CHAPTER 8

정준 상관 분석

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 있다면 어떤 관계가 있는지 분석하는 것이다.

($X_1, X_2, ..., X_m$) 변수 군과 ($Y_1, Y_2, ..., Y_n$) 변수 군의 선형 관계를 분석한다. X's 변수 군과 Y's 변수 군간 $V = a_1 X_1 + a_2 X_2 + ... + a_m X_m$, $W = b_1 Y_1 + b_2 Y_2 + ... + b_n Y_n$ 1)상관 관계를 가장 크게 하는 계수를 구하고 2)그것을 이용하여 얻어내어 $Corr(U,V)$ 해석하는 방법이다. p 개 변수를 2 개의 변수 군으로 나누었다고 하자.

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

다음은 정준 상관 분석 의 특수한 예이다. 1)벡터 변수($\underline{x}_1, \underline{x}_2$)에 변수가 하나이면 단순 상관 계수가 되고 2)하나의 벡터 변수만 변수가 하나이면 이는 다중회귀 모형에서 결정계수

((determination coefficient= SSR / SST)의 제곱근($\sqrt{R^2}$)이다. 다중 회귀의 결정 계수는 종속 변수(변수가 하나인 벡터)와 설명변수의 선형결합($a_1X_1 + a_2X_2 + ... + a_pX_p$)간 상관 계수가 된다.

8.1. 정준 변수 구하기

8.1.1. 제일 정준 변수

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

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

위의 조건을 만족하는 $\underline{a}_1, \underline{b}_1$를 제일 정준 변수(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.

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

8.1.2. 제이 정준 변수

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

(1)V_2와 W_2은 각각 V_1과 W_1들과 독립이다.

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

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

다른 정준 변수도 같은 방법으로 구하면 된다. 일반적으로 현실에서는 해석의 어려움이 있어 정준 변수의 수는 2-3 개를 넘지 않는다.

8.1.3. 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.

8.1.4. 정준 상관 계수 검정

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. $H_{01}:\Sigma_{12}\neq 0$

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

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

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

8.2. 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.

8.2.1. 프로그램

SAS 데이터 만드는 방법은 이전과 동일하다.

```
DATA wheat;
     INFILE 'C:\TEMP\wheat.TXT';
     INPUT ID loc $ TYPE $ GROUP
          D_A D_P D_L D_B R_A R_P R_L R_B;
RUN;
```

우선 두 변수 군의 변수들간 상관 관계를 구해 보자. NOSIMPLE 옵션은 변수들의 기초 통계량(평균, 분산 등)을 출력하지 말라는 명령이다.

```
PROC CORR DATA=WHEAT NOSIMPLE;
    VAR D_A D_P D_L D_B ;
    WITH R_A R_P R_L R_B;
RUN;
```

피어슨 상관 계수, N = 172
H0: Rho=0 검정에 대한 Prob > |r|

	D_A	D_P	D_L	D_B
R_A	0.64006 <.0001	0.74907 <.0001	0.40169 <.0001	-0.00329 0.9659
R_P	0.67153 <.0001	0.85636 <.0001	0.57903 <.0001	-0.04050 0.5978
R_L	0.57441 <.0001	0.72169 <.0001	0.59331 <.0001	-0.07667 0.3175
R_B	0.43131 <.0001	0.39036 <.0001	0.06040 0.4312	0.04386 0.5678

오른쪽 면의 길이와 아래쪽 면의 폭은 상관 관계가 존재하지 않는다. 또한 오른쪽 면적은 폭은 아래쪽 면적은 길이와 폭의 상관 관계는 존재하지 않는다. 이처럼 쌍체(pair-wise) 상관 계수는 부분적인 상관 관계만을 알 수 있다.

```
PROC CANCORR DATA=WHEAT OUT=SCORE NCAN=2
    VPREFIX=DOWN WPREFIX=RIGHT CORR;
    VAR D_A D_P D_L D_B ;
    WITH R_A R_P R_L R_B;
RUN;
```

(1)OUT 옵션은 결과를 SAS 데이터 SCORES 에 저장하라는 명령이다.

(2)NCAN=2 옵션은 정준 상관 계수를 2 개만 구하라는 명령이다. (V1, W1), (V2,W2)

(3)CORR 은 변수들의 단순 상관 계수 값도 출력되게 한다. 위의 PROC CORR 과 같은 결과를 얻는다.

(4)VPREFIX=DOWN 는 V 대신 DOWN 이름으로 출력되게 한다. WPREFIX=RIGHT 는 W 대신 RIGHT 로 이름으로 출력되게 한다.

(5)이 옵션들을 사용하지 않으면 VAR 변수 그룹에는 V, WITH 변수 그룹에는 W 로 출력된다.

8.2.2. 출력 결과 해석

▣원 변수 상관 계수

Correlations Among the VAR Variables

	D_A	D_P	D_L	D_B
D_A	1.0000	0.8076	0.4535	0.2724
D_P	0.8076	1.0000	0.6139	0.1184
D_L	0.4535	0.6139	1.0000	0.3268
D_B	0.2724	0.1184	0.3268	1.0000

Correlations Among the WITH Variables

	R_A	R_P	R_L	R_B
R_A	1.0000	0.8631	0.6803	0.7262
R_P	0.8631	1.0000	0.8277	0.4571
R_L	0.6803	0.8277	1.0000	0.1975
R_B	0.7262	0.4571	0.1975	1.0000

Correlations Between the VAR Variables and the WITH Variables

	R_A	R_P	R_L	R_B
D_A	0.6401	0.6715	0.5744	0.4313
D_P	0.7491	0.8564	0.7217	0.3904
D_L	0.4017	0.5790	0.5933	0.0604
D_B	-0.0033	-0.0405	-0.0767	0.0439

변수 그룹 내의 변수들간의 상관 계수, 변수 그룹간 변수들의 상관 계수가 된다. 정준 상관 분석의 개략적인 결과를 예상할 수 있다.

▣CANONICAL 상관 계수

Canonical Correlation Analysis

$\hat{\rho}_1$ →
$\hat{\rho}_2$
$\hat{\rho}_3$
…

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation
1	0.881804	0.877840	0.017009	0.777579
2	0.398430	0.368710	0.064332	0.158747
3	0.249597	0.240671	0.071708	0.062299
4	0.003717	.	0.076471	0.000014

정준 상관 계수의 수는 4 개이다. (각 그룹내의 변수의 개수가 각각 4 개이므로) 정준 상관 계수는 $Corr(V1,W1), Corr(V2,W2)\cdots$ 상관 계수이다. 그럼 $Corr(V1,W2)$ 는 얼마인가? 당연히 0 이다.

▣CANONICAL 상관 계수 유의성 검정

Test of H0: The canonical correlations in the current row and all that follow are zero

	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.17545295	24.07	16	501.67	<.0001
2	0.78883314	4.57	9	401.72	<.0001
3	0.93768820	2.71	4	332	0.0300
4	0.99998618	0.00	1	167	0.9617

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 정준 변수

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 the VAR Variables

	DOWN1	DOWN2
D_A	0.0797	-0.2867
D_P	0.7768	-0.6688
D_L	0.2544	1.2224
D_B	-0.2537	-0.4587

Standardized Canonical Coefficients for the WITH Variables

	RIGHT1	RIGHT2
R_A	-0.0165	-0.8764
R_P	0.8941	0.1905
R_L	0.1601	0.8096
R_B	-0.0407	-0.4412

아래 면 변수 그룹의 제일 정준 변수

$$V1 = DOWN1 = 0.0797 * Z\_DA + 0.7768 * Z\_DP + 0.2544 * Z\_DL - 0.2537 * Z\_DB$$

아래 면 변수 그룹의 제이 정준 변수

$$V1 = DOWN2 = -0.2867 * Z\_DA - 0.6688 * Z\_DP + 1.2224 * Z\_DL - 0.4587 * Z\_DB$$

오른쪽 면 변수 그룹의 제일 정준 변수

$$W1 = RIGHT1 = -0.0165 * Z\_RA + 0.8941 * Z\_RP + 0.1601 * Z\_RL - 0.0407 * Z\_RB$$

오른쪽 면 변수 그룹의 제이 정준 변수

$$W1 = RIGHT2 = -0.8764 * Z\_RA + 0.1905 * Z\_RP + 0.8096 * Z\_RL - 0.4412 * Z\_RB$$

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

RAW 는 변수의 원래 값으로 구한 것이고 STANDADIZED 는 변수를 표준화하여 구한 것이다. 원 변수는 측정 단위와 분산(변동)에 의해 영향을 받으므로 표준화 변수를 사용하는 것이 좋다.

정준 변수의 이름은 선형 계수의 크기에 의해 이름을 붙일 수 있다. V1 은 아래면 원주, V2 는 아래면 길이, W1 은 오른쪽 면 원주, W2 는 오른쪽 면 면적과 길이(부호는 반대)로 붙일 수 있다. 여전히 이름을 부여하는 것은 다소 주관적이다.

▣정준 변수와 동일 군집 원 변수간의 상관 관계

그러나 정준 변수의 이름은 정준 변수와 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.

Correlations Between the VAR Variables and Their Canonical Variables

	DOWN1	DOWN2
D_A	0.7533	-0.3974
D_P	0.9673	-0.2042
D_L	0.6845	0.5319
D_B	-0.0569	-0.2165

Correlations Between the WITH Variables and Their Canonical Variables

	RIGHT1	RIGHT2
R_A	0.8345	-0.4817
R_P	0.9938	-0.0975
R_L	0.8809	0.2839
R_B	0.3876	0.8308

아래면 변수 그룹 제일 정준 변수 DOWN1 은 면적, 원주, 길이와 상관 관계가 높으므로 크기로 아래면 제이 정준 변수 DOWN2 는 길이로 이름 붙이면 적절할 것 같다. 오른쪽 면 제일 정준 변수(RIGHT1)도 크기, 제이 정준 변수(RIGHT2)는 폭으로 명명하면 적당하다.

▣정준 변수와 다른 군집 원 변수간의 상관 관계

Correlations Between the VAR Variables and the Canonical Variables of the WITH Variables

	RIGHT1	RIGHT2
D_A	0.6643	-0.1583
D_P	0.8530	-0.0814
D_L	0.6036	0.2119
D_B	-0.0502	-0.0863

Correlations Between the WITH Variables and the Canonical Variables of the VAR Variables

	DOWN1	DOWN2
R_A	0.7359	-0.1919
R_P	0.8763	-0.0389
R_L	0.7768	0.1131
R_B	0.3418	-0.3310

오른쪽 면의 크기(RIGHT1)는 아래 면의 면적, 원주, 길이와 양의 상관 관계가 있다. 즉 오른쪽 면의 크기가 클수록 아래 면의 면적, 원주, 길이는 커진다. 오른쪽 면의 폭(RIGHT2)은 아래 면의 길이, 원주, 폭과 음의 상관 관계가 있고 길이와는 음의 상관 관계가 있다.

아래 면의 크기(DOWN1)는 오른쪽 면의 면적, 원주, 길이, 폭과 양의 상관 관계가 존재하고 길이(DOWN2)는 오른쪽 면의 면적, 원주, 폭과는 음의 상관 관계가 존재한다.

▣정준 변수간 상관 계수 해석 및 산점도

정준 변수간의 상관 관계는 이미 알고 있다. 확인하기 위하여 다음 프로그램을 돌려보자. 앞 프로그램에서 정준 변수는 OUT 옵션에 의해 SCORE 에 저장되어 있다.

```
PROC CORR DATA=SCORE NOSIMPLE;
    VAR DOWN1 DOWN2;
    WITH RIGHT1 RIGHT2;
RUN;
```

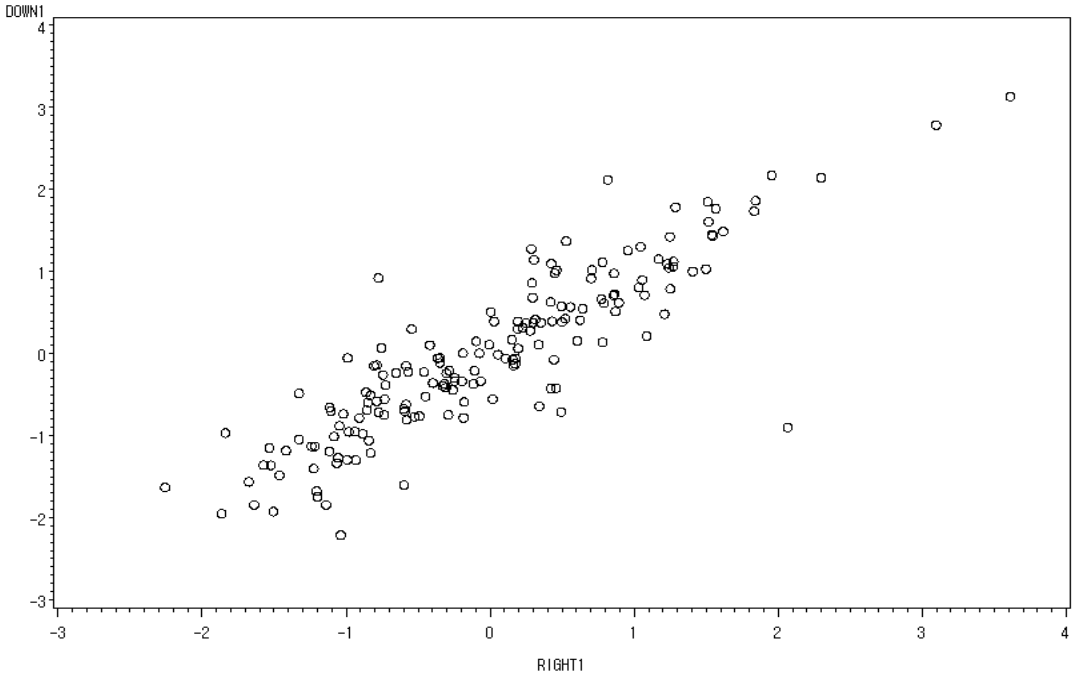
피어슨 상관 계수, N = 172
H0: Rho=0 검정에 대한 Prob > |r|

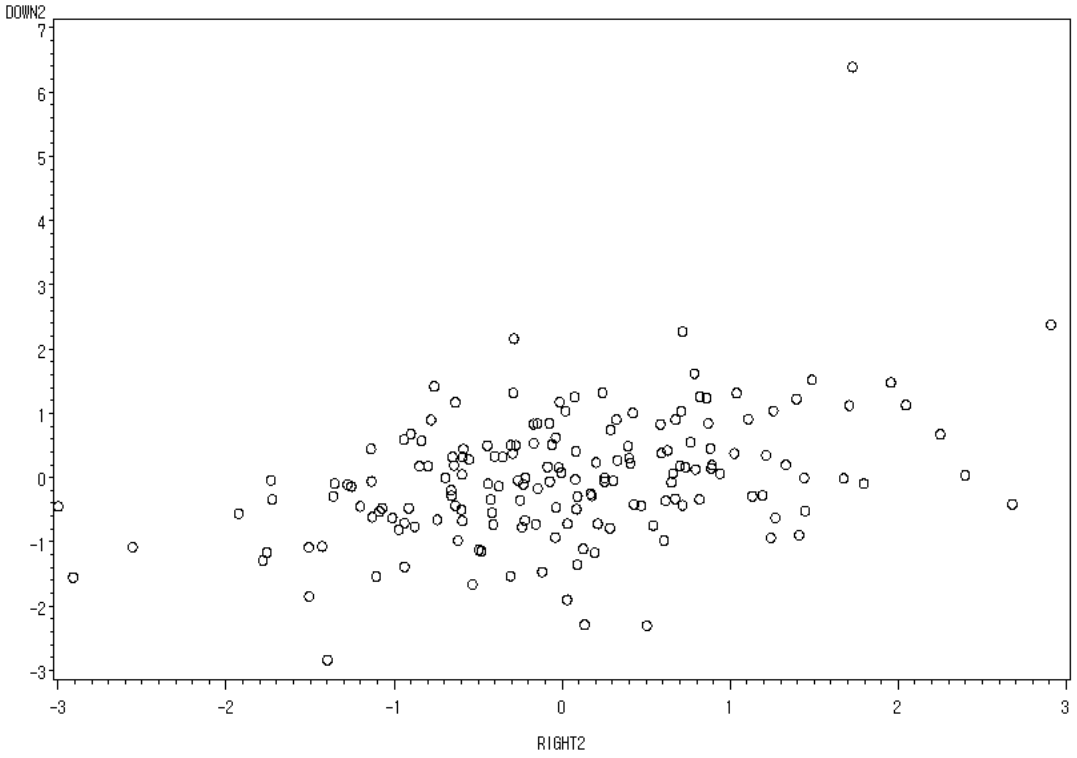
	DOWN1	DOWN2
RIGHT1	0.88180 <.0001	0.00000 1.0000
RIGHT2	0.00000 1.0000	0.39843 <.0001

(DOWN1, RIGHT1)이 (DOWN2, RIGHT2)보다 높고 (DOWN1, RIGHT2), (DOWN2, RIGHT1)의 각 상관 계수는 0 이다. 우리는 V1, W1, V2, W2 를 구할 때 이렇게 되는 것을 구했다. 산점도를 그려보자.

```
PROC GPLOT DATA=SCORE;
    SYMBOL V=CIRCLE;
    PLOT DOWN1*RIGHT1;
    PLOT DOWN2*RIGHT2;
RUN;
```

제일 정준 상관 계수가 0.88 로 제이 정준 상관 계수 0.39 에 비해 1 에 많이 가까우므로 점들이 직선에 모여 있다. 이 산점도는 정준 상관 계수를 시각화한 것에 지나지 않는다.




DOWN2
7
6
5
4
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3
RIGHT2

[EXERCISE]

◈체육관을 찾는 중년 남자들 대상으로 6 개 항목을 측정하였다.

키 가슴둘레 맥박 턱걸이 윗몸 일으키기 줄넘기 (n=20)

191	36	50	5	162	60
189	37	52	2	110	60
193	38	58	12	101	101
162	35	62	12	105	37
189	35	46	13	155	58
182	36	56	4	101	42
211	38	56	8	101	38
167	34	60	6	125	40
176	31	74	15	200	40
154	33	56	17	251	250
169	34	50	17	120	38
166	33	52	13	210	115
154	34	64	14	215	105
247	46	50	1	50	50
193	36	46	6	70	31
202	37	62	12	210	120
176	37	54	4	60	25
157	32	52	11	230	80
156	33	54	15	225	73
138	33	68	2	110	43

체격 조건(키, 가슴둘레, 맥박)과 운동 능력간 정준 상관 분석을 실시하시오. 8 장에 설명된 순서대로 결과를 해석하시오.