	2	10	8	5	5	2
39	10	6	9	10	9	10	10	10	10	10	8	10	10	10	10
40	10	6	9	10	9	10	10	10	10	10	10	10	10	10	10
41	10	7	8	0	2	1	2	0	10	2	0	3	0	0	10
42	10	3	8	0	1	1	0	0	10	0	0	0	0	0	10
43	3	4	9	8	2	4	5	3	6	2	1	3	3	3	8
44	7	7	7	6	9	8	8	6	8	8	10	8	8	6	5

(2)회귀 분석에 이용

다중 회귀($Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + e_i$) 모형에서 설명 변수들간의 상관 관계가 매우 유의하면 (높으면) $|X'X| \approx 0$ 이 된다. ($|X'X| \approx 0$ 에 대하여: 자료 행렬 X 의 열은 각 설명 변수가 된다. 설명 변수가 상관 관계가 높다는 것은 한 설명 변수를 다른 설명변수의 선형 함수로 표시할 수 있다는 것이다. ($X_k \approx aX_j$) 행렬의 성질에 의하면 한 열이 다른 열의 선형 함수로 표현되면 행렬식의 값은 0 이다.) $(X'X)^{-1} = \frac{1}{|X'X|} adj(X'X)$ 이므로 $(X'X)^{-1}$ 가 매우 커지게 된다. 회귀 계수의 추정치

$\hat{\beta} = (X'X)^{-1}X'Y$, 추정치의 분산 $s^2_{\hat{\beta}} = MSE(X'X)^{-1}$ 이므로 다중 공선성 문제가 발생하면 추정치가 불안해지고(계수 부호까지도 반대가 되는 경우 발생) 이는 t-검정이나 잔차 분석에 의해서도 발견되지 않는다.

다중 공선성의 발견 방법은 다음과 같다.

①산점도 행렬이나 상관 계수(두 변수간 상관 관계만 존재할 때는 유용하고 편리하지만 두 변수와 다른 변수간 상관 관계가 존재하는 것을 진단할 수 없다.)

②VIF(Variance Inflation Index)와 Condition Index 를 이용, 책마다 차이는 있지만 대충 10 이상이다.

다중 공선성의 해결 방법으로는 다음과 같다.

①상관 관계가 높은 변수 제외

②주성분 분석 이용

③능형 회귀(Ridge Regression: 다중 공선성은 회귀 계수의 분산을 증가시키므로 불편성을 포기하는 대신 MSE(Mean Square of Error)를 최소화 하는 편기 추정량을 구하는 계수 추정 방법) 등이 있다.

회귀 분석에 주성분 분석을 이용한다는 것은 주성분을 설명 변수로 사용한다는 것이다. 주성분을 구하고 이를 회귀 모형의 설명 변수로 사용하는 것이다. 주성분은 서로 독립이므로 설명 변수의 상관 관계로부터 발생하는 다중 공선성 문제를 해결할 수 있다. 설명 변수의 공분산 행렬(설명 변수들의 측정 단위가 다르면 상관 계수 행렬)로부터 주성분을 얻는다. 주성분이 새로운 설명 변수로 이용되고 주성분 점수가 새로운 설명 변수의 측정치가 된다. 새로운 회귀 모형은 다음과 같다.

$$Y_i = \beta_0 + \beta_1 \mathrm{Prin1}_{1i} + \beta_2 \mathrm{Prin2}_{2i} + \cdots + \beta_p \mathrm{Prinp}_{pi} + e_i$$

PRIN1, PRIN2 … 는 OUT= 옵션에 의해 저장된 주성분 점수이다. 다중공선성 문제 해결로 주성분 분석을 사용하려면 다음 두 조건이 만족할 때만 하기 권한다. 그 외에는 변수 제거 방법이 가장 적절한 방법이다.

①다중 공선성 문제를 발생하는 설명변수가 꼭 필요하다고 판단되거나 제외하면 설명 변수 수가 너무 적은 경우

②주성분 이름을 부여하기 용이한 경우, 주성분이 새로운 설명 변수이기 때문이다.

③주성분을 회귀 분석에 이용하는 경우 2~3 개 주성분을 선택할 필요는 없다. 모든 주성분을 설명 변수로 사용하고 변수 선택 방법(stepwise, backward, forward)에 의해 선택하면 된다. 일반적으로 원 변수의 변동을 잘 설명하는 제일 주성분, 제이 주성분 등은 유의한 설명 변수로 선택된다. 물론 원 변수들의 종속 변수에 대한 설명력이 유의할 경우이지만

④주성분에 대한 적절한 이름을 부여할 수 없으면 회귀 분석 결과에 대한 해석이 어렵다.

회귀 분석의 주목적이 예측치($\hat{y}$)를 구하는데 있지 않다면 다중 공선성 문제 해결 방안으로 주성분을 설명 변수로 이용하는 방법을 사용하지 말기를 권한다.

4.9. 주성분 분석 절차

주성분 분석은 다음 절차에 의한다.

(1)원 변수 단위 점검

원 변수의 단위를 살펴 단위가 유사하면 주성분 분석을 위하여 공분산 행렬(covariance matrix)을 이용한다. 단위가 다른 경우 맞출 수 있으면 (예: kg ► pound, 단위: 원 ► 단위: 천원) 맞춘 후 공분산 행렬을 사용하고 불가능한 경우 상관 계수(correlation matrix) 행렬을 사용한다.

(2)공분산 행렬/상관 계수 행렬로부터 고유치, 고유 벡터 계산]

```
DATA APPLICANT;

     INFILE "C:\TEMP\APPLICANT.TXT";
     INPUT ID X1-X20;
  RUN;
  PROC PRINCOMP DATA=APPLICANT OUT=SCORE OUTSTAT=OUT1 COVARIANCE;
     VAR X1-X20;
RUN;
```

(3)고유치 변동 설명 비율 이용하여 적절한 고유치 개수 정하기

공분산 행렬 사용한 경우에는 누적 변동 설명 비율이 80%, 상관 계수 행렬 사용하였으면 고유치가 1 이상인 고유치만 선택한다. 고유 벡터는 $\underline{y} = L\underline{x}$ 에서 계수 L 을 구성한다. 고유 벡터 값은 OUT1 에 저장된다. 고유 벡터의 값은 PRIN1, PRIN2.....에 저장된다.

(4)고유 벡터를 이용하여 주성분 이름 부여

고유 벡터의 계수 값의 크기나 고유 벡터 산점도 그리기(아래)를 이용하여 주성분에 적절한 이름을 부여한다. 이름을 부여할 때는 다소 주관적일 수 있다.

```
TITLE "PLOT OF PRIN1*PRIN2";
%PLOTIT(DATA=OUT2, LABELVAR=_NAME_, PLOTVARS=PRIN1 PRIN2,
        COLOR=BLACK, COLORS=BLUE);
```

(5)주성분 점수 이용하기

주성분 점수는 주성분 관측치를 의미하며 $\underline{y} = L\underline{x}$ 에 각 원 변수 관측치 $\underline{x}$ 의해 계산된 $\underline{y}$ 의 값이며 OUT=SCORE 옵션에 의해 SCORE 에 저장된다. 변수 명은 PRIN1, PRIN2, ...이다.

①정규성 검정

원 변수가 정규 분포를 따르는지 선택된 주성분 PRIN1, PRIN2, (혹은 PRIN3)에 의해 검정할 수 있다.

②이상치 진단

선택된 주성분 산점도, PRIN1, PRIN2, (혹은 PRIN3)간 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. 물론 산점도를 이용하거나 PRIN1(상, 하), PRIN2(상, 하) 4 그룹으로 나눌 수 있다.

④회귀 분석

설명 변수간 다중공선성 문제가 발생하면 그 해결책으로 PRIN1, PRIN2..... 모든 주성분을 설명 변수로 사용하여 회귀 분석을 실시한다.

◈행렬을 SAS data 로 만들기

```
proc iml;
    x={1 2 3,
       4 5 6,
       7 8 9,
       0 1 2};
    create one from x
    append from x;
    show contents;
    close one;
run;

proc print data=one;
run;
```

DATASET : WORK.ONE.DATA

VARIABLE	TYPE	SIZE
COL1	num	8
COL2	num	8
COL3	num	8

Number of Variables : 3
Number of Observations: 4

Obs	COL1	COL2	COL3
1	1	2	3
2	4	5	6
3	7	8	9
4	0	1	2

[EXERCISE]

(1)1990 년 미 해군에서 학사 장교들의 관사에 필요한 인력을 추정하기 (MMH) 위하여 25 개 지역에 대해 7 개 분야에 (ADO—RMS) 대해 조사한 것이다. [Applied Multivariate Methods for Data Analysts, Dallas E. Johnson, 1998, p94] **NAVY.txt**(변수에 대한 정보는 자료 참조)

①7 개 분야를 설명 변수로 하고 MMH 를 종속 변수로 하여 회귀분석을 실시한다고 하자. 다중 공선성 문제가 있음을 보이시오.

②다중 공선성 문제를 일으키는 설명 변수를 제외하고 회귀분석 하시오(해결 방법 1). 어떤 설명 변수가 가장 영향을 많이 미치는지 알아보시오(표준화 회귀 계수).

③다중 공선성 문제를 해결하기 위한 방법으로 설명 변수를 주성분 변수로 하자 (해결 방안 2). 주성분 설명 변수의 의미가 무엇인지 해석하시오. 공분산 행렬로 구할까? 상관계수 행렬로 구할까?

④주성분 분석 결과를 이용하여 이상치를 밝히시오.

⑤②번 추정 회귀 모형과 ③번 추정 회귀 모형 중 어느 방법이 더 좋은지 잔차 변동 합 $\sum(y_i - \hat{y}_i)^2$ 을 이용하여 해석하시오.

(2)다음은 1994 년 미국 BIG8 리그 8 개 대학의 football 팀에 관한 자료이다.

[http://lib.stat.cmu.edu/DASL]]**BIG8.TXT**

학교명, 경기 수, Rushing 공격 야드(RO_YDS), Rushing 수비 야드, Passing 공격 야드, Passing 수비 야드, 총공격 야드, 총 수비 야드(TD_YDS), 공격 점수(Scoring Offence), 실점(SD), 게임당 실수 마진(Turn-over margin per game), 이긴 회수(WINS)

SCHOOL	GAMES	RO_YDS	RD_YDS	PO_YDS	PD_RAT	TO_YDS	TD_YDS	SO	SD	TOM	WINS
"COLORADO"	11	291.5	114.2	203.8	125.2	495.3	343.7	36.2	19.2	0.55	10
"IOWA STATE"	11	178.0	272.8	137.1	137.1	315.1	460.7	17.5	33.0	-0.64	0
"KANSAS"	11	247.1	171.2	140.9	135.8	363.3	400.7	28.5	22.0	0.73	6
"KANSAS STATE"	11	125.6	167.5	237.6	94.3	363.3	312.5	27.7	14.2	1.18	9
"MISSOURI"	12	107.9	235.3	202.5	138.5	310.4	414.9	17.3	27.1	0.17	3.5
"NEBRASKA"	12	340.0	79.3	137.8	96.7	477.8	258.8	36.3	12.1	0.08	12
"OKLAHOMA"	11	182.2	148.5	173.9	107.5	356.1	295.7	19.8	21.6	-0.18	6
"OKLAHOMA STATE"	11	204.6	192.5	133.5	130.3	338.1	385.9	16.4	23.3	-0.45	3.5

박스 안의 6 개 변수만을 이용하여 다음 절차에 따라 주성분 분석을 시행하시오.

①분산-공분산 행렬을 이용해야 하나 상관계수 행렬을 이용해야 하나? 이유는?

②고유치와 고유 벡터를 구하시오.

③총 변동 설명 비율에 의해 적절한 주성분 개수를 구하시오.

④③에서 선택된 주성분에 대해 주성분 계수를 이용하여 주성분에 적당한 이름을 붙이시오.

⑤주성분 계수간 산점도를 그리고 4)의 결과와 일치함을 보이시오.

⑥변수들이 다변량 정규 분포를 따르는지 분석하시오.

⑦선택된 주성분들의 주성분 점수를 산점도로 나타내(ID=학교별) 점수를 이상치가 있는가? 있으면 어떤 대학이고 어떤 면에서 이상치인가?

⑧제 1 주성분 점수, 제 2 주성분 점수, 그리고 선택된 주성분의 점수 합에 의해 각 팀의 등수를 매겨보시오. 그리고 성적과 비교하시오. 차이가 있다면 이유를 설명하시오.