Ch 1: Aspects of Multivar Analysis

I. Intro

multivariate analysis
the analysis of the relationships bet many vars
analysis of data gathered by experimentation or observation will usually suggest mod explanation of phenomenon
: vars often (+)ed or (-)ed
complexities of most phenomena require an investigator to collect obs on many diff vars

Book: emphasis
analysis of measurements obtained about actively controlling/manipulating any of vars on which measurements are made
Ch 6, 7
few experimental plans (designs)...
generating data that describe active manipulation of important vars

most important pt of scientific investigation

Note: experimental design is frequently impossible to control generation of appropriate data in certain disciplines

Note: many multivar methods are based upon underlying prob model known as multivar norm distr

multivar techniques
a. to study interdependent relationships
b. classify according to # of pops & of sets of vars being studied

In book, choice of methods & types of analyses employed are largely determined by objectives of investigation
- inference abt
  - treatment means
  - cov structure
- techniques for sorting/grouping

Objs of scientific investigations

1. Data reduction / Structural simplification
   Phenomenon being studied is represented as simply as possible w/out sacrificing valuable info
   - hoped it'll make interpretation easier

2. Sorting & Grouping
   groups of "similar" objs/vars are created, based upon measured characteristics
   - rules for classifying objs into well-defined groups may be required

3. Investigation of dependence among vars
   Relationships among vars
   - are all vars mutually indep?
   - are 1+ vars dependent on others
     - if so, how

I. Intro

Objs of scientific investigations

4. Prediction
vars' relationships must be determined for purpose of predicting values of 1+ vars on basis of obs on other vars

5. Hypothesis construction & testing
Testing specific stat hypotheses, formulated in terms of parameters of multivar pops
may be done to validate assumptions or to reinforce prior convictions

II. Apps of Multivar Techniques

Data (R)tion/Simplification examples:

1. using data on several vars related to cancer patient responses to radiotherapy
simple measure of patient response to radiotherapy obs constructed

2. Track records from many nations were used to develop index of performance for both ♀ & ♂ athletes

Sorting & Grouping

1. IRS uses tax return data to sort taxpayers audited vs not audited

II. Apps of Multivar Techniques

Investigation of dependence among vars

1. Data on several vars were used to ID factors that were responsible for client success in hiring external consultants

Prediction

1. Predictors of success in college
   associations bet
   - test scores
   - several HS performance vars
   - several college performance vars

Ho Testing

$H_0: \mu = 0$
$H_1: \mu \neq 0$

1. pollution-related vars
   - Wks for very metropolitan areas were either
     - roughly constant throughout week
     - noticeable diff bet weekdays & ends

III Organization of Data

A. Arrays

$p \geq 1$ of vars/characters to record

↓

multivar data

x_{jk} = measurement of kth var on jth item

	Var I	Var II	. . .	Var k	Var p
Item I	x_{11}	x_{12}	. . .	x_{1k}	x_{1p}
Item II	x_{21}	x_{22}	. . .	x_{2k}	x_{2p}
⋮	⋮	⋮	. . .	⋮	⋮
Item j	x_{j1}	x_{j2}	. . .	x_{jk}	x_{jp}
Item n	x_{n1}	x_{n2}	. . .	x_{nk}	x_{np}

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & x_{2p} \\ \vdots & \vdots & & \vdots & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & x_{jp} \\ x_{n1} & x_{n2} & \cdots & x_{nk} & x_{np} \end{bmatrix}$$

B. Descriptive Stats

descriptive stats: summary #'s of large dataset most heavily rely on ones that measure:

location

&

linear association

B. Descriptive Stats

Let $x_{11}, x_{21}, \ldots, x_{n1}$ be n measurements on 1st var

$$\bar{x}_1 = \frac{1}{n}\sum_{j=1}^{n} x_{j1}$$ arithmetic avg

also called sample mean

if n measurements rep a subset of full set of measurements that might've been observed

sample mean

$$\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk} \quad k = 1, 2, \ldots, p$$

can be computed from n measurements on each of p vars

sample VV

$$s_1^2 = \frac{1}{n}\sum_{j=1}^{n} (x_{j1} - \bar{x}_1)^2$$ for n measurements on 1st var

$\bar{x}_1$ = sample mean of x_{j1}'s

$$s_k^2 = \frac{1}{n}\sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2 \quad k = 1, 2, \ldots, p$$

Comment: many authors define sample VV w/ divisor of (n-1) rather than n

theoretical reasons & appropriate if n is small

3. Descriptive Stats

Comment: s^2 is traditionally used to indicate sample VV

double subscripts to denote positions in array

for arrays where sample VVs lie along diagonal

$$s_{kk} = s_{kk} = \frac{1}{n}\sum_{j=1}^{n}(x_{jk} - \bar{x}_k)^2 \quad k = 1, 2, \ldots, p$$

sample std $\sqrt{s_{kk}}$

Consider n pairs of measurements on each of vars 1 & 2:

$$\begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix}, \begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix}, \ldots, \begin{bmatrix} x_{n1} \\ x_{n2} \end{bmatrix}$$

sample cov: measure of linear association bet measurements of vars 1 & 2

or avg product of devs from their respective means

(large)(large) = s_{12}+ | (large)(small) = s_{12}-

(small)(small) = s_{12}+ | no assoc = $s_{12} \approx 0$

$$s_{ik} = \frac{1}{n}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) \quad \begin{matrix} i = 1, 2, \ldots p \\ k = 1, 2, \ldots p \end{matrix}$$

note: cov ↓ → sample VV when $i = k$

moreover, $s_{ik} = s_{ki}$ $\forall i \& k$

B. Descriptive Stats

sample corr coeff / Pearson's product-moment corr coeff

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = \frac{\sum_{j=1}^{n}(x_{ji}-\bar{x}_i)(x_{jk}-\bar{x}_k)}{\sqrt{\sum_{j=1}^{n}(x_{ji}-\bar{x}_i)^2}\sqrt{\sum_{j=1}^{n}(x_{jk}-\bar{x}_k)^2}}$$

for $i = 1, 2, \ldots, p$ & $k = 1, 2, \ldots, p$; $r_{ik} = r_{ki}$ $\forall i \& k$

standardized vers of sample cov'll, where product of $\sqrt{\ }$'s of sample V provides standardizations

note: r_{ik} has same val whether n / $n-1$ is chosen as common divisor for s_{ii}, s_{kk}, & s_{ik}
r_{ik} can also be viewed as sample cov'll

sample corr r
easier to interpret b/c its magnitude is bounded

1. value of r must be $(-1, +1)$ inclusive
2. r = strength of linear association
   $r = 0 \Rightarrow$ lack of linear assoc bet components
   $r < 0 \Rightarrow$ tendency for larger value avg when other is small avg
   $r > 0 \Rightarrow$ tendency for both values to be large/small together

3. Descriptive Stats

sample corr (r)

3. value of r_{ik} remains unchanged if measurements of ith var are changed to $y_{ji} = ax_{ji} + b$ $(j = 1, 2, \ldots, n)$ & values of kth var are changed to $y_{jk} = cx_{jk} + d$ $(j = 1, 2, \ldots, n)$
provided that constants a & c have same sign

Note: quantities s_{ik} & r_{ik} don't (in gen) convey all there is to know abt association bet 2 vars
nonlinear assocs can exist that aren't revealed by the desc stats

cov & corr
provide measures of linear assocs (assoc along line)
values < informative for other kinds of assoc
can be very sensitive to "wild" obs ("outliers") & may indicate assoc when little exists

despite this, they're routinely calc & analyzed
provide cogent numerical summaries of assoc when data don't exhibit obvious nonlinear patterns of assoc & when outliers aren't present

suspect obs must be accounted for by:
correcting obvious recording mistakes
taking actions consistent w/ presumed causes
s_{ik} & r_{ik} values must be noted b4 & after

B. Descriptive Stats

Sum of $\square$s of devs from mean

$$w_{kk} = \sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2 \quad k = 1, 2, \ldots, p$$

sum of cross-product devs

$$w_{ik} = \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) \quad \begin{matrix} i = 1, 2, \ldots, p \\ k = 1, 2, \ldots, p \end{matrix}$$

C. Arrays of Basic Desc Stats

Sample means $\bar{x} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$

Sample Vs coVs
n = divisor for elements s_{ik}

$$S_n = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

Sample coors

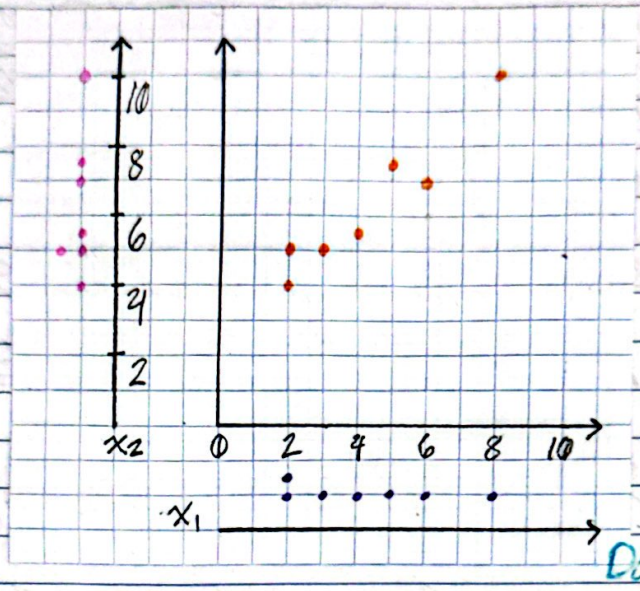
$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

symmetric: $s_{ik} = s_{ki}$ & $r_{ik} = r_{ki}$ $\forall i, k$, entries in symmetric positions abt main NW-SE diagonals in arrays S_n & R are same

1.2 Graphical Techniques

scatter diagram/plot

x_1	x_2
3	5
4	5.5
2	4
6	7
8	10
2	5
5	7.5



Dot diagram

single-var dot diagrams
calc sample means $\bar{x}_1$ & $\bar{x}_2$
sample VVs s_{11} & s_{22}

scatter diagram
coordinates to calc sample covVs

p-Dim Scatter Plot
n pts in p dims

Consider natural extension of scatter plot to p dims $(x_{j1}, x_{j2}, \ldots, x_{jp})$ on jth item rep coordinates of pt in p-dim space

can reveal group structure

D. Graphical Techniques

p Pts in n Dims

$\begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ni} \end{bmatrix}$ ith column, consisting of all n measurements on ith var., determines ith pt

III. Data Displays & Pictoral Reps

A. Linking Multi 2D Scatterplots

R studio

pairs(function, data = data file)

brushing: operation of highlighting pts corres to selected range of one of vars

↑-dim data ⇒ slices of various 3D perspective

spinning & rotating slices allows to note things like stiffness of data & outliers

B. Graphs of Growth Curves

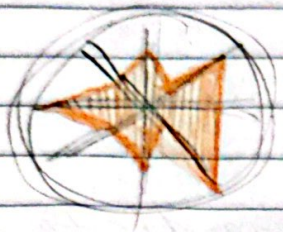
Example: height measured @ each birth

pts can be plotted → connected by lines → graph

growth curve: repeated measurements of same characteristics on same unit / subject

C. Stars

Suppose each data unit consists of:
non-(-) obs on $p \geq 2$ vars



In 2D...
circle w/ fixed (ref) radius w/ p equally spaced rays emanating from circle's center

length of ray = value of var
star = multivar obs

D. Chernoff Faces

- p-dim obs → 2D face
  - face shape
  - mouth curve
  - nose length
  - etc...

  ] determined by measures on p vars

Abt Chernoff Faces: up to 18 vars
assign var → facial ft
done by experimentor
diff choice → diff results

most useful for verifying:
initial groupings suggested by subject-matter knowledge & intuition

final groupings produced by clustering algorithms

II. Distance

multivar techniques are based on:
concept of distance
straight-line/Euclidean

$d(O,P) = \sqrt{x_1^2 + x_2^2}$
from $P \to O$
distance according to Pythagorean Theorem

pt P
(x_1, x_2)
$d(O,P)$
straight-line distance
$O = (0,0)$
origin

$P = (x_1, x_2, \ldots, x_p)$
$d(O,P) = \sqrt{x_1^2 + x_2^2 + \cdots + x_p^2}$

if pts lie a constant ⬭ed distance

$d^2(O,P) = x_1^2 + x_2^2 + \cdots + x_p^2 = c^2$ hypersphere

$P = (x_1, x_2, \ldots, x_p)$
$Q = (y_1, y_2, \ldots, y_p)$ $d(P,Q) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_p - y_p)^2}$

Euclidean distance is unsatisfactory for most stat purposes
each coordinate in it contributes (=)ly to calculate it

when coordinates rep measurements that are subject to rand fluctuations of diff magnitudes
often desirable to weigh ↑-Vbility coordinates < ↓-Vbility ones
suggests diff measure

II. Distance

Statistical distance

fundamental to multivar analysis

accounts for diff in Vbtion & presence of corr

does depend on sample Vbl & cobl

take fixed set of obs graphed as p-dim scatter plot

construct measure of distance from O to P

Standardizing coordinates

if Vbility differs ref. coordinates

(÷) each coordinate by sample st dev

$x_1^* = x_1 / \sqrt{s_{11}}$

$x_2^* = x_2 / \sqrt{s_{22}}$

$$d(O,P) = \sqrt{(x_1^*)^2 + (x_2^*)^2} = \sqrt{\left(\frac{x_1}{\sqrt{s_{11}}}\right)^2 + \left(\frac{x_2}{\sqrt{s_{22}}}\right)^2}$$

$$= \sqrt{\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}}}$$

Note: if Vbility in x_1 direction = Vbility in x_2 direction & x_1 values vary independently of x_2 values, Euclidean distance is appropriate

$\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}} = c^2$ Eq of ellipse centered @ origin whose major & minor axes coincide w/ coordinate axes

What if $P = (x_1, x_2)$ & $Q = (y_1, y_2)$

$$d(P,Q) = \sqrt{\frac{(x_1 - y_1)^2}{s_{11}} + \frac{(x_2 - y_2)^2}{s_{22}}} \qquad d(P,Q) = \sqrt{\frac{(x_1 - y_1)^2}{s_{11}} + \cdots + \frac{(x_p - y_p)^2}{s_{pp}}}$$

II. Distance

What is meaningful measure of distance when Vbility in x_1 direction is diff from Vbility in x_2 direction & vars x_1 & x_2 are corr?

rotate OG coordinate system thru angle θ

keep scatter fixed

axes are now: $\tilde{x}_1$, $\tilde{x}_2$

$$d(O,P) = \sqrt{\frac{\tilde{x}_1^2}{\tilde{s}_{11}} + \frac{\tilde{x}_2^2}{\tilde{s}_{22}}}$$

$\tilde{x}_1 \approx x_1 \cos\theta + x_2 \sin\theta$

$\tilde{x}_2 \approx -x_1 \sin\theta + x_2 \cos\theta$

$P = (\tilde{x}_1, \tilde{x}_2)$

$O = (0, 0)$

$d(O,P) = \sqrt{a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2}$

a = #s s.t. distance is non(-) & possible x_1, x_2 vals determined by angle θ

$2a_{12}x_1x_2$ necessitated by non0 corr r_{12}

statistical distance

$$d(P,Q) = \sqrt{a_{11}(x_1-y_1)^2 + 2a_{12}(x_1-y_1)(x_2-y_2) + a_{22}(x_2-y_2)^2}$$

$$a_{11}(x_1-y_1)^2 + 2a_{12}(x_1-y_1)(x_2-y_2) + a_{22}(x_2-y_2)^2 = c^2$$

II. Distance

$P = (x_1, x_2, \cdots, x_p)$
$Q = (y_1, y_2, \cdots, y_p)$
$\theta = (\theta_1, \theta_2, \cdots, \theta_p)$

$$a_{11} = \frac{\cos^2(\theta)}{\cos^2(\theta)s_{11} + 2\sin(\theta)\cos(\theta) + \sin^2(\theta)s_{22}} +$$

$$\frac{\sin^2\theta}{\cos^2(\theta)s_{22} - 2\sin(\theta)(\cos\theta)s_{12} + \sin^2(\theta)s_{11}}$$

$$a_{22} = \frac{\sin^2\theta}{(\cos^2\theta)s_{11} + 2\sin\theta(\cos\theta)s_{12} + \sin^2\theta(s_{22})} +$$

$$\frac{\cos^2\theta}{\cos^2\theta(s_{22}) - 2\sin\theta(\cos\theta)s_{12} + \sin^2\theta(s_{11})}$$

$$a_{12} = \frac{\cos\theta(\sin\theta)}{\cos^2\theta(s_{11}) + 2\sin\theta(\cos\theta)s_{12} + \sin^2\theta(s_{22})} -$$

$$\frac{\sin\theta(\cos\theta)}{\cos^2\theta(s_{22}) - 2\sin\theta(\cos\theta)s_{12} + \sin^2\theta(s_{11})}$$

$$d(0,P) = \sqrt{a_{11}x_1^2 + a_{22}x_2^2 + \cdots + a_{pp}x_p^2 + 2a_{12}x_1x_2 + 2a_{13}x_3 + \cdots +}$$

$$\sqrt{2a_{p-1,p}x_{p-1}x_p}$$

$$d(P,Q) = \sqrt{a_{11}(x_1-y_1)^2 + a_{22}(x_2-y_2)^2 + \cdots + a_{pp}(x_p-y_p)^2 +}$$

$$\sqrt{2a_{12}(x_1-y_1)(x_2-y_2) + 2a_{13}(x_1-y_1)(x_3-y_3) + \cdots +}$$

$$\sqrt{2a_{p-1,p}(x_{p-1}-y_{p-1})(x_p-y_p)}$$

distances are determined by coeffs a_{ik} $(i,k = 1, 2, \ldots, p)$

II. Distance

Coefficient array

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{12} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1p} & a_{2p} & \cdots & a_{pp} \end{bmatrix}$$

a_{ik}'s w/ $i \neq k$ are displayed twice
$\times 2$ in distance formulas

Note: Cannot be arbitrary #s
must be s.t. computed distance is non(-)
for every pair of pts
entries specify distance $f(x)$s

any distance measure $d(P, Q)$ bet P & Q is valid, provided it satisfies:

$d(P, Q) = d(Q, P)$
$d(P, Q) > 0$ if $P \neq Q$
$d(P, Q) = 0$ if $P = Q$
$d(P, Q) \leq d(P, R) + d(R, Q)$ Δ inequality

R is any other intermed pt