CHAPTER 3.

다변량 분포

통계 분석의 시작은 적합한 그래프를 그려 대략적인 결과를 미리 예상하는 것이다. 변수가 하나인 경우에는 상자 그림(box-whisker plot)이나 줄기-잎 그림(stem and leaf plot)을 그려 분포 형태, 중앙 위치, 산포 정도를 파악한다. 물론 분포가 정규 분포를 따르는지(W-통계량), 정확한 중앙 위치(평균, 중앙값)나 산포도(표준 편차) 값을 알아 보기 위해서는 계산을 해야 한다. 변수가 2 개 이상인 경우에는 변수들간의 관계를 알아보는 산점도(scatter plot) 그래프를 그린다. 두 변수의 관계는 상관 분석이나 회귀 분석을 이용하면 된다.

그럼 변수가 3 개 이상인 경우에는 어떤 그래프가 적절할까? 3 차원 그래프를 그릴 수 있으나 그래프를 이해하고 정보를 얻는데 한계가 있고 4 차원이상의 다차원 그래프를 그릴 수 있는 방법이 없어 여러 개의 산점도를 그려 변수들간의 퍼즐 맞추기 식 해석에 의존한다. 이를 산점도 행렬이라고 한다. 그러나 산점도 행렬만으로는 변수들 간의 부분 관계만 알아 볼 수 있을 뿐 전체적인 관계나 모든 관계를 고려했을 때 이상치의 존재 여부를 아는데 어려움이 있다. 이런 이유로 $p(\geq 3)$ 개 변수를 축약하여 2 차원 그래프에 표현하는 여러 방법이 제안되어 사용되고 있다. 이 절에서는 다변량 정규분포의 개념과 다변량 데이터를 2 차원에 공간에 표현하는(마치 산점도처럼) 방법을 살펴보기로 하자.

3.1. 다변량 정규 분포

변수의 개수가 p 인 경우 변수 벡터는 다음과 같이 정의하며 각 열의 x_i 는 관측치가 아니라 변수이다.

$$\underline{x}_p = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

차수가 p 인 변수 벡터가 평균이 $\underline{\mu}$ 이고 공분산이 Σ 인 다변량 정규분포 (Multivariate Normal Distribution)을 따른다면, 즉 $\underline{x}_p \sim N_p(\underline{\mu}_p, \Sigma_{p\times p})$ 이면

$$f_{\underline{x}}(\underline{x};\underline{\mu},\Sigma) = \frac{1}{(2\pi)^{p/2}\,|\Sigma|^{1/2}} \exp[-1/2(\underline{x}-\underline{\mu})'\Sigma^{-1}(\underline{x}-\underline{\mu})]$$

$$\underline{\mu}_p = \begin{bmatrix} E(x_1) \\ E(x_2) \\ \vdots \\ E(x_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}, \quad \Sigma = Cov(\underline{x}) = E(\underline{x}-\underline{\mu})(\underline{x}-\underline{\mu})' = \begin{bmatrix} \sigma_{11}\ \sigma_{12}\cdots\sigma_{1p} \\ \sigma_{21}\ \sigma_{22}\cdots\sigma_{2p} \\ \vdots \\ \sigma_{p1}\ \sigma_{p2}\cdots\sigma_{pp} \end{bmatrix}$$

공분산 $\sigma_{ij} = Cov(x_i, x_j) = E(x_i-\mu_i)(x_j-\mu_j) \quad for\ i \neq j$

분산 $\sigma_{ii} = Cov(x_i, x_i) = Var(x_i) = E(x_i-\mu_i)^2$

►참고

상관 계수 $\rho_{ij} = \dfrac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}\ for\ i \neq j$ or $\rho_{ii} = 1\ for\ i = j$

상관 계수 행렬 $R = \begin{bmatrix} \rho_{11}\ \rho_{12}\cdots\rho_{1p} \\ \rho_{21}\ \rho_{22}\cdots\rho_{2p} \\ \vdots \\ \rho_{p1}\ \rho_{p2}\cdots\rho_{pp} \end{bmatrix}$

모평균 추정치($\hat{\mu}_j$)는 표본 평균 $\bar{x}_j$, 모집단 표본 분산 추정치($\hat{\sigma}_{jj}$)는 표본 분산 s_{jj} 이다.

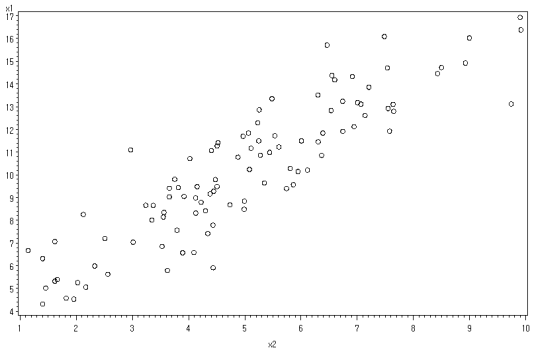
차수가 2 인 변수 벡터가 평균이 $\underline{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ 이고, 공분산이 $\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$ 인 이변량 정규분포를 따르는 경우 분포함수는 다음과 같다.

$$f_{\underline{x}}(\underline{x}_2;\underline{\mu},\Sigma)=\frac{1}{(2\pi)^{p/2}\,|\,\Sigma\,|^{1/2}}\exp[-1/2(\underline{x}-\underline{\mu})'\Sigma^{-1}(\underline{x}-\underline{\mu})]\,,\quad \underline{\mu}_2=\begin{bmatrix}E(x_1)\\E(x_2)\end{bmatrix}=\begin{bmatrix}\mu_1\\\mu_2\end{bmatrix},\quad \Sigma=\begin{bmatrix}\sigma_{11}\ \sigma_{12}\\\sigma_{21}\ \sigma_{22}\end{bmatrix}$$

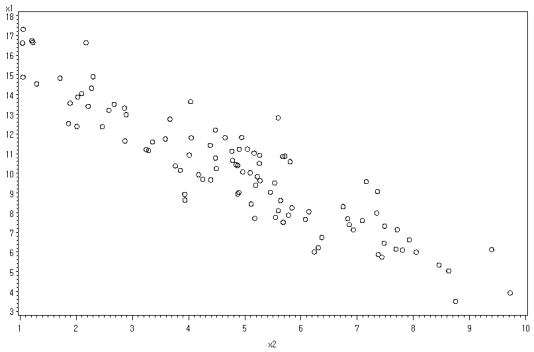
3.2. 상관 계수

3.2.1. 계산식

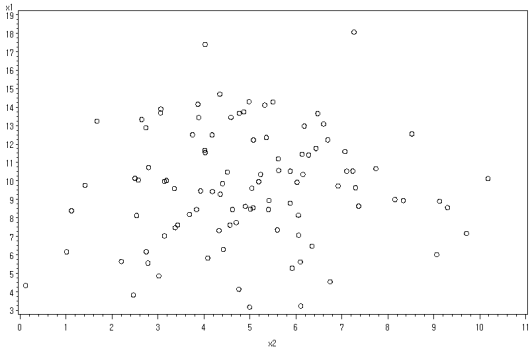
$$r=\frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}}=\frac{\sum(X-\bar{X})(Y-\bar{Y})/(n-1)}{\sqrt{\sum(X-\bar{X})^2/(n-1)}\sqrt{\sum(Y-\bar{Y})^2/(n-1)}}=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2}\sqrt{\sum(Y-\bar{Y})^2}}$$



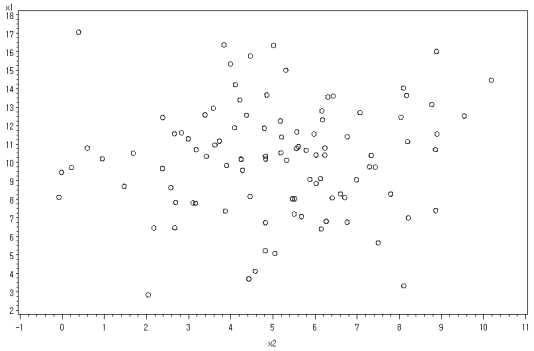
$r = 0.9$



$r = -0.9$



$r = 0.1$



$r = -0.1$

상관 계수는 두 변수간의 선형 관계를 (linear association) 측정한 값이다.

(1)–1 과 1 사이의 값이다.

(2)1 에 가까우면 양의 선형 상관 관계가 존재한다. 한 변수의 값이 증가(감소)하면 다른 변수 값도 증가(감소)한다.

(3)–1 에 가까우면 음의 선형 상관 관계가 존재한다. 한 변수의 값이 증가(감소)하면 다른 변수 값은 감소(증가)한다.

(4)0 이면 선형 상관 관계가 없다는 것이다.

두 변수의 상관 관계가 높다는 것은 두 변수가 동일한(comparable) 개념을 측정한다는 의미도 담고 있다. 그러므로 변수를 축약하거나 개체를 분류하는데 사용되는 다변량 분석이 공분산, 혹은 상관 계수 개념을 사용하게 된다.

3.2.2. 추정과 검정

◈귀무가설 $H_0 : \rho = 0$ 에 대한 가설 검정

(1)검정통계량

$$T = \frac{r}{\sqrt{(1-r^2)/(n-2)}} \sim t(n-2) \quad \text{where} \quad r = \frac{E[(X-\bar{X})(Y-\bar{Y})]}{\sqrt{E(X-\bar{X})^2}\sqrt{(Y-\bar{Y})^2}} \text{ (표본 상관 계수)}$$

(2)귀무가설 $H_0 : \rho = \rho_0 \neq 0$ 에 대한 가설 검정 및 신뢰 구간

$z = 0.5 \ln \frac{1+r}{1-r} \overset{app}{\sim} Normal(0.5 \ln \frac{1+\rho}{1-\rho}, \frac{1}{n-3})$ 이를 이용하여 가설을 검정하면 된다.

(3)신뢰구간 계산하기

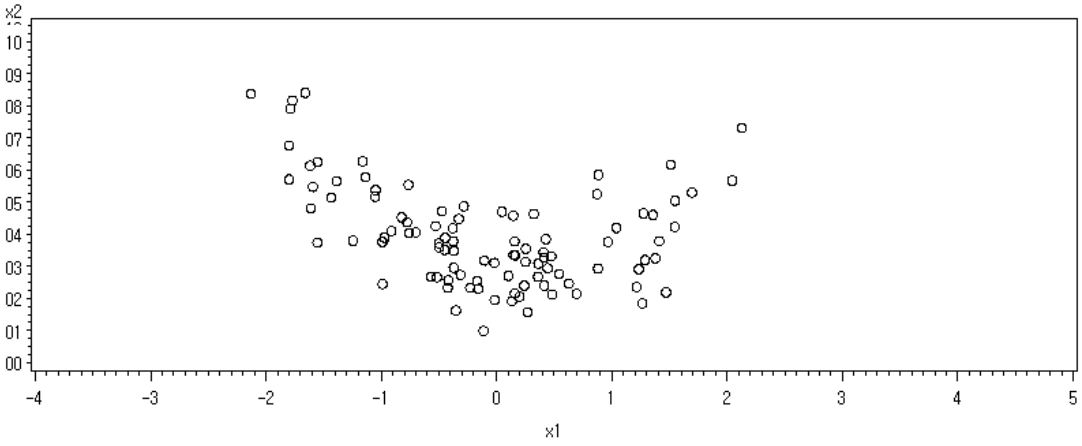
$z = 0.5 \ln \frac{1+\rho}{1-\rho} \pm z_{\alpha/2} \frac{1}{\sqrt{n-3}}$ 이용하여 (L,U) 를 계산한다.

그러므로 모상관 계수 ρ 의 신뢰구간은 $L_\rho = (e^{2U}-1)/(e^{2U}+1),\ U_\rho = (e^{2L}-1)/(e^{2L}+1)$

$n=50,\ r=0.527$ 인 경우 예를 들어 보자. $z=0.5\ln\frac{1+r}{1-r}=0.5\ln\frac{1+0.527}{1-0.527}=0.586$ 이므로 95% 신뢰구간은 $(L=0.3, U=0.872)$ 이다.

3.2.3. 상관 관계 의미

상관 계수가 0 에 가깝다는 것은 선형 상관 관계가 없다는 것이지 함수 관계가 없다는 것은 아니다. 다음 산점도를 살펴보면 상관 계수는 0 이지만 두 변수는 이차식에 의한 ($X_2=100+X_1^2-0.4X_1$) 상관 관계가 존재한다.



상관 계수의 크기는 자료의 크기가 커지면 증가하므로 값 자체의 크기가 의미가 있는 곳이 아니라 자료로부터 검정 시 계산되는 유의 확률(p-값)의 크기를 이용하면 된다. 다음은 유의성을 검정하지 않고도 유의한 정도를 알아볼 수 있는 기준이다.

(1)실험실 자료와 같이 연구자가 자료 수집을 control 할 수 있는 경우는 0.9 이다.

(2)연구자가 control 하기 어려운 경우는 0.7 정도이다.

(3)일반적으로 자료의 수가 20-30 정도인 경우 0.6 정도를 생각한다.

(4)설문 조사의 리커드 척도와 같이 변수가 가질 수 있는 값이 한정된 경우 (1-5 점, 물론 여러 문항을 합쳐 평균을 이용하는 경우에는 다소 문제가 해결되지만) 상관 계수는 매우 낮다. 같은 이유로 결정 계수($R^2=SSR/SST$)도 매우 낮다. 그러므로 이런 경우는

비모수 상관 계수를 구하는 것을 권한다. Spearman 순위 상관 계수, Kendall's Tau 는 비모수 상관 계수 분석 방법이다.

3.2.4. SAS 프로그램

```
DATA CLASS;
    INFILE "D:\TEMP\CLASS.txt" delimiter='09'x MISSOVER DSD;
    INPUT NAME $ AGE GENDER $ HEIGHT WEIGHT;
RUN;

PROC CORR DATA=CLASS NOPRINT OUTP=OUT1;
    VAR HEIGHT WEIGHT;
RUN;

PROC PRINT DATA=OUT1;
RUN;
```

delimiter='0.9'x MISSOVER DSD 옵션은 메모장 데이터 입력 시 Tab 키를 사용해 코딩한 데이터를 읽어올 때 사용한다.

(1)NOPRINT 옵션으로 output 이 출력하지 않는다. NOPRINT 옵션을 없애면 각 변수들의 기초 통계량(평균, 표준 편차 등)과 상관 계수 값, 유의 확률(p-값)등이 출력된다. 만약 기초 통계량을 출력하지 않으려면 NOSIMPLE 옵션을 쓰면 된다.

(2)OUTP=OUT1 은 상관 계수 분석 결과 중 필요한 통계량을 OUT1 이라는 SAS data 에 저장하라는 것이다. 아래 출력 결과를 보자.

Obs	_TYPE_	_NAME_	HEIGHT	WEIGHT
1	MEAN		62.6410	105.256
2	STD		4.2580	22.432
3	N		39.0000	39.000
4	CORR	HEIGHT	1.0000	0.708
5	CORR	WEIGHT	0.7077	1.000

3.3. 상관 계수를 이용한 변수 분류

3.3.1. 문제 제기

A 회사에서는 48 명의 지원자에 대해 그들의 능력을 10 점 만점으로 15 개 영역에 대해 조사하여 가장 점수가 높은 지원자 6 명을 선발하려고 한다. [APPLICANT.TXT]/ [Applied Multivariate Methods for Data Analysts, Dallas E. Johnson, p. 101]

- ID(지원자 번호)
- Letter(이력서 X_1)
- Appearance(외모 X_2)
- Academic Ability(학교 성적 X_3)
- Likeability(친밀감 X_4)
- Self-Confidence(자신감 X_5)
- Lucidity(명석 X_6)
- Honest(진실 X_7)
- Experience(경험 X_9)
- Drive(추진력 X_{10})
- Ambition(야망 X_{11})
- Potential(장래성 X_{13})
- Keenness to Join(화합 X_{14})
- Suitability(업무 적합성 X_{15})
- Salesmanship(마케팅 능력 X_8)
- Grasp Concept(개념 파악 능력 X_{12})

3.3.2. 평균 이용

15 개 항목 점수의 평균을 계산하여 $AVG=(L+AP+AA+\ldots)/15$ 점수가 가장 높은 6 명을 선발하면 된다. 그러나 이는 문제가 있다. 평균 점수가 계산될 때 모든 측정 항목들이 동일한 가중치(1/15)로 반영되므로 유사한 능력을 측정하는 항목이 여러 개 있다면 이 분야 점수가 높은 지원자가 선발될 것이다.

아래 프로그램을 이용하여 HOMEWORK#2-2 을 하시오. 단 변수 명은 적절하게 사용하시오.

```
data app;
    input id x1-x15;
    avg=mean(of x1-x15); /*mean(x1, x2, ..., x15)*/
    cards;
```

/* */은 주석문으로 안에 있는 내용은 실행되지 않으며, /* */안의 내용은 mean 함수의 또 다른 사용 방법을 나타낸 것이다.

```
proc sort data=app out=out1;
    by descending avg;
run;

proc print data=out1;
    var avg id x1-x15;
run;
```

Obs	avg	id	x1	x2	x3
1	9.60000	40	10	6	9
2	9.46667	20	10	6	9

3.3.3. 가중 평균

회사가 자체적으로 각 변수(측정 항목)에 가중치(weight)를 부여하거나 통계 분석 방법에 의해 가중 평균을 계산한다. ($Avg = w_1 L + w_2 AP + ... + w_{15} SU$, where $\sum_i w_i = 1$)

주관적인 가중치 방법은 경험을 바탕으로 설정한다.

A 회사에서는 마케팅 부서에 적합한 지원자를 뽑기 위하여 외모, 친밀감, 진실, 마케팅 능력, 경험에 높은 가중치를 주려고 한다. 5 개 능력에 2 배의 가중치를 주고 지원자의 능력 점수를 계산하는 방법을 살펴 보자. 5 개 항목에 2 배의 가중치를 주므로 가중치 분모는 20 이다.

```
data app;
input id x1-x15;
weight=(x1+2*x2+x3+2*x4+x5+x6+2*x7+2*x8+2*x9+x10+x11+x12+x13+x14+x15)/20;
cards;
```

```
proc sort data=app out=out1;
    by descending weight;
run;

proc print data=out1;
    var weight id x1-x15;
run;
```

Obs	weight	id	x1	x2	x3	x4
1	9.50	40	10	6	9	10
2	9.40	39	10	6	9	10
3	8.90	8	9	9	9	8
4	8.60	7	9	9	8	8
5	8.55	23	7	10	7	9
6	8.45	2	9	10	5	8

3.3.4. 상관 계수 이용

가중치를 부여하는 것보다 객관적인 방법을 생각해 보자. 상관 관계가 높은 변수는 유사 개념을 측정한다는 것을 의미하므로 상관 계수를 이용하여 변수를 분류(grouping)하고 이를 가중치로 이용하는 방법이다. 다음은 15 개의 변수의 상관 계수 크기만으로 분류한 것이다. 여러분의 결과와 상이할 가능성이 높다. 변수가 많아지면 변수들간 상관 계수 관계가 매우 복잡해져 분류가 다소 주관적이다. 그러므로 변수를 분류하는 방법으로 요인 분석, 변수를 축약하는 방법으로 주성분 분석을 다루게 된다.

```
proc corr data=app;
    var x1-x15;
run;
```

Group 1	X_5, X_6, X_8, X_{10}, X_{11}, X_{12}, X_{13}
Group 2	X_1, X_9, X_{15}
Group 3	X_4, X_7, X_{14}
Group 4	X_2
Group 5	X_3

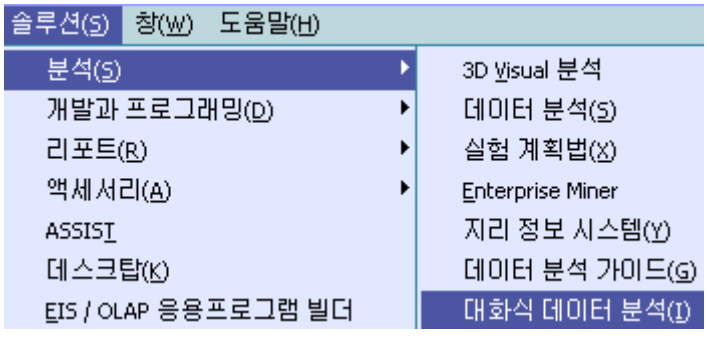
위의 결과에 의하면 15 개 변수를 (개념, 능력) 측정하였지만 실제는 5 개의 그룹 능력을 측정한 것과 동일하다. group 1 능력(어떤 능력인지는 항목을 자세히 살펴야겠지만)은 7 개 항목이나 측정되고 있으므로 방법 1 을 사용하면 이 분야에 점수가 높은 학생이 선택될 가능성이 높으므로 다음과 같이 가중 평균을 구한다.

$$AVG_w = [(X5 + X6 + \ldots + X13)/7 + (X1 + X9 + X15)/3 + \ldots + X14]/5$$

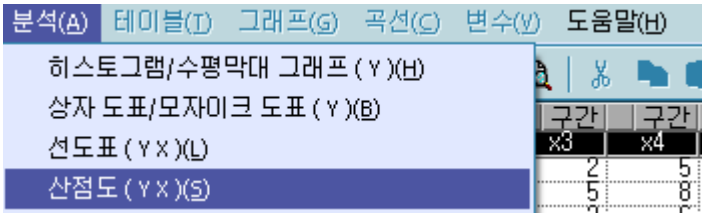
3.4. 다변량 데이터 그림

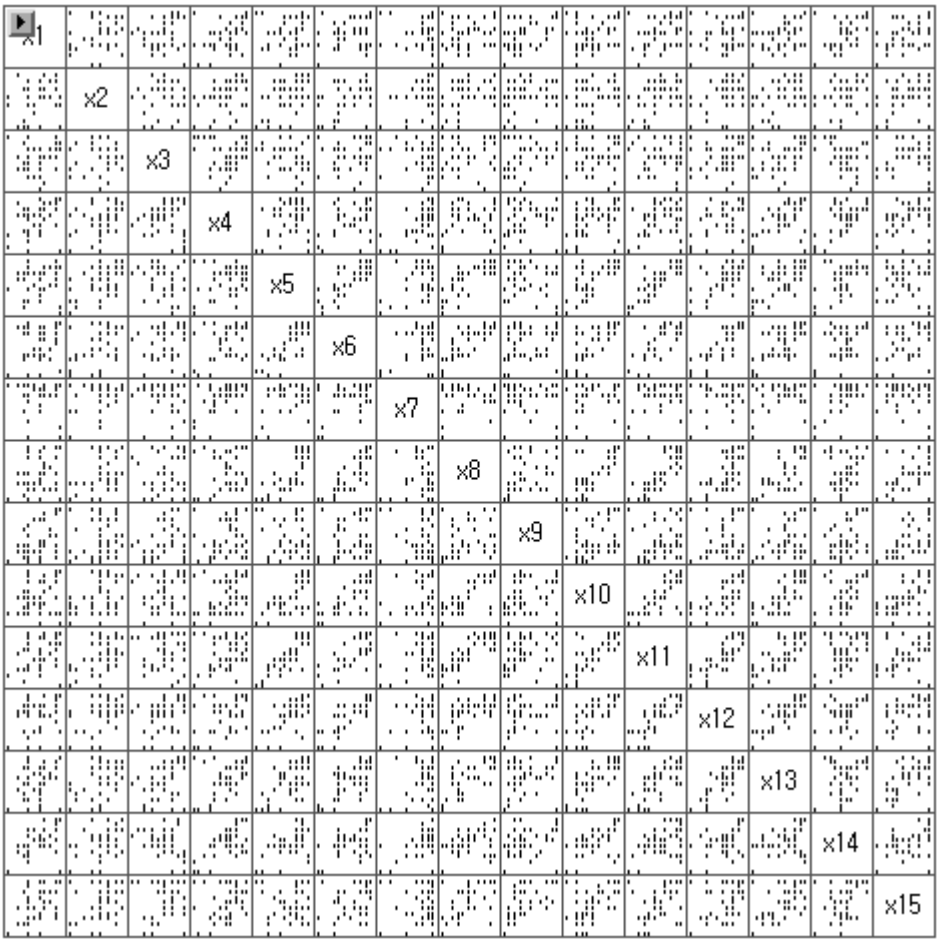
3.4.1. 산점도 행렬

변수가 두 개인 경우 변수들 간의 관계나 이상치를 알아볼 적절한 그래프는 산점도(scatter plot)이다. 변수가 3 개 이상인 경우는 각 변수의 쌍에 대해 산점도를 그리면 된다. 이를 산점도 행렬(scatter plot matrix)이라 한다. SAS 데이터가 만들어지면 SAS/INSIGHT 를 이용하여 산점도 행렬을 그릴 수 있다.



산점도를 그리려는 변수를 선택한다. 변수 선택 시 CTRL 키를 누르고 변수를 마우스로 클릭하면 된다. 그런 후 분석에서 산점도 메뉴를 선택하시오.





산점도는 두 변수 간의 함수 관계를 시각적으로 표현할 수 있어 관계가 높은 변수를 쉽게 찾아 낼 수 있다는 장점이 있다. 또한 두 변수 간의 관계에서 이상치가 존재하는지

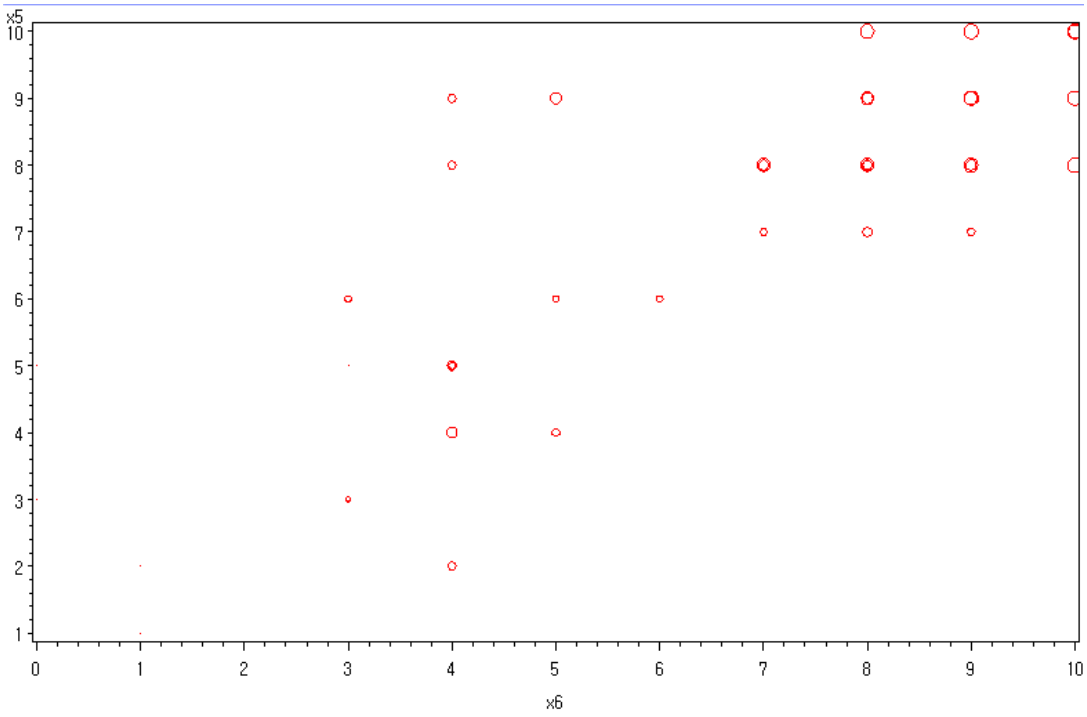
알 수 있다. 그러나 변수 전체를 고려한 상관 관계나 이상치를 판단할 수는 없으므로 주성분 분석에 의해 변수의 차원을 줄이게 된다.

3.4.2. Bubble 그림

변수 2 개만을 한 그림에 나타내는 산점도와 유사한 개념으로 변수 3 개를 한 그림에 표현한 그래프를 Bubble 그림(혹은 blob 그림)이라 한다. 2 개의 변수로 산점도를 그리고 점을 다른 변수의 크기로 나타낸 그래프이다. 다음 프로그램은 GC 를 y 축, PO 변수를 x 축으로 하고 방울 크기를 AM 변수로 하여 Bubble plot 을 그리는 것이다.

```
GOPTIONS RESET=ALL;
PROC GPLOT DATA=APP;
    BUBBLE X5*X6=X8 /BCOLOR=RED;
RUN;
```

	x5	x6	x8
x5	1.00000	0.80755 <.0001	0.79963 <.0001
x6	0.80755 <.0001	1.00000	0.81802 <.0001
x8	0.79963 <.0001	0.81802 <.0001	1.00000



방울의 크기가 클수록 관측치의 값이 크다는 것을 의미한다. X_5 (y 축)와 X_6 (x 축)의 관계는 점의 크기와 관계없이 점들의 흩어진 경향만 보면 된다. X_5 와 X_6 는 양의 상관

관계가(X_5 의 값이 커지면 X_6 의 값도 커진다. 상관 계수=0.80755) 존재한다. X_5 와 X_8 의 관계를 보면 X_5 값이 커질수록(y 축으로 이동) 원의 크기가 커지므로 양의 상관 관계($r = 0.79963$)가 있다. X_6 와 X_8 의 관계를 보면 X_6 값이 커질수록(x 축으로 이동) 원의 크기가 커지므로 양의 상관 관계($r = 0.81802$)가 있다.

Chernoff 얼굴 그림은 사람의 얼굴의 눈, 미간, 눈썹, 입, 코, 귀 등의 위치, 넓이, 크기 등으로 변수 값을 나타냄으로써 측정 변수가 3 개 이상인 다변량 데이터의 개체를 표현하는데 사용된다.

[EXERCISE]

(1)다음 물음에 답하여라. [CLASS.txt]

①키, 몸무게 상관 계수가 유의한지 검정하시오. (유의수준=0.05)

②키와 몸무게 상관 계수의 95% 신뢰구간을 구하시오.

(2)주어진 자료를 이용하여 물음에 답하여라. [APPLICANT.txt]

①평균을 이용하여 6 명을 선발하시오.

②여러분 자신이 정한 가중치에 의해 가중 평균을 계산하고 점수가 가장 좋은 6 명을 선택하시오.

③상관 계수에 의해 변수를 그룹화 하고 가중 평균을 구하고 6 명을 선발하시오.

(3)경찰에 지원한 50 명의 신체적 특성 15 개를 측정한 것이다. [POLICE.txt]

▪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: 러닝 머신에서 (treadmill) 제자리 달리고 5분 후 맥박	
▪SPEED: 러닝 머신에서 제자리 달리기 최대 속력	
▪ENDUR: 러닝 머신에서 달릴 수 있는 최대 시간 (분)	
▪FAT: 비만도	

①15 개 변수의 산점도 행렬을 그리고 해석하시오.

②가슴 넓이, 턱걸이 회수, 비만도 변수에 대해 Bubble plot 을 그리고 해석하시오.