MULTIVARIATE ANALYSIS

1. The following table shows for each of 18 cinchona plants the yield of dry bark (in oz.), the height (in inches) and the girth (in inches) at a height of $6''$ from the ground. Fit a multiple linear regression model to the data. Also calculate the multiple correlation coefficient and the partial correlation coefficients.

Plant no.	Yield of dry bark (oz.)	Height (in.)	Girth at a height of $6''$
1	19	8	4
2	51	15	5
3	30	11	3
4	42	21	3
5	25	7	2
6	18	5	1
7	44	10	4
8	56	13	6
9	38	12	3
10	32	13	4
11	25	5	2
12	10	6	3
13	20	4	4
14	27	8	4
15	13	7	3
16	49	12	5
17	27	6	3
18	55	16	7

2. The following constraints are obtained from measurements on length in mm. (x_1), volume in c.c. (x_2) and weight in gm. (x_3) of 300 eggs:

$$\begin{array}{lll} \bar{x}_1 = 55.95 & s_1 = 2.26 & r_{12} = 0.578 \\ \bar{x}_2 = 51.48 & s_2 = 4.39 & r_{13} = 0.581 \\ \bar{x}_3 = 56.03 & s_3 = 4.41 & r_{23} = 0.974 \end{array}$$

(a) Obtain the linear regression equation of egg-weight on egg-length and egg-volume. Hence estimate the weight of an egg whose length is 58.0 mm. and volume is 52.5 c.c.

(b) Give a measure of the usefulness of the above regression equation as a predicting formula.

(c) Compute the partial correlation coefficient of weight and volume, eliminating the effect of length.

3. Three important measurements from which the crenial capacity (C) of humans may be predicted are the glabella-occipital length (L), the maximum parietal breadth (B) and the basis-bregmatic height (H). Since the magnitude to be estimated is a volume, it is thought proper to use a regression formula of the type,

$$C = \alpha L^{b_2} B^{b_3} H^{b_4}.$$

which takes the linear form

$$x_1 = a + b_2 x_2 + b_3 x_3 + b_4 x_4.$$

Using the measurements on 86 male skulls, one finds the mean vector

$$(3.1685,\ 2.2752,\ 2.1523,\ 2.1128)$$

and the matrix of corrected sums of squares and sums of products

$$\begin{pmatrix} 0.12692 & 0.03030 & 0.04410 & 0.030629 \\ \ldots\ldots & 0.01875 & 0.00848 & 0.00684 \\ \ldots\ldots & \ldots\ldots & 0.02904 & 0.00878 \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & 0.02886 \end{pmatrix}$$

for x_1, x_2, x_3 and x_4.

Get the multiple regression equation of x_1 on x_2, x_3 and x_4. Obtain the multiple correlation coefficient $r_{1.234}$ and comment. Decide on the basis of the partial correlation coefficients, if any one of the independent variables may be omitted.

4. Suppose $\underset{\sim}{\boldsymbol{X}} = (X_1, X_2, X_3, X_4)'$ has a multivariate normal distribution with mean vector $\underset{\sim}{\boldsymbol{\mu}} = (2, 4, -1, 3)'$ and dispersion matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} 4 & -1 & \frac{1}{2} & -\frac{1}{2} \\ & 3 & 1 & -1 \\ & & 6 & 1 \\ & & & 4 \end{pmatrix}.$$

Find

(i) the conditional distribution of $X_1 | X_2, X_3, X_4$

(ii) a linear combination of X_2, X_3, X_4 having maximum correlation with X_1. Also determine the correlation.

(iii) A 2×2 matrix C such that the random vectors $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ and $\begin{pmatrix} X_3 \\ X_4 \end{pmatrix} - C \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ are independently distributed.

(iv) the conditional correlation between X_1 and X_2 given X_3 and X_4.

(v) $\rho_{12.34}$. Is it same with the correlation as found in (iv) ?

5. Suppose $\underset{\sim}{\boldsymbol{X}} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim N_3(\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$ where $\underset{\sim}{\boldsymbol{\mu}} = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 3 & 0 \\ 1 & 0 & 1 \end{pmatrix}$.

Find the joint distribution of $Y_1 = X_1 + X_2 + X_3$ and $Y_2 = X_1 - X_2$.

6. Consider the density of a multivariate normal distribution :

$$\frac{\sqrt{|A|}}{(2\pi)^{p/2}} \cdot exp\left\{ -\frac{1}{2}(\underset{\sim}{x} - \underset{\sim}{b})' A (\underset{\sim}{x} - \underset{\sim}{b}) \right\}$$

Let $\underset{\sim}{b} = \underset{\sim}{0}$ and $A = \begin{pmatrix} 7 & 3 & 2 \\ 3 & 4 & 1 \\ 2 & 1 & 2 \end{pmatrix}$.

Find the dispersion matrix.

7. The electric power consumed each month by a chemical plant (y) is thought to be related to the average ambient temperature (x_1), the number of days in the month (x_2), the average product purity (x_3) and the tons of product produced (x_4). The observed correlation matrix based on a sample of size 12 is

$$R = \begin{pmatrix} 1 & 0.8025 & 0.8270 & 0.0929 & -0.1327 \\ & 1 & 0.6605 & -0.2876 & -0.0236 \\ & & 1 & 0.1127 & -0.0253 \\ & & & 1 & 0.0789 \\ & & & & 1 \end{pmatrix}$$

(a) Test whether the partial correlation between y and x_1, eliminating the effects of x_2, x_3, x_4 is significant.

(b) Test the hypothesis that the electric power consumed is independent of the average product purity and the tons of product produced.