Chapter 3. 요인분석

요인분석(FA: Factor Analysis 혹은 인자 분석이라고도 함)은 사람의 지적 능력을 측정하고 이에 연관된 변수들을 이해하려는 노력의 일환으로 Galton(회귀분석 창시자, 1888)에 의해 제안되었고 수학적인 모형은 Spearman(1904 상관 계수 제안자)에 의해 발전되었다. 요인분석은 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.

요인분석의 예를 보면, 설문 조사에서 동일한 개념을 측정하기 위해 설계된 리커트(Likert) 척도 문항들이 정말 그런지 알아보기 위한 분석 방법으로 요인분석이 사용된다. 물론 그 문항들의 신뢰도(혹은 내적 일치도)는 Cronbach α에 의해 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 요인분석이 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 분류하고자 할 때 사용한다.

3.1 요인분석 개념

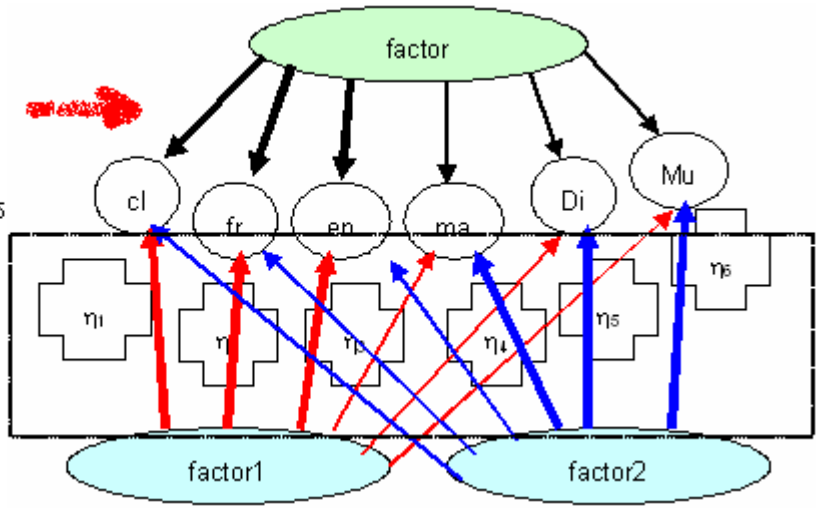
3.1.1 Spearman (1904)

Spearman 은 학생들의 6 과목 성적에 대한 상관 계수를 구한 후 상관 계수 값을 살펴보아 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.

$$classic = \lambda_1 f + \eta_1$$
$$french = \lambda_2 f + \eta_2$$
$$english = \lambda_3 f + \eta_3$$
$$math = \lambda_4 f + \eta_4$$
$$discover = \lambda_5 f + \eta_5$$
$$music = \lambda_6 f + \eta_6$$



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 경우 도식화 한 것이다. 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. 이것이 요인분석이다.

3.1.2 주성분분석과 비교

주성분분석은 p개의 원 변수를 1-2 개의 주성분으로 축약하는데 사용한다면 요인분석은 p개의 변수들이 상호 어떤 관계가 있는지 결정하여 $m(< p)$개 변수 그룹으로 나누는데 목적이 있다. 요인분석 결과 묶여진 변수 그룹을 살펴 보면 그룹 내 변수들 간에는 상관 계수가 높고 다른 그룹의 변수 간에는 상관 관계는 낮다. 요인을 구하는 방법으로 주성분을 이용한 방법을 가장 많이 사용하므로 주성분분석과 유사해 보인다. 다음은 주성분분석과 요인분석을 비교한 표이다.

<table>
<tr><th>주성분분석</th><th>요인분석</th></tr>
<tr><td>주성분은 원 변수의 직교 선형 결합으로 표현 $\underline{Y} = L\underline{X}$ ► l_{ij} 는 선형 결합 함수의 계수</td><td>인자들의 직교 선형 결합으로 원 변수들을 표현하며 인자는 관측될 수 없다. $\underline{X} = L\underline{F} + \eta$ ► l_{ij} 을 loading이라 (부하) 함</td></tr>
<tr><td>$y_1 = l_{11}x_1 + l_{12}x_2 \ldots + l_{1p}x_p$
$y_2 = l_{21}x_1 + l_{22}x_2 \ldots + l_{2p}x_p$
...
$y_p = l_{p1}x_1 + l_{p2}x_2 \ldots + l_{pp}x_p$</td><td>$x_1 = l_{11}f_1 + l_{12}f_2 \ldots + l_{1p}f_p + \eta_1$
$x_2 = l_{21}f_1 + l_{22}f_2 \ldots + l_{2p}f_p + \eta_2$
...
$x_p = l_{p1}f_1 + l_{p2}f_2 \ldots + l_{pp}f_p + \eta_p$
η_{ij} =오차항</td></tr>
<tr><td>주성분은 변수들의 변동을 설명</td><td>요인은 변수들의 분산-공분산 구조 설명</td></tr>
<tr><td colspan="2">요인분석이나 주성분분석의 l_{ij}를 구하는 방법 유사하다.
► 공분산 행렬, 상관행렬로부터 고유치 그에 대응하는 고유 벡터를 이용한다.
► (주성분분석) $l_{ij} = e_{i(j)}$ (요인분석) $l_{ij} = \sqrt{\lambda_i} e_{i(j)}$</td></tr>
<tr><td>변수의 개수 축약하는데 사용되며 l_{ij} 는 주성분의 이름을 붙이는데 사용</td><td>변수에 내재된 관계를 알아보는데 사용되며 l_{ij} 는 변수들을 그룹화 하는데 사용한다.</td></tr>
<tr><td>적절한 주성분의 수를 구하고 주성분의 이름을 부여한 후 2차 분석에 사용</td><td>적절한 인자의 수를 구하고 이를 이용하여 변수들을 그룹화 변수 그룹으로 2차 분석</td></tr>
</table>

3.1.3 요인분석 모형

요인분석의 목적은 다음과 같다.

(1)원 변수에 내재된 관계를 설명할 공통 개념을 (요인) 구한다.

(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.

p 개의 원 변수 $\underline{x}' = (x_1, x_2, ..., x_p)$ 가 평균 벡터가 $\underline{\mu}$, 분산-공분산 행렬이 Σ 이라 하면 일반적인 요인분석 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) ; unknown

(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; unknown

(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. $Cov(\underline{\eta}, \underline{f}) = 0$

$\underline{x} = L\underline{f} + \underline{\eta}$ 모형에 대해 다음을 설명할 수 있다.

(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 공통인자들은 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 인자들에 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.

3.2 요인분석하기

3.2.1 요인 구하기

상관 계수 행렬(R)을 이용하여 요인을 구하는 방법을 살펴보자. (수학적 전개 용이)

(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.

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 요인분석이다. 그래서 요인분석이 주성분분석과 유사해 보인다.

(1)주성분 방법

변수의 상관 계수 행렬 R에 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.

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

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

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

원 변수의 변동은 공분산 행렬을 사용할 경우 $tr(\hat{\Sigma}) = s_{11} + s_{22} + \cdots + s_{pp}$ 이고 $tr(\hat{R}) = p$ 이다.

j번째 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 j번째 요인에 의한 원 변수 변동의 설명 비율은 다음과 같다.

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

이제 요인 값을 구해 보자. 상관 계수 행렬로부터 구한 주성분을 $y_1, y_2, ..., y_p$ 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. $\underline{x} = L\underline{f} + \underline{\eta}$

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

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

(2)최대우도 추정법

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. SPSS 는 이 방법의 요인분석은 제공하지 않는다.

3.2.2 부하 값과 공통성

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

주성분 방법에 의해 구한 요인의 부하 값은 상관 계수 행렬로부터 구한 고유 벡터(e_i : 주성분분석의 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)

부하 값은 원 변수를 그룹화 하는데 사용된다. 요인(사실 이것의 개념은 모른다. 그래도 상관 없다) 이 원 변수를 설명하는 값의 크기이므로 임의의 요인에서 부하 값이 큰 변수들은 동일한 개념을 나타나는 것으로 간주하여 묶을 수 있을 것이다. 3.11 절 그림 참고

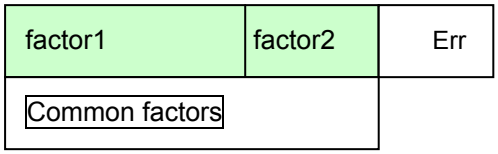
$Var(x_j) = \sigma_{jj} = l_{j1}^2 + l_{j2}^2 + ... + l_{jm}^2 + \Psi_j = \sum_{k=1}^{m} l_{jk}^2 + \psi_j$ 쓸 수 있는데 $\sum_{k=1}^{m} l_{jk}^2$ 을 공통성 (communality)이라 하고 ψ_j 는 특정(specific) 분산이라 한다. 1(100%)에 가까운 값이면 그

변수의 변동이 선택된 요인에 의해 거의 모두 설명 된다는 의미이고 낮으면 다른 요인이 존재한다는 것이다. SPSS 에서는 공통성에 대한 출력 결과는 없다. 이는 실제 분석에서는 사용되지 않는 통계량이기 때문으로 생각된다.

3.2.3 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) 상관행렬 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.

(3)SCREE 그림 사용: 주성분 방법에서 사용되었던 SCREE 그림을 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: Σ 는 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. SPSS 에서는 제공되지 않는 통계량이다.

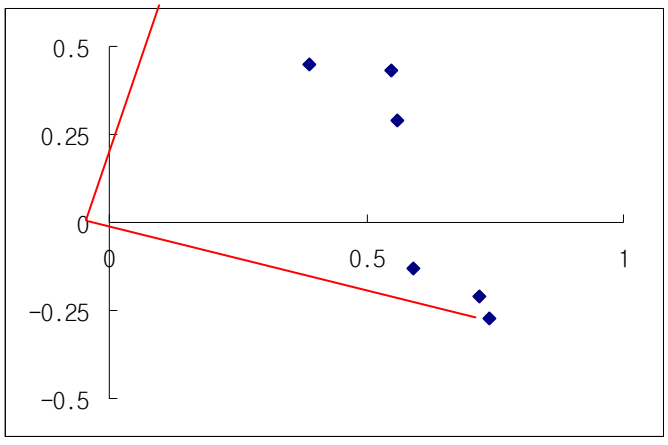
3.2.4 요인 회전

요인분석에서 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 0을 중심으로 ±의 작은 값이 있는 경우 부하 값만으로 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 직교변환 $L^* = LP$ (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° 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



3.2.5 요인 점수

요인분석은 $x_1, x_2, ..., x_p$ 의 변수를 공통인자인 요인을 이용하여 원 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 요인점수(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 요인점수는 바로 계산될 수 없어 다음 2 가지 방법을 사용한다.

▣Bartlett 방법(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. $\underline{f}_r = (\hat{L}\hat{\psi}^{-1}\hat{L})^{-1}\hat{L}\hat{\psi}^{-1}\underline{z}_r$

▣Thompson 방법(Regression Method)

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

주성분 점수와는 달리 요인 점수는 필요하지 않다. 요인분석은 원 변수를 그룹화 하고 그룹화 된 변수들의 평균(혹은 합)이 새로운 변수가 되기 때문이다.

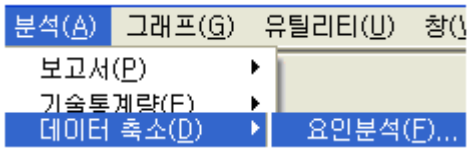
3.3 요인분석 예제

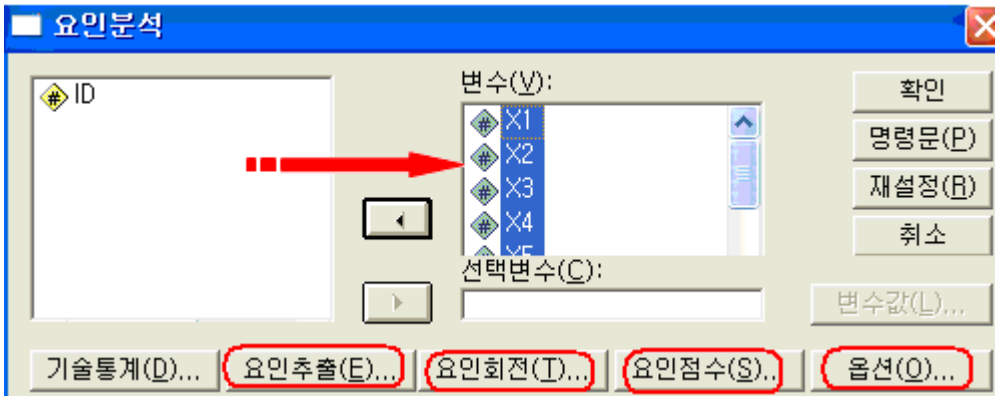
3.3.1 예제 1

앞 절에서 살펴본 주성분 방법에 의해 요인을 구하는 방법을 보면 주성분분석의 주성분을 구하는 방법 유사함을 알 수 있다. 요인과 주성분은 관계는 $f_i = \frac{y_i}{\sqrt{\lambda_i}}$ 이다. 주성분분석은 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. 요인분석은 변수의 내재된 관계를 이용하여 변수를 분류하는 방법이므로 상관행렬로부터 고유치, 고유 벡터를 구한다.

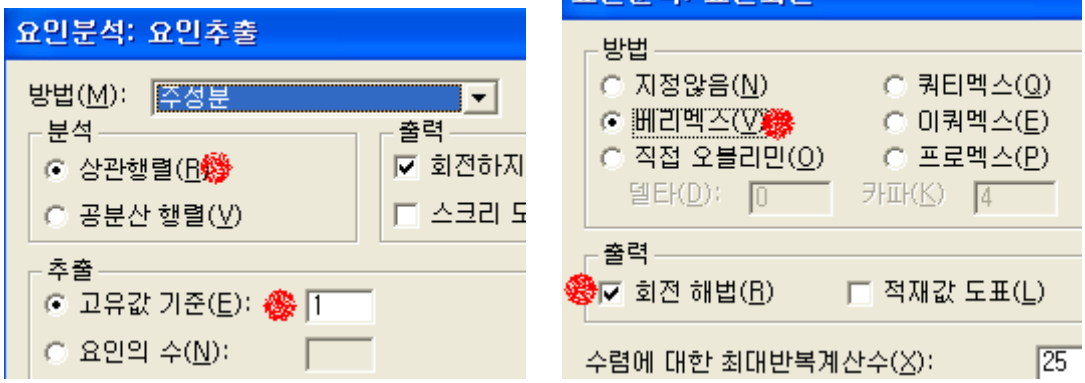
EXAMPLE 3-1 주성분과 요인분석 비교

A 회사에서는 48 명의 지원자에 대해 그들의 능력을 10 점 만점으로 15 개 영역을 조사하여 가장 점수가 높은 지원자 6 명을 선발하려고 한다. ■APPLICANT.SAV■ 주성분분석 절차와 동일하다. 차이가 있는 부분만 언급하기로 한다.

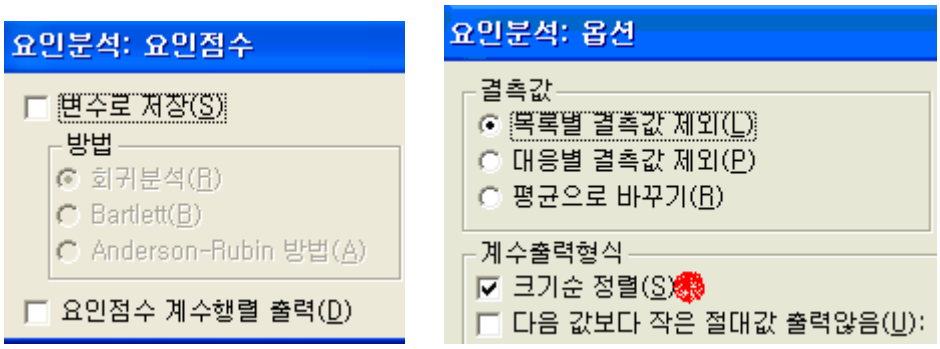




요인 구하는 방법 중 가장 많이 사용되는 주성분 방법을 사용하였다. 앞에서 언급하였듯이 상관 계수 행렬을 사용해야 한다. 요인 개수는 고유치=1 기준으로 하였다. 부하 값을 보다 더 잘 구별하기 위하여 요인 회전 방법을 VERIMAX 방법을 선택하였고 부하 값을 출력하도록 설정하였다.



요인 점수는 사용하지 않을 것이므로 저장할 필요는 없다. 부하(계수) 값을 크기 순으로 출력하면 변수를 그룹화 하는데 도움이 되므로 “크기 순 정렬” 옵션을 선택하였다.



상관 계수 행렬이 사용되었으므로 고유치의 합은 원 변수 개수 15 이다. 고유치가 1 이상 공통 요인의 원 변수 변동의 설명력은 80% 이상이다.

설명된 총분산

성분	초기 고유값		
	전체	% 분산	% 누적
1	7,514	50,092	50,092
2	2,056	13,709	63,801
3	1,456	9,705	73,506
4	1,198	7,986	81,492
5	,739	4,928	86,420
6	,495	3,297	89,717
7	,351	2,342	92,059
8	,310	2,066	94,125
9	,257	1,713	95,838
10	,185	1,233	97,071
11	,153	1,018	98,088
12	,098	,650	98,739
13	,089	,592	99,331
14	,065	,431	99,762
15	,036	,238	100,000

추출 방법: 주성분 분석.

부하 값의 크기를 이용하여 원 변수를 묶는다. 원 변수 앞의 부하 값이 얼마나 되어야 같은 그룹에 넣을 수 있을까? 일정한 기준은 없고 크기가 비슷하면 같은 그룹으로 묶는다. 경계 부분에 있는 변수를 그룹에 포함할지 여부는 그 변수가 그룹에 적절한가에 따라 분석자가 판단하면 된다.

그룹1: (자신감 X_5), (명석 X_6), (마케팅 능력 X_8), (추진력 X_{10}), (야망 X_{11}),

(개념 파악 능력 X_{12}), (장래성 X_{13})

그룹 2: (이력서 X_1), (경험 X_9), (업무 적합성 X_{15})

그룹 3: (친밀감 X_4), (진실 X_7)

나머지 변수들은 개별적으로 사용하게 된다.

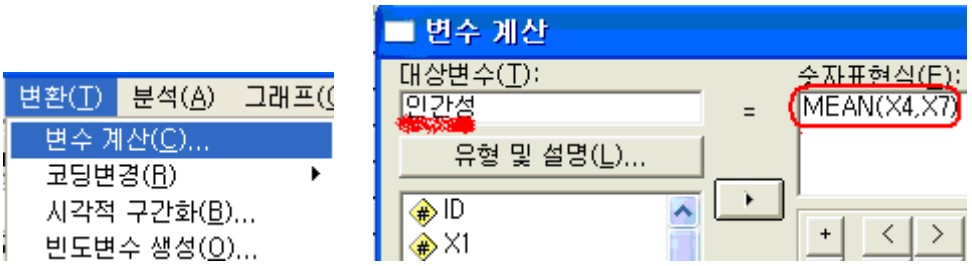
회전된 성분행렬[a]

	성분			
	1	2	3	4
X11	,918	,159	,100	-,041
X5	,916	-,107	,163	-,065
X8	,910	,223	,103	-,041
X6	,863	,097	,255	,002
X12	,811	,255	,331	,143
X10	,800	,349	,156	-,052
X13	,747	,326	,413	,224
X2	,440	,151	,399	,227
X9	,087	,851	-,055	,211
X1	,116	,830	,109	-,136
X15	,383	,797	,076	,084
X4	,220	,245	,871	-,081
X7	,219	-,242	,863	,001
X14	,440	,363	,534	-,524
X3	,064	,128	,007	,928

요인추출 방법: 주성분 분석.
회전 방법: Kaiser 정규화가 있는 베리멕스.

각 그룹의 이름 부여는 속해 있는 원 변수를 참고한다. 그룹 3 은 인간성으로 하면 적절해 보인다. 향후에는 친밀감, 진실 점수를 개별적으로 사용하는 것이 아니라 두 점수의 평균(합보다는 평균 사용을 권한다. 점수 단위 일치) 점수를 인간성 점수로 사용하면 된다.

인간성이 좋은 지원자를 뽑으려면 “인간성” 점수 크기에 의해 정렬한 후 높은 지원자 6 명을 선발하면 된다. 원 변수 15 개의 영향이 모두 들어 있는 주성분과는 달리 요인분석에 의해 구한 인간성 변수에는 2 개 원 변수 점수만 들어 있다.



ID	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	인간성
1	6	7	2	5	8	7	8	8	3	8	9	7	5	7	10	6.50
2	9	10	5	8	10	9	9	10	5	9	9	8	8	8	10	8.50
3	7	8	3	6	9	8	9	7	4	9	9	8	6	8	10	7.50
4	5	6	8	5	6	5	9	2	8	4	5	8	7	6	5	7.00
5	6	8	8	8	4	4	9	5	8	5	5	8	8	7	7	8.50

3.3.2 예제 1

경찰에 지원한 50 명의 신체적 특성 15 개를 측정하였다. [Applied Multivariate Methods for Data Analysts, Dallas E. Johnson, 1998, p160] ■🖫POLICE.SAV■

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

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

THIGH: 허벅지 피부 두께 (mm) / PULSE: 맥박 / DIAST: 심장 혈압 / CHNUP: 턱걸이 회수

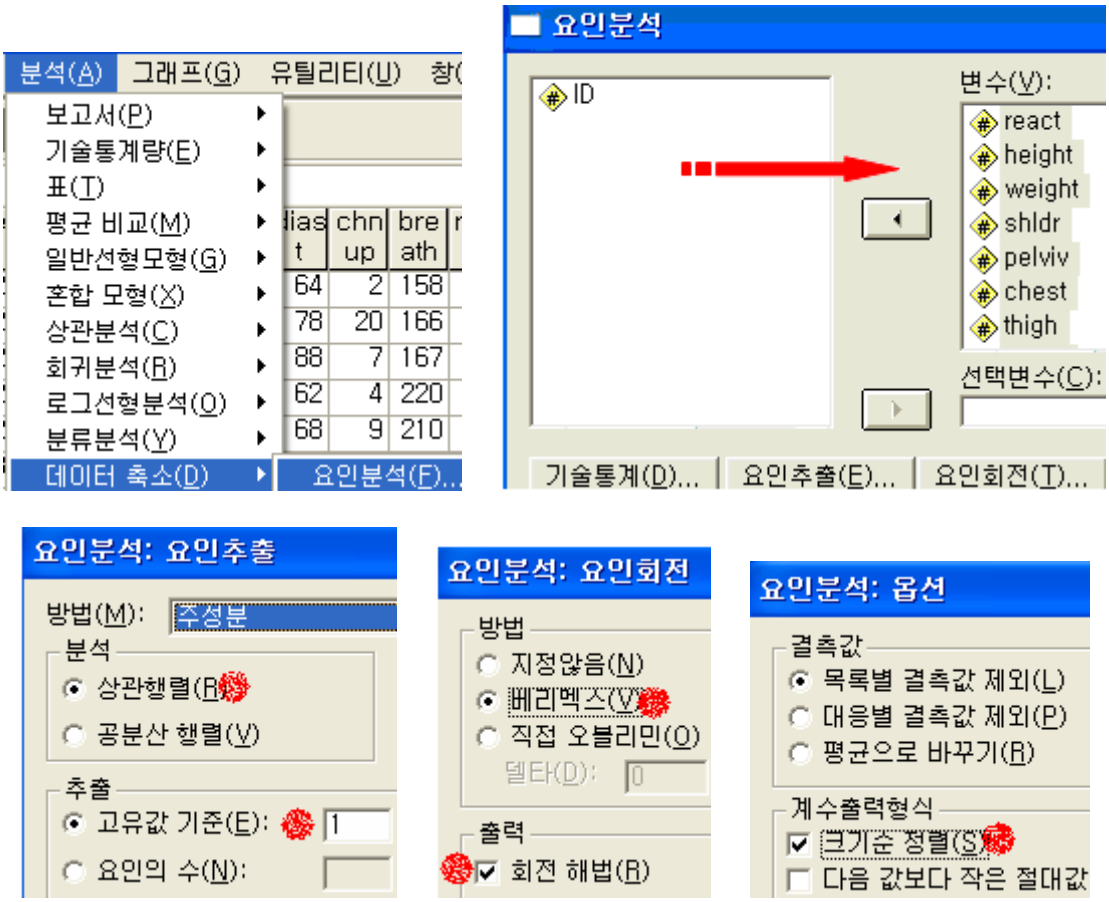
BREATH: 폐활량 (liter) / RECVR: 런닝머신에서 제자리 달리고 5분 후 맥박

SPEED: 런닝머신에서 제자리 달리기 최대 속력

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

ID	react	height	weight	shldr	pelviv	chest	thigh	pulse	diast	chnup	breath	recvr	speed	endur	fat
1.0	.3	180	74.2	41.7	27.3	82.4	19.0	64	64	2	158	108	5.5	4.0	12
2.0	.3	176	62.0	37.5	29.1	84.1	5.5	88	78	20	166	108	5.5	4.0	3.1
3.0	.3	166	73.0	39.4	26.8	88.1	22.0	100	88	7	167	116	5.5	4.0	17
4.0	.3	174	85.9	41.2	27.6	97.6	19.5	64	62	4	220	120	5.5	4.0	20

다음은 요인 방정식 해를 구하는 방법으로는 주성분 방법을 사용하고 요인 회전 방법은 VARIMAX 방법을 사용하여 요인분석을 실시하는 절차이다.



고유치 1 이상인 요인의 수가 5 개이다. 예상과는 달리 원 변수 변동을 76%만 설명하고 있다. 그렇다고 요인의 개수를 늘릴 필요는 없다.

설명된 총분산

성분	초기 고유값		
	전체	% 분산	% 누적
1	5,219	34,790	34,790
2	2,407	16,045	50,835
3	1,313	8,751	59,587
4	1,231	8,207	67,794
5	1,204	8,026	75,819
6	,848	5,653	81,472
7	,705	4,698	86,170
8	,578	3,856	90,027

회전된 성분행렬[a]

	성분				
	1	2	3	4	5
fat	,898	,304	-,017	,056	,046
thigh	,865	,074	,111	-,028	,057
chnup	-,830	-,106	,004	,074	-,146
chest	,607	,572	-,143	,118	-,178
endur	-,390	-,265	-,167	,369	,016
height	,115	,824	-,099	-,207	,295
shldr	,146	,821	-,043	-,132	-,170
pelviv	,161	,795	-,239	,268	-,105
weight	,653	,685	-,173	,025	-,041
breath	,191	,607	,205	-,330	,307
recvr	,107	-,045	,884	,012	-,197
pulse	-,144	-,130	,785	,117	,196
speed	-,383	,169	-,493	-,463	-,067
diast	-,016	,010	,185	,868	,093
react	,119	-,004	-,007	,113	,935

요인추출 방법: 주성분 분석.
회전 방법: Kaiser 정규화가 있는 베리멕스.

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

(2)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. 그러나 변수 WEIGTH 는 요인 1 에 의해 FAT, THIGH 등과 같이 분류해도 무방하다. 이처럼 부하 값 요인별 크로스 체크도 필요하다.

(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. 요인분석 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 의해 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 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? 앞에서 언급한 것처럼 변수가 1-2 개 정도 묶이는 요인을 제외한다면 15 개 변수를 요인분석에 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)요인분석 후 변수는 비만 변수(FAT_INDEX), 신체 골격 변수(BODY), 그리고 나머지 6 개 개별 원 변수 총 8 개 변수이다. 15 개 원 변수를 개별적으로 사용하여 2 차 분석(회귀분석, 분산 분석, 판별 분석 등)을 실시해도 되나 향후 분석에서는 8 개의 변수를 이용하여 분석을 하는 것이 요인분석을 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.

3.4 설문 분석

요인분석이 어떻게 설문 분석에 이용될까? 리커트(Likert) 척도로 조사된 문항들을 그룹화하는데 사용된다. 몇 문항들을 합쳐 하나의 지표(index) 점수로 사용할 수 있느냐를 알아볼 때 요인분석이 사용된다.

원 변수 $x_1, x_2, ..., x_p$ 가 설문 조사의 각 리커드 척도 문항에 해당된다. 아래 예제 설문에서 시설물 관련 만족 정도를 묻는 문항이 Q1-Q9 으로 아홉 문항이다. 이 9 문항들을 어떻게 그룹화 할 수 있는가? 알아보는 것을 요인분석이라 한다. 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 수에 의해 결정되고 그룹에 어떤 문항이 묶여지느냐는 부하 값에 의해 결정된다.

설문 분석에서 요인분석이 가능 하려면 다음 2 조건이 만족되어야 한다.

(1)리커트 척도 문항이어야 한다.

(2)여러 문항들이 합쳐져 하나의 지표로 나타내는지 알아보려는 목적이 있어야 한다.

3.4.1 예제

※학생들의 시설물 만족도에 대한 설문 조사의 일부이다. (n=130) ■CODING.SAV■

학년은? 1 학년( ), 2 학년( ), 3-4 학년( )

Q1①경상대학 건물 안의 공간은?

매우 쾌적하다 7 6 5 4 3 2 1 매우 답답하다

Q2②경상대학 건물 안팎의 휴식 공간은?

매우 충분하다 7 6 5 4 3 2 1 매우 부족하다

Q3③강의실 공간은 수업을 하는데 있어~

매우 여유 있다 7 6 5 4 3 2 1 매우 비좁다

Q4④강의실 안의 시설 및 비품은 수업을 하기에~

매우 잘 갖추어져 있다 7 6 5 4 3 2 1 매우 부족하다

Q5⑤강의시간에 보조기자재를 이용하는 것은?

매우 편리하다 7 6 5 4 3 2 1 매우 불편하다

Q6⑥경상대학 내에 외국어 공부를 하기 위한 시설은?

매우 적절하다 7 6 5 4 3 2 1 매우 부족하다

Q7⑦경상대학 내에 컴퓨터 실습을 위한 시설은?

매우 적절하다 7 6 5 4 3 2 1 매우 부족하다

Q8⑧경상대학 내에 도서관 시설은?

매우 적절하다 7 6 5 4 3 2 1 매우 부족하다

Q9⑨경상대학 화장실 시설은?

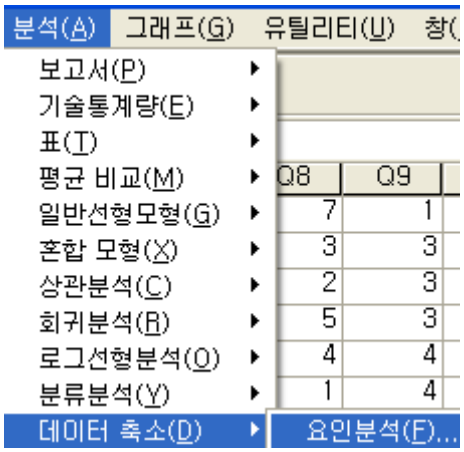
매우 청결하다 7 6 5 4 3 2 1 매우 불결하다

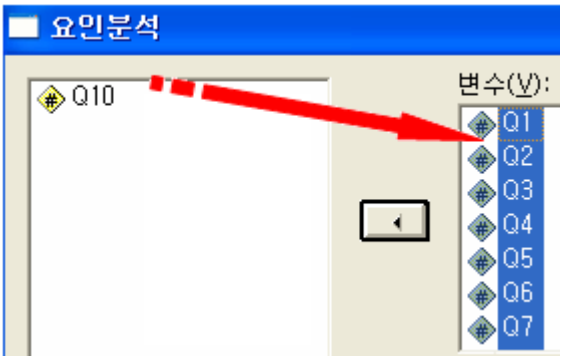
Q10⑩경상대학 시설에 대한 전체적인 만족 정도는?

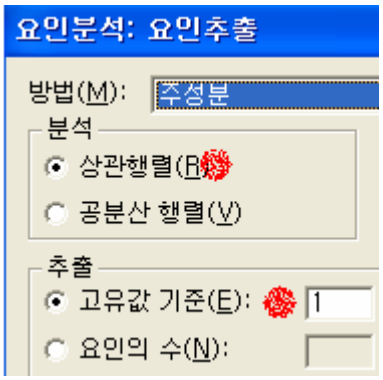
매우 만족하다 7 6 5 4 3 2 1 매우 불만족하다

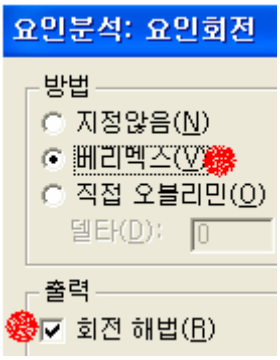
학년	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
1	1	5	2	3	3	2	1	7	1	3
2	1	1	1	1	3	1	.	3	3	3
1	4	1	1	2	1	1	1	2	3	1
1	3	1	3	3	2	5	3	5	3	3

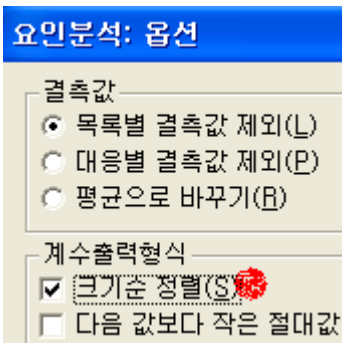
시설물에 대한 Q1-Q9 번 문항을 그룹화 하기 위하여 요인분석을 실시하자.











고유치 1 이상인 것이 2 개이므로 요인의 개수는 2 개가 적당하다. 원 변수 설명 비율이 54% 밖에 안 된다. 이유는 변수가 리커트 척도로 동일 관측치가 많이 반복되기 때문이다.

설명된 총분산

성분	초기 고유값		
	전체	% 분산	% 누적
1	3,901	43,349	43,349
2	1,029	11,435	54,784
3	,938	10,419	65,204
4	,803	8,926	74,130
5	,658	7,307	81,437
6	,498	5,537	86,974
7	,489	5,438	92,413
8	,431	4,794	97,207
9	,251	2,793	100,000

추출 방법: 주성분 분석.

크기 순으로 정렬되어 출력되었다. Q1-Q5(건물 및 강의시설 만족도, 성분 값: 0.63~0.73), Q6-Q9(시설물 만족도, 성분 값: 0.65~0.778)라 이름 붙일 수 있다. 그룹 내의 변수들의 평균을 구하여 새로운 변수로 2 차 분석에 사용하면 된다. Q6-Q9 문항들의 평균을 시설물 만족도라 할 수 있는가? 이 문항들의 내적 일치도는 얼마인가? 해답은 문항 신뢰도이다.

회전된 성분행렬[a]

	성분	
	1	2
Q3	,733	,194
Q2	,663	,151
Q4	,652	,004
Q1	,644	,390
Q5	629	420
Q6	,193	778
Q9	,453	724
Q7	,314	658
Q8	-,018	652

요인추출 방법: 주성분 분석.
회전 방법: Kaiser 정규화가 있는 베리멕스.

3.4.2 문항 내적 일치도

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

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

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

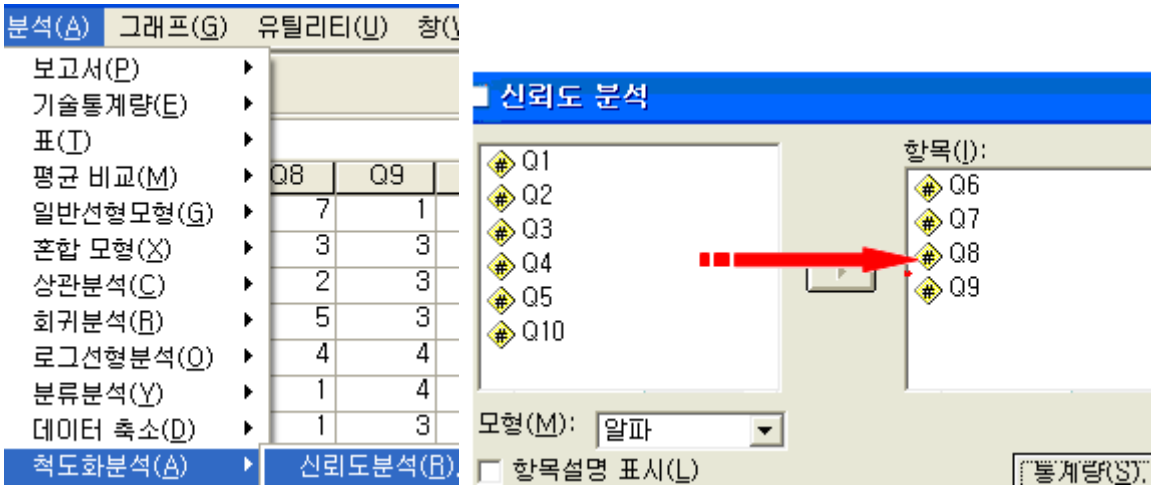
$$\alpha = \left(\frac{p}{p-1}\right)\frac{\sum_{i\neq j}\operatorname{cov}(Y_i, Y_j)}{\operatorname{var}(Y_O)} = \left(\frac{p}{p-1}\right)\left(1 - \frac{\sum_j \operatorname{var}(Y_j)}{\operatorname{var}(Y_O)}\right)$$

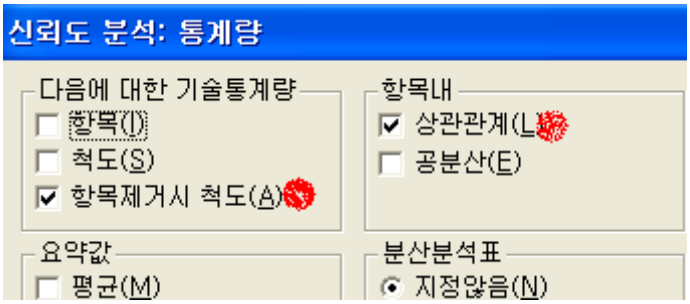
크론바흐 α는 0 과 1 사이의 값이고 1 에 가까울수록 내적 일치도가 높다. 얼마면 높다고 할 수 있는가? 0.6 이상? 0.7 이상? 그러나 이런 기준에 나는 수긍할 수 없다. 왜냐하면 Cronbach α 값은 문항의 수가 많을수록, 응답자 수가 많을수록 높아지는 경향이 있기 때문이다. 그러므로 값의 크기가 판단의 근거가 되는 것이 아니라 한 문항을 제외했을 때 Cronbach α 값이 적어지느냐, 커지느냐를 를 보고 그 문항을 제외하느냐 그대로 두느냐를 판단하기 바란다. 그러나 보고서나 논문 작성과 같이 내적 일치도 값을 제시해야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

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

EXAMPLE 3-2 신뢰도 계수 구하기

요인분석 결과 Q6-Q9 문항을 하나로 묶었다. 이 문항을 합쳐 시설물 만족도라 해도 되는가? 이들 문항의 내적 일치도는? 신뢰도 계수는 얼마인가?





신뢰도 계수는 0.733 이다. 문항을 표준화 하여 구한 신뢰도 계수가 조금 높으므로 이것을 사용하면 된다. Q8 을 삭제하면 신뢰도 계수가 0.752 로 높아진다. 삭제할까? 판단은 분석자의 몫이다.

만약 리커트 척도 문항 중에 부정 문항(역문항)이 있으면 요인분석의 부하는 음이다. 그리고 삭제된 크론바흐 알파를 보면 음의 값이 되어 있을 것이다. 이런 경우는 그 문항은 역으로 변환하여 사용하면 된다.

신뢰도 통계량

Cronbach의 알파	Cronbach's Alpha Based on Standardized Items	항목 수
,719	,733	4

항목 총계 통계량

	항목이 삭제된 경우 척도 평균	항목이 삭제된 경우 척도 분산	수정된 항목-전체 상관관계	제곱 다중 상관관계	항목이 삭제된 경우 Cronbach 알파
Q6	8,48	9,889	,558	,387	,626
Q7	9,23	11,767	,491	,318	,668
Q8	8,58	11,269	,366	,224	,752
Q9	8,73	10,827	,669	,449	,576

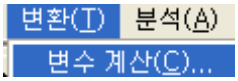
EXAMPLE 3-3 요인분석 후 2 차 분석하기

요인분석에 의해 9 개의 문항이 2 개의 그룹으로 묶여졌다. 이제 무엇을 할 수 있을까?

①인구학적 문항에 따른 시설물 만족도의 차이? 문항의 범주가 2 개이면 독립인 두 모집단 t-검정, 3 개 이상이면(예제 데이터에서 학년) 일원 분산분석을 실시하면 된다. 인구학적 문항을 2 개 이상 고려하는 다원 분산분석일 필요한가? 그럴 필요는 없다. 이것은 실험 설계가 아니기 때문이다.

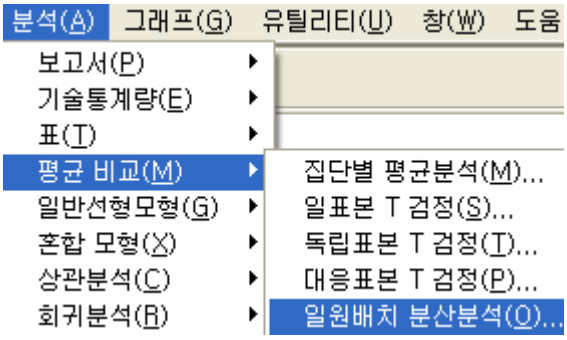
②건물 만족도와 시설물 만족도가 전체적인 만족도에 영향을 미치는가? (회귀분석)

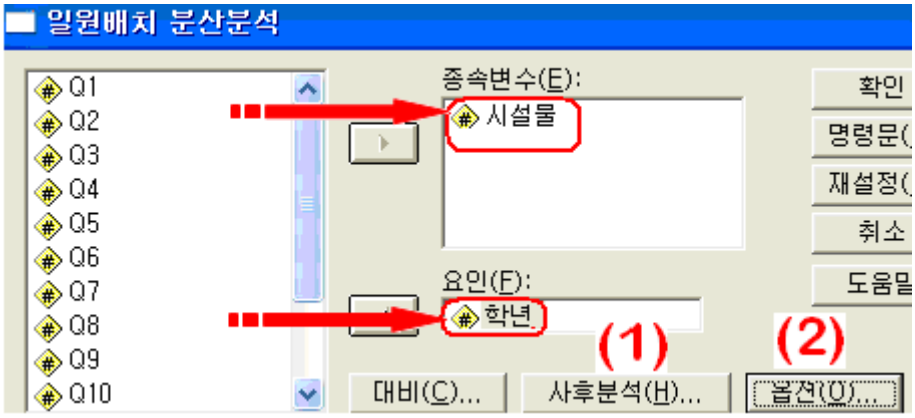
먼저 그룹 변수를 만든다. Q1-Q5 을 묶어 “건물” 만족도, Q6-Q9 을 묶어 “시설물” 만족도라 하자.

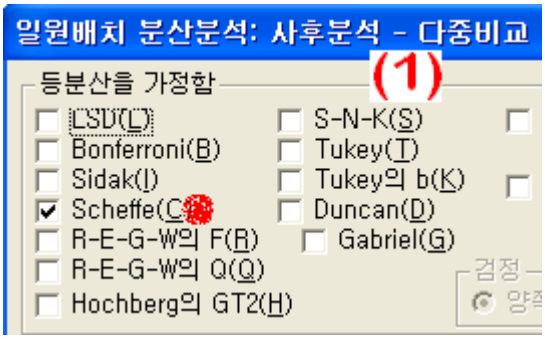


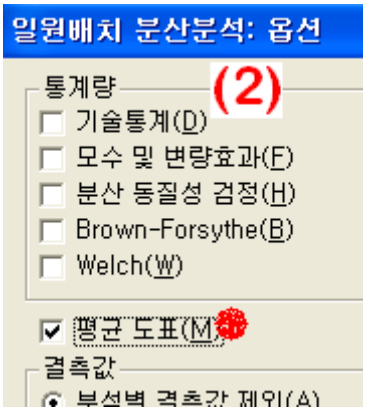
학년	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	건물	시설물
1	1	5	2	3	3	2	1	7	1	3	2.80	2.75
2	1	1	1	1	3	1	.	3	3	3	1.40	2.33
1	4	1	1	2	1	1	1	2	3	1	1.80	1.75

▣일원분산분석









평균도표를 보면 "3 학년" 학생들의 시설물 만족도가 가장 높다. 그러나 F-검정 결과 학년별 차이도 없고(유의확률=0.695) 사후검정 결과 두 학년간 차이도 유의하지 않음을 알 수 있다(모든 집단간 유의 확률이 0.05 보다 크다).

분산분석

시설물

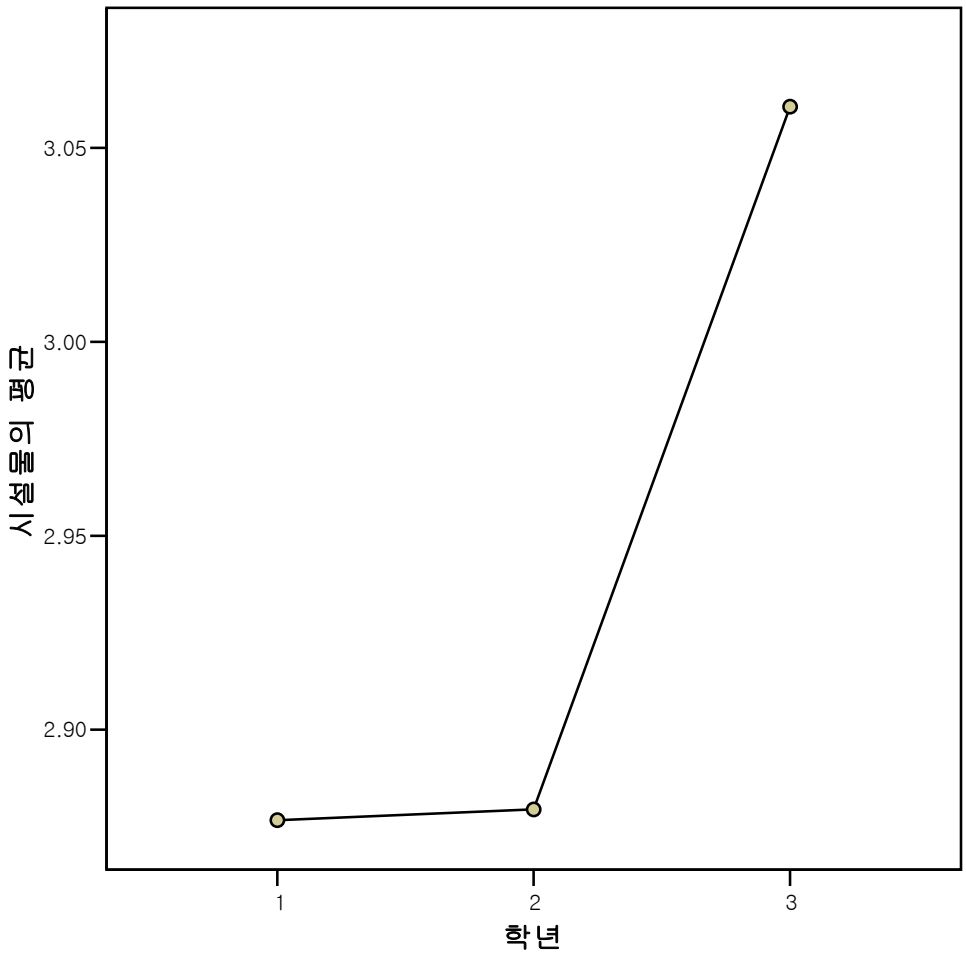
	제곱합	자유도	평균제곱	F	유의확률
집단-간	,821	2	,411	,366	,695
집단-내	142,657	127	1,123		
합계	143,478	129			

다중 비교

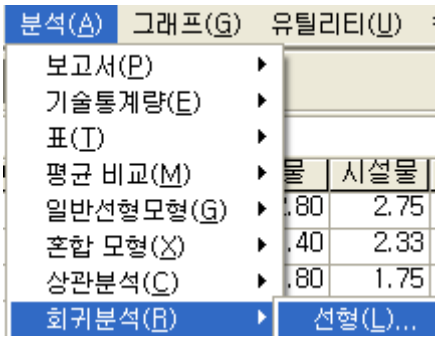
종속변수: 시설물

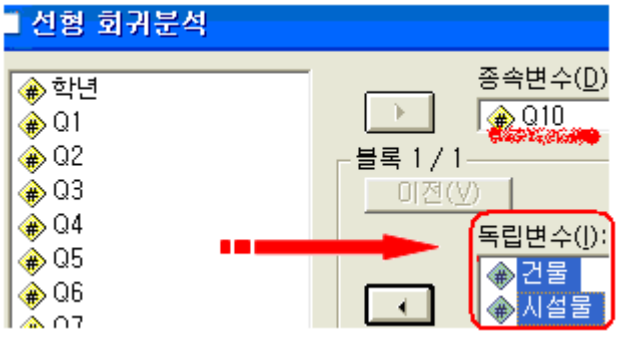
Scheffe

(I) 학년	(J) 학년	평균차 (I-J)	표준오차	유의확률	95% 신뢰구간	
					하한값	상한값
1	2	-,00277	,21533	1,000	-,5361	,5306
	3	-,18394	,23771	,742	-,7727	,4048
2	1	,00277	,21533	1,000	-,5306	,5361
	3	-,18117	,24070	,754	-,7774	,4150
3	1	,18394	,23771	,742	-,4048	,7727
	2	,18117	,24070	,754	-,4150	,7774



▣회귀분석





계수[a]

모형		비표준화 계수		표준화 계수	t	유의확률
		B	표준오차	베타		
1	(상수)	2.579	.339		7.610	.000
	건물	-.087	.132	-.067	-.661	.510
	시설물	.476	.117	.412	4.085	.000

a. 종속변수: Q10

시설물 만족도만 전체 만족도에 영향을 미친다. 회귀계수 부호가 양이므로 시설물 만족도가 올라가면 전체 만족도도 올라간다. 시설물 만족도가 1 점 올라가면 전체 만족도는 0.37 만큼 올라간다. 만약 건물 설명변수가 유의했다면 표준화 회귀계수에 의해 시설물 만족도가 건물 만족도에 비해 전체만족도에 영향을 더 많이 미친다고 결론 내릴 수 있다.

계수[a]

모형		비표준화 계수		표준화 계수	t	유의확률
		B	표준오차	베타		
1	(상수)	2.468	.294		8.385	.000
	시설물	.432	.095	.374	4.556	.000

a. 종속변수: Q10

EXERCISE

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

[http://lib.stat.cmu.edu/DASL] ■JOB.SAV■

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 요인분석과 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