Matrix Algebra (II)

Chapter 1 Introduction

Chapter 2 Equations and Algebra

Chapter 3 Optimization

Chapter 4 Special Topics

Department of Statistics
Professor Yong-Seok Choi

Chapter 1 Introduction

Key-words

- Matrix
- Addition(Matrix sum)
- Scalar multiplication
- Product of matrices
- Data matrix, Centred data matrix, Covariance matrix, Correlation matrix

1.1 Definition of a matrix

■ Definition of a matrix

A matrix is a rectangular or square array of numbers arranged in rows and columns.
X has n rows and p columns ;

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

(x_{ij} : The element in the ith row and jth column of a matrix X)

$$X = \{x_{ij}\} \text{ , for } i = 1, 2, \dots, n \text{ and } j = 1, 2, \dots, p$$

: An alternative and briefer form

Note: X has n rows and p columns. The size of X is $n \times p$.

1.2 Necessity of matrix

In statistics, why do you need some matrix representations?

Data matrix X, Covariance Matrix S, Correlation Matrix R, Distance Matrix D, etc.

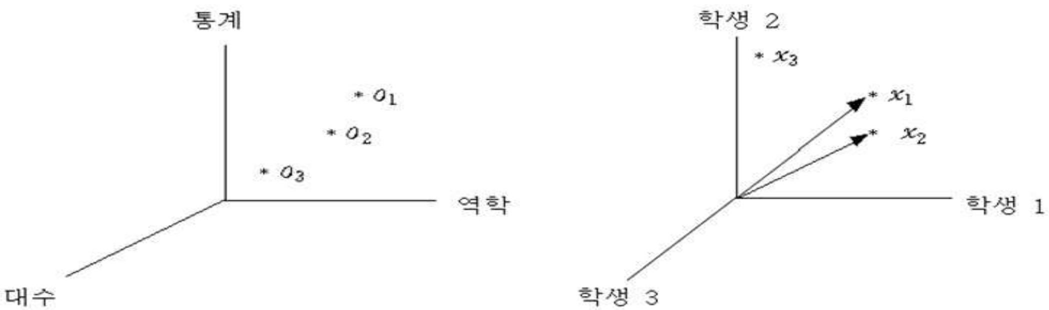
■ Multivariate data

Consider data containing observations on two or more variables.

In general, data contain the n observations $O_1, \ldots, O_n$ and p variables $x_1, \ldots, x_n$.

❖ Geometrical Representations of 3-dimensinal space

Students	Mechanics	Algebra	Statistics
1	77	67	81
2	63	80	81
3	50	50	50



a) $n = 3$ points in p-space

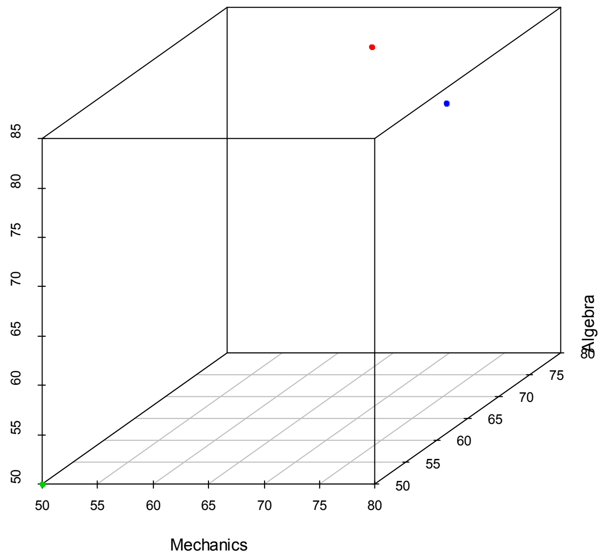
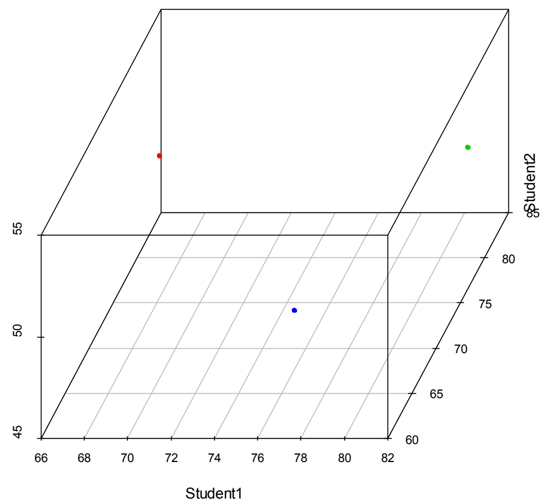
b) $p = 3$ points in n-space

$D = \{d_{ik}\}\ i,k = 1,2,3$

$R = \{r_{jk}\}\quad j,k = 1,2,3$

3×3 data matrix

$$X = \begin{bmatrix} 77 & 67 & 81 \\ 63 & 80 & 81 \\ 50 & 50 & 50 \end{bmatrix}$$

<table>
<tr><td colspan="2">3-Dimensional Scatterplot Using the scatterplot3d in R</td></tr>
<tr><td colspan="2">X<-matrix(c(77, 67, 81,
63, 80, 81,
50, 50, 50), byrow=T, nrow=3)
library(scatterplot3d)</td></tr>
<tr><td>Mechanics<-X[,1]
Algebra<-X[,2]
Statistics<-X[,3]
scatterplot3d(Mechanics,Algebra, Statistics,
color=c("blue", "red", "green3"), pch=20)</td><td>Student1<-X[1,]
Student2<-X[2,]
Student3<-X[3,]
scatterplot3d(Student1,Student2, Student3,
color=c("blue", "red", "green3"), pch=20)</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>library(rgl)
plot3d(Mechanics, Algebra, Statistics,
col=c("blue", "red", "green3"), size=3)</td><td>library(rgl)
plot3d(Student1,Student2, Student3,
col=c("blue", "red", "green3"), size=3)</td></tr>
</table>

❖ [Data 1] Examination marks on 5 subjects (Mardia et al., 1979, pp. 3-4)

C: closed-book

O: open-book

Obs	Mechanics	Vectors	Algebra	Analysis	Statistics
개체	역학(c)	벡터(c)	대수(o)	해석(o)	통계(o)
1	77	82	67	67	81
2	63	78	80	70	81
3	75	73	71	66	81
4	55	72	63	70	68
5	63	63	65	70	63
6	53	61	72	64	73
7	51	67	65	65	68
8	59	70	68	62	56
9	62	60	58	62	70
10	64	72	60	62	45
			⋮		
80	30	24	43	33	25
81	3	9	51	47	40
82	7	51	43	17	22
83	15	40	43	23	18
84	15	38	39	28	17
85	5	30	44	36	18
86	12	30	32	35	21
87	5	26	15	20	20
88	0	40	21	9	14

Questions:

- How to combine or average these marks?
- Relationship between open-book and closed-book ?

❖ [Data 2] Protein consumption in European 25 countries

Protein Source: Meat Pigs Eggs Milk Fish Cereals Starchy foods Nuts/ Oil-seeds Fruits/Vegetables

	소고기	돼지·닭	계란	우유	생선	곡식	전분	콩·견과	과일·채소
알바니아	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
오스트리아	8.9	14.0	4.3	19.9	2.1	28.0	3.6	1.3	4.3
벨기에-룩셈	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4.0
불가리아	7.8	6.0	1.6	8.3	1.2	56.7	1.1	3.7	4.2
체코	9.7	11.4	2.8	12.5	2.0	34.3	5.0	1.1	4.0
덴마크	10.6	10.8	3.7	25.0	9.9	21.9	4.8	0.7	2.4
동독	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
핀랜드	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1.0	1.4
프랑스	18.0	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
그리스	10.2	3.0	2.8	17.6	5.9	41.7	2.2	7.8	6.5
					⋮				
포르투갈	6.2	3.7	1.1	4.9	14.2	27.0	5.9	4.7	7.9
루마니아	6.2	6.3	1.5	11.1	1.0	49.6	3.1	5.3	2.8
스페인	7.1	3.4	3.1	8.6	7.0	29.2	5.7	5.9	7.2
스웨덴	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2.0
스위스	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9
영국	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
소련	9.3	4.6	2.1	16.6	3.0	43.6	6.4	3.4	2.9
서독	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8
유고슬라비아	4.4	5.0	1.2	9.5	0.6	55.9	3.0	5.7	3.2

Questions:

- Which countries have high consumption of each kind of protein?
- Relationship between protein sources ?

1.3 Basic operations

■ Addition

For the two matrices of the same size

$$A=\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{np} \end{pmatrix} \quad \text{and} \quad B=\begin{pmatrix} b_{11} & b_{12} & \dots & b_{1p} \\ b_{21} & b_{22} & \dots & b_{2p} \\ \vdots & \vdots & & \vdots \\ b_{n1} & b_{n2} & \dots & b_{np} \end{pmatrix},$$

the sum of two matrices A and B : $A+B=\begin{pmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \dots & a_{1p}+b_{1p} \\ a_{21}+b_{21} & a_{22}+b_{22} & \dots & a_{2p}+b_{2p} \\ \vdots & \vdots & & \vdots \\ a_{n1}+b_{n1} & a_{n2}+b_{n2} & \dots & a_{np}+b_{np} \end{pmatrix}$

: The matrix sum A plus B

It is the matrix formed by adding the matrices A and B element by element.

$C \equiv A+B$, $\{c_{ij}\}=\{a_{ij}+b_{ij}\}$, for $i=1,2,\ \dots, n$ and $j=1,2,\ \dots, p$

■ Scalar multiplication

For

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}, \quad k \text{ : any scalar}$$

the product of a scalar and a matrix :

$$kX\ (= Xk) = \begin{pmatrix} kx_{11} & kx_{12} & \dots & kx_{1p} \\ kx_{21} & kx_{22} & \dots & kx_{2p} \\ \vdots & \vdots & & \vdots \\ kx_{n1} & kx_{n2} & \dots & kx_{np} \end{pmatrix}$$

$$= \{kx_{ij}\} \text{ for } i = 1, 2, \dots, n \text{ and } j = 1, 2, \dots, p$$

: The matrix X multiplied by the scalar k.

■ Product of two matrices

For

$$A=\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ \vdots & \vdots & & \vdots \\ a_{i1} & a_{i2} & \dots & a_{ip} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{np} \end{pmatrix} \text{ and } B=\begin{pmatrix} b_{11} & \dots & b_{1j} & \dots & b_{1m} \\ b_{21} & \dots & b_{2j} & \dots & b_{2m} \\ \vdots & & \vdots & & \vdots \\ b_{p1} & \dots & b_{pj} & \dots & b_{pm} \end{pmatrix},$$

$$\begin{aligned} c_{ij} &= a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{ip}b_{pj} \\ &= \sum_{k=1}^{p} a_{ik}b_{kj}, \text{ for } i=1, \cdots, n \text{ and } j=1, \cdots, m \end{aligned}$$

: The element of the product $C = A_{n\times p}B_{p\times m} = \{c_{ij}\}$

<u>The required condition</u> :

The ith row of A(and hence all rows) must have the same number of elements as does the jth column of B(and hence all columns).

This means that there must be exactly as many columns in A as there are rows in B.

1.4 Matrix Representation of Multivariate Data

■ Data Matrix

The $n \times p$ data matrix :

$$X = \begin{bmatrix} x_{11} \cdots x_{1j} \cdots x_{1p} \\ \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \\ x_{i1} \cdots x_{ij} \cdots x_{ip} \\ \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \\ x_{n1} \cdots x_{nj} \cdots x_{np} \end{bmatrix} = (x_{ij}) = [\underline{x}_1, \ldots, \underline{x}_j, \ldots, \underline{x}_p] \text{: Column Representation}$$

The mean of jth variable: $$\bar{x}_j = \frac{(x_{1j} + \cdots + x_{nj})}{n} = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = \frac{1_n^t \underline{x}_j}{n}$$

The centred data matrix: $$Y = (x_{ij} - \overline{x_j}), i = 1, \ldots, n; \; j = 1, \ldots, p.$$

$$Y = X - \frac{1}{n} JX = HX$$

where $H = I - \frac{1}{n} J$: centering matrix, J : unit matrix

■ Summary Statistics : mean vector, covariance matrix, correlation matrix

(1) mean vector : $\bar{\boldsymbol{x}} = [\bar{\boldsymbol{x}}_1, .., \bar{\boldsymbol{x}}_j, ..., \bar{\boldsymbol{x}}_p]^t = \dfrac{X^t 1_n}{n}$

(2) $p \times p$ covariance matrix:

$$S = (s_{kj}) = \begin{bmatrix} s_{11} & s_{12} & \cdots s_{1j} \cdots & s_{1p} \\ & s_{22} & \cdots s_{2j} \cdots & s_{2p} \\ & & \cdot & \cdot \\ & & \cdot & \cdot \\ & & s_{jj} \cdots & s_{jp} \\ & (\text{대칭}) & & \\ & & & s_{pp} \end{bmatrix}$$

- variance of j th variable: $s_{jj} = s_j^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$

=[Sum of squared x_j deviation]/$(n-1)$

- standard deviation of j th variable: $s_j = \sqrt{s_{jj}}$

- covariance of k th variable and j th variable:

$$s_{kj} = s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)(x_{ij} - \bar{x}_j)$$

=[Sum of cross product of x_j deviation and x_k deviation]/$(n-1)$

(3) $S = \frac{1}{n-1} Y^t Y = \frac{1}{n-1} X^t HX.$

(4) standardized data matrix:

$$Z = (z_{kj}) = ((x_{kj} - \overline{x_j})/s_j)$$

(5) $p \times p$ correlation matrix:

$$R = (r_{kj}) = \begin{bmatrix} 1 & r_{12} & \cdots r_{1j} \cdots & r_{1p} \\ & 1 & \cdots r_{2j} \cdots & r_{2p} \\ & & . & . \\ & & . & . \\ & & 1 & \cdots r_{jp} \\ & \text{(대칭)} & & . \\ & & & 1 \end{bmatrix}$$

correlation coefficient of k variable and j th variable:

$$r_{kj} = \frac{s_{kj}}{\sqrt{s_{kk}}\sqrt{s_{jj}}} = \frac{\sum_{i=1}^{n}(x_{ik} - \overline{x_k})(x_{ij} - \overline{x_j})}{\sqrt{\sum_{i=1}^{n}(x_{ik} - \overline{x_k})^2}\sqrt{\sum_{i=1}^{n}(x_{ij} - \overline{x_j})^2}} \qquad \text{if } k \neq j$$

$$= 1 \qquad \text{if } k = j$$

(6) $R = D_s^{-1/2} S D_s^{-1/2}$ where $D_s^{-1/2} = (D_s^{1/2})^{-1}$, $D_s^{1/2} = diag(\sqrt{s_{11}}, ..., \sqrt{s_{pp}})$: standard deviation matrix (squared root matrix)

1.5 Creating Matrices X, Y, S, R in R and SAS/IML

<table>
<tr><th>R</th><th>SAS/IML</th></tr>
<tr><td><pre>
 X <- matrix(c(77, 67, 81,
               63, 80, 81,
                50, 50, 50), byrow=T, nrow=3)
</pre></td><td><pre>
proc iml;
X={77 67 81,
    63 80 81,
    50 50 50};
</pre></td></tr>
<tr><td><pre>
> X=matrix(scan("c:/work/home/matrix_eng/subject.dat"), ncol=3)
Read 9 items
> X
     [,1] [,2] [,3]
[1,]   77   63   50
[2,]   67   80   50
[3,]   81   81   50
</pre></td><td></td></tr>
</table>

<table>
<tr><th>R</th><th>SAS/IML</th></tr>
<tr><td><pre>
n <- nrow(X)
p <- ncol(X)
Xbar <- matrix(0,1,p)
for (i in 1:p) {
        Xbar[,i] <- mean(X[,i])
}
Y <- matrix(0,n,p)
J <- matrix(1,n,1)
# Centred Matrix
Y <- X - J%*% Xbar
# Covariance Matrix
S <- t(Y)%*%Y/(n-1)
D<- sqrt(diag(S))
# Correlation Matrix
R<- diag(1/D)%*%S%*%diag(1/D)
</pre></td><td><pre>
n=nrow(X) ;
u=J(n,1,1) ;
mean=u`*X/n ;
X_bar=repeat(mean,n,1) ;
Y=X-X_bar ;   /* Centred Matrix */
S=Y`*Y /(n-1);   /* Covariance Matrix */
D=sqrt(diag(S)) ;
R=inv(D)*S*inv(D) ;   /* Correlation Matrix */
print Y D S R ;
</pre></td></tr>
</table>