CHAPTER 5

요인 분석

5.1. 개요

요인 분석(FA: Factor Analysis 혹은 인자 분석이라고도 함)은 사람의 지적 능력을 측정하고 이에 연관된 변수들을 이해하려는 노력의 일환으로 Galton(회귀 분석 창시자)에 (1888) 의해 제안되었고 수학적인 모형은 Spearman(1904 상관 계수 제안자)에 의해 발전되었다. 요인 분석은 변수들의 내재된 상관 관계를 이용하여 요인을 구하고 이를 이용하여 (1)변수들을 분류하고 (변수 그룹에는 원 변수 일부만 포함되어 있다) (2)그룹에 적절한 의미를 부여하는(그룹 이름 부여) 분석 방법이다.

요인 분석의 예를 보면, 설문 조사에서 동일한 개념을 측정하기 위해 설계된 리커드(Likert) 척도 문항들이 정말 그런지 알아보기 위한 분석 방법으로 요인 분석이 사용된다. 물론 그 문항들의 신뢰도(혹은 내적 일치도)는 Cronbach α로 측정된다. 예를 들어 학생들의 학교 만족도를 측정하기 위하여 교수 강의, 조교, 행정 인력, 강의실, 도서실, 전산실습실, 체육 시설, 건물 만족도를 조사하였다고 하자. 8 개의 만족도 항목을 그룹화 할 수 있을까? 이에 대한 해답을 요인 분석이 제공한다. A 기업 지원자 48 명의 능력에 대해 측정한 15 개 항목 점수들을 분류(그룹)하고자 할 때 사용된다. 또한 기업

관련 지표에 관해 20 개의 항목을 (매출액, 종업원 수, 부채비율, ...) 유사한 항목끼리 분류하고 할 때 사용한다.

5.1.1. Spearman (1904)

Spearman 은 학생들의 6 과목 성적에 대한 상관 계수를 구한 후 상관 계수 값을 살펴 보아 각 학생들의 과목 성적은 다음과 같이 두 부분으로 나눌 수 있을 것이라 생각했다.(언어, 수리) 그러나 상관 계수 값으로 과목을 분류하는데 한계에 부딪히게 된다. (아래 상관 계수 결과를 보고 변수를 나누어 보라)

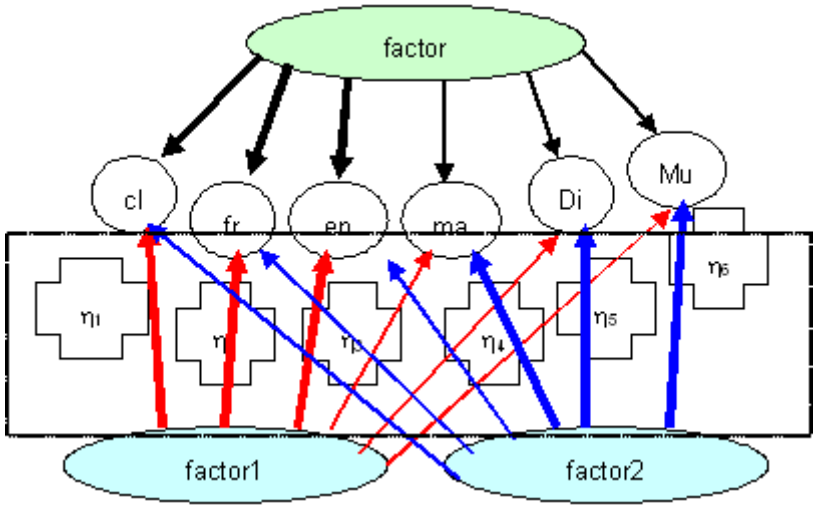
	Classic	French	English	Math	Discover	Music
Classic	1	.83	.78	.7	.66	.63
French		1	.67	.67	.65	.57
English			1	.64	.54	.51
Math				1	.45	.51
Discover					1	.4
Music						1

이에 Spearman 은 각 시험 점수(변수)는 다음과 같이 변수 간에 내재된 공통 개념(*f*: 이를 factor 라 함) 부분과 랜덤 부분에 해당하는(η) 부분으로 나눌 수 있을 것이라 생각했다. 물론 f 와 η_j 들은 서로 독립이라고 가정하였다. 또한 그는 학생들의 과목 성적은 일반적 재능으로 해석되는 인자 f 와 과목에 대한 특별 재능으로 나눌 수 있다고 믿었다.

$$
\begin{aligned}
classic &= \lambda_1 f + \eta_1 \\
french &= \lambda_2 f + \eta_2 \\
english &= \lambda_3 f + \eta_3 \\
math &= \lambda_4 f + \eta_4 \\
dis\,cov\,er &= \lambda_5 f + \eta_5 \\
music &= \lambda_6 f + \eta_6
\end{aligned}
$$

아래 그림의 위 부분(그림 맨 위의 fact 부분)은 공통 개념이 하나인 경우를 도식화 한 것이다. 그림의 아래 부분(그림 맨 아래 factor1, factor2 부분)은 6 개 과목에 2 개의 공통

개념이 존재할 때 도식화 한 그림이다. 굵은 선은 영향을 많이 미치는 것을 의미하므로 공통 개념(요인)이 무엇인지는 모르지만 공통 개념이 영향을 주는 정도가 같은 과목끼리(변수끼리) 묶으면 될 것이다. 즉 고전 (classic), 불어(French), 영어(English)를 하나로 묶고 수학(Math), 과학(Discovery), 음악(Music) 하나로 묶을 수 있을 것이다.



5.1.2. 주성분 분석과 비교

주성분 분석은 p 개의 원 변수를 2~3 개의 주성분으로 축약하는데 사용한다면 요인 분석은 p 개의 변수들이 상호 어떤 관계가 있는지 결정하여 $m(< p)$ 개 변수 그룹으로 나누는데 목적이 있다. 요인 분석 결과 묶여진 변수 그룹을 살펴 보면 그룹 내 변수들 간에는 상관계수가 높고 다른 그룹의 변수 간에는 상관 관계는 낮다. 요인을 구하는 방법으로 주성분을 이용한 방법을 가장 많이 사용하므로 주성분 분석과 유사해 보인다. 다음은 주성분 분석과 요인 분석을 비교한 표이다.

변수 벡터 $\underline{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$ 이고 공분산 행렬 Σ, 상관 계수 행렬은 R 이라 하자.

주성분 분석	요인 분석
주성분은 원 변수의 직교 선형 결합으로 표현 $\underline{Y} = L\underline{X}$ ► l_{ij} 는 선형 결합 함수의 계수	인자들의 직교 선형 결합으로 원 변수들을 표현하며 인자는 관측될 수 없다. $\underline{X} = L\underline{F} + \eta$ ► l_{ij} 을 loading이라 (부하) 함
$y_1 = l_{11}x_1 + l_{12}x_2 \ldots + l_{1p}x_p$ $y_2 = l_{21}x_1 + l_{22}x_2 \ldots + l_{2p}x_p$ ... $y_p = l_{p1}x_1 + l_{p2}x_2 \ldots + l_{pp}x_p$	$x_1 = l_{11}f_1 + l_{12}f_2 \ldots + l_{1p}f_p + \eta_1$ $x_2 = l_{21}f_1 + l_{22}f_2 \ldots + l_{2p}f_p + \eta_2$ ... $x_p = l_{p1}f_1 + l_{p2}f_2 \ldots + l_{pp}f_p + \eta_p$ η_{ij} =오차항
주성분은 변수들의 변동을 설명한다. 공분산 행렬 사용	요인은 변수들의 분산-공분산 구조 설명한다. 상관 계수 행렬 사용
요인분석이나 주성분 분석의 l_{ij} 를 구하는 방법 유사하다. ► 공분산 행렬, 상관 행렬로부터 고유치 그에 대응하는 고유 벡터를 이용한다. ► (주성분 분석) $l_{ij} = e_{i(j)}$ (요인 분석) $l_{ij} = \sqrt{\lambda_i} e_{i(j)}$	
변수의 개수 축약하는데 사용되며 l_{ij} 는 주성분의 이름을 붙이는데 사용	변수에 내재된 관계를 알아보는데 사용되며 l_{ij} 는 변수들을 그룹화 하는데 사용한다.
적절한 주성분의 수를 구하고 주성분의 이름을 부여하고 주성분들간 산점도로 이상치 발견하거나 각 주성분 점수에 의해 개체 순위	적절한 인자의 수를 구하고 이를 이용하여 변수들을 그룹화 하고 그룹을 이용하여 변수에 내재된 관계를 알아본다.

5.2. 요인 분석 모형

5.2.1. 목적

(1)원 변수에 내재된 관계를 설명할 공통 개념을 (요인) 살펴 본다.

(2)공통 개념(요인)의 개수를 결정한다.

(3)요인의 부하 값을 이용하여 원 변수를 그룹화 하고 적절한 이름을 부여한다.

(4)새 변수를 이용하여 개체들을 평가하고 향후 연구에 이 변수들을 이용한다.

5.2.2. 모형 및 가정

p 개의 원 변수 $\underline{x}' = (x_1, x_2, ..., x_p)$ 가 평균 벡터가 $\underline{\mu}$, 분산-공분산 행렬이 Σ 이라 하면 일반적인 요인 분석 모형은 다음과 같다.

$$\underline{x} = L\underline{f} + \underline{\eta} \quad \Leftrightarrow \quad \begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_p \end{pmatrix} = \begin{pmatrix} l_{11}\, l_{12} \cdots l_{1m} \\ l_{21}\, l_{22} \cdots l_{2m} \\ \cdots \\ l_{p1}\, l_{p2} \cdots l_{pm} \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ \cdots \\ f_m \end{pmatrix} + \begin{pmatrix} \eta_1 \\ \eta_2 \\ \cdots \\ \eta_p \end{pmatrix}$$

(1) f_k , f_2 , ..., f_m 들은 공통 인자 (요인: common factor)

(2) l_{ij} (인자 부하: factor loading)

i 번째 변수에 j 번째 요인이 미치는 영향이므로 L 을 요인 부하 행렬(factor loading matrix)이라 한다.

(3) $\eta_1, \eta_2, ..., \eta_p$ (특정 인자: specific factor)

η_j 는 j 번째 변수에 한정된 오차 변동

L 를 요인 부하 행렬(factor loading matrix)이라 한다. 이 모형에 대한 가정을 다음과 같다.

(1) f_k 들은 상호 독립이고 평균이 0, 분산이 1 인 동일 분포를 따른다. ($k = 1,2,..,m$)
$\underline{f} \sim (\underline{0}, I)$

(2) η_j 들은 상호 독립이고 평균이 0, 분산이 ψ_j 인 동일 분포를 따른다. ($j = 1,2,..,p$)

$\underline{\eta} \sim (\underline{0}, \psi = diag(\psi_1, \psi_2, .., \psi_p)$ 즉, Ψ 는 대각 행렬이다.

(3) f_k 들과 η_j 들은 상호 독립이다. $Cov(\underline{\eta}, \underline{f}) = 0$

5.2.3. 모형 해석

$\underline{x} = L\underline{f} + \underline{\eta}$ 모형에 대해

(1) $\Sigma = Cov(\underline{x}) = Cov(L\underline{f} + \underline{\eta}) = LCov(\underline{f})L' + \psi = LL' + \psi$ 이고 $\underline{\eta} \sim (\underline{0}, \psi = diag(\psi_1, \psi_2, .., \psi_p))$ 이므로 공통 인자들은 변수($x_1, x_2, ..., x_p$)들의 공분산을 완전히 설명한다. 왜냐하면 ψ 은 대각 행렬이므로 대각을 제외하고는 0 이다.

(2) j 번째 원 변수 분산은 $Var(x_j) = \sigma_{jj} = l_{j1}^2 + l_{j2}^2 + ... + l_{jm}^2 + \Psi_j = \sum_{k=1}^{m} l_{jk}^2 + \psi_j$ 쓸 수 있는데 $\sum_{k=1}^{m} l_{jk}^2$ 을 공통성(communality)이라 하고 ψ_j 는 특정(specific) 분산이라 한다. 즉 $\sum_{k=1}^{m} l_{jk}^2$ 는 원 변수 x_i 은 변동 중 공통 인자(common factor)들에 의해 설명되는 부분이다. 공분산 행렬 대신 상관 계수 행렬을 이용한다면 대각 원소가 1 이므로 $\sum_{k=1}^{m} l_{jk}^2 + \psi_j = 1$ 이 성립한다.

(3) i 번째 변수와 j 번째 변수의 공분산은 $\text{cov}(x_i, x_j) = \sum_{k=1}^{m} l_{ik} l_{jk}$ 이고 x_j (j 번째 변수)와 f_k (k 번째 요인) 공분산 $\text{cov}(x_j, f_k)$ 는 l_{jk} 이다.

5.3. 요인 구하기

5.3.1. 요인 방정식 풀기

상관 계수 행렬(R)을 이용하여 요인 분석을 한다고 하자. (수학적 전개가 용이하므로)

(1)상관 계수 행렬 R 에 대해 $R = LL' + \psi$ 을 만족하는 L, Ψ 가 존재한다고 하자. 또 다른 직교 행렬 P 에 대해 다음이 성립하므로 $R = LIL' + \psi = R = (LP)(LP)' + \psi = P_* P_*' + \psi$ 요인 부하 행렬 L 은 무수히 존재한다.

(2) L , ψ 의 미지수 개수를 보면 $(pm + p)$ 이고 행렬 P 로부터 얻을 수 있는 값의 개수는 $p(p+1)/2$ 이므로 방정식 수보다 미지수 개수가 많으므로 해가 무수히 많이 존재한다.

예를 들어 원 변수가 3 개인 경우 상관 계수 행렬로부터 얻을 수 있는 값은 6 개(대각 원소 3 개, 상위 원소 3 개)이나 미지수 9 개이다. 만약 $m = p$ 인 경우에는 $\Sigma = LL'$ 로 유일하게 분해되고 $\Psi = 0$ 이다.

(3)그럼 $m(< p)$ 인 경우 어떤 L 을 이용할 것인가? 얻어진 요인을 해석하는데 용이하도록 요인 변환(factor rotation)을 실시하여 그 값을 이용한다. (예: SAS 에서 ROTATE=VARIMAX 옵션)

5.3.2. 해를 구하는 방법

요인 방정식을 푸는 방법으로는 principal factoring w/ or w/o iteration, Rao's canonical factoring, alpha factoring, image factoring, maximum likelihood, un-weighted least square factor analysis, Harris factoring 등이 있다. 어느 방법이 가장 좋은지는 알 수 없으나 가장 많이 사용하고 기초적인 방법이 반복 있는/없는 주성분 이용한 인자 분석이다. (principal factoring w/ or w/o iteration) 그래서 요인 분석이 주성분 분석과 유사해 보인다.

(1)주성분 방법(Principal Component (factor) method)

변수의 상관 계수 행렬 R 에 대한 고유치, 고유 벡터를 구하여 그것을 각각 $\lambda_1 \geq \lambda_2 \geq ... \lambda_p$, $\underline{e}_1, \underline{e}_2, ..., \underline{e}_p$ 라고 하자. 이 경우 $\Sigma = \lambda_1 e_1 e_1' + \lambda_2 e_2 e_2' + \cdots + \lambda_p e_p e_p'$ 로 분해 가능하다.

$$\Sigma = [\sqrt{\lambda_1} e_1 \mid \sqrt{\lambda_2} e_2 \mid \cdots \mid \sqrt{\lambda_p} e_p] \begin{bmatrix} \sqrt{\lambda_1} e_1' \\ \sqrt{\lambda_2} e_2' \\ \cdots \\ \sqrt{\lambda_p} e_p' \end{bmatrix} = LL'$$

$m < p$ 인 경우 $(\lambda_1, e_1), (\lambda_2, e_2), \cdots, (\lambda_p, e_p)$ 이를 이용하여 요인 부하 행렬 L 을 구하면 다음과 같다.

$$\Sigma = [\sqrt{\lambda_1} e_1 \mid \sqrt{\lambda_2} e_2 \mid \cdots \mid \sqrt{\lambda_m} e_m] \begin{bmatrix} \sqrt{\lambda_1} e_1' \\ \sqrt{\lambda_2} e_2' \\ \cdots \\ \sqrt{\lambda_m} e_m' \end{bmatrix} + \begin{bmatrix} \Psi_1 & 0 & \cdots & 0 \\ 0 & \Psi_1 & \cdots & 0 \\ --- & & & \\ 0 & \cdots & 0 & \Psi_p \end{bmatrix} = LL' + \Psi \quad , \quad \Psi_i = s_{ii} - \sum_{i=1}^{m} l_{ij}^2$$

원 변수의 변동은 공분산 행렬을 사용할 경우 $tr(\hat{\Sigma}) = s_{11} + s_{22} + \cdots + s_{pp}$ 이고 $tr(\hat{R}) = p$ 이다. j 번째 공통 요인에 의한 변동 설명 부분은 $l_{j1}^2 + l_{j2}^2 + \cdots + l_{jp}^2 = \sqrt{\lambda_j} e_j' \sqrt{\lambda_j} e_j = \lambda_j$ 이다. 그러므로 j 번째 요인에 의한 원 변수 변동의 설명 비율은 다음과 같다.

공분산 행렬: $\frac{\lambda_i}{s_{11} + s_{22} + \cdots + s_{pp}}$, 상관 계수 행렬: $\frac{\lambda_i}{p}$

이제 요인 값을 구해 보자. 상관 계수 행렬로부터 구한 주성분을 $y_1, y_2, \ldots, y_p$ 라 하면 $Var(y_1) = \lambda_1$, $Var(y_2) = \lambda_2$ …… $Var(y_p) = \lambda_p$ 임을 이용하여 요인을 다음과 같이 정의해 보자.

$$f_1 = y_1 / \sqrt{\lambda_1}, \quad f_2 = y_2 / \sqrt{\lambda_2}, \ldots, \quad f_p = y_p / \sqrt{\lambda_p}$$

요인을 위와 같이 정의하면 요인의 분산(변동)은 1 이고 서로 독립이므로 페이지 75 의 가정을 만족한다. 원 변수를 요인 변동 행렬과 요인으로 나타내면 아래와 같다. $\underline{x} = L\underline{f} + \underline{\eta}$

$$\begin{aligned}
x_1 &= \sqrt{\lambda_1} e_{11} f_1 + \sqrt{\lambda_2} e_{12} f_2 + \ldots + \sqrt{\lambda_p} e_{1p} f_p \\
x_2 &= \sqrt{\lambda_1} e_{21} f_1 + \sqrt{\lambda_2} e_{22} f_2 + \ldots + \sqrt{\lambda_p} e_{2p} f_p \\
&\vdots \\
x_p &= \sqrt{\lambda_1} e_{p1} f_1 + \sqrt{\lambda_2} e_{p2} f_2 + \ldots + \sqrt{\lambda_p} e_{pp} f_p
\end{aligned}$$

요인이 원 변수만큼 존재하면 오차항 $\underline{\eta}$ 은 0 이고 오차항의 분산도 $\psi_j^2 = \sigma_{jj} - (l_{j1}^2 + l_{j2}^2 + \ldots + l_{jm}^2) = 0$. 만약 인자의 개수 $m(< p)$ 이 정해지면 나머지 요인 항은 오차항이 된다.

(2)최대 우도 추정법(Maximum Likelihood Method)

원 변수의 다변량 정규 분포를 따른다면 MLE 방법에 의해 L 과 $\underline{f}$ 을 다음 방법에 의해 구할 수 있다. f_j 와 η_j 가 서로 독립이고 x_j 가 다변량 정규 분포를 따른다고 가정하고 다음에 의해 L, Ψ 을 구한다,

$$\max_{L,\Psi} L(\underline{\mu}, \Sigma \mid \underline{x}) = L(\underline{\mu}, LL' + \Psi \mid \underline{x})$$

MLE 추정치의 해를 구하기 위해서는 반복 추정 과정을 거치게 된다. 이 경우 L 의 초기치로 다중 상관계수 제곱을 취하므로 0 과 1 사이의 값이다. MLE 방법에서는 큰 공통성을 가진 변수에는 큰 가중치를 주게 되므로 공통성의 추정치가 1 이상이 되는 Heywood 가 발생한다. 이 상황에서는 Ψ 의 추정치가 음이 된다. Heywood 상황이 발생하면 다른 추정 방법을 사용하기 바란다.

5.4. 맛보기

앞 절에서 살펴본 principal factoring 방법에 의해 요인(factor)을 구하는 방법을 보면 주성분 분석의 주성분을 구하는 방법과 요인 분석의 요인을 구하는 방법이 유사함을 알 수 있다. 요인과 주성분은 관계는 $f_i = \frac{y_i}{\sqrt{\lambda_i}}$ 이다.

주성분 분석은 원 변수의 변동을 잘 설명하는 주성분을 찾는 것이므로 원 변수 단위의 차이가 많지 않다면 공분산 행렬(S)을 이용하여 고유치, 고유 벡터를 구하는 것이 바람직하다. 요인 분석은 변수의 내재된 관계를 이용하여 변수를 분류하는 방법이므로 상관 행렬로부터 고유치, 고유 벡터를 구한다. 만약 주성분에서 상관 행렬로부터 고유치, 고유벡터를 구하면 요인 분석의 고유치, 고유 벡터가 동일하다. 지원자 데이터 예제 자료의 15 개 변수에 대해 주성분 분석(상관 행렬 이용)와 요인 분석 결과를 비교해 보자. SAS 는 방정식 해를 지정하지 않으면(default) 해를 구하는 방법으로 principal factoring 방법을 사용한다. 만약 원 변수들이 다변량 정규분포를 따르면 추정 방법으로 Maximum Likelihood 방법을 사용한다. PROC FACTOR DATA=APPLICANT METHOD=ML; 사용하면 된다.

```
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;
    VAR L--SU;
RUN;

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

The PRINCOMP Procedure

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

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8
L	0.162440	0.428846	0.315375	-.094347	0.114181	0.621238	0.171116	0.155368
ap	0.213108	-.035266	-.022878	0.262175	0.870203	-.037767	-.009998	-.009056
aa	0.040184	0.236919	-.430470	0.636274	-.212812	0.223389	0.310998	-.043216
li	0.225078	-.129796	0.465825	0.345375	-.128784	0.111960	-.130674	-.308273
sc	0.290481	-.248896	-.241026	-.172804	0.004916	0.019939	0.143008	0.386488
lc	0.314870	-.130990	-.150037	-.071033	-.206810	0.174643	-.514547	0.023612
ho	0.158117	-.405450	0.283928	0.416491	-.063642	-.303949	0.144365	0.343708
sm	0.324256	-.029492	-.185975	-.198227	0.037393	-.117958	0.010157	-.141580
ex	0.134068	0.553139	0.082591	0.067752	-.103091	-.367209	-.112752	0.584307
dr	0.315071	0.046243	-.079635	-.155987	-.200942	-.250153	0.489632	-.255634
am	0.318024	-.068155	-.208651	-.199291	0.163090	0.113408	0.201079	0.041316
gc	0.331497	-.023150	-.117142	0.074726	-.082414	0.147997	-.408244	0.106225
po	0.333289	0.022257	-.072544	0.188140	-.127323	0.058988	-.016186	-.149205
kj	0.259208	-.082272	0.467206	-.201376	-.111522	0.075376	0.246975	0.051937
SU	0.236037	0.420662	0.089152	-.019913	0.080921	-.414317	-.172807	-.382172

주성분 분석의 고유 벡터는 원 변수($\underline{x}$)로부터 주성분($\underline{y}$)을 구하기 위한 계수로 사용되어 $\underline{y} = L\underline{x}$ 가 되면 고유 벡터 내의 값의 크기를 이용하여 주성분 이름을 부여하게 된다. 결과는 주성분 결과와 동일하다.

The FACTOR Procedure

	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

Factor Pattern

	Factor1	Factor2	Factor3	Factor4
L	0.44527	0.61496	0.38052	-0.10326
ap	0.58416	-0.05057	-0.02760	0.28695
aa	0.11015	0.33974	-0.51939	0.69639
ii	0.61697	-0.18612	0.56205	0.37801
sc	0.79625	-0.35691	-0.29082	-0.18913
lc	0.86310	-0.18784	-0.18103	-0.07774
ho	0.43342	-0.58141	0.34258	0.45584
sm	0.88883	-0.04229	-0.22439	-0.21696
ex	0.36750	0.79319	0.09965	0.07415
dr	0.86365	0.06631	-0.09609	-0.17073
am	0.87175	-0.09773	-0.25175	-0.21812
gc	0.90868	-0.03320	-0.14134	0.08179
po	0.91359	0.03192	-0.08753	0.20592
kj	0.71052	-0.11798	0.56372	-0.22040
SU	0.64701	0.60322	0.10757	-0.02179

i번째 요인이 j번째 원 변수에 대한 Loading(부하) 값은 $f_{ij} = \sqrt{\lambda_i} e_{ij}$ 이다. 주성분 y_1의 원 변수 ap의 계수에 고유치 제곱근을 곱하면($\sqrt{7.5138} \times 0.2131$) 요인 1 의 ap 부하 값과 동일하다. 요인 분석의 고유 벡터는 $\underline{x} = L\underline{f} + \underline{\eta}$ 방정식의 L 이다. 각 값을 부하(Loading)라 한다. 부하의 의미는 공통 개념(요인)이 원 변수에 미치는 영향 정도를 나타내는 값이다. 그러므로 부하가 크다는 의미는 공통 개념에 의해 설명이 잘되고 있음을 의미한다.

요인 부하 값은 변수를 그룹화 하는데 사용된다. 요인 1 의 경우 부하 값이 상대적으로 큰 것을 묶으면 (빨간 표시 변수: 요인 1 이 원 변수에 영향을 많이 준다는 의미) 그 변수들이 같은 그룹(동일 개념을 측정)에 속할 수 있게 된다. (LC, SM, DR, AM, GC, PO) 변수를 설명하는 공통 개념은 무엇인가? 즉 요인 1 의 이름은? (LC=명석, SM=마케팅, DR=추진력, AM=야망, GC=개념 파악 능력, PO=잠재력)이 함께 있으므로 요인 1 은 마케팅 능력이라 할 수 있을 것이다.

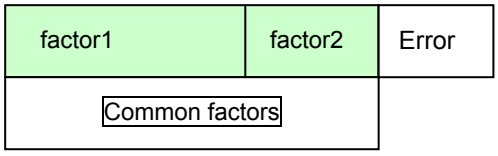
마케팅 능력(Marketing Ability)=(LC+SM+DR+AM+GC+PO)/6 (새로운 변수)

마케팅 능력이 뛰어난 사람을 뽑으려면 (LC, SM, DR, AM, GC, PO) 평균 점수가 높은 사람을 뽑으면 된다. 평균 점수를 사용하는 이유는 원 변수의 측정 단위와 맞추기 위함이다.

5.5. 요인 개수 구하기

부하(loading) 값의 의미는 각 요인이 원 변수를 설명하는 정도(크기)를 나타내며 요인은 변수들에 내재된 관계에서 공통 부분에 해당된다.

$$\begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_p \end{pmatrix} = \begin{pmatrix} l_{11}\, l_{12} \ldots l_{1m} \\ l_{21}\, l_{22} \ldots l_{2m} \\ \cdots \\ l_{p1}\, l_{p2} \ldots l_{pm} \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ \cdots \\ f_m \end{pmatrix} + \begin{pmatrix} \eta_1 \\ \eta_2 \\ \cdots \\ \eta_p \end{pmatrix}$$



그러므로 각 요인에서 부하 값의 절대값이 큰 것들만 (음의 부호는 동일 개념의 반대 척도) 선택하여 변수들을 그룹화 하면 된다. 요인의 개수는 다음 사항을 고려하여 결정한다.

(1)trivial 한 요인은 제외하자. 원 변수 1-2 개에만 부하 값이 큰 요인은 제외하자. 이 요인에 의해 묶을 수 있는 변수는 1-2 개이므로 그룹의 의미가 없기 때문이다.

(2)Kaiser 판단(가장 많이 이용)

변수들의 상관 관계가 0 이면(관계가 없으면) 상관 계수 행렬은 R 은 항등 행렬 I 이다 이 경우 원 변수의 개수와 주성분의 개수가 같아지고 주성분의 분산은 모두 1 이므로 각 주성분이 가지는 분산 평균도 1 이다. 그러므로 상관 계수 행렬로부터 구한 고유치가 평균인 1 이상인 되어야 한다는 판단 하에 고유치가 1 이상인 것만으로 요인의 개수를 정한다. SAS 도 이 방법에 의해 요인의 개수 출력한다. (5.4 절 예제에서 요인은 4 개만 출력되었다.)

(3)SCREE 그림 사용

주성분 방법에서 사용되었던 SCREE 그림(페이지 47)을 사용하여 인자의 개수를 예상한다. 총변동 80%에 연연하지 말고 주성분 분산 설명 변동의 크기(고유치)가 갑자기 줄어들기 바로 전까지의 개수로 적절한 인자 개수로 사용하면 된다. APPLICAT 예제 데이터 경우 Kaiser 판단에 의하면 4 개 필요하였지만 고유치가 7.51➔2.05 ➔1.46 ➔1.19 로 떨어지므로 인자는 1 개 혹은 2 개로 하면 된다.

(4)Large-sample Test(χ^2 -검정)

MLE 방법에 의해 요인 방정식 해를 구하는 경우 요인 개수를 결정하기 위한 검정 방법으로 χ^2 -적합성 검정을 실시한다. 원 변수의 개수를 p , 적절한 요인의 개수를 m 이라 하자.

①귀무가설

$$H_0 : \Sigma_{p\times p} = L_{p\times m} L'_{m\times p} + \Psi_{p\times p}$$

②대립 가설:

Σ는 임의의 양정치(positive definite) 행렬

$S_n = \frac{(n-1)S}{n}$ 라 하면 $-2\ln L = -\ln \frac{\max L \, under \, H_0}{\max \mathrm{L}} = n\ln[\frac{|\hat{\Sigma}|}{|S_n|}] \sim (app)\chi^2$ 을 이용하여 요인 개수를 정한다. 사용 방법은 5.9 절에서 예제 중심으로 살펴보기로 하자.

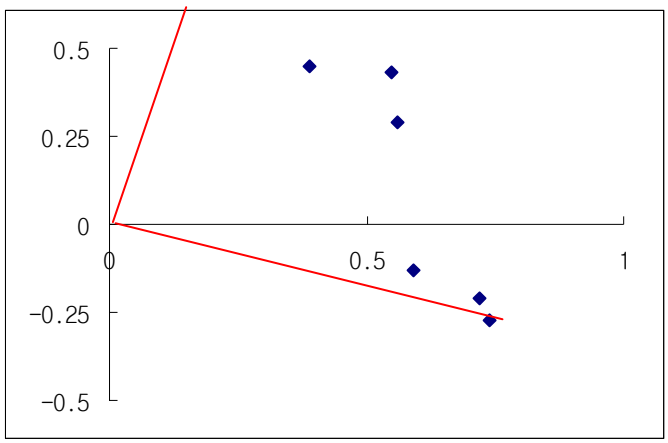
5.6. 요인 회전

요인 분석에서 요인의 부하 값은 요인(공통 개념)과 원 변수의 상관 관계 정도를 나타내는 크기로 해석될 수 있으므로 부하 값에 의해 원 변수를 그룹화 한다. 그러나 (1)요인의 복합성: 하나의 원 변수에 부하 값이 큰 요인이 2 개 이상 존재하거나 (2)인자의 크기가 0 을 중심으로 ±의 작은 값이 있는 경우 부하 값으로 변수를 그룹화 하는 것은 불가능하다. 요인 회전은 각 요인이 상대적으로 큰 부하 값을 갖도록 요인을 회전(rotate)하는 것으로 QUARTIMAX rotation, OBLIQUE rotation, PROMAX rotation 방법이 있는데 가장 많이 사용되는 것은 직교 회전 방법인 VARIMAX 방법이다. VARIMAX 방법은 Kaiser 가 제안한 것으로 간단한 구조의 측정치로 요인 행렬의 각 열 내의 부하 제곱의 분산의 합을 제안하고 이 분산을 최대화 하는 회전 방법이다.

요인 회전이 가능한 것은 앞에서 언급하였듯이 인자의 개수 m 가 원 변수의 개수 p 보다 적은 경우 $\Sigma = LL' + \Psi$ 을 만족하는 행렬 L 은 무수히 많이 존재한다. 이 성질로 인하여 요인의 회전이 가능하게 된다. 부하의 값들이 잘 구별되도록 요인을 회전하여도 요인 방정식을 만족하는 해가 존재하는 것이다. 즉, $\Sigma = LL' + \Psi$ 을 만족하는 행렬을 L 이라 하면 직교 변환 $L^* = LP$ (P 는 직교 행렬)도 $\Sigma = LL' + \Psi$ 을 만족한다.

다음은 두 요인(f_1, f_2)의 부하 값의 산점도를 그린 것이다. 빨간 선은 축을 오른쪽으로 20^o 회전한 것이다.

변수	요인1	요인2
X1	0.55	0.43
X2	0.56	0.29
X3	0.39	0.45
X4	0.74	-0.27
X5	0.72	-0.21
X6	0.59	-0.13



5.7. 요인 분석 예제

5.7.1. 자료 설명 [POLICE.txt]

경찰에 지원한 50 명의 신체적 특성 15 개를 측정한 것이다. [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: 런닝 머신에서(treadmill) 제자리 달리고 5 분 후 맥박

SPEED: 런닝 머신에서 제자리 달리기 최대 속력

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

```
DATA POLICE;
     INFILE "C:\TEMP\POLICE.TXT";
     INPUT ID REACT HEIGHT WEIGHT SHLDR PELVIC CHEST THIGH
           PULSE DIAST CHNUP BREATH RECVR SPEED ENDUR FAT;
RUN;
```

5.7.2. SAS 프로그램

다음 프로그램은 요인 방정식 해를 구하는 방법으로는 principal factoring 방법을 사용하였고 (default) 요인 회전 방법은 VARIMAX 방법을 사용한 예이다. (가장 일반적인 방법이다)

```
PROC FACTOR DATA=POLICE ROTATE=VARIMAX;
VAR REACT HEIGHT WEIGHT SHLDR PELVIC CHEST THIGH
    PULSE DIAST CHNUP BREATH RECVR SPEED ENDUR FAT;
RUN;
```

아래 프로그램은 위의 프로그램과 동일하다. REACT 변수부터 FAT 변수까지 모두 사용하므로……

```
PROC FACTOR DATA=POLICE ROTATE=VARIMAX;
VAR REACT--FAT;
RUN;
```

원 변수가 다변량 정규 분포를 따른다는 가정 하에 요인 방정식 해를 구하는 방법인 Maximum Likelihood 방법을 사용 할 경우 프로그램은 다음과 같다.

```
PROC FACTOR DATA=POLICE ROTATE=VARIMAX METHOD=ML;
```

5.7.3. 요인 개수

Eigenvalues of the Correlation Matrix: Total = 15 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	5.21852549	2.81172746	0.3479	0.3479
2	2.40679803	1.09411978	0.1605	0.5084
3	1.31267825	0.08160302	0.0875	0.5959
4	1.23107523	0.02722958	0.0821	0.6779
5	1.20384565	0.35594143	0.0803	0.7582
6	0.84790423	0.14315618	0.0565	0.8147
7	0.70474804	0.12633945	0.0470	0.8617
8	0.57840860	0.18486182	0.0386	0.9003
9	0.39354677	0.02535508	0.0262	0.9265
10	0.36819169	0.04160632	0.0245	0.9510
11	0.32658537	0.13971542	0.0218	0.9728
12	0.18686996	0.04806356	0.0125	0.9853
13	0.13880640	0.09492503	0.0093	0.9945
14	0.04388137	0.00574644	0.0029	0.9975
15	0.03813492		0.0025	1.0000

5 factors will be retained by the MINEIGEN criterion.

(1)PROC FACTOR 라인에 COVARIANCE 옵션을 따로 사용하지 않으면 default 로 상관 계수 행렬을 이용하여 고유치를 구한다. 상관 행렬을 사용하면 고유치가 설명하는 분산 변동의 합은 변수 개수 p (즉 위의 예제에서 고유치의 합은 15 이다)이고 각 고유치의 평균 설명력은 1 이다.

(2)공통 인자의 개수 선택? Kaiser 제안 방법을 가장 많이 사용한다. SAS 도 default 로 고유치가 1 이상인 요인만을 출력한다. 공통 인자는 변수들의 내재된 관계를 설명하는 것이고 고유치는 각 요인들의 설명력에 해당되는 것이므로 평균 설명력이 1 이상인 것을 선택한다.

(3)(MINEIGEN= minimum eigen value)의 의미는 다른 옵션(NFACTOR=)이 설정되지 않았으므로 default 로 고유치가 1 이상인 요인들만 출력한다는 의미이다. 이렇게 선택하면 대략적으로 전체 변동의 80%가 설명된다. 그러나 앞에서도 언급하였듯이 이것에 얽매일 필요 없다. [주관적 판단 가능] 고유치가 1 이상인 5 개 요인만 선택되었다. 이것은 15 개의 변수를 5 개의 그룹으로 나눌 수 있다는 의미이고 나누는 방법은 요인의 부하 값의 크기로 나눈다.

(4)요인 패턴은 (요인 부하 값) 특별한 옵션이 없으면 크기가 1 이상인 고유치 개수만큼만 출력된다. 그러므로 출력 결과를 보면 요인 부하 값이 5 개만 출력되어 있다. 만약 요인 패턴을 원하는 만큼 출력하려면 다음과 같이 적어 주어야 한다.

```
PROC FACTOR DATA=POLICE ROTATE=VARIMAX NFACTORS=10;
```

5.7.4. 부하 값

$\begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_p \end{pmatrix} = \begin{pmatrix} l_{11}\, l_{12} \cdots l_{1m} \\ l_{21}\, l_{22} \cdots l_{2m} \\ \cdots \\ l_{p1}\, l_{p2} \cdots l_{pm} \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ \cdots \\ f_m \end{pmatrix} + \begin{pmatrix} \eta_1 \\ \eta_2 \\ \cdots \\ \eta_p \end{pmatrix}$ 에서 $\begin{pmatrix} l_{11}\, l_{12} \cdots l_{1m} \\ l_{21}\, l_{22} \cdots l_{2m} \\ \cdots \\ l_{p1}\, l_{p2} \cdots l_{pm} \end{pmatrix}$ 이 부하 값이다.

Principal factoring 방법에 의해 구한 요인의 부하 값은 상관 계수 행렬로부터 구한 고유 벡터(e_i : 주성분 분석의 고유 벡터)에 $\sqrt{\lambda_i}$ 곱하여 얻는다. 각 요인에 의해 설명된 원 변수 변동은 고유치와 동일하다. 요인 1 의 원 변수 변동 설명 비율은 5.21 인데 이는 가장 큰 고유치와 동일하다. 각 요인에 의해 설명된 원 변수 변동을 합하면 변수의 개수인 $p = 15$ 이다. (이는 상관 계수 행렬을 사용했기 때문이다)

Factor Pattern

	Factor1	Factor2	Factor3	Factor4	Factor5
REACT	0.11577	0.23649	0.12762	0.06168	0.90082
HEIGHT	0.69783	-0.33725	0.41902	0.05972	0.22820
WEIGHT	0.95187	-0.04446	-0.07678	0.10093	-0.06101
SHLDR	0.68565	-0.32752	0.31506	0.11806	-0.23196
PELVIC	0.67123	-0.29937	0.08798	0.48711	-0.10913
CHEST	0.82416	0.00558	-0.14664	0.15902	-0.18894
THIGHP	0.64905	0.47592	-0.26352	-0.22796	0.01317
ULSE	-0.27258	0.59160	0.52991	0.02023	-0.00286
DIAST	-0.08173	0.42916	-0.04908	0.77472	0.05289
CHNUP	-0.66679	-0.36578	0.27478	0.24007	-0.12565
BREATH	0.57632	-0.05036	0.50353	-0.17149	0.17768
RECVR	-0.05822	0.65763	0.45810	-0.11492	-0.41729
SPEED	-0.06704	-0.76236	0.03135	-0.21987	0.04457
ENDUR	-0.46683	-0.08312	-0.14959	0.36143	0.09197
FAT	0.84105	0.34734	-0.26738	-0.07236	0.01546

Variance Explained by Each Factor

Factor1	Factor2	Factor3	Factor4	Factor5
5.2185255	2.4067980	1.3126783	1.2310752	1.2038457

5.7.5. 공통성

$Var(x_j) = \sigma_{jj} = l_{j1}^2 + l_{j2}^2 + ... + l_{jm}^2 + \Psi_j = \sum_{k=1}^{m} l_{jk}^2 + \psi_j$ 쓸 수 있는데 $\sum_{k=1}^{m} l_{jk}^2$ 을 공통성(communality)이라 하고 ψ_j 는 특정(specific) 분산이라 한다. 앞의 예제에서 요인이 5 개 선택되었으므로 $m = 5$ 이고 REACT 변수의 공통성 값은 REACT 의 요인 1~요인 5(5 개가 선택되었으므로) 부하 값의 제곱합이다. 즉, $0.11577^2 + 0.23649^2 + 0.12762^2 + 0.06168^2 + 0.90082^2 = 0.9009$ 이다. 다음 그림의 공통성 출력 결과와 동일하다.

1(100%)에 가까운 값이면 그 변수의 변동이 선택된 요인에 의해 거의 모두 설명 된다는 의미이고 낮으면 다른 요인이 존재한다는 것이다. SPEED 나 ENDUR 변수를 제외하고는 우리가 선택한 5 개의 요인(고유치가 1 이상)에 의해 변동이 80% 이상이다. 특히 ENDUR 변수는 5 요인들에 의해 설명되는 정도가 낮은데 이 변수는 아마 5 개의 그룹 어디에도 포함되지 않을 가능성이 높다.

Final Communality Estimates: Total = 11.372923

REACT	HEIGHT	WEIGHT	SHLDR	PELVIC
0.90090440	0.83192324	0.92783038	0.74439622	0.79709996

CHEST	THIGHP	ULSE	DIAST	CHNUP
0.76176311	0.76936042	0.70551551	0.79625170	0.72733481

BREATH	RECVR	SPEED	ENDUR	FAT
0.64920781	0.83305264	0.63699092	0.38630835	0.90498318

5.7.6. 변수 그룹 하기 (Rotate 이용)

요인의 개수는 원 변수의 개수(p)만큼 존재하지만 Kaiser 규칙에 의해 고유치가 1 이상인 요인만 선택하면 요인의 개수는 줄어들게 되고(이를 m 이라 하자) 원 변수의 변동은 선택된 요인에 의해 설명되는 공통성 부분과 나머지 부분은 specific 분산인 오차로 나뉘어진다.

$$\begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_p \end{pmatrix} = \begin{pmatrix} l_{11}\, l_{12} \ldots l_{1m} \\ l_{21}\, l_{22} \ldots l_{2m} \\ \cdots \\ l_{p1}\, l_{p2} \ldots l_{pm} \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ \cdots \\ f_m \end{pmatrix} + \begin{pmatrix} \eta_1 \\ \eta_2 \\ \cdots \\ \eta_p \end{pmatrix}$$

요인 분석은 원 변수를 $\underline{x} = L\underline{f} + \Psi$ 으로 나타낼 수 있는 $L, \underline{f}$ 를 구하는데 행렬 L 은 부하 값 행렬이라 하고 벡터 $\underline{f}$ 는 요인 벡터이다. 요인의 부하 행렬 L 은 상관 계수 행렬의 고유 벡터로부터 구하고 $\underline{f}$ 는 주성분 변수를 이용하여 구하게 되는데 부하 값에 의해 원 변수를 분류할 수 있다.

원 변수($x_1, x_2, \ldots, x_p$)는 선택된 부하($f_1, f_2, \ldots, f_m$)의 선형 결합으로 표현되므로 부하는 선형 계수에 해당한다. 그러므로 원 변수가 표현될 때 요인의 부하가 크다는 것은 그 요인에 의해 가장 영향을 많이 받는 것을 의미한다. 요인은 원 변수의 내재된 관계를 설명하는 변수이므로 각 요인의 부하 값이 큰 원 변수를 묶을 수 (그룹) 있다는 것을 의미한다.

요인을 회전하는 방법을 사용하면 부하 값의 크기를 쉽게 비교할 수 있다. 가장 많이 사용되는 VARIMAX 방법을 사용하여 설명하겠다. REORDER 옵션은 부하의 크기 순으로 정렬하여 주므로 변수 분류에 용이하다.

```
PROC FACTOR DATA=POLICE ROTATE=VARIMAX REORDER;
VAR REACT--FAT;
RUN;
```

Rotated Factor Pattern

	Factor1	Factor2	Factor3	Factor4	Factor5
FAT	0.89834	0.30408	-0.01710	0.05600	0.04550
THIGH	0.86470	0.07378	0.11051	-0.02796	0.05664
CHEST	0.60652	0.57244	-0.14309	0.11808	-0.17828
ENDUR	-0.38966	-0.26455	-0.16683	0.36931	0.01610
CHNUP	-0.83023	-0.10598	0.00362	0.07366	-0.14626
HEIGHT	0.11446	0.82407	-0.09857	-0.20697	0.29527
SHLDR	0.14604	0.82138	-0.04338	-0.13253	-0.17015
PELVIC	0.16044	0.79465	-0.23908	0.26790	-0.10469
WEIGHT	0.65304	0.68489	-0.17334	0.02459	-0.04056
BREATH	0.19065	0.60670	0.20462	-0.32989	0.30673
RECVR	0.10678	-0.04500	0.88357	0.01179	-0.19694
PULSE	-0.14414	-0.12994	0.78471	0.11661	0.19618
SPEED	-0.38288	0.16941	0.49297	-0.46286	-0.06663
DIAST	-0.01615	0.01011	0.18516	0.86776	0.09270
REACT	0.11922	-0.00396	-0.00664	0.11263	0.93485

(1)직교 변환된 요인들을 이용하여 변수들을 묶을 수 있는데 묶여진 변수 그룹에 적절한 이름을 부여하기 위해서는 변수에 대한 지식과 경험이 필요하다.

(2)각 요인 내에 부하 값 크기 순으로 정렬될 뿐 아니라 각 요인 별로 크로스 체크하여 가장 큰 요인에 넣는다. 변수 WEIGTH 의 경우 요인 1(Factor1)의 0.65 보다 요인 2 의 0.68 이 크므로 요인 2 에서 크기 순서대로 정렬되어 있다. 변수 WEIGTH 는 요인 1 에 의해 FAT, THIGH 등과 같이 분류해도 무방하다.

(3)(FAT 비만도, THIGHP 허벅지두께, CHEST 가슴둘레, CHNUP 턱걸이) 하나의 그룹으로 묶을 수 있다. 이 그룹 이름을 부여하기 어렵지 않을 것이다. 몸집 비대(obesity)면 적당하다. 그럼 턱걸이 변수는? 몸이 비대할수록 턱걸이 회수는 줄어들 것이므로 부호가 음이다. 몸집 비대 그룹에서는 다른 변수들과 반대 특성을 측정하고 있음을 알 수 있다. 여전히 같은 그룹에 묶이기는 한다. 요인 분석 결과 하나의 그룹으로 묶인 변수들의 평균을 구할 때는 부하 값이 –인 변수는 –값을 이용한다. 즉 (FAT+THIGH+CHEST-CHNUP)/4 가 새로운 변수(비만도 지수)가 된다. 평균을 사용하는 이유는 원 변수와 단위를 맞추기 위해서이다.

(4)한편 ENDURE 변수는 –0.38966 으로 절대 크기가 낮음에도 불구하고 요인 1 에 의해 분류된 것처럼 보이는 것은 다른 요인(요인 2~요인 5)들의 부하 값에 비해 크기 때문이다. 그러나 공통성 값에서 알 수 있듯이 ENDURE 변수의 변동이 요인 1~요인 5 에 의

해 38% 밖에 설명되지 않으므로 당연하다. ENDURE 변수는 요인 1~요인 5 에 의해 분류되지 못한 변수이다.

(5)(HEIGHT 키, SHLDR 어깨 넓이, PELVIC 골반 넓이, WEIGHT 몸무게, BREATH 폐활량) 하나의 그룹으로 묶고 신체 골격 구조 변수라 이름 붙일 수 있다. 몸무게 요인 1 에서 부하 값이 0.65 이고 요인 2 에서는 0.68 로 서로 비슷하므로 이 그룹에 묶어도 되고 요인 1 에 묶어도 된다. (앞에서 언급)

(6)WEIGHT 변수는 요인 1 에서도 0.65 로 부하 값이 크므로 그룹 이름 붙이기 적당한 요인 1 에 넣어도 무방하다. 그러면 요인 2 그룹에서는 빠진다.

(7)(PULSE 맥박, RECVR 런닝 머신에서 제자리 달리고 5 분 후 맥박)은 심장 지구력이라 할 수 있다.

(8)DIAST 심장 혈압, REACT 반응 시간은 각각 분류된다. 하나씩 분류되므로 요인 4-5 는 변수 분류에 의미가 없다. 요인 3 은 두 변수만 묶이므로 요인 2 개 정도면 족하지 않을까? 페이지 82 에서 언급한 것처럼 변수가 1-2 개 정도 묶이는 요인을 제외한다면 15 개 변수를 요인 분석에 의해 그룹화한 결과는 다음과 같다. 나머지 원 변수들은 한 개씩 개별적으로 사용하게 된다.

▪비만 변수=(FAT 비만도, THIGHP 허벅지두께, CHEST 가슴둘레, CHNUP 턱걸이)
▪신체 골격=(HEIGHT 키, SHLDR 어깨 넓이, PELVIC 골반 넓이, WEIGHT 몸무게, BREATH 폐활량)

(9)요인 분석 후 변수는 비만 변수(FAT_INDEX), 신체 골격 변수(BODY), 나머지 6 개 개별 원 변수 8 개가 된다. 15 개 원 변수를 개별적으로 사용하여 2 차 분석(회귀 분석, 분산 분석, 판별 분석 등)을 실시해도 되나 향후 분석에서는 8 개의 변수를 이용하여 분석을 하는 것이 요인 분석을 사용하는 이유이다. 묶이는 문항은 변수들의 평균을 계산하여 새로운 변수로 사용하면 된다. 합을 이용하는 것보다는 평균을 이용하는 것이 단위 동일화 면에서 유리하다. CHNUP 변수의 부하 값은 음이므로 값에 –을 붙여준 후 평균을 구해야 한다.

```
DATA POLICE1;
     SET POLICE;
     CHNUPO=-CHNUP;
     FAT_INDEX=MEAN(FAT,THIGH,CHEST,CHNUP);
     BODY=MEAN(HEIGHT,SHLDR,PELVIC, WEIGHT, BREATH);
RUN;
```

회전된 요인에 의해 원 변수의 변동을 설명하는 부분이 회전되지 않은 경우와 다소 다르다. 이는 축을 회전했기 때문이다. 그러나 선택된 요인들에 의해 설명되는 각 변수의 공통성은 같다. ROTATE 옵션을 사용하지 말고 요인 분석을 실시하여 결과를 비교해보자.

Variance Explained by Each Factor

Factor1	Factor2	Factor3	Factor4	Factor5
3.4799455	3.3769250	1.8753118	1.3949543	1.2457860

Final Communality Estimates: Total = 11.372923

REACT	HEIGHT	WEIGHT	SHLDR	PELVIC
0.90090440	0.83192324	0.92783038	0.74439622	0.79709996

CHEST THIGHP ULSE DIAST CHNUP

5.7.7. 부하 값 산점도 그리기

다음 프로그램은 요인 분석의 통계량 값을 OUTSTAT 에 의해 F_STAT 이름의 SAS data 에 저장하고 출력한 것이다.

```
PROC FACTOR DATA=POLICE ROTATE=VARIMAX REORDER OUTSTAT=F_STAT;
    VAR REACT HEIGHT WEIGHT SHLDR PELVIC CHEST THIGHP
    ULSE DIAST CHNUP BREATH RECVR SPEED ENDUR FAT;
RUN;

PROC PRINT DATA=F_STAT;
RUN;
```

F_STAT 에서 _TYPE_이 UNROTATE 는 회전되지 않은 요인, PATTERN 은 회전된 요인의 부하 값이다.

Obs	_TYPE_	_NAME_	REACT	HEIGHT	WEIGHT	SHLDR	PELVIC
31	TRANSFOR	Factor5	0.0457	0.152	0.2296	-0.0108	0.9602
32	PATTERN	Factor1	0.1192	0.114	0.6530	0.1460	0.1604

필요한 데이터를 SUBSET 하고 전치를 한다. _TYPE_="PATTERN" ➔ 데이터 이름: TEMP

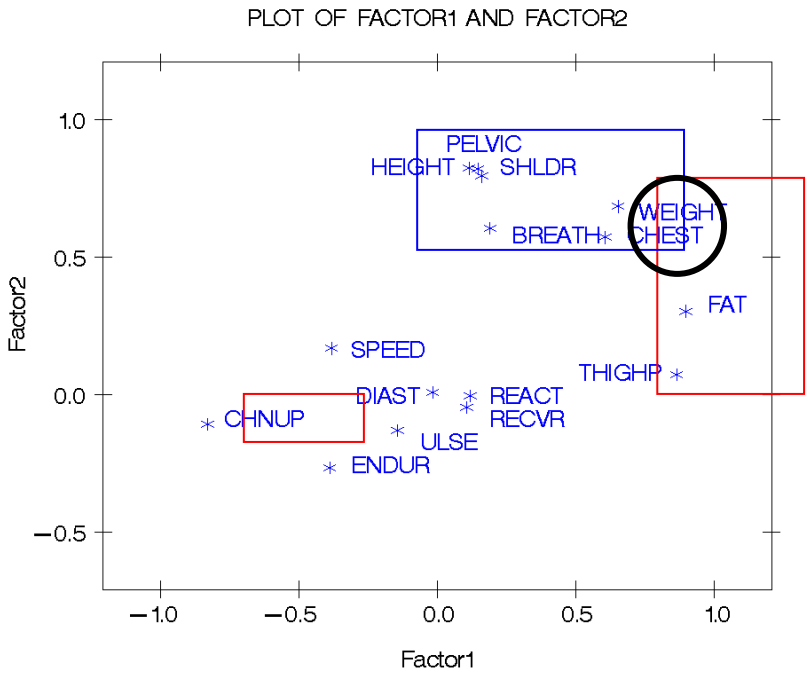
```
DATA TEMP;
     SET F_STAT;
     IF (_TYPE_='PATTERN');
  RUN;

PROC TRANSPOSE DATA=TEMP OUT=TEMPO;
RUN;
PROC PRINT DATA=TEMPO;
RUN;
```

Obs	_NAME_	Factor1	Factor2	Factor3	Factor4	Factor5
1	REACT	0.11922	-0.00396	-0.00664	0.11263	0.93485
2	HEIGHT	0.11446	0.82407	-0.09857	-0.20697	0.29527
3	WEIGHT	0.65304	0.68489	-0.17334	0.02459	-0.04056
4	SHLDR	0.14604	0.82138	-0.04338	-0.13253	-0.17015

요인 1 과 요인 2 의 부하 값에 대한 산점도를 그리면 다음과 같다. 요인 3 은 원 변수 2 개, 요인 4 와 요인 5 는 각 한 개만 분류되므로 요인을 2 개로만 선택하는 것이 적당하다. 산점도 그리는 프로그램의 MACRO 함수 PLOTIT 을 사용하자. 까만 원 부분은 요인 1 이나 요인 2 의 어느 쪽에도 분류될 수 있는 변수를 나타낸 것이다.

```
TITLE H=1 "PLOT OF FACTOR1 AND FACTOR2";
%PLOTIT(DATA=TEMPO, LABELVAR=_NAME_,
    PLOTVARS=FACTOR2 FACTOR1, COLOR=BLACK, COLORS=BLUE);
RUN;
```



5.8. 요인 점수(Factor score)

요인 분석은 $x_1, x_2, ..., x_p$ 의 변수를 공통인자인 요인을 이용하여 변수를 그룹화 하는 방법이다.

$$\underline{x} = L\underline{f} + \underline{\eta} \leftrightarrow \begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_p \end{pmatrix} = \begin{pmatrix} l_{11}\, l_{12} \ldots l_{1m} \\ l_{21}\, l_{22} \ldots l_{2m} \\ \cdots \\ l_{p1}\, l_{p2} \ldots l_{pm} \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ \cdots \\ f_m \end{pmatrix} + \begin{pmatrix} \eta_1 \\ \eta_2 \\ \cdots \\ \eta_p \end{pmatrix}$$

위의 식을 살펴 보면 원 변수들을 공통 인자 $\underline{f}$ 의 선형 결합으로 표현한 형태이지만 이를 역으로 생각하면 공통인자 $\underline{f}$ 는 변수들의 선형 결합으로 표현할 수 있을 것이다. 요인분석은 변수의 개수보다 적은 개수의 공통 인자가 구해지므로 공통인자를 변수들의 결합으로 표현할 수 있다는 것은 변수를 줄일 수 있음을 의미한다. 각 개체의 요인 값을 요인 점수라 (인자 점수: factor score) 한다. 이 점수는 주성분 점수와 매우 유사하다. 이 요인 점수를 이용하여 이차 분석을 (그룹 변수를 이용하여 개체를 분류하는데 사용) 시행하면 된다. 주성분은 변수들의 선형 결합이므로 고유 벡터에 의해 바로 계산할 수 있으나

요인분석의 경우는 오차항이 $\underline{\eta}$ 있으므로 요인 점수는 바로 계산될 수 없어 다음 2 가지 방법을 사용한다.

5.8.1. Bartlett's Method (Weighted Least Square Method)

r 번째 관측치에 대한 표준화를 $\underline{z}_r = (\underline{x}_r - \underline{\mu})$ 라 하자. $(\underline{z}_r - \hat{L}\underline{f}_r)\hat{\psi}^{-1}(\underline{z}_r - \hat{L}\underline{f}_r)$ 를 최소화하는 $\underline{f}$ 를 구하면 이것이 r 번째 개체의 요인 점수이다. $\underline{f}_r = (\hat{L}\hat{\psi}^{-1}\hat{L})^{-1}\hat{L}\hat{\psi}^{-1}\underline{z}_r$

5.8.2. Thompson's Method (Regression Method)

$$\begin{bmatrix}\underline{z}\\ \underline{f}\end{bmatrix} \sim N\left(\begin{bmatrix}\underline{0}\\ \underline{0}\end{bmatrix}, \begin{bmatrix}P & L\\ L' & \mathrm{I}\end{bmatrix}\right) \;\rightarrow\; E(\underline{f}\mid\underline{z}) = L'P^{-1}\underline{z} \;\rightarrow\; \underline{f}_r = L'R^{-1}\underline{z}_r$$

5.8.3. SAS 이용하기

(1)Default 옵션은 Regression 방법에 의해 요인 점수 계수를 얻는다. SCORE 옵션을 사용하면 인자 점수를 구할 때 사용되는 계수가 출력된다.

```
PROC FACTOR DATA=POLICE SCORE OUT=SCORE1 NFACTORS=2 ROTATE=VARIMAX;
VAR REACT--FAT;
RUN;

PROC PRINT DATA=SCORE1;
RUN;
```

(2)요인의 개수를 반드시 지정해 주어야 한다. (NFACTORS=2) 요인의 개수를 2 로 지정해 주었으므로 요인 5 개를 사용했을 때의 부하 값과 동일하지 않다. 그러나 각 요인에 의해 설명되는 원 변수 변동은 고유치의 크기와 같으므로 동일하다.

(3)요인을 2 개만 사용하였으므로 묶이는 변수는 다소 다르다. 요인 1 은 부하 값이 .8 이상인 변수에 비해 다소 떨어지므로 WEIGHT, FAT, CHEST 변수와 함께 묶어도 되고 아니면 따로 사용해도 된다.

Rotated Factor Pattern

	Factor1	Factor2
WEIGHT	0.95012	-0.07276
FAT	0.85101	0.32217
CHEST	0.82396	-0.01894
HEIGHT	0.68749	-0.35786
SHLDR	0.67560	-0.34777
THIGH	0.66293	0.45640
PELVIC	0.66203	-0.31921
BREATH	0.57457	-0.06749
ENDUR	-0.46910	-0.06919
CHNUP	-0.67738	-0.34578
RECVR	-0.03863	0.65907
PULSE	-0.25486	0.59945
DIAST	-0.06893	0.43141
REACT	0.12276	0.23294
SPEED	-0.08969	-0.76002

Variance Explained by Each Factor

Factor1	Factor2
5.2160368	2.4092867

(4) $\underline{f}_r = L'R^{-1}\underline{z}_r$ 에서 계수 부분 $L'R^{-1}$ 표준화 점수 계수 (standardized scoring coefficient)가 다음과 같이 출력된다. 이를 이용하여 요인 점수가 계산되고 OUT 옵션에 의해 SAS data 로 저장된다.

Standardized Scoring Coefficients

	Factor1	Factor2
WEIGHT	0.18177	-0.02389
FAT	0.16539	0.13946
CHEST	0.15793	-0.00238
HEIGHT	0.12949	-0.14404
SHLDR	0.12728	-0.13993
THIGH	0.13020	0.19395
PELVIC	0.12487	-0.12816
BREATH	0.10977	-0.02420
ENDUR	-0.09044	-0.03186
CHNUP	-0.13224	-0.14811
RECVR	-0.00302	0.27345
PULSE	-0.04490	0.24725
DIAST	-0.01035	0.17870
REACT	0.02510	0.09756
SPEED	-0.02226	-0.31623

(5)인자 점수를 출력하려면 일단 SAS data 로 만들어야 하는데 이 옵션이 OUT 이라는 것이다. 위의 예제에서는 인자 점수들이 SCORE1 이라는 SAS data 에 저장되므로 PRINT procedure 사용하여 출력하면 된다. 인자 점수를 얻으려면 반드시 NFACTOR 라는 옵션과 같이 사용해야 한다. 이 옵션은 인자의 수를 지정해 주는 것이다. 인자 점수 SAS data 를 원하지 않으면 이 옵션을 사용하지 않아도 SAS 가

자동으로 고유치가 1 이상인 요인만 출력하게 되지만 인자 점수 SAS data 를 원하면 반드시 같이 사용해야 한다.

Obs	CHNUP	BREATH	RECVR	SPEED	ENDUR	FAT	Factor1	Factor2
1	2	158	108	5.5	4.0	11.91	-0.19401	-0.39645
2	20	166	108	5.5	4.0	3.13	-1.67745	-0.44907
3	7	167	116	5.5	4.0	16.89	-0.65264	1.53083

(7)인자 점수가 필요한가?

인자 점수를 주성분 점수와 유사한 개념으로 사용할 수 있다. 즉 (요인 1, 요인 2)를 주성분 y_1, y_2의 이용 방법과 동일하게 활용할 수 있다. 그러나 요인 분석의 주 목적은 변수 분류에 있으므로 굳이 요인 점수까지 구할 필요는 없으며 요인 점수를 사용하기 보다는 주성분 점수를 사용하여 2 차 분석(회귀 분석, 다중 공선성 해결, 판별 분석 등)을 하는 것이 일반적이다.

5.9. Comment

5.9.1. 요인 개수 검정

요인 분석에서 인자는 고유치가 1 이상인 요인만 선택되고, 각 요인들의 부하 값에 의해 원 변수들이 분류된다. 경찰 데이터 예제에서 고유치가 1 이상인 인자는 5 개 있으므로 요인이 5 개 선택되었다. EDURE 변수의 경우는 5 개의 요인 어느 것에도 공통 인자로 역할을 하지 못하므로 더 많은 요인이 필요한지를 알아볼 필요가 있다. 요인 개수에 대한 검정 방법으로는 χ^2-검정 방법을 사용하게 된다. 그러므로 우도 함수가 정의될 수 있어야 하므로 원 변수가 정규 분포를 따른다는 가정 하에서 요인 방정식을 구하는 ML (Maximum Likelihood) 방법에 의해서만 가능하다.

```
PROC FACTOR DATA=POLICE ROTATE=VARIMAX NFACTOR=5 METHOD=ML HEYWOOD;
    VAR REACT--FAT;
RUN;
```

HEYWOOD 옵션을 사용하는 이유는 공통성 값이 1 이상인 되는 경우를 방지하기 위한 것이다. ML 방법을 이용하여 추정하는 경우 반드시 함께 사용해야 한다. (앞에서 언급)

Significance Tests Based on 50 Observations

Test	DF	Chi-Square	Pr > ChiSq
H0: No common factors HA: At least one common factor	105	473.1958	<.0001
H0: 5 Factors are sufficient HA: More factors are needed	40	53.7839	0.0714

귀무가설을 (5 개 요인이면 충분하다) 기각하지 못하므로 5 개 요인으로 충분하다. 이는 고유치가 1 이상인 것만 고른 경우와 대부분 일치하므로 자주 사용되지 않는다.

반대로 요인의 수를 임의로 줄이려 한다면 어떻게 할 것인가? 예를 들어 경찰 자료의 경우 요인 5 개가 적정 수준이지만 2 개로 줄일 수 있는가 검정하려 한다면 다음과 같이 분석하면 된다.

```
PROC FACTOR DATA=POLICE SCORE OUT=SCORE1 ROTATE=VARIMAX
            NFACTORS=2  METHOD=ML HEYWOOD;
    VAR REACT--FAT;
RUN;
```

Test	DF	Chi-Square	Pr > ChiSq
H0: No common factors HA: At least one common factor	105	473.1958	<.0001
H0: 2 Factors are sufficient HA: More factors are needed	76	140.4278	<.0001

ML 방법에 의해 요인의 수에 대한 검정을 실시하면 귀무가설(요인이 2 개이면 충분)이 기각된다. 그러므로 요인이 2 개로 충분하지 않다. NFACTORS=3, 4 로 변경하면서 최적 요인 개수를 찾으면 5 개로 Kaiser 규칙에 의해 찾는 것과 동일하다. 일반적으로 대부분의 경우 이와 같이 요인 개수의 검정은 실효성이 없다.

5.9.2. 상관 계수와 요인 분석 관계

변수를 그룹화 한다는 것은 결국은 변수들간의 상관 관계 정도가 큰 것끼리 묶는다는 것을 의미한다. 그러나 상관 계수만 의지하여 묶다 보면 겹치는 부분이 많아 어려움이 있다. 이에 대한 해결책이 바로 요인 분석이고 요인 분석도 시작은 상관 계수 행렬이다. 다음은 POLICE 예제에서 요인 1 과 요인 2 에 묶인 변수들의 상관 계수를 구한 것이다.

	FAT	THIGH	CHEST	CHNUP	HEIGHT
FAT	1.00000	0.84421 <.0001	0.72462 <.0001	-0.69117 <.0001	0.36511 0.0091
THIGH	0.84421 <.0001	1.00000	0.39780 0.0042	-0.66953 <.0001	0.22319 0.1192
CHEST	0.72462 <.0001	0.39780 0.0042	1.00000	-0.45359 0.0009	0.42592 0.0020
CHNUP	-0.69117 <.0001	-0.66953 <.0001	-0.45359 0.0009	1.00000	-0.27601 0.0524
HEIGHT	0.36511 0.0091	0.22319 0.1192	0.42592 0.0020	-0.27601 0.0524	1.00000
	SHLDR	PELVIC	WEIGHT	BREATH	
FAT	0.33063 0.0190	0.41322 0.0029	0.80951 <.0001	0.29870 0.0351	
THIGH	0.20457 0.1541	0.20749 0.1482	0.55422 <.0001	0.20652 0.1502	
CHEST	0.55449 <.0001	0.52205 0.0001	0.88869 <.0001	0.34730 0.0135	
CHNUP	-0.27387 0.0543	-0.15816 0.2727	-0.57578 <.0001	-0.35762 0.0108	
HEIGHT	0.65429 <.0001	0.58589 <.0001	0.63534 <.0001	0.58783 <.0001	

WEIGHT 변수는 요인 1 과 요인 2 에 모두 묶이므로 모든 변수들과 상관 관계가 유의하다. 요인 1, 2 에 의해 묶인 변수들은 그룹 안 변수 간에는 상관 관계가 높고 그룹 외부 변수와는 상관 관계가 다소 낮다. 요인 분석을 통하여 변수를 그룹화 한 후 상관 관계를 살펴 보면 이런 현상이 쉽게 파악되지만 상관 계수만으로 변수는 분류하기는 어려움이 있다.

또한 요인 분석 결과 묶인 변수들간 상관 관계를 보면 낮아야 한다. FAT_INDEX, BODY 상관 계수가 유의하지 않아야 하지만 그룹 내 변수가 상관 관계가 높기 때문에 이런 결과가 발생했다.

	FAT_INDEX	BODY	REACT	PULSE	DIAST
FAT_INDEX	1.00000	0.52135 0.0001	0.07825 0.5891	-0.10531 0.4667	0.05587 0.7000
BODY	0.52135 0.0001	1.00000	0.15310 0.2885	-0.16964 0.2389	-0.15540 0.2812
REACT	0.07825 0.5891	0.15310 0.2885	1.00000	0.16311 0.2577	0.14732 0.3073
PULSE	-0.10531 0.4667	-0.16964 0.2389	0.16311 0.2577	1.00000	0.23409 0.1018
DIAST	0.05587 0.7000	-0.15540 0.2812	0.14732 0.3073	0.23409 0.1018	1.00000

5.10. 설문 분석에 요인 분석 이용

요인 분석이 언제 설문 분석에 이용될 수 있을까? 리커드(Likert) 척도로 조사된 문항들을 그룹화하는데 사용된다. 몇 문항들을 합쳐 하나의 지표(index) 점수로 사용할 수 있느냐를 알아볼 때 요인 분석이 사용된다. 위에서 원 변수 $x_1, x_2, ..., x_p$ 가 설문 조사의 각 리커드 척도 문항에 해당된다. 예제 설문에서 시설물 관련 만족 정도를 묻는 문항이 Q4-Q13 으로 열 문항이다. 이 10 문항을 하나로 혹은 2-3 그룹으로 묶어 어떤 항목을 측정하는 점수로 사용할 수 있느냐가 궁금할 것이다. 만약 하나로 묶어진다면 그 10 개 문항의 (평균) 점수가 응답자들의 시설물 만족도 점수가 되는 것이다. 만약 2 개 이상으로 묶어진다면 각 그룹을 구성하는 문항을 고려하여 조사자가 이름을 부여하면 된다.

문항을 몇 개의 그룹으로 묶을 수 있느냐는 고유치가 1 이상인 **요인의 수**에 의해 결정되고 그룹에 어떤 문항이 묶여지느냐는 **부하 값에** 의해 결정된다. 설문 분석에서 요인 분석이 가능 하려면 다음 2 조건이 만족되어야 한다.

(1)리커드 척도 문항이어야 한다.

(2)여러 문항들을 몇 개의 그룹으로 묶으려는 목적에서 실시해야 한다.

5.10.1. 예제

※다음은 ○○ 대학학생들의 시설물 만족도에 대한 설문 조사의 일부이다.[**CODING.TXT**]

Q4①경상대학 건물 안의 공간은?

매우 쾌적하다 7 6 5 4 3 2 1 매우 답답하다

Q5②경상대학 건물 안팎의 휴식 공간은?

매우 충분하다 7 6 5 4 3 2 1 매우 부족하다

Q6③강의실 공간은 수업을 하는데 있어~

매우 여유 있다 7 6 5 4 3 2 1 매우 비좁다

Q7④강의실 안의 시설 및 비품은 수업을 하기에~

매우 잘 갖추어져 있다 7 6 5 4 3 2 1 매우 부족하다

Q8⑤강의시간에 보조기자재를 이용하는 것은?

매우 편리하다 7 6 5 4 3 2 1 매우 불편하다

Q9⑥경상대학 내에 외국어 공부를 하기 위한 시설은?

매우 적절하다 7 6 5 4 3 2 1 매우 부족하다

Q10⑦경상대학 내에 컴퓨터 실습을 위한 시설은?

매우 적절하다 7 6 5 4 3 2 1 매우 부족하다

Q11⑧경상대학 내에 도서관 시설은?

매우 적절하다 7 6 5 4 3 2 1 매우 부족하다

Q12⑨경상대학 화장실 시설은?

매우 청결하다 7 6 5 4 3 2 1 매우 불결하다

예제 설문 시설물에 대한 Q4-Q12 번 문항을 어떻게 그룹화 할지 요인 분석하여 보자. 요인 분석은 그룹으로 나눌 필요가 있는 리커드(Likert) 척도 문항에(4 점, 5 점, 7 점 척도) 대해 가능하다. 다음 프로그램은 리커드 척도 문항에 대한 요인 분석을 실시할 때 사용되는 전형적인 프로그램이다. 다른 부분은 그대로 사용하고 DATA = ~ 부분과 VAR = ~의 ~ 부분만 적절히 고쳐주면 된다.

```
PROC FACTOR DATA=SURVEY ROTATE=VARIMAX REORDER;
    VAR Q4-Q12;
RUN;
```

ROTATE=VARIMAX 옵션은 요인을 직교 변환하는 방법 중 VARIANCE 를 최대한 방법을 사용하라는 것으로 부하 값을 잘 구별할 수 있다.

REORDER 옵션은 부하 값의 크기 순서대로 출력하라는 명령으로 변수(문항)를 그룹화 하는데 편리하다.

COVARIANCE 변수(문항)들의 공분산 행렬을 이용하여 요인을 추정한다. 변수들의 측정 단위가 다를 경우는 상관 행렬(COVARIANCE 를 사용하지 않으면 된다. default)을 사용 해야 하나 설문 분석에서 문항들은 리커드 척도이고 같은 점수 척도이므로 공분산 행렬을 권한다.

요인 추정 방법(METHOD)은 default=주성분 방법을 선택하였다.

Eigenvalues of the Correlation Matrix: Total = 9 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	3.75509467	2.75103818	0.4172	0.4172
2	1.00405650	0.05050753	0.1116	0.5288
3	0.95354897	0.16659613	0.1059	0.6347
4	0.78695284	0.13707952	0.0874	0.7222
5	0.64987332	0.14284924	0.0722	0.7944
6	0.50702408	0.02290537	0.0563	0.8507
7	0.48411871	0.04536948	0.0538	0.9045
8	0.43874923	0.01816755	0.0487	0.9533
9	0.42058168		0.0467	1.0000

2 factors will be retained by the MINEIGEN criterion.

실험실 자료나 측정 자료인 경우 고유치가 1 이상인 경우만 택해도 누적 설명 비율이 80% 이지만 리커드 척도 문항과 같이 1, 2, 3, 4, 5 이산형 데이터인 경우 누적 설명 비율이 매우 낮다.(53%) 그러나 설문 조사에서 요인 분석은 리커드 척도 문항 분류하는데 사용되므로 걱정하지 말자.

Rotated Factor Pattern

	Factor1	Factor2
Q7	0.73815	0.21534
Q8	0.69564	-0.03666
Q6	0.64055	0.20458
Q5	0.62266	0.37486
Q9	0.56059	0.49437
Q10	0.13259	0.80343
Q11	0.21992	0.75404
Q12	0.08964	0.55255
Q4	0.51461	0.53426

← 이 부분 결과를 이용하면 된다

부하의 값의 크기가 0.6 이상인 (크기 값이 유사하고) 변수(문항)를 묶으면(분류하면) 된다. 요인 1(factor 1)이 주로 설명하는 변수는 Q5-Q8 이므로 묶으면 되고, 요인 2 의 부하 값에 의해서는 Q10-Q11 을 하나로 묶을 수 있을 것이다. Q4, Q12 도 요인 2 에 의해 묶을 수 있을 것 같으나 0.75 에 비해 0.55 면 차이가 많으므로 안 묶는 것이 좋다. 그러므로 문항은 (강의실 만족도: Q5, Q6, Q7, Q8), (정보 시설 만족도: Q10, Q11)을 묶고 나머지 문항들은 개별 문항으로 간주하여 분석한다.

5.10.2. 보고서 작성

리커드 척도 문항들에 대한 요인 분석 결과는 다음과 같이 정리하면 된다.

	Factor1	Factor2
Q7	**0.74**	**0.22**
Q8	**0.70**	**-0.04**
Q6	**0.64**	**0.20**
Q5	**0.62**	**0.37**
Q9	**0.56**	**0.49**
Q10	**0.13**	**0.80**
Q11	**0.22**	**0.75**
Q12	**0.09**	**0.55**
Q4	**0.51**	**0.53**
신뢰도 계수	**0.69**	**0.68**

이제 분류된 리커드 척도 문항 사용에 대해 설명해 보자. 요인 1(factor 1)에 의해 Q5-Q8 을 묶고 강의실 만족도라 하고 요인 2 에 의해 Q10, Q11 을 묶어 정보 시설 만족도라 하였다. 향후 분석(회귀 분석, 분산 분석, 기초 통계량 분석)에서는 묶은 변수 집단(묶은 변수는 합보다는 평균을 이용하는 것이 바람직하다. 이유는 (1)단위 맞추기 (2)결측치 있는 경우)을 하나로 사용하는 것이 좋다. 물론 개별적으로 보기도 하지만.....

```
DATA SURVEY0;
    SET SURVEY;
        LECTURE=MEAN(OF Q5-Q8);
        INFORMATION=MEAN(Q10, Q11);
RUN;
```

요인을 2 개 선택하였음에도 불구하고 누적 변동이 55% 밖에 되지 않는 것은 Q8 이 빠져 있고 요인 1 에서 Q4, Q6, Q8 의 부하 값이 다른 문항에 비해 작기 때문이다. 그러나 이 부분에 대해서는 굳이 언급할 필요가 없다. 설문 분석 시 요인 분석을 이용하는 주된 이유는 문항 분류라는 사실을 잊지 말아야 할 것이다. 그리고 리커드 척도 문항과 같이 실험실 측정 자료가 아니면 누적 변동이 낮을 수 밖에 없다. 마지막 행에는 묶은 문항의 신뢰도 계수(내적 일치도)를 적어 주면 된다. 이 값을 구하는 방법은 다음 절에 있다.

5.10.3. 문항 내적 일치도

문항이 요인 분석에 문항이 그룹화 되면 문항들이 하나의 개념(index)을 얼마나 잘 표현하는지를 알아보는 것을 내적 일치도(internal consistency)를 알아본다고 하는데 이 개념을 계산한 값이 Cronbach alpha(α)라 한다. 이를 문항의 신뢰도라 하기도 한다. 응답자로부터 얻은 설문 응답 결과(측정치: observed value)는 실제 응답자의 만족 점수와 측정 오차(measurement error)로 구성되어 있다. $Y = T + E,\ \mathrm{cov}(T, E) = 0$. 그러므로 측정치의 신뢰계수(reliability coefficient)는 다음과 같이 정의된다.

$$\begin{aligned}\sigma^2(Y,T) &= \frac{\mathrm{cov}(Y,T)^2}{\mathrm{var}(Y)\,\mathrm{var}(T)} \\ &= \frac{\mathrm{var}(T^2)}{\mathrm{var}(Y)\,\mathrm{var}(T)} \\ &= \frac{\mathrm{var}(T)}{\mathrm{var}(Y)}\end{aligned}$$

위의 측정치 신뢰 계수는 변수가 하나인 경우인데, 이를 변수가 여러 개인 경우(문항이 여러 개)로 일반화 시킨 것이 Cronbach α 값이다. p 개 문항이 있을 경우 $Y_j = T_j + E_j\,(j = 1,2,\ldots,p)$ 이고 $Y_O = \sum Y_j, T_0 = \sum T_j$ 라고 놓으면 다음이 성립한다.

$$\begin{aligned}\alpha &= \left(\frac{p}{p-1}\right)\frac{\sum_{i\neq j}\mathrm{cov}(Y_i, Y_j)}{\mathrm{var}(Y_O)} \\ &= \left(\frac{p}{p-1}\right)\left(1 - \frac{\sum_j \mathrm{var}(Y_j)}{\mathrm{var}(Y_O)}\right)\end{aligned}$$

Cronbach α 는 0 과 1 사이의 값이고 1 에 가까울수록 내적 일치도가 높다. 얼마면 높다고 할 수 있는가? 0.6 이상? 0.7 이상? 그러나 이런 기준에는 나는 수긍할 수 없다. 왜냐하면 Cronbach α 값은 문항의 수가 많을수록, 응답자 수가 많을수록 높아지는 경향이 있기 때문이다. 그러므로 값의 크기가 판단의 근거가 되는 것이 아니라 한 문항을 제외했을 때 Cronbach α 값이 적어지느냐, 커지느냐를 를 보고 그 문항을 제외하느냐 그대로 두느냐를 판단하기 바란다. 그러나 보고서나 논문 작성과 같이 내적 일치도 값을 제시해야 하는 경우에는 전체 내적 일치도 값(Cronbach α)을 제시할 수 밖에는 없다. 다시 강조하지만 이 값

의 크기가 중요한 것이 아니라 문항을 제외하였을 때 CRONBACH 값의 변화가 더 중요하다.

문항의 보기가 2 개(binary, dichotomous (0,1)) 한 경우 Cronbach α 신뢰 계수는 Kuder-Richardson 20 (KR-20) 신뢰 계수가 된다.

요인 분석에 의해 묶은 리커드 척도 문항에 대해서만 내적 일치도 Cronbach α 를 구하면 된다.

```
PROC CORR DATA=SURVEY NOSIMPLE NOCORR ALPHA;
    VAR Q5-Q8;
RUN;

PROC CORR DATA=SURVEY NOSIMPLE NOCORR ALPHA;
    VAR Q10 Q11;
RUN;
```

▪NOCORR 은 변수(문항)들의 상관 계수 값을 출력하지 말라는 옵션이다.

▪NOSIMPLE 은 변수들의 기초 통계량(평균, 표준 편차)을 출력하지 말라는 옵션이다.

▪ALPHA 는 CRONBACH 값을 계산하라는 옵션이다.

Cronbach의 α계수

변수	α계수
원데이터	0.686085
표준화	0.690973

변수를 제외했을때의 Cronbach 계수

	데이터 변수		표준화된 변수	
삭제한 변수	합계와의 상관 계수	α계수	합계와의 상관	α계수
Q5	0.513210	0.596535	0.515436	0.599651
Q6	0.446627	0.634924	0.452792	0.639815
Q7	0.555856	0.574055	0.552425	0.575144
Q8	0.386110	0.670726	0.381783	0.683350

일반적으로 변수를 표준화 시킨 후 구한 신뢰도 계수 값이 크므로 이를 이용한다. Q5-Q8 4 개 문항 모두 사용할 경우 신뢰도 계수는 0.69 이다. Q5 를 제외하고 Q6-Q8 만 사용하면 신뢰도 계수가 0.59 로 떨어진다.(생략) Q8 를 제외하고 Q5-Q7 만 사용하면 신뢰도 계수가

0.68 로 떨어진다. 그러므로 4 개 변수(문항)를 묶는 것이 옳으며 신뢰도 계수(내적 일치도)는 0.68 이다.

Cronbach의 α계수

변수	α계수
원데이터	0.670557
표준화	0.684502

변수를 제외했을때의 Cronbach 계수

	데이터 변수		표준화된 변수	
삭제한 변수	합계와의 상관 계수	α계수	합계와의 상관	α계수
Q10	0.512626	.	0.520337	.
Q11	0.512626	.	0.520337	.

변수가 2 개인 경우는 제외 신뢰도 계수가 계산될 수 없다. (Q10, Q11) 문항의 신뢰도 계수는 0.68 이다.

5.10.4. 부정 문항이 들어 있는 경우의 예

(1)Q5. 당신은 외모에 만족합니까? (5 점 척도)

(2)Q6. 당신은 혼자 있는 것이 편하십니까? (5 점 척도)

(3)Q7. 당신은 능력이 있다고 생각합니까? (5 점 척도)

(4)Q8. 당신은 친구 관계가 원만하다고 생각합니까? (5 점 척도)

	Q5	Q6	Q7	Q8
Q5	1.00000	-0.33808 <.0001	0.46952 <.0001	0.34472 <.0001
Q6	-0.33808 <.0001	1.00000	-0.44755 <.0001	-0.25000 0.0041
Q7	0.46952 <.0001	-0.44755 <.0001	1.00000	0.30147 0.0005
Q8	0.34472 <.0001	-0.25000 0.0041	0.30147 0.0005	1.00000

유사 개념에 대한 만족도를 측정한 4 개 문항 Q5-Q8 을 묶을 수 있는지 알아 보려면 요인분석(factor analysis)을 실시해 보자.

```
PROC FACTOR DATA=ZZZ ROTATE=VARIMAX;
    VAR Q5-Q8;
RUN;
```

Factor Pattern

	Factor1
Q5	0.75881
Q6	-0.70522
Q7	0.79112
Q8	0.62252

▶Q5, Q6, Q7를 하나의 문항으로 묶는 것이 적당하다.

Q6 번 문항은 다른 문항과 개념이지만 반대 점수로 측정되고 있다. 즉 Q6=1(매우 만족), ..., 5(매우 불만족)이다. 그러므로 이를 다른 문항과 합쳐(그룹화) 사용할 때는 다음과 같이 한다.

```
DATA ZZZ;
    set ZZZ;
    q6=6-q6;
    GROUP1=MEAN(Q5, Q6, Q7);
RUN;
```

Q5, Q6, Q7 문항의 내적 일치도(크론바흐 알파, 신뢰도 계수)를 구하면 다음과 같다.

```
PROC CORR DATA=ZZZ NOSIMPLE NOCORR ALPHA;
    VAR Q5 Q6 Q7;
RUN;
```

Cronbach의 α계수

변수	α계수
원데이터	0.670726
표준화	0.683350

만약 부정 문항이 있다면 신뢰도 계수가 –가 나오는 경우가 발생한다. 이래와 같이 나오면 Q6 가 부정 문항임을 인지한다. 그런데 이는 요인 분석에서 밝혀질 것이다.

변수를 제외했을때의 Cronbach 계수

삭제한 변수	데이터 변수		표준화된 변수	
	합계와의 상관 계수	α계수	합계와의 상관	α계수
Q5	0.097730	-1.61297	0.125043	-1.62027
Q6	-.447373	0.621031	-.458268	0.639011
Q7	0.105136	-.985364	0.019090	-1.02153

[EXERCISE]

(1)경찰 데이터에서 원 변수가 다변량 정규 분포를 따르는지 주성분 분석을 이용하여 살펴보고 따르지 않더라도 Maximum Likelihood 방법에 의해 요인 분석을 실시하고 원 변수를 묶고 해석하시오. 회전된 요인들의 부하 값에 대한 산점도를 그리시오.

(2)다음은 유럽 26 개국 9 개 업종 종사자 비율을 측정한 데이터이다.

[http://lib.stat.cmu.edu/DASL]**JOB.txt**

①종사자 비율 변수 9 개에 의해 국가를 적절히 분류해 보자.
②종사자 비율 변수 9 개를 이용하여 이상치 국가가 있는지 살펴보시오.
③9 개 업종 종사자 비율 변수를 분류해 보고 그룹에 적절한 이름을 붙이자.
④변수들간(요인 변수로 묶인 것은 묶은 것 사용) 상관 계수 구하고 요인 분석과 비교하여 해석하시오.

Country: 국가 명
Agr: 농업 종사자 비율
Min: 광업 종사자 비율
Man: 제조업 종사자 비율
PS: 전력 종사자 비율
Con: 건축 종사자 비율
SI: 서비스 종사자 비율
Fin: 금융 종사자 비율
SPS: 사회 및 개인 복지 서비스 종사자 비율
TC: 교통 통신 종사자 비율