4강. 인자 분석

- 인자분석이란
- 주성분분석과 인자분석
- 인자분석 모형
- 인자모형 추정
- 인자회전
- 인자점수
- R을 이용한 인자분석 (5강)
- 파이썬을 이용한 인자분석(5강)

1. 인자 분석이란?

◆ 주성분분석

서로 관련이 있는(즉, 상관계수가 0이 아닌) 변수들의 선형결합을 이용하여
새로운 변수를 만드는 과정. 즉, 원래 변수들이 가지고 있는 정보의 일정수준을
확보하도록 소수의 새로운 변수들을 만드는 방법으로 새로 만들어진 변수들이
서로 직교인 특성이 있음

◆ 인자분석

여러 개의 서로 관련이 있는 변수들로 측정된 자료에서
그 변수들을 설명할 수 있는 새로운 공통변수를 파악하는 통계적 분석방법

1. 인자 분석이란?

인자분석 예)

고등학교 학생들 100명을 대상으로 국어, 영어, 수학, 일반사회, 지리, 역사, 물리, 화학, 생물 등 9개 과목의 시험을 실시

⇒ 9개의 과목들을 공통적으로 설명할 수 있는 공통 인자(변수)들을 유도하여 분석

⇒ 공통 인자는 추상적인 개념, 예를 들면 이해력, 분석력 등을 의미하는 변수임.
이와 같은 인자들이 주어지면 각각의 과목들은 인자들의 선형결합에 의하여 표현됨.
여기서 인자(factor)는 관측되지 않는 가상의 변수임

⇒ 인자분석에서는 이러한 인자들을 생성하게 되는데, 이때 생성된 인자들에
대한 해석은 주관적. 주어진 자료에 가장 적절하도록 해석하는 것이 필요

2. 주성분 분석과 인자분석

◆ 주성분분석과 인자분석은 모두 관측된 여러 개의 변수들로부터 소수의 새로운 변수들을 생성하는 통계분석방법, 분석과정이 유사하지만 두 분석방법은 기본적인 접근방법이 다르다고 할 수 있음

주성분 분석의 목적	인자 분석의 목적
(i) 여러 개의, (ii) 변이에 관한 구조적 해석이 어려운 (iii) 서로 상관되어 있는 변수들을 적절히 선형변환시켜 (I) 소수 몇 개의 (II) 개념적 의미를 부여할 수 있는 (III) 서로 직교하는 주성분을 유도하여, 다음 단계의 통계분석에 이용	(i) 상관관계를 맺고 있어, (ii) 직접 해석하기 어려운, (iii) 여러 변수들간의 구조적 연관관계를, (I) 상대적으로 독립이면서, (II) 변수들의 저변구조에 관한 개념적 의미를 부여할 수 있는, (III) 원래 변수보다 훨씬 적은 공통인자들을 유도하여 이를 통해 분석하고자 하는 통계적 방법

2. 주성분 분석과 인자분석

※ 공통인자

변수들이 구조적 측면에서 서로 공유하고 있는 확률적 인자로서, 변수들 간의 상관관계를 생성시키는 가설적인, 이론적인, 관찰할 수 없는, 저변에 깔려있는 변수를 의미

◆ 두 분석방법의 차이점

① 주성분 분석이 관측된 변수들의 선형결합에 의하여 새로운 변수들을 만드는데 비하여, 인자분석에서는 가공의 인자들을 만든 후에 관측된 변수들을 가공의 인자들의 선형 결합식으로 표현

② 주성분 분석에서는 주성분들이 가지고 있는 정보의 크기에 따라 순서가 주어지나 인자분석의 인자들은 순서의 의미가 없음

③ 주성분 분석에서는 관측된 변수들의 선형 결합식이므로 오차항(error)이 없으나, 인자분석에서는 관측된 변수들을 인자들의 선형식으로 설명하며, 설명되지 않는 부분을 오차항(error) 또는 특수인자(specific factor)로 정의

3. 인자 분석 모형

- 인자분석에서 p개의 변수들을 $X = (X_1, X_2, \cdots, X_p)^T$ 와 같이 표현할 때, X의 분산 · 공분산 행렬은 다음과 같이 표현

$$Var(X) = \Sigma_{pp} = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ & \sigma_{22}^2 & \cdots & \\ & \vdots & & \\ & & & \sigma_{pp}^2 \end{pmatrix}$$

- 인자모형 : 각각의 변수에서 그 평균을 뺀 값 $(X_1 - \mu_1 \cdots, X_p - \mu_p)$이 q개의 가공인자 $(f_1, f_2, \cdots, f_q)$들의 선형결합과 오차항 ε의 선형결합으로 표현되는 모형을 말함

3. 인자 분석 모형

◆ 분석모형 p개의 변수를 설명하는 q개의 인자가 있을때

$$
\begin{aligned}
X_1 - \mu_1 &= l_{11} \cdot f_1 + l_{12} \cdot f_2 + \cdots + l_{1q} \cdot f_q + \varepsilon_1 \\
X_2 - \mu_2 &= l_{21} \cdot f_1 + l_{22} \cdot f_2 + \cdots + l_{2q} \cdot f_q + \varepsilon_2 \\
&\vdots \qquad\qquad\qquad \vdots \\
X_p - \mu_p &= l_{p1} \cdot f_1 + l_{p2} \cdot f_2 + \cdots + l_{pq} \cdot f_q + \varepsilon_p
\end{aligned}
$$

$$
\begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \\ \vdots \\ X_p - \mu_p \end{bmatrix} = \begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1q} \\ l_{21} & l_{22} & \cdots & l_{2q} \\ \vdots & & & \\ l_{p1} & l_{p2} & \cdots & l_{pq} \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_q \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix}
$$

$$
\boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{L} \cdot \boldsymbol{F} + \boldsymbol{\varepsilon}
$$

l_{ij} : 인자부하값(factor loadings) i=1, 2 , …, p , j = 1, 2 , … ,q

L : 인자부하행렬(matrix of factor loadings)

f_j : 인자(common factor), j=1,2,…,q

F : 인자벡터 (Factor vector)

ε_i : 특수인자(specific factor), i=1,2 …,p

3. 인자 분석 모형

- 위의 모형에서 " $X_1 - \mu_1 = l_{11} \cdot f_1 + l_{12} \cdot f_2 + \cdots + l_{1q} \cdot f_q + \varepsilon_1$ "은 변수 X_1 에서 평균을 뺀 값 $X_1 - \mu_1$은 q개의 인자 $(f_1, f_2, \cdots, f_q)$ 들의 선형결합 $l_{11} \cdot f_1 + l_{12} \cdot f_2 + \cdots + + l_{1q} \cdot f_q$ 과 오차항(특수인자) ε_1의 결합된 형태로 표현된다는 의미임

- 이와 같은 식을 모든 X_i , $i = 1, 2, \cdots, p$ 에 대하여 구하는 분석방법이 인자분석, 즉 각 변수 X_i에 대하여 가공의 변수(인자) $f_1, \cdots, f_q$ $(q \leq p)$를 이용하는 회귀모형을 구하는 분석방법임

- 관심의 초점 : 계수 l_{ij} , $i = 1, 2, \cdots, p,\ q = 1, 2, \cdots, q$ 의 추정과 각 변수 X_i 들이 인자들의 선형결합에 의하여 설명되는 수준(%)

3. 인자 분석 모형

◆ 인자분석의 기본가정

① 변수벡터 $X = (X_1, X_2, \cdots, X_p)^T$ 는 다변량 정규분포를 따른다.

② 인자 f 와 특수인자 ε 의 평균은 모두 0이고 인자들 $f_1 \cdots, f_q$ 의 분산은 모두 1이다.

③ 각 인자쌍 (f_i, f_j) 의 공분산은 모두 0 이고, 인자 f 와 특수인자 ε 는 서로 독립이다.

④ 특수인자 ε_i 의 분산은, $\phi_i,\ i = 1 \cdots p$ 이나 특수인자쌍 $(\varepsilon_i, \varepsilon_j)$ 의 공분산은 0이다.

3. 인자 분석 모형

◆ 인자분석의 기본가정

$$X \sim N(\mu, \Sigma)$$

$$E(f_j) = 0,\ j = 1,2,\cdots,q$$

$$Var(f_j) = 1,\ \ Cov(f_i, f_j) = 0,\ \ i,j = 1,2,\cdots,q$$

$$E(F) = \underline{0},\ Cov(F) = I_{q*q}$$

$$E(\varepsilon_i) = 0,\ i = 1, 2, \cdots, p$$

$$Var(\varepsilon_i) = \phi,\ Cov(\varepsilon_i, \varepsilon_j) = 0$$

$$E(\varepsilon) = \underline{0},\ Cov(\varepsilon) = \Phi = \begin{bmatrix} \phi_1 & \cdots & 0 \\ \vdots & & \\ 0 & \cdots & \phi_p \end{bmatrix}$$

$Cov(\varepsilon, F) = 0_{p*q}$: ε와 F는 서로 독립

3. 인자 분석 모형

◆ 공통성 (Communality)

① 공통성 수식

$$X - \mu = L \cdot F + \varepsilon \quad \Rightarrow \quad Cov(X - \mu) = Cov(L \cdot F + \varepsilon)$$

여기에서,

$$Cov(X - \mu) = Cov(X) = \Sigma, \quad Cov(L \cdot F) = L \cdot Cov(F) L^T = LL^T, \quad Cov(\varepsilon) = \Phi$$

이고, 인자 $\underline{F}$ 와 특수인자 $\underline{\varepsilon}$는 서로 독립이므로

$$Cov(X - \mu) = \Sigma = LL^T + \Phi$$

⇒ 따라서, 변수 X_i 의 분산 σ_{ii}^2 는 LL^T의 i 번째 대각원소 $\ell_{i1} + \ell_{i2} + \cdots + \ell_{iq}$ 와 ε_i 의 분산 ϕ_i 의 합. 즉,

$$\sigma_{ii}^2 = \ell_{i1} + \ell_{i2} + \cdots + \ell_{iq} + \phi_i$$

여기에서 LL^T의 i 번째 대각원소 $\ell_{i1} + \ell_{i2} + \cdots + \ell_{ip}$ 를 공통성(Communality)이라고 정의하고 h_i^2으로 표현 $h_i^2 = \ell_{i1} + \ell_{i2} + \cdots + \ell_{ip}$, $i = 1, 2, \cdots, p$

3. 인자 분석 모형

◆ 공통성 (Communality)

② 공통성 의미

공통성이란 변수 X_i의 분산 중에서 q개의 인자 ($f_1, \cdots, f_q$)들에 의하여 확보되는 부분 X_i의 분산 σ_{ii}^2 은

$$\sigma_{ii}^2 = h_i^2 + \phi_i$$

⇒ 따라서 공통성은 변수 X_i가 가지고 있는 정보 중에서 q개의 인자들에 의하여 확보될 수 있는 정보의 비율을 측정하는 척도가 됨

⇒ 공통성 h_i^2 의 값은 $0 \le h_i^2 \le 1$사이의 값을 갖는데, h_i^2 값이 1에 가까울수록 변수 X_i 가 가지고 있는 정보 중에서 q개의 인자 $f_1, \cdots, f_q$ 가 확보하는 비율이 크다는 것을 의미

3. 인자 분석 모형

◆ 인자분석에서 분석의 초점

- 변수 X_i 에 대한 인자부하값 $(l_{i1}\, l_{i2}, \cdots l_{iq})$ 을 추정하여
  변수 X_i 와 q 개의 인자 $f_1, \cdots, f_q$ 사이의 관계를 추정
- 공통성 h_i^2 을 구하여 변수 X_i 가 가지고 있는 정보들 중에서 q 개의 인자들에 의하여
  어느 정도 확보되는가를 추정하는 것

◆ 인자의 수

- 인자의 수 q 는 변수의 수 p 보다 작아야 함.
  이는 p 개의 변수가 가지고 있는 정보를 보다 작은 q 개의 인자들을 이용하여
  최대한 확보하는 것이 인자분석의 초점이기 때문임
- 적절한 인자의 수 q 는 확률변수 벡터 X의 분산 · 공분산행렬 Σ 의 고유근의 크기에 의하여 결정
- 인자의 수 q 의 디폴트 값은 Σ 의 고유근이 1보다 큰 개수임. 원하는 경우에는 q 를 특정 값으로 지정

4. 인자 모형 추정

◆ 인자분석모형에서는 관측된 자료를 이용하여 인자부하값과 특수분산, 그리고 공통성 등을 추정

- 추정방법으로 가장 기초적인 방법이 주성분분석으로부터 유도되는 주성분방법(principal component method) 임. 최우추정법도 많이 이용

① 주성분방법(principal component method)

- 관측값 X의 분산·공분산행렬 Σ 또는 상관계수행렬 R 의 고유근과 고유벡터를 이용하여 인자부하값과 특수분산을 추정하는 방법임
- 여기서 상관계수행렬 R 의 대각선 값 1 대신 공통성 추정치로 대치한 방법을 주성분인자법(principal factor method) 이라고 하며, 인자모형추정에 널리 이용됨

② 최우추정법(maximum likelihood method)

- X가 다변량 정규분포를 따른다는 가정 하에서 우도함수(likelihood function)을 구하고, 이를 최대화하는 최우추정법으로 인자부하값과 특수분산을 추정하는 방법임. 추정의 신뢰성이 높아 많이 이용되는 방법임

5. 인자의 수와 인자부하값의 유의성

◆ 인자의 수는 최대로 변수의 수 p 와 같을 수 있음. 그러나 인자분석의 목적이 최소의 인자를 구하는 것으로 많은 경우 인자의 수가 3개 또는 4개를 선택

① 인자의 수는 상관계수행렬 R 의 고유근이 1보다 큰 경우만 채택 - 이는 주성분분석에서 보유주성분의 개수결정을 위한 Kaiser의 규칙을 인자분석에 적용한 것으로, 유의미한 인자로 보유되는 것들의 개수는 R 의 고유값 중 1보다 큰 것들의 개수로 정함. Jolliffe(1972)는 고유값 기준으로서 Kaiser의 1 대신 0.7을 제안함

② 인자부하값의 유의성 : 수학적 근거보다는 통상적 관념으로 $n \geq 50$ 인 경우, 절대값 기준으로

인자부하값 〉 0.3 : 유의함
〉 0.4 : 좀 더 유의함
〉 0.5 : 아주 유의함

과 같이 의미를 부여

6. 인자회전 (Factor rotation)

- 인자분석은 인자부하 값들의 크기에 의하여 각 변수들을 유사한 것끼리 묶거나 공통적인 요인을 찾음. 이 때, 인자부하값들을 이용하여 인자를 찾을 때, 해석을 쉽게 하기 위해 인자회전을 실시

6. 인자회전 (Factor rotation)

예1) 고등학교 학생 100명을 대상으로 국어(X_1), 영어(X_2), 수학(X_3), 물리(X_4), 사회(X_5) 등 5개 과목의 시험을 치른 후에, 인자분석을 실시하여 구한 2개 유의한 인자에 대한 인자부하행렬이 다음과 같은 경우

〈 5개과목의 인자부하 행렬 〉

과목 \ 인자	F_1	F_2
국어(X_1)	0.05	<u>0.97</u>
영어(X_2)	0.10	<u>0.88</u>
수학(X_3)	<u>0.96</u>	0.20
물리(X_4)	<u>0.85</u>	0.12
사회(X_5)	0.18	<u>0.76</u>

6. 인자회전 (Factor rotation)

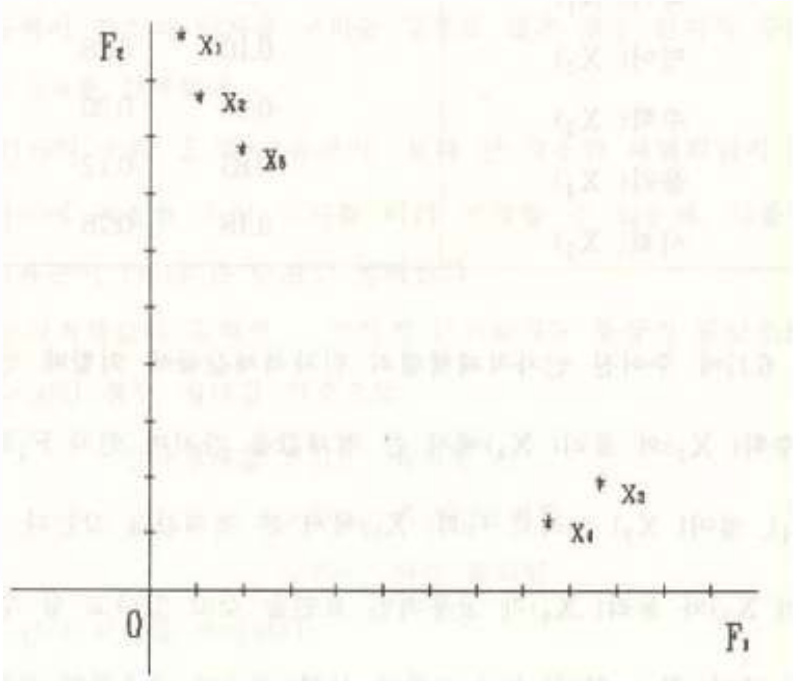
⇒ 인자 F_1은 수학(X_3)와 물리(X_4)에서 큰 적재값을 가지며(하단에 밑줄을 표시함),
인자 F_2는 국어(X_1), 영어(X_2) 그리고 사회(X_5)에서 큰 적재값을 가짐.

⇒ 따라서 수학(X_3)와 물리(X_4)가 공통적인 요인을 갖고 있다고 할 수 있고,
또한 국어 (X_1), 영어(X_2) 그리고 사회(X_5)가 공통적인 요인을 갖고 있다고 생각할 수 있음.
즉, 인자 F_1은 수학과 물리를 대표하는 인자이고, F_2는 국어, 영어, 사회를 대표하는 인자.

⇒ 이에 의하여 인자 F_1은 "분석력"이라고 정의하고 F_2는 "이해력"이라고 정의.

6. 인자회전 (Factor rotation)

※ 인자 F_1과 인자 F_2를 직교축으로 하는 평면상에 점으로 표현한 그림



<그림> (F_1, F_2) 좌표

6. 인자회전 (Factor rotation)

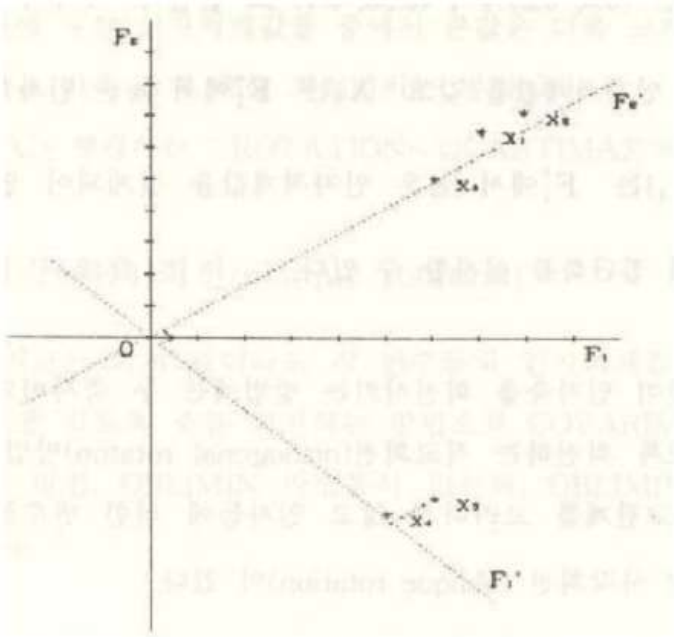
예2) 5개 변수의 2개 인자에 대한 인자부하값이 다음과 같은 경우

인자 / 과목	F_1	F_2
국어(X_1)	0.70	0.65
영어(X_2)	0.73	0.67
수학(X_3)	0.60	−0.50
물리(X_4)	0.50	−0.55
사회(X_5)	0.60	0.50

⇒ 인자부하 값들에 의할 때 각 과목들은 두 인자 모두에서
높은 인자부하값을 갖고 있으므로 두 인자 F_1과 F_2의 특성이 쉽게 구분되지 않음

6. 인자회전 (Factor rotation)

※ 주어진 5개 변수들을(F_1, F_2) 평면상에 표현한 결과

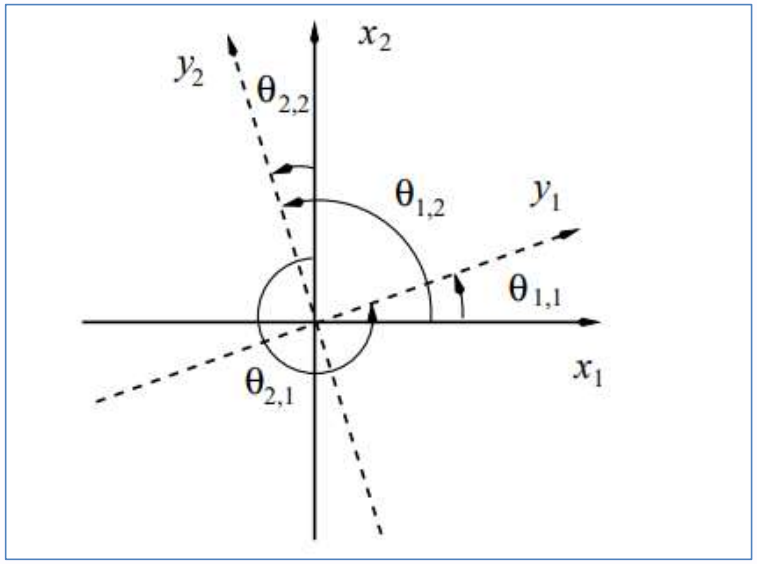


<그림> (F_1, F_2)좌표

⇒ 인자회전은 점선으로 주어진 것과 같이
각 변수들이 하나의 인자에서는 큰 인자부하값을 갖고
다른 인자에서는 작은 인자부하값을 갖도록 인자축을 회전.

⇒ 인자축을 (F_1, F_2) 에서 (F_1^*, F_2^*) 으로 회전하였을 때
인자들에 의한 변수들의 집단화를 실시하는 것이 용이함.

참고 : 인자회전 (Factor rotation) 이론적 배경



$$T = \begin{pmatrix} \cos\theta_{1,1} & -\sin\theta_{1,1} \\ \sin\theta_{1,1} & \cos\theta_{1,1} \end{pmatrix}$$

$$X - \mu = LF + \varepsilon$$

$$= LTT'F + \varepsilon \qquad \Leftarrow TT' = T'T = I$$

$$= L^*F^* + \varepsilon$$

6. 인자회전 (Factor rotation)

◆ 인자회전 방법

- 인자축을 회전시키는 방법에는 두 축 사이의 직교관계를 유지하도록 회전하는 직교회전 (orthogonal rotaton) 방법과, 두 축 사이의 직교관계를 고려하지 않고 인자들에 의한 변 수들의 집단화를 실시하는 사각회전 (oblique rotation)이 있음

① 직교회전 (orthogonal rotation)

- 인자축이 직교하도록 축을 회전하는 방법으로 VARIMAX방법과 QUARTIMAX방법 등이 이용

② 사각(**斜角**)회전(oblique rotation)

- 인자축이 직교가 되지 않더라도 각 변수들의 인자부하값이 한 인자에만 큰 값을 갖도록 축을 회전하는 방법으로 COVARIMIN 방법, QUARTIMIN 방법, OBLIMIN 방법 등이 있는데, OBLIMIN 방법이 많이 이용됨

7. 인자점수

◆ 인자점수

- 인자모형은 변수 $X_1, X_2, \cdots, X_p$ 를 소수의 인자 $F_1, F_2, \cdots, F_q$ 로 나타낸 모형
- 각각의 인자 $F_1, F_2, \cdots, F_q$ 을 변수 $X_1, X_2, \cdots, X_p$ 의 함수식으로 표현.
  이를 인자점수(Factor Score)라고 함
- 각 케이스에 대한 인자점수를 추정하여 산출하는 것이 바람직함

차원의 축소를 통해 추정된 인자점수들은 2차원 등의 저차원 그림 위에 각 케이스를 점으로 나타낼 수 있어 각 케이스의 특성을 파악하는데 효과적으로 활용될 수 있음.
또한 인자점수를 이용하여 회귀분석, 판별분석 등의 다음 단계의 분석에 이용

인자점수의 추정방법 : 회귀분석(R), Bartlett(B)과 Anderson-Rubin 방법(A) 등이 있음

※ 만약 원 자료가 아닌 상관행렬 또는 공분산행렬이 입력자료로 사용된 경우에는
인자점수를 구할 수 없음

다음시간에는

5강 인자분석 (2)

수고했습니다.