Chapter 6. 정준상관분석

6.1 정준상관분석

정준 상관 분석(Canonical Correlation Analysis)은 변수들의 군집간 선형 상관 관계를 파악하는 분석 방법이다. 예를 들어 신체적 조건(키, 몸무게, 가슴둘레)과 운동력(달리기, 윗몸 일으키기, 턱걸이) 사이의 선형 상관 관계가 있는지 알아 보고, 관계가 있다면 어떤 관계가 있는지 분석하는 것이다.

정준상관분석은 $(X_1, X_2, ..., X_m)$ 변수 군과 $(Y_1, Y_2, ..., Y_n)$ 변수 군의 선형 관계를 분석한다. p 개 원 변수를 2 개의 변수 군으로 나눌 수 있다고 가정하자.

$$\underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \cdots \\ x_p \end{bmatrix} = \begin{bmatrix} \underline{x}_1 \\ \underline{x}_2 \end{bmatrix} \sim Normal\left(\begin{bmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

다음은 정준 상관 분석 의 특수한 예이다.

1)벡터 변수$(\underline{x}_1, \underline{x}_2)$에 변수가 하나이면 단순 상관 계수가 된다.

2)하나의 벡터 변수만 변수가 하나이면 이는 다중회귀 모형에서 결정계수 R^2 이다. 다중회귀의 결정 계수는 종속변수(변수가 하나인 벡터)와 설명변수의 선형결합 $(a_1 X_1 + a_2 X_2 + ... + a_p X_p)$간 상관 계수가 된다.

6.1.1 정준변수 구하기

▣제일 정준변수

두 변수 군의 선형 결합간 상관 계수를 가장 크게 하는 선형 결합을 생각해 보자.

$$\rho_1 = \max_{\underline{a}=\underline{b}\neq 0} corr(V_1, W_1) \quad \text{where} \quad V_1 = \underline{a}_1' \underline{x}_1, W_1 = \underline{b}_1' \underline{x}_2$$

위의 조건을 만족하는 $\underline{a}_1, \underline{b}_1$를 제일 정준변수(first canonical variate)라 하고 그 중 다음 식을 만족하는 $\underline{a}_1, \underline{b}_1$ 을 구하면 된다. 이때 ρ_1 을 제일 정준 상관 계수(first canonical correlation)라 한다.

$$\mathrm{var}(V_1) = \mathrm{var}(W_1) = 1 \;\rightarrow\; \underline{a}_1' \Sigma_{11} \underline{a}_1, \;\; \underline{b}_1' \Sigma_{22} \underline{b}_1$$

▣제이 정준변수

$V_2 = \underline{a}_2' \underline{x}_1, W_2 = \underline{b}_2' \underline{x}_2$ 이라 놓고 다음 조건을 만족하는 $\underline{a}_2, \underline{b}_2$를 제이 정준변수라 한다.

(1)V_2와 W_2은 각각 V_1과 W_1들과 독립이다.

(2) $\mathrm{var}(V_2) = \mathrm{var}(W_2) = 1$

$\rho_2 = corr(V_2, W_2)$ 을 제이 정준 상관 계수라 한다.

다른 정준변수도 같은 방법으로 구하면 된다. 해석의 어려움이 있어 실제 사용되는 정준 변수의 수는 2 개를 넘지 않는다.

6.1.2 정준 상관 계수 개수

두 벡터 변수의 차수 중 낮은 차수 수만큼 존재한다. 즉 변수 군을 형성하는 변수의 수가 적은 변수 군의 변수 수만큼 정준 상관 계수 값이 존재한다. 한 변수 군의 변수 수가 p 이면 다른 변수 군의 변수 수는 q 이면 정준 상관 계수의 수는 $\min(p,q)$ 이다.

정준 상관 계수의 유의성 검정은 다음과 같이 실시하면 된다.

(1) $H_{01} : \rho_1 = 0$ vs. $H_{01} : \rho_1 \neq 0$ ←→ $H_{01} : \Sigma_{12} = 0$ vs. $H_{01} : \Sigma_{12} \neq 0$

$$\text{검정 통계량 } T = \frac{|\hat{\Sigma}|}{|\hat{\Sigma}_{11}||\hat{\Sigma}_{22}|} = \prod_{i=1}^{k} (1 - \hat{\rho}_i^2), \;\; k = \min(q, p-q)$$

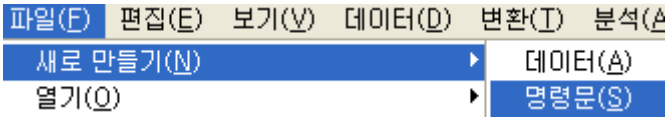
(2) $H_{0r} : \rho_r = 0$ vs. $H_{0r} : \rho_r \neq 0$

$$\text{검정 통계량 } T_r = \prod_{i=r}^{k} (1 - \hat{\rho}_i^2), \text{ 검정 통계량 분포 } \alpha \log(T_r) \sim \chi^2_{\alpha,(q-r+1)(p-q-r+1)}$$

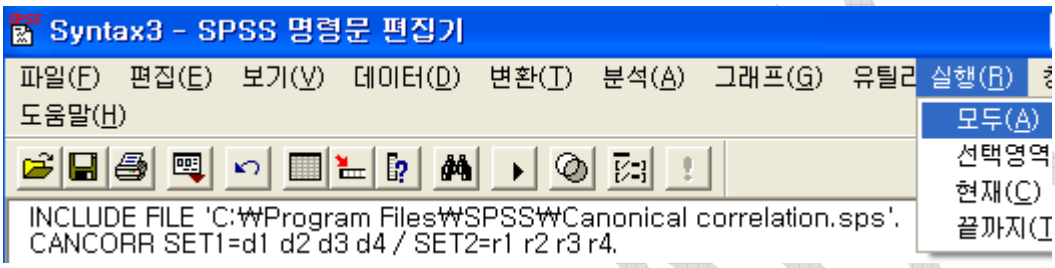
6.1.3 예제

밀 예제 자료(WHEAT.txt)에서 밀의 오른쪽 면의 측정 변수(면적, 원주, 길이 폭)와 아래쪽 면의 측정 변수(면적, 원주, 길이 폭)간에 상관 관계를 분석해 보자.

SPSS 에는 정준상관 분석을 위한 메뉴가 없다. 대신 매크로 프로그램을 실행할 수 있도록 했다. 우선 WHEAT.SAV 데이터를 열고 매크로 프로그램 작성을 위해 편집기를 연다.



편집기 창이 나타나면 아래 프로그램을 작성하고 실행한다. Canonical Correlation.sps 파일은 SPSS 가 설치된 루트 파일에 있다. SET1, SET2 는 집단 내 변수를 지정해 주면 된다. 마침표(.)는 프로그램 문장이 끝났음을 알려주는 것이다.



프로그램이 실행되면 출력 창에 엄청나게 많은 결과가 출력되고 데이터에는 정준 변수가 저장된다.

▣원 변수 상관 계수

변수 그룹 내의 변수들간의 상관 계수, 변수 그룹간 변수들의 상관 계수가 된다. 정준상관 분석의 개략적인 결과를 예상할 수 있다. SET1 군에서는 D4 가 다른 변수와 상관 관계가 낮고, SET2 에서는 R4 가 군 내 다른 변수와 상관 관계가 낮음을 알 수 있다.

Correlations for Set-1

	D1	D2	D3	D4
D1	1.0000	.8631	.6803	.7262
D2	.8631	1.0000	.8277	.4571
D3	.6803	.8277	1.0000	.1975
D4	.7262	.4571	.1975	1.0000

Correlations for Set-2

	R1	R2	R3	R4
R1	1.0000	.8076	.4535	.2724
R2	.8076	1.0000	.6139	.1184
R3	.4535	.6139	1.0000	.3268
R4	.2724	.1184	.3268	1.0000

Correlations Between Set-1 and Set-2

	R1	R2	R3	R4
D1	.6401	.7491	.4017	-.0033
D2	.6715	.8564	.5790	-.0405
D3	.5744	.7217	.5933	-.0767
D4	.4313	.3904	.0604	.0439

▣CANONICAL 상관 계수

Canonical Correlations

1	.882
2	.398
3	.250
4	.004

$\hat{\rho}_1$
$\hat{\rho}_2$
$\hat{\rho}_3$
...

정준 상관 계수의 수는 4 개이다. (각 그룹내의 변수의 개수가 각각 4 개이므로) 정준 상관 계수는 $Corr(V1,W1)=0.882, Corr(V2,W2)=0.398, Corr(V3,W3)=0.25, Corr(V4,W4)=0.004$ 상관 계수이다. 그럼 $Corr(V1,W2)$ 는 얼마인가? 당연히 0 이다.

▣CANONICAL 상관 계수 유의성 검정

Test that remaining correlations are zero

	Wilk's	Chi-SQ	DF	Sig.
1	.175	289.774	16.000	.000
2	.789	39.494	9.000	.000
3	.938	10.712	4.000	.030
4	1.000	.002	1.000	.962

각 열은 정준 상관 계수의 유의성을 검정한다. 귀무가설은 “현재 열 포함 이후 정준 상관 계수는 0 이다”이다. 그러므로 귀무가설이 기각된다는 것은 그 열의 정준 상관 계수는 0 이 아니라는 것을 포함하고 있다. 3 번째 열의 유의확률이 0.03 으로 일반적인 유의수준 0.05 보다 작으므로 귀무가설이 기각된다. 그러므로 제삼 정준 상관 계수는 유의하다. 4 열의 유의확률은 0.9617 이므로 제사 정준 상관 계수는 유의하지 않다.

▣제일, 제이 정준변수

RAW(원 점수)와 STANDARDIZED(표준화 점수) 2 개의 출력 결과가 나타나는데 RAW 는 변수의 원래 값으로 구한 것이고 STANDADIZED 는 원 변수를 표준화하여 구한 것이다. 밀 예제의 경우 원 변수는 측정 단위 다르므로 표준화 변수를 사용하는 것이 좋다. 다음 출력 결과는 $V_1 = \underline{a}_1' \underline{x}_1, W_1 = \underline{b}_1' \underline{x}_2$, $V_2 = \underline{a}_2' \underline{x}_1, W_2 = \underline{b}_2' \underline{x}_2$ 의 $\underline{a}_1, \underline{b}_1$, $\underline{a}_2, \underline{b}_2$ 이다.

Standardized Canonical Coefficients for Set-1

	1	2	3	4
D1	.016	.876	1.032	-2.700
D2	-.894	-.190	1.192	2.177
D3	-.160	-.810	-1.737	-.412
D4	.041	.441	-1.332	1.169

Standardized Canonical Coefficients for Set-2

	1	2	3	4
R1	-.080	.287	-1.791	.111
R2	-.777	.669	1.747	-.270
R3	-.254	-1.222	-.455	.331
R4	.254	.459	.522	.864

아래 면 변수 그룹의 제일 정준변수

$$V1 = DOWN1 = 0.016 * Z\_D1 - 0.894 * Z\_D2 - 0.16 * Z\_D3 - 0.041 * Z\_D4$$

오른쪽 면 변수 그룹의 제일 정준변수

$$W1 = RIGHT1 = -08 * Z\_D1 - 0.777 * Z\_D2 - 0.254 * Z\_D3 + 0.254 * Z\_D4$$

단. $Z\_* = \frac{* - \text{평균}}{\text{표준편차}}$ 로 각 변수의 표준화 값이다.

이 계수를 이용하여 정준변수 이름을 붙일 수 있다. SET1 군의 제일 정준변수는 D2, 제이 정준변수 (D1, D3), SET2 군의 제일 정준변수는 R2, 제이 정준변수는 R3 영향이 크므로 이를 고려하여 이름을 부여할 수 있다. 이름을 부여하는 것은 주성분 이름 부여처럼 다소 주관적이다.

▣정준변수와 동일 군집 원 변수간의 상관 관계

계수를 이용하기보다는 정준변수의 이름은 정준변수와 그 그룹 변수들간의 상관 계수 값을 이용하여 명명하는 것이 좋다. 다시 한 번 강조하지만 V1 과 V2, W1 과 W2 는 서로 독립이다. 공통된 정보가 없다.

Canonical Loadings for Set-1

	1	2	3	4
D1	-.835	.482	-.088	-.253
D2	-.994	.098	.037	.040
D3	-.881	-.284	-.311	-.216
D4	-.388	.831	-.380	.122

Canonical Loadings for Set-2

	1	2	3	4
R1	-.753	.397	-.444	.279
R2	-.967	.204	.083	.125
R3	-.685	-.532	-.024	.498
R4	.057	.217	.092	.970

아래면 변수 그룹 제일 정준변수는 면적, 원주, 길이와 상관 관계가 높으므로 크기로 아래면 제이 정준변수는 길이로 이름 붙이면 적절할 것 같다. 오른쪽 면 제일 정준변수도 크기로 이름을 붙일 수 있다. 제일 정준변수와 제이 정준변수의 상관 계수는 0.882 이었다. 즉 오른쪽 면의 크기가 커지면 아래면 크기도 커진다고 해석할 수 있다.

▣정준변수와 다른 군집 원 변수간의 상관 관계

Cross Loadings for Set-1

	1	2	3	4
D1	-.736	.192	-.022	-.001
D2	-.876	.039	.009	.000
D3	-.777	-.113	-.078	-.001
D4	-.342	.331	-.095	.000

Cross Loadings for Set-2

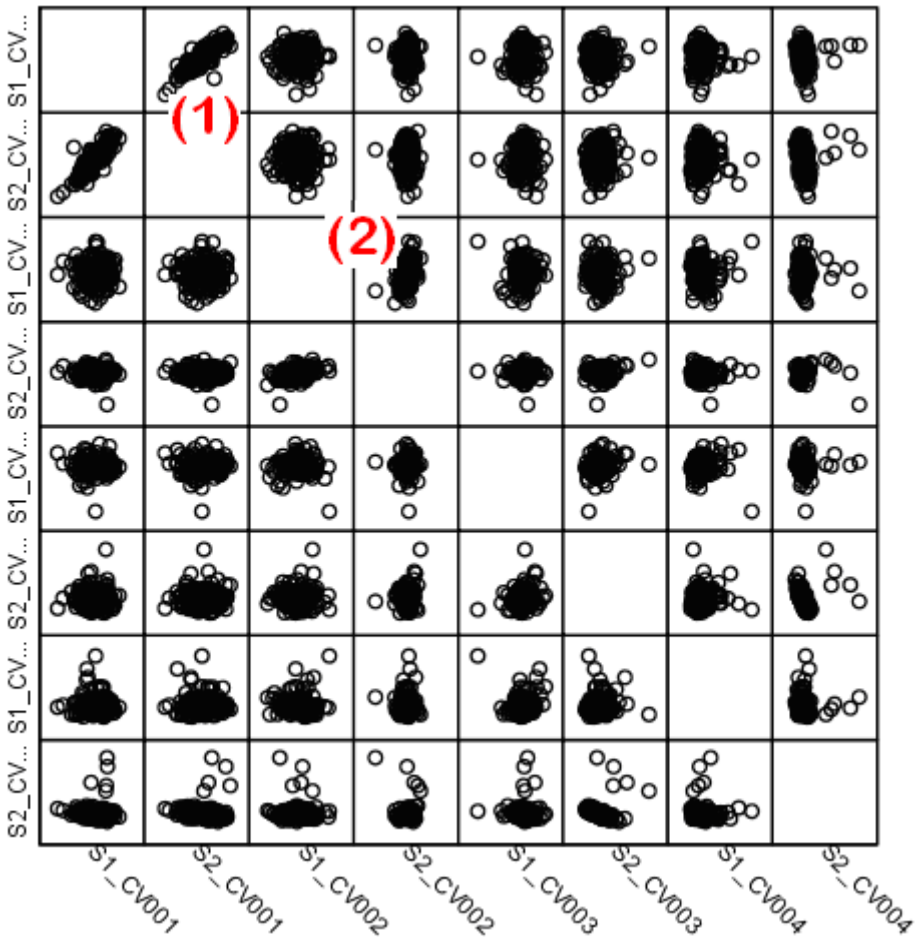
	1	2	3	4
R1	-.664	.158	-.111	.001
R2	-.853	.081	.021	.000
R3	-.604	-.212	-.006	.002
R4	.050	.086	.023	.004

아래 면의 크기(제일 주성분)는 오른쪽 면의 면적, 원주, 길이, 폭과 양의 상관 관계가 존재한다. 상관 계수는 부호가 음인 이유는 정준변수가 반대 개념으로 계산되었기 때문이다. 제일 주성분과 군내 다른 변수들간의 상관 계수를 보라. 음이다(-0.835, -0.994, -0.881, -0.388). 정준변수가 계산될 때 계수가 음인 것의 영향을 많이 받았기 때문이다. 오른쪽 면의 크기(제일 주성분)는 아래 면의 면적, 원주, 길이와 양의 상관 관계가 있다.

데이터에는 정준변수들이 저장되어 있다. S1_CV001 은 SET1 의 제일 정준변수, S2_CV001 은 SET2 의 제일 정준변수를 의미한다. 산점도 행렬을 그려보자.

S1_CV001	S2_CV001	S1_CV002	S2_CV002	S1_CV003	S2_CV003	S1_CV004	S2_CV004
-23.62	-24.39	9.90	7.64	-.05	-2.73	-4.10	6.89
-23.51	-24.57	9.30	7.66	-.72	-2.80	-4.74	7.38
-23.46	-24.75	8.54	7.73	-.44	-1.47	-4.39	6.93
-22.60	-23.66	8.61	7.10	-1.08	-2.37	-4.12	6.61
-21.98	-22.72	9.67	8.68	.73	-2.03	-3.76	6.33
-21.67	-22.54	8.57	7.27	-.01	-2.36	-3.47	6.51

제일 정준변수간 상관 관계가 가장 높고(“1”) 그 다음은 제이 정준변수간 상관 관계(“2”)이다. 제일과 제이 정준변수간 상관 관계는 0 이다. “(1)”, “(2)” 산점도에서 떨어진 한 두 개의 점들은 변수들 간의 상관 관계 면에서 이상치이다.



다음은 정준변수들간의 Pearson 상관 계수를 구한 결과이다. 위의 상관 관계를 값으로 나타낸 것이다. 제일, 제이, 제삼, 제사 정분 변수간 상관 계수는 앞의 결과와 동일하다.

상관계수

	S1_CV001	S2_CV001	S1_CV002	S2_CV002	S1_CV003	S2_CV003	S1_CV004	S2_CV004
S1_CV001	1	,882**	,000	,000	,000	,000	,000	,000
S2_CV001	,882**	1	,000	,000	,000	,000	,000	,000
S1_CV002	,000	,000	1	,398**	,000	,000	,000	,000
S2_CV002	,000	,000	,398**	1	,000	,000	,000	,000
S1_CV003	,000	,000	,000	,000	1	,250**	,000	,000
S2_CV003	,000	,000	,000	,000	,250**	1	,000	,000
S1_CV004	,000	,000	,000	,000	,000	,000	1	,004
S2_CV004	,000	,000	,000	,000	,000	,000	,004	1