M.Sc. Course
in
Applied Mathematics with Oceanology
and
Computer Programming

Vidyasagar University

Semestar-IV

Paper-MTM 403/II **Paper Name: Probability and Statistics**

Unit 3 Unit Name: **Multiple Correlation and Regression**

by

Prof. Madhumangal Pal

Department of Applied Mathematics, Vidyasagar University

Midnapore-722102

email: mmpalvu@gmail.com

Unit Structure:

3.1 Introduction

A regression model that involves more than one regressor variable is called a multiple regression model. Fitting of multiple regression equation and analysis of it are discussed in this unit. Multiple regression gives a relationship among the multiple variables and multiple and partial correlation coefficients give the measure of relationship in different situations.

Objectives:

Gone through this unit the students will learn the following:

- Multiple regression
- Multiple correlation
- Partial correlation
- Regression coefficients
- Linear estimation
- Gauss-Markov linear model.

3.2 Multiple Regression

In bivariate regression there is a linear relation between two variables one is taken as dependent variable and another is taken as independent variable. In multiple regression, the linear relation may exists among more than two variables. Here we consider p variables $x_1, x_2, \ldots, x_p$ which are connected by a linear relation.

Now our object is to build up a relationship between the 'dependent variable'(called regressand), x_1 and the 'independent variables'(called regressors), $x_1, x_2, \ldots, x_p$, with the idea of using this relationship for predicting the value of the regressand from a knowledge of the values of the regressors. Thus, in estimating the rainfall at a place in a year, it is appropriate to consider the effects of the latitude, the longitude and the altitude of the place on rainfall. Similarly, in estimating the yield of a crop in a year, it is proper to take into account the effects of, say, rainfall average temperature and average humidity, during the period between the sowing and the harvesting of the crop.

Let us assume that the relationship between x_1 and $x_2, x_3, \ldots, x_p$ is, at least in an appropriate sense, given by an equation of the form

$$X_1 = a + b_2x_2 + b_3x_3 + \ldots + b_px_p \tag{3.1}$$

Our data here will consist of p values, corresponding to the p variables, for each individuals. The values of the variables for the α^{th} individual may be denoted by $(x_{1\alpha}, x_{2\alpha}, \ldots, x_{p\alpha}),\ \alpha = 1, 2, \ldots, n$.

We apply the least square method to determine the constants $a, b_2, b_3, \ldots, b_p$.

Let X_1 be the predicted value of x_1 obtained from the equation (3.1). The difference $x_{1\alpha} - X_{1\alpha}$ is the error of estimate corresponding to the α^{th} individual. Thus the sum of square of all errors is $\sum\limits_{\alpha}(x_{1\alpha} - X_{1\alpha})^2$ and let it be

$$E_1 = \sum_{\alpha}(x_{1\alpha} - a - b_2x_{2\alpha} - \ldots - b_px_{p\alpha})^2 \tag{3.2}$$

The values of the constants $a, b_2, b_3, \ldots, b_p$ are to be determined such that E_1 is minimum.
The normal equations are

$$\frac{\partial E_1}{\partial a} = 0, \frac{\partial E_1}{\partial b_2} = 0, \frac{\partial E_1}{\partial b_3} = 0, \ldots, \frac{\partial E_1}{\partial b_p} = 0.$$

Differentiating (3.2) partially w.r.t a we get

$$\begin{aligned} \frac{\partial E_1}{\partial a} = & -2\sum_{\alpha}(x_{1\alpha} - a - b_2x_{2\alpha} - \ldots - b_px_{p\alpha}) = 0 \\ \text{or,} \quad & \sum_{\alpha}(x_{1\alpha} - a - b_2x_{2\alpha} - \ldots - b_px_{p\alpha}) = 0 \\ \text{or,} \quad & \sum_{\alpha}x_{1\alpha} = na + b_2\sum_{\alpha}x_{2\alpha} + \ldots + b_p\sum_{\alpha}x_{p\alpha} \end{aligned} \tag{3.3}$$

Differentiating (3.2) partially w.r.t b_2 we get

$$\begin{aligned} & \frac{\partial E_1}{\partial b_2} = -2\sum_{\alpha}x_{2\alpha}(x_{1\alpha} - a - b_2x_{2\alpha} - \ldots - b_px_{p\alpha}) = 0 \\ \text{or,} \quad & \sum_{\alpha}x_{2\alpha}x_{1\alpha} = a\sum_{\alpha}x_{2\alpha} + b_2\sum_{\alpha}x_{2\alpha}^2 + \ldots + b_p\sum_{\alpha}x_{2\alpha}x_{p\alpha}. \end{aligned} \tag{3.4}$$

Similarly,

$$\sum_\alpha x_{3\alpha}x_{1\alpha} = a\sum_\alpha x_{3\alpha} + b_2\sum_\alpha x_{3\alpha}x_{2\alpha} + \ldots + b_p\sum_\alpha x_{3\alpha}x_{p\alpha}$$

and so on.

Thus the set of normal equations are

$$\left.\begin{aligned}
\sum_\alpha x_{1\alpha} &= na + b_2\sum_\alpha x_{2\alpha} + b_3\sum_\alpha x_{3\alpha} + \ldots + b_p\sum_\alpha x_{p\alpha}\\
\sum_\alpha x_{2\alpha}x_{1\alpha} &= a\sum_\alpha x_{2\alpha} + b_2\sum_\alpha x_{2\alpha}^2 + b_3\sum_\alpha x_{2\alpha}x_{3\alpha} + \ldots + b_p\sum_\alpha x_{2\alpha}x_{p\alpha}\\
\sum_\alpha x_{3\alpha}x_{1\alpha} &= a\sum_\alpha x_{3\alpha} + b_2\sum_\alpha x_{3\alpha}x_{2\alpha} + b_3\sum_\alpha x_{3\alpha}^2 + \ldots + b_p\sum_\alpha x_{3\alpha}x_{p\alpha}\\
&\ldots \qquad\qquad \ldots\\
\sum_\alpha x_{p\alpha}x_{1\alpha} &= a\sum_\alpha x_{p\alpha} + b_2\sum_\alpha x_{p\alpha}x_{2\alpha} + b_3\sum_\alpha x_{p\alpha}x_{3\alpha} + \ldots + b_p\sum_\alpha x_{p\alpha}^2
\end{aligned}\right\} \tag{3.5}$$

Dividing (3.3) by n and denoting $\frac{1}{n}\sum x_{1\alpha} = \bar{x}_1,\ \frac{1}{n}\sum x_{2\alpha} = \bar{x}_2$ and so on, we get

$$\bar{x}_1 = a + b_2\bar{x}_2 + b_3\bar{x}_3 + \ldots + b_p\bar{x}_p, \tag{3.6}$$

which shows incidently that the mean point $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p)$ necessarily satisfies the prediction equation.

Multiplying (3.6) by $n\bar{x}_2$ and subtracting from the second equation of (3.5) we get

$$\begin{aligned}
\sum_\alpha x_{2\alpha}x_{1\alpha} - \bar{x}_1 n\bar{x}_2 \ &= b_2\sum_\alpha (x_{2\alpha}^2 - n\bar{x}_2^2) + b_3\sum_\alpha (x_{2\alpha}x_{3\alpha} - n\bar{x}_2\bar{x}_3)\\
&\quad + \ldots + b_p\sum_\alpha ((x_{2\alpha}x_{p\alpha} - n\bar{x}_2\bar{x}_p))\\
\text{or, } S_{21} = \ & b_2S_{22} + b_3S_{23} + \ldots + b_pS_{2p}, \qquad (3.7)\\
\text{where } S_{ij} = \ & \sum \alpha x_{i\alpha}x_{j\alpha} - n\bar{x}_i\bar{x}_j = \sum_\alpha (x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j) \qquad (3.8)
\end{aligned}$$

Similarly, multiplying (3.6) by $n\bar{x}_3, n\bar{x}_4, \ldots, n\bar{x}_p$ and subtracting from the third, fourth,..., p^{th} equation, respectively, of (3.5), we have $(p-2)$ equations determining the $b's$. Thus

$$\left.\begin{aligned}
S_{21} &= b_2S_{22} + b_3S_{23} + \ldots + b_pS_{2p}\\
S_{31} &= b_2S_{32} + b_3S_{33} + \ldots + b_pS_{3p}\\
&\ldots \quad \ldots \quad \ldots\\
&\ldots \quad \ldots \quad \ldots\\
S_{p1} &= b_2S_{p2} + b_3S_{p3} + \ldots + b_pS_{pp}
\end{aligned}\right\} \tag{3.9}$$

We denote

$$\frac{1}{n} \times S_{ij} = \frac{1}{n}\sum_\alpha (x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j) \text{ by } s_{ij}. \tag{3.10}$$

Then

$$s_{ij} = \begin{cases} Cov(x_i, x_j) & \text{if } i \neq j \\ Var(x_i) & \text{if } i = j \end{cases} \tag{3.11}$$

Dividing all equations of (3.12) by n and using (3.10) we obtain

$$\left.\begin{aligned} s_{21} &= b_2 s_{22} + b_3 s_{23} + \ldots + b_p s_{2p} \\ s_{31} &= b_2 s_{32} + b_3 s_{33} + \ldots + b_p s_{3p} \\ &\ldots \quad \ldots \quad \ldots \\ &\ldots \quad \ldots \quad \ldots \\ s_{p1} &= b_2 s_{p2} + b_3 s_{p3} + \ldots + b_p s_{pp} \end{aligned}\right\} \tag{3.12}$$

This system of equations can be written as

$$\begin{pmatrix} s_{21} \\ s_{22} \\ \vdots \\ s_{p1} \end{pmatrix} = \begin{pmatrix} s_{22} & s_{23} & \ldots & s_{2p} \\ s_{32} & s_{33} & \ldots & s_{3p} \\ \ldots & \ldots & \ldots & \ldots \\ s_{p2} & s_{p3} & \ldots & s_{pp} \end{pmatrix} \begin{pmatrix} b_2 \\ b_3 \\ \vdots \\ b_p \end{pmatrix} \tag{3.13}$$

We denote the matrix $\begin{pmatrix} s_{22} & s_{23} & \ldots & s_{2p} \\ s_{32} & s_{33} & \ldots & s_{3p} \\ \ldots & \ldots & \ldots & \ldots \\ s_{p2} & s_{p3} & \ldots & s_{pp} \end{pmatrix}$ by S. This matrix is called the variance-covariance or dispersion matrix of $x_1, x_2, \ldots x_p$.

If the matrix S is non-singular then the values of $b_2, b_3, \ldots, b_p$ can be determined from the equation (3.13) by Cramer's rule. Therefore,

$$b_j = \frac{\begin{vmatrix} s_{22} & s_{23} & \ldots & s_{2(j-1)} & s_{21} & s_{2(j+1)} & \ldots & s_{2p} \\ s_{32} & s_{33} & \ldots & s_{3(j-1)} & s_{31} & s_{3(j+1)} & \ldots & s_{3p} \\ \ldots & & \ldots & & \ldots & & \ldots & \\ s_{p2} & s_{p3} & \ldots & s_{p(j-1)} & s_{p1} & s_{p(j+1)} & \ldots & s_{pp} \end{vmatrix}}{\begin{vmatrix} s_{22} & s_{23} & \ldots & s_{2p} \\ s_{32} & s_{33} & \ldots & s_{3p} \\ \ldots & & \ldots & \\ s_{p2} & s_{p3} & \ldots & s_{pp} \end{vmatrix}}, \; j = 2, 3, \ldots, p. \tag{3.14}$$

The correlation coefficient r_{ij} between x_i and x_j is $r_{ij} = \frac{s_{ij}}{s_i s_j}$, where s_i, s_j are the standard deviation of x_i

and x_j. Then

$$b_j = \frac{\begin{vmatrix} r_{22}s_2s_2 & r_{23}s_2s_3 & \dots & r_{2(j-1)}s_2s_{j-1} & r_{21}s_2s_1 & r_{2(j+1)}s_2s_{j+1} & \dots & r_{2p}s_2s_p \\ r_{32}s_3s_2 & r_{33}s_3s_3 & \dots & r_{3(j-1)}s_3s_{j-1} & r_{31}s_3s_1 & r_{3(j+1)}s_3s_{j+1} & \dots & r_{3p}s_3s_p \\ \dots & & \dots & & \dots & & \dots & \\ r_{p2}s_ps_2 & r_{p3}s_ps_3 & \dots & r_{p(j-1)}s_ps_{j-1} & r_{p1}s_ps_1 & r_{p(j+1)}s_ps_{j+1} & \dots & r_{pp}s_ps_p \end{vmatrix}}{\begin{vmatrix} r_{22}s_2s_2 & r_{23}s_2s_3 & \dots & r_{2p}s_2s_p \\ r_{32}s_3s_2 & r_{33}s_3s_3 & \dots & r_{3p}s_3s_p \\ \dots & & \dots & \\ r_{p2}s_ps_2 & r_{p3}s_ps_3 & \dots & r_{pp}s_ps_p \end{vmatrix}}$$

$$= \frac{s_1}{s_2} \frac{\begin{vmatrix} r_{22} & r_{23} & \dots & r_{2(j-1)} & r_{21} & r_{2(j+1)} & \dots & r_{2p} \\ r_{32} & r_{33} & \dots & r_{3(j-1)} & r_{31} & r_{3(j+1)} & \dots & r_{3p} \\ \dots & & \dots & & \dots & & \dots & \\ r_{p2} & r_{p3} & \dots & r_{p(j-1)} & r_{p1} & r_{p(j+1)} & \dots & r_{pp} \end{vmatrix}}{\begin{vmatrix} r_{22} & r_{23} & \dots & r_{2p} \\ r_{32} & r_{33} & \dots & r_{3p} \\ \dots & & \dots & \\ r_{p2} & r_{p3} & \dots & r_{pp} \end{vmatrix}}$$

$$= (-1)^{j-2}\frac{s_1}{s_j} \frac{\begin{vmatrix} r_{21} & r_{22} & r_{23} & \dots & r_{2(j-1)} & r_{2(j+1)} & \dots & r_{2p} \\ r_{31} & r_{32} & r_{33} & \dots & r_{3(j-1)} & r_{3(j+1)} & \dots & r_{3p} \\ & \dots & & \dots & & \dots & & \dots \\ r_{p1} & r_{p2} & r_{p3} & \dots & r_{p(j-1)} & r_{p(j+1)} & \dots & r_{pp} \end{vmatrix}}{\begin{vmatrix} r_{22} & r_{23} & \dots & r_{2p} \\ r_{32} & r_{33} & \dots & r_{3p} \\ \dots & & \dots & \\ r_{p2} & r_{p3} & \dots & r_{pp} \end{vmatrix}} \tag{3.15}$$

We write R for the matrix $\begin{pmatrix} r_{11} & r_{12} & \dots & r_{1p} \\ r_{21} & r_{22} & \dots & r_{2p} \\ \dots & & \dots & \\ r_{p1} & r_{p2} & \dots & r_{pp} \end{pmatrix}$ which is the correlation matrix of $x_1, x_2, x_3 \dots, x_p$; $|R|$ for the determinant of R and R_{ij} for the cofactor of r_{ij} in R. It may be noted that R is symmetric i.e., $r_{ij} = r_{ji}$ moreover $r_{ii} = 1$.

The numerator of (3.15) is the minor of r_{1j} in R and hence it is $(-1)^{1+j}\times$ cofactor of r_{1j}. Also, the determinant in the denominator is the minor (and also the cofactor) of r_{11} in R.

Hence

$$b_j = (-1)^{2j-1} \times \frac{s_1}{s_j} \times \frac{R_{1j}}{R_{11}} = -\frac{s_1}{s_j} \times \frac{R_{1j}}{R_{11}},\ j = 2, 3, \ldots, p \tag{3.16}$$

From (3.6),

$$a = \bar{x}_1 + \sum_{j=2}^{p} \frac{R_{1j}}{R_{11}} \frac{s_1}{s_j} \bar{x}_j. \tag{3.17}$$

Thus the prediction equation called the multiple regression equation of x_1 on $x_2, x_3, \ldots, x_p$ becomes

$$X_1 = \bar{x}_1 - \frac{R_{12}}{R_{11}} \frac{s_1}{s_2}(x_2 - \bar{x}_1) - \frac{R_{13}}{R_{11}} \frac{s_1}{s_3}(x_3 - \bar{x}_3) - \ldots - \frac{R_{1p}}{R_{11}} \frac{s_1}{s_p}(x_p - \bar{x}_p) \tag{3.18}$$

The coefficient $b_j = -\frac{R_{1j}}{R_{11}} \frac{s_1}{s_j}$ is called the partial regression coefficient of x_1 on x_j for fixed $x_2, x_3, \ldots, x_{j-1}$, $x_{j+1}, \ldots, x_p$ and is often written in the form

$$b_{1j23\ldots(j-1)(j+1)\ldots p} \tag{3.19}$$

It gives the amount by which the predicted value X_1 increses when x_j is creased by a unit amount, the other independent variables being kept fixed.

3.2.1 Multiple regression for three variables

The multiple regression equation for the independent variables x_2 and x_3 on x_1

$$\begin{aligned} X_1 &= \bar{x}_1 - \frac{R_{12}}{R_{11}} \frac{s_1}{s_2}(x_2 - \bar{x}_2) - \frac{R_{13}}{R_{11}} \frac{s_1}{s_3}(x_3 - \bar{x}_3) \\ &= \bar{x}_1 + b_{12.3}(x_2 - \bar{x}_2) + b_{13.2}(x_3 - \bar{x}_3), \end{aligned} \tag{3.20}$$

$$\text{or, } X_1 = a + b_{12.3} x_2 + b_{13.2} x_3, \tag{3.21}$$

$$\text{where} \quad b_{12.3} = -\frac{R_{12}}{R_{11}} \frac{s_1}{s_2}, b_{13.2} = -\frac{R_{13}}{R_{11}} \frac{s_1}{s_3}$$

Now,

$$\begin{aligned} R_{11} &= \begin{vmatrix} r_{22} & r_{23} \\ r_{32} & r_{33} \end{vmatrix} = r_{22} r_{33} - r_{23} r_{32} \\ &= 1 - r_{23}^2 \text{ as } r_{ii} = 1 \text{ and } r_{ij} = r_{ji}. \\ R_{12} &= -\begin{vmatrix} r_{21} & r_{23} \\ r_{31} & r_{33} \end{vmatrix} = -r_{21} r_{33} + r_{23} r_{31} = r_{23} r_{13} - r_{21} \\ \text{and } R_{13} &= \begin{vmatrix} r_{21} & r_{22} \\ r_{31} & r_{32} \end{vmatrix} = r_{21} r_{32} - r_{22} r_{31} = r_{21} r_{32} - r_{31} \end{aligned}$$

Thus $b_{12.3} = \dfrac{r_{12} - r_{23} r_{13}}{1 - r_{23}^2} \dfrac{s_1}{s_2}$ and $b_{13.2} = \dfrac{r_{13} - r_{12} r_{23}}{1 - r_{23}^2} \dfrac{s_1}{s_3}$.

Example 3.2.1 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.

Plant No.	Yield of dry bark(oz)	Height(in.)	Girth at a height of 6"(in)
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

Plant No.	Yield of dry bark(oz)	Height(in.)	Girth at a height of 6"(in)
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

Solution. We denote these variables by x_1, x_2 and x_3 respectively. Here we find the dependence of x_1 on x_2 and x_3, i.e., the multiple regression equation of x_1 on x_2 and x_3.

x_1	x_2	x_3	x_1^2	x_2^2	x_3^2	x_1x_2	x_1x_3	x_2x_3
19	8	4	361	64	16	152	76	32
51	15	5	2601	225	25	765	255	75
30	11	3	900	121	9	330	90	33
42	21	3	1764	441	9	882	126	63
25	7	2	625	49	4	175	50	14
18	5	1	324	25	1	50	18	5
44	10	4	1936	100	16	440	176	40
56	13	6	3136	169	36	728	336	78
38	12	3	1444	144	9	456	114	36
32	13	4	1024	169	16	416	128	52
25	5	2	625	25	4	125	50	10
10	6	3	100	36	9	60	30	18
20	4	4	400	16	16	80	80	16
27	8	4	729	64	16	216	108	32
13	7	3	169	49	9	91	39	21
49	12	5	2401	144	25	588	245	60
27	6	3	729	36	9	162	81	18
55	16	7	3025	256	49	880	385	112
Total 581	179	66	22293	2133	278	6636	2387	715

Now

$$\bar{x}_1 = \frac{\sum x_{1\alpha}}{n} = \frac{518}{18} = 32.28 \text{ oz}$$
$$\bar{x}_2 = \frac{\sum x_{2\alpha}}{n} = \frac{179}{18} = 9.94 \text{ in}$$
$$\bar{x}_3 = \frac{\sum x_{3\alpha}}{n} = \frac{66}{18} = 3.67 \text{ in}$$
$$s_1 = \frac{1}{n}\sqrt{n\sum x_{1\alpha}^2 - \Big(\sum x_{1\alpha}\Big)^2} = \frac{\sqrt{63713}}{18} = 14.02 \text{ oz}$$
$$s_2 = \frac{1}{n}\sqrt{n\sum x_{2\alpha}^2 - \Big(\sum x_{2\alpha}\Big)^2} = \frac{\sqrt{6353}}{18} = 4.43 \text{ in}$$
$$s_3 = \frac{1}{n}\sqrt{n\sum x_{3\alpha}^2 - \Big(\sum x_{3\alpha}\Big)^2} = \frac{\sqrt{648}}{18} = 1.41 \text{ in}$$
$$r_{12} = \frac{n\sum x_{1\alpha}x_{2\alpha} - \big(\sum x_{1\alpha}\big)\big(\sum x_{2\alpha}\big)}{\sqrt{n\sum x_{1\alpha}^2 - \big(\sum x_{1\alpha}\big)^2}\sqrt{n\sum x_{2\alpha}^2 - \big(\sum x_{2\alpha}\big)^2}} = \frac{15449}{\sqrt{63713}\sqrt{6353}} = 0.768$$
$$r_{13} = \frac{n\sum x_{1\alpha}x_{3\alpha} - \big(\sum x_{1\alpha}\big)\big(\sum x_{3\alpha}\big)}{\sqrt{n\sum x_{1\alpha}^2 - \big(\sum x_{1\alpha}\big)^2}\sqrt{n\sum x_{3\alpha}^2 - \big(\sum x_{3\alpha}\big)^2}} = \frac{4620}{\sqrt{63713}\sqrt{648}} = 0.719$$
$$r_{23} = \frac{n\sum x_{2\alpha}x_{3\alpha} - \big(\sum x_{2\alpha}\big)\big(\sum x_{3\alpha}\big)}{\sqrt{n\sum x_{2\alpha}^2 - \big(\sum x_{2\alpha}\big)^2}\sqrt{n\sum x_{3\alpha}^2 - \big(\sum x_{3\alpha}\big)^2}} = \frac{1056}{\sqrt{6353}\sqrt{648}} = 0.520$$

If the multiple regression equation is

$$X_1 = a + b_{12.3}x_2 + b_{13.2}x_3$$
$$\text{Then} \quad b_{12.3} = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\frac{s_1}{s_2} = \frac{0.394}{0.730}\frac{14.02}{4.43} = 1.71$$
$$b_{13.2} = \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\frac{s_1}{s_3} = \frac{0.320}{0.730}\frac{14.02}{1.41} = 4.36$$
$$\text{and} \quad a = \bar{x}_1 - b_{12.3}\bar{x}_2 - b_{13.2}\bar{x}_3 = -0.72.$$

Hence, the multiple regression equation of x_1 on x_2 and x_3 is $X_1 = -0.72 + 1.71x_2 + 4.36x_3$.

3.3 Multiple Correlation

In studying the dependence of x_1 on a set of independent variables, we may want to know to what extent x_1 is influenced by the independent variables. In the case of two variables, x and y, we have seen that r_{xy} serves as a measure of the strength of the interdependence of x and y or, if y may be looked upon as dependent on x, of the extent to which x influences y. Generalizing this approach, we may take the simple correlation between x_1 and X_1, i.e., the value of x_1 given by the multiple regression equation of x_1 on $x_2, \ldots, x_p$, as a measure of the joint influence of $x_2, x_3, \ldots, x_p$ on x_1. It is called the multiple correlation

coefficient of x_1 on $x_2, x_3, \ldots, x_p$ and is denoted by $r_{1.23\ldots,p}$. Then

$$r_{1.23\ldots,p} = \frac{Cov(x_1, X_1)}{\sqrt{Var(x_1)}\sqrt{Var(X_1)}}. \tag{3.22}$$

Again, the mean of the predicted value X_1 is

$$\begin{aligned}
\bar{X}_1 &= \frac{1}{n}\sum_{\alpha} X_{1\alpha} \\
&= \bar{x}_1 - \frac{R_{12}}{R_{11}}\frac{s_1}{s_2}\frac{1}{n}\sum_{\alpha}(x_{2\alpha} - \bar{x}_2) - \frac{R_{13}}{R_{11}}\frac{s_1}{s_3}\frac{1}{n}\sum_{\alpha}(x_{3\alpha} - \bar{x}_3) \\
&\qquad - \ldots - \frac{R_{1p}}{R_{11}}\frac{s_1}{s_p}\frac{1}{n}\sum_{\alpha}(x_{p\alpha} - \bar{x}_p) \\
&= \bar{x}_1
\end{aligned} \tag{3.23}$$

$\big[$ As $\frac{1}{n}\sum(x_{2\alpha} - \bar{x}_2) = \frac{1}{n}\sum x_{2\alpha} - \bar{x}_2 = \bar{x}_2 - \bar{x}_2 = 0$ etc.$\big]$
The error e_1 is $e_1 = x_1 - X_1$. Then $\bar{e}_1 = \bar{x}_1 - \bar{X}_1 = 0$.

Now,

$$\begin{aligned}
Cov(x_1, X_1) &= \frac{1}{n}\sum_{\alpha}(x_{1\alpha} - \bar{x}_1)(x_{1\alpha} - \bar{X}_1) \\
&= \frac{1}{n}\sum_{\alpha}(e_{1\alpha} + X_{1\alpha} - \bar{X}_1)(X_{1\alpha} - \bar{X}_1) = \frac{1}{n}\sum_{\alpha} e_{1\alpha}(X_{1\alpha} - \bar{X}_1) + \frac{1}{n}\sum_{\alpha}(X_{1\alpha} - \bar{X}_1)^2 \\
&= \frac{1}{n}\sum_{\alpha} e_{1\alpha}(X_{1\alpha} - \bar{X}_1) + Var(X_1)
\end{aligned} \tag{3.24}$$

Now,

$$\begin{aligned}
&\frac{1}{n}\sum_{\alpha} e_{1\alpha}(X_{1\alpha} - \bar{X}_1) = \frac{1}{n}\sum_{\alpha} e_{1\alpha}X_{1\alpha} - \bar{X}_1\frac{1}{n}\sum_{\alpha} e_{1\alpha} \\
&= \frac{1}{n}\sum_{\alpha} e_{1\alpha}\Big\{\bar{x}_1 + b_2(x_{2\alpha} - \bar{x}_2) + b_3(x_{3\alpha} - \bar{x}_3) + \ldots + b_p(x_{p\alpha} - \bar{x}_p)\Big\} - 0 \\
&\qquad\qquad [\because \frac{1}{n}\sum e_{1\alpha} = \bar{e}_1 = 0] \\
&= 0 \text{ [Using normal equations] .}
\end{aligned}$$

Therefore,

$$Cov(x_1, X_1) = Var(X_1) \tag{3.25}$$

Now,

$$\begin{aligned}
Cov(x_1, X_1) &= \frac{1}{n}\sum_{\alpha}(x_{1\alpha} - \bar{x}_1)(X_1 - \bar{x}_1) \\
&= \frac{1}{n}\sum_{\alpha}(x_{1\alpha} - \bar{x}_1)\left\{-\frac{R_{12}}{R_{11}}\frac{s_1}{s_2}(x_{2\alpha} - \bar{x}_2) - \frac{R_{13}}{R_{11}}\frac{s_1}{s_3}(x_{3\alpha} - \bar{x}_3) - \ldots - \frac{R_{1p}}{R_{11}}\frac{s_1}{s_p}(x_{p\alpha} - \bar{x}_p)\right\} \\
&= -\frac{R_{12}}{R_{11}}\frac{s_1}{s_2}s_{12} - \frac{R_{13}}{R_{11}}\frac{s_1}{s_3}s_{13} - \ldots - \frac{R_{1p}}{R_{11}}\frac{s_1}{s_p}s_{1p} \\
&= -\frac{s_1^2}{R_{11}}\left(r_{12}R_{12} + r_{13}R_{13} + \ldots + r_{1p}R_{1p}\right) \\
&= -\frac{s_1^2}{R_{11}}\left(|R| - r_{11}R_{11}\right) = \left(1 - \frac{|R|}{R_{11}}\right)s_1^2. \qquad (3.26)
\end{aligned}$$

Therefore, $Var(X_1) = Cov(x_1, X_1) = \left(1 - \frac{|R|}{R_{11}}\right)s_1^2$.

Hence

$$r_{1.23\ldots p} = \frac{\left(1 - \frac{|R|}{R_{11}}\right)s_1^2}{\sqrt{s_1^2\left(1 - \frac{|R|}{R_{11}}\right)s_1^2}} = \left(1 - \frac{|R|}{R_{11}}\right)^{\frac{1}{2}}. \qquad (3.27)$$

The multiple correlation coefficient basically a simple correlation coefficient and so must lie between -1 and 1. But, $Cov(x_1, X_1) = Var(x_1) > 0$. Thus

$$r_{1.23\ldots p} = \frac{Cov(x_1, X_1)}{\sqrt{Var(x_1)}\sqrt{Var(X_1)}} > 0 \text{ and hence } 0 \le r_{1.23\ldots p} \le 1. \qquad (3.28)$$

3.3.1 Some results on multiple regression and multiple correlation

(i) $\bar{X}_1 = \bar{x}_1$ and $\bar{e}_1 = 0$ (3.29)

(ii) $Var(X_1) = \left(1 - \frac{|R|}{R_{11}}\right)^{\frac{1}{2}} s_1^2 = r_{1.23\ldots p}^2 s_1^2$ (3.30)

(iii) Using normal equations it can be shown that $Cov(x_i, e_1) = 0, i = 2, 3, \ldots, p$ (3.31)

(iv) $Var(x_1) = Var(X_1) + Var(e_1)$ since $x_1 = X_1 + e_1$ and $Cov(X_1, e_1) = 0$. (3.32)

Hence

$$Var(e_1) = s_1^2 - Var(X_1) = \frac{|R|}{R_{11}}s_1^2. \qquad (3.33)$$

The term $Var(e_1)$ being the standard error of estimate and we have

$$Var(e_1) = \left(1 - r_{1.23\ldots p}^2\right)s_1^2 \qquad (3.34)$$

Using (3.30) and (3.34) we may write

$$r_{1.23\ldots p}^2 = \frac{Var(X_1)}{Var(x_1)} = 1 - \frac{Var(e_1)}{Var(x_1)}. \qquad (3.35)$$

Example 3.3.1 For the data of example 12.2.1, the multiple correlation coefficient of weight of dry bark (x_1) on height (x_2) and girth at a height of $6''$ (x_3) may be computed. We have $r_{12} = 0.768,\ r_{13} = 0.719$ and $r_{23} = 0.520$.

The multiple correlation coefficient is

$$r_{1.23} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} \tag{3.36}$$
$$= \sqrt{\frac{0.5325}{0.7296}} = 0.854.$$

It indicates that x_2 and x_3 have considerable influence on x_1. It indicates that the multiple regression equation obtained in example 12.2.1 serves as an excellent formula for predicting x_1 from given value of x_2 and x_3.

3.4 Partial Correlation

Sometimes the correlation between two variables say x_1 and x_2, may be partly (or wholly) due to the influence of a group of variables , say $x_3, x_4, \ldots, x_p$ on both x_1 and x_2. In such a situation one may want to know what the correlation between x_1 and x_2 would be if the effects of $x_3, x_4, \ldots, x_p$ on each of them were eliminated. This correlation is called the partial correlation or net correlation between x_1 and x_2, eliminating the effects of $x_,x_4, \ldots, x_p$, as opposed to their simple or total correlation.

Consider the multiple regression equations of x_1 on $x_3, x_4, \ldots, x_p$ and of x_2 on $x_3, x_4, \ldots, x_p$. Then we write

$$x_1 = X_1^{'} + e_1^{'} \text{ and } x_2 = X_2^{'} + e_2^{'},$$

where $X_1^{'}$ and $X_2^{'}$ are the predicted values of x_1 and x_2, $e_1^{'}$ and $e_2^{'}$ being the errors of estimation. Since $e_1^{'}$ and $e_2^{'}$ are uncorrelated with $x_3, x_4, \ldots, x_p$ these may be looked upon as the parts of x_1 and x_2 respectively, which are unaffected by this group of variables. Hence the simple correlation coefficient between $e_1^{'}$ and $e_2^{'}$ may be used to measure the partial correlation of x_1 and x_2, eliminating the effects of $x_3, x_4, \ldots, x_p$, in so far as this can be done with the help of linear regression equations. This is known as a partial correlation coefficient and is denoted by $r_{12.34\ldots p}$.

Thus assuming $Var(e_1^{'}) > 0$ and $Var(e_2^{'}) > 0$, so that R_{11} and R_{22} are both positive, we have

$$r_{12.34\ldots p} = \frac{Cov(e_1^{'}, e_2^{'})}{\sqrt{Var(e_1^{'})Var(e_2^{'})}} \tag{3.37}$$

Now,

$$e_1^{'} = x_1 - X_1^{'} = (x_1 - \bar{x}_1) + \frac{R_{13}^{(2)}}{R_{11}^{(2)}}\frac{s_1}{s_3}(x_3 - \bar{x}_3) + \frac{R_{14}^{(2)}}{R_{11}^{(2)}}\frac{s_1}{s_4}(x_4 - \bar{x}_4) + \ldots + \frac{R_{1p}^{(2)}}{R_{11}^{(2)}}\frac{s_1}{s_p}(x_p - \bar{x}_p)$$

where $R_{14}^{(2)}$ is the cofactor of r_{ij} in $R^{(2)}$, the determinant obtained from R by deleting the second row and the second column. Now, putting

$$u_i = x_i - \bar{x}_i \tag{3.38}$$

$$\text{and} \quad e_{u_i} = e_1^{'} - \bar{e}_1^{'} = e_1^{'} \tag{3.39}$$

Therefore,

$$e_{u_1} = u_1 + \frac{R_{13}^{(2)}}{R_{11}^{(2)}}\frac{s_1}{s_3}u_3 + \frac{R_{14}^{(2)}}{R_{11}^{(2)}}\frac{s_1}{s_4}u_4 + \ldots + \frac{R_{1p}^{(2)}}{R_{11}^{(2)}}\frac{s_1}{s_p}u_p.$$

Similarly, putting

$$e_{u_2} = e_2^{'} - \bar{e}_2^{'} = e_2^{'} \tag{3.40}$$

We have

$$e_{u_2} = u_2 + \frac{R_{23}^{(2)}}{R_{22}^{(2)}}\frac{s_2}{s_3}u_3 + \frac{R_{24}^{(2)}}{R_{22}^{(2)}}\frac{s_2}{s_4}u_4 + \ldots + \frac{R_{2p}^{(2)}}{R_{22}^{(2)}}\frac{s_2}{s_p}u_p.$$

Thus, we have,

$$Var(e_1^{'}) = \frac{R^{(2)}}{R_{11}^{(2)}}s_1^2 \tag{3.41}$$

$$\text{and} \quad Var(e_2^{'}) = \frac{R^{(1)}}{R_{22}^{(1)}}s_2^2 \tag{3.42}$$

Also,

$$nCov(e_1^{'}, e_2^{'}) = \sum_\alpha u_{1\alpha}u_{2\alpha} + \frac{R_{23}^{(1)}}{R_{22}^{(1)}}\frac{s_2}{s_3}\sum_\alpha u_{1\alpha}u_{3\alpha} + \frac{R_{24}^{(1)}}{R_{22}^{(1)}}\frac{s_2}{s_4}\sum_\alpha u_{1\alpha}u_{4\alpha} + \ldots + \frac{R_{2p}^{(1)}}{R_{22}^{(1)}}\frac{s_2}{s_p}\sum_\alpha u_{1\alpha}u_{p\alpha}$$

$$\text{or,} \quad Cov(e_1^{'}, e_2^{'}) = s_1 s_2\Big(r_{12} + r_{13}\frac{R_{23}^{(1)}}{R_{22}^{(1)}} + r_{14}\frac{R_{24}^{(1)}}{R_{22}^{(1)}} + \ldots + r_{1p}\frac{R_{2p}^{(1)}}{R_{22}^{(1)}}\Big) \tag{3.43}$$

[Since $\sum_\alpha u_{i\alpha}u_{j\alpha} = nCov(x_i, x_j) = nr_{ij}s_i s_j$]

Now,

$r_{12}R_{22}^{(1)} + r_{13}R_{23}^{(1)} + r_{14}R_{24}^{(1)} + \ldots + r_{1p}R_{2p}^{(1)}$

$=$ determinant obtained from $R^{(1)}$ by replacing its first row $(r_{22}r_{23}\ldots r_{2p})$ with $(r_{12}r_{13}\ldots r_{1p})$

$$= \begin{vmatrix} r_{12} & r_{13} & \ldots & r_{1p} \\ r_{32} & r_{33} & \ldots & r_{3p} \\ \vdots & \vdots & & \vdots \\ r_{p2} & r_{p3} & \ldots & r_{pp} \end{vmatrix}$$

$=$ minor of r_{21} in R =minor of r_{12} in R

$= -R_{12}$.

Therefore, from (3.43) we have

$$Cov(e_1^{'}, e_2^{'}) = -\frac{R_{12}}{R_{12}^{(1)}} s_1 s_2. \tag{3.44}$$

Thus, in terms of the simple or total correlation coefficients r_{ij},

$$r_{12.34...p} = \frac{-\frac{R_{12}}{R_{12}^{(1)}} s_1 s_2}{\left[\frac{R^{(2)}}{R_{11}^{(2)}}\right]^{\frac{1}{2}} \left[\frac{R^{(1)}}{R_{22}^{(1)}}\right]^{\frac{1}{2}} s_1 s_2} = -\frac{R_{12}}{\sqrt{R_{11} R_{22}}} \tag{3.45}$$

since $R^{(1)} = R_{11},\ R^{(2)} = R_{22}$ and $R_{11}^{(2)} = R_{22}^{(1)}$.

3.4.1 Case of three variables

For the case of three variables x_1, x_2, x_3

$$R = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{pmatrix}$$

Then $-R_{12} = \begin{vmatrix} r_{21}\ r_{23} \\ r_{31}\ r_{33} \end{vmatrix} = r_{12} - r_{13} r_{23}$.

$R_{11} = \begin{vmatrix} r_{22}\ r_{23} \\ r_{32}\ r_{33} \end{vmatrix} = 1 - r_{23}^2$. and $R_{22} = \begin{vmatrix} r_{11}\ r_{13} \\ r_{31}\ r_{33} \end{vmatrix} = 1 - r_{13}^2$.

Thus

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}. \tag{3.46}$$

The value of partial correlation coefficient $r_{12.34...p}$ lies between -1 and 1. i.e., $-1 \leq r_{12.34...p} \leq 1$.

Example 3.4.1 Let us consider the data of example 12.2.1. The partial correlation coefficient of x_1 (yield of dry bark) and x_2 (height of plant), the effect of x_3 (girth at a height of $6^{''}$) being accounted for, is

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}} = \frac{0.394}{\sqrt{0.4830}\sqrt{0.7296}} = 0.663$$

The partial correlation coefficient of x_1 and x_3, eliminating the effect of x_2, is

$$r_{13.2} = \frac{r_{13} - r_{12} r_{23}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{23}^2}} = \frac{0.320}{\sqrt{0.4102}\sqrt{0.7296}} = 0.585$$

These values may be considered together together with the total correlation coefficients $r_{12} = 0.768$ and $r_{13} = 0.719$.

Since r_{12} is quite large, one will naturally take x_2 as an independent variable for predicting x_1. The partial correlation $r_{13.2}$, being equal to 0.585, indicates that the inclusion of x_3 as an independent variable, in addition to x_2 would be worth while as it would considerably increase the accuracy of prediction.

3.4.2 Some results

Property 3.4.1 $1 - r^2_{1.23} = (1 - r^2_{12})(1 - r^2_{13.2})$.

Proof.

$$\begin{aligned}
1 - r^2_{1.23} &= 1 - \left(\frac{r_{13} - r_{12}r_{23}}{\sqrt{1 - r^2_{12}}\sqrt{1 - r^2_{23}}} \right)^2 \\
&= 1 - \frac{(r_{13} - r_{12}r_{23})^2}{(1 - r^2_{12})(1 - r^2_{23})} \\
&= \frac{1 - r^2_{12} - r^2_{23} - r^2_{13} + 2r_{12}r_{23}r_{13}}{(1 - r^2_{12})(1 - r^2_{23})} \\
\text{or, } (1 - r^2_{12})(1 - r^2_{13.2}) &= \frac{1 - r^2_{12} - r^2_{23} - r^2_{13} + 2r_{12}r_{23}r_{13}}{(1 - r^2_{23})} \\
&= 1 - \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{23}r_{13}}{(1 - r^2_{23})} \\
&= 1 - r^2_{1.23}
\end{aligned}$$

Hence $1 - r^2_{1.23} = (1 - r^2_{12})(1 - r^2_{13.2})$.

Property 3.4.2 *The correlation coefficients* r_{12}, r_{13} *and* r_{23} *must satisfy the inequality*

$$r^2_{12} + r^2_{23} + r^2_{13} - 2r_{12}r_{23}r_{13} \leq 1.$$

Proof. We know $r_{1.23} \leq 1$ That is

$$\begin{aligned}
\frac{r^2_{12} + r^2_{13} - 2r_{12}r_{23}r_{13}}{(1 - r^2_{23})} &\leq 1 \\
\text{or,} \quad r^2_{12} + r^2_{13} - 2r_{12}r_{23}r_{13} &\leq 1 - r^2_{23} \\
\text{or,} \quad r^2_{12} + r^2_{13} + r^2_{23} - 2r_{12}r_{23}r_{13} &\leq 1.
\end{aligned}$$

3.5 Linear Estimation

Here we are concerned with point estimation under a special set-up. This will be based on linear models for the expectation and finiteness of first and second order moments of the observation in the sample. The estimation with which we shall be concerned here are, linear functions of the observations and they are known as linear estimators. Further, we shall consider only unbiased linear estimators of linear functions of parameters.

We know that the sample mean $\bar{Y}$ is an unbiased estimator for the population mean μ, and $\bar{Y} = \frac{1}{n}\sum_i Y_i$ is a linear estimator. Also, we have seen that $S^2 = \frac{1}{n-1}\sum_i (Y_i - \bar{Y})^2$ is an unbiased estimator of the population variance σ^2 but it is not a linear estimator; S^2 is quadratic estimator of σ^2. Here we consider only unbiased linear estimators of estimable linear functions of the parameters occurring in the expressions

for the expectations of the random variables. We may have more than one unbiased linear estimator e.g., $\bar{Y}$ and $\sum_i a_i Y_i$ with $\sum_i a_i = 1$, are both unbiased linear estimators of the population mean. So we have the problem of selection the one that may be taken to be the best in some suitable sense. One widely used principle to make this selection is to choose that linear estimator from amongst all unbiased linear estimators which has the smallest variance. The sampling distribution of that estimator will have the maximum concentration around the unknown true parametric function. such an estimator is known as a minimum variance unbiased linear estimator or best linear unbiased estimator (BLUE). These minimum variance unbiased linear estimators have variances and covariances which, again themselves unbiased estimation.

3.5.1 Gauss-Markov linear model

Consider a set of n independent random variables $Y_1, Y_2, \ldots, Y_n$ with a common variance σ^2, whose expectations are linear functions with known coefficients (a'_{ij}s) of p unknown parameter $\beta_1, \beta_2, \ldots, \beta_p$ (with $p < n$). Thus

$$\left.\begin{aligned} E(Y_i) = a_{i1}\beta_1 + a_{i2}\beta_2 + \ldots + a_{ip}\beta_p \\ \text{and } Var(Y_i) = \sigma^2 \text{ for } i = 1, 2, \ldots, n \\ Cov(Y_i, Y_j) = 0 \text{ for } i \neq j \end{aligned}\right\} \tag{3.47}$$

The above system of equations is called the Gauss-Markov linear model. With the help of the following column vectors of the random variables and parameters

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix},$$

and the matrix of the known coefficients:

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1p} \\ a_{21} & a_{22} & \ldots & a_{2p} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{np} \end{pmatrix}$$

the equation (3.47) may be written completely as

$$E(Y) = A\beta,\ D(Y) = \sigma^2 I, \tag{3.48}$$

where $D(Y)$ is called dispersion matrix and I is the identity matrix of order n.

An alternative representation of (3.48), using the column vector e of independent vectors $e_1, e_2, \ldots, e_n$ is

$$\left.\begin{aligned} Y = A\beta + e \\ E(e) = O \text{ and } D(e) = \sigma^2 I \end{aligned}\right\} \tag{3.49}$$

where O is null vector.

The unknown parameters $\beta_j's$ in the model are called effects. In linear estimation the effects are all fixed quantities (parameters) and such a model where all effects are known parameters is also called a fixed effect model or Model I. Sometimes one of the β_j's is a constant with $a_{ij} = 1$ for that j and all $i = 1, 2, \ldots, n$. Such an effect is called a general effect or the additive constant.

In linear estimation, we find unbiased linear estimators of estimable linear parametric functions starting from model of the form (3.47). Among unbiased linear estimators, again, we find one having the minimum possible variance. The estimator is considered to be the best, in the sense of having the maximum concentration of the distribution of the estimator around the unknown true value of the parametric function. The value of the estimator in a particular sampling situation will be the corresponding estimate. Thus we obtain the estimate if we put the observed values $y_1, y_2, \ldots, y_n$ for the random variables $Y_1, Y_2, \ldots, Y_n$ in the estimator.

3.5.2 Least-square estimators and normal equations

Let $b_1, b_2, \ldots, b_p$ denote any set of p known quantities which may be used as estimates of $\beta_1, \beta_2, \ldots, \beta_p$. Then for such a $b' = (b_1, b_2, \ldots, b_p)$ we may form the sum of squares $(Y - Ab)'(Y - Ab)$, which will provide us with a measure of how well the estimate b for β fits the model (3.48). The smaller the above measure the better the corresponding estimate. And, from this argument, we get the following set of least-square estimators.

A set of measurable functions of Y, say $\hat{\beta}_1 = \hat{\beta}_1(Y), \hat{\beta}_2 = \hat{\beta}_1(Y), \ldots, \hat{\beta}_p = \hat{\beta}_p(Y)$, such that the values $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ minimize the sum of squares of the deviations of $Y_1, Y_2, \ldots, Y_n$ from their expectations, i.e., $S = (Y - A\beta)'(Y - A\beta)$, is called a set of least-square (LS) estimators of the unknown parameters $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ of the linear model (3.48).

We next show that minimum value of S is attained when $\hat{\beta}$ is a solution of a set of equations which are called the normal equations.

We have $\hat{\beta}A'Y = Y'AB = \sum_{\alpha,j} a_{\alpha j}\beta_j Y_\alpha$ where $j = 1, 2, \ldots, p$ and $\alpha = 1, 2, \ldots, n$.

Hence

$$\frac{d}{d\beta_j}(\beta' B' A' Y) = \frac{d}{d\beta_j}(Y' A\beta) = \sum_\alpha a_{\alpha j} Y_\alpha = d_j' Y$$

where $d_1, d_2, \ldots, d_p$ are the column vectors of A.

Let $A'A = C = (C_{ij})$.

Clearly, C is a symmetric matrix of order p.

Now, $\beta' A' A\beta = \beta' C\beta = \sum_{ij} C_{ij}\beta_i\beta_j$, with $i, j = 1, 2 \ldots, p$.

Hence

$$\frac{d}{d\beta_j}(\beta' A' A\beta) = \frac{d}{d\beta_j}(\beta' C\beta) = 2\sum_i C_{ij}\beta_i = 2C_j'\beta,$$

where $C_j' = (C_{1j}, C_{2j}, \ldots, C_{pj})$.

Differentiating $S = (Y - A\beta)'(Y - A\beta)$ w.r.t. $\beta_1, \beta_2, \ldots, \beta_p$ and equating the derivatives to zero, we have then

$$\frac{1}{2}\frac{d}{d\beta_j}S = -d_j'Y + C_j'\beta = 0 \text{ for } j = 1, 2, \ldots, p. \tag{3.50}$$

The equation (3.50) are called the normal equations and equivalent to

$$\left.\begin{aligned} A'A\beta &= A'Y \\ \text{or,} \quad C\beta &= A'Y \end{aligned}\right\} \tag{3.51}$$

where $C = A'A$.

The normal equations always admit a solution since $A'Y$ lies in the vector space generated by the column of C. Let $\hat{\beta}$ be a solution of these equations.

Every solutions of the normal equations is a set of LS estimators and every set of LS estimators satisfies the normal equations.

The solution of the normal equations $\beta = \hat{\beta}$ gives an extreme value of S. To show that this extreme values is the minimum value os S, we produce as follows:

$$\begin{aligned} (Y - A\beta)'(Y - A\beta) &= \left[Y - A\hat{\beta} + A(\hat{\beta} - \beta)\right]' \left[Y - A\hat{\beta} + A(\hat{\beta} - \beta)\right] \\ &= (Y - A\hat{\beta})'(Y - A\hat{\beta}) + (\hat{\beta} - \beta)'A'A(\hat{\beta} - \beta) \\ &\geq (Y - A\hat{\beta})'(Y - A\hat{\beta}) \end{aligned} \tag{3.52}$$

since the quadratic form $\left[A(\hat{\beta} - \beta)\right]' \left[A(\hat{\beta} - \beta)\right]$ cannot be negative. The equality holds only when $\beta = \hat{\beta}$. Thus $\beta = \hat{\beta}$ minimizes S.

Further, if $\hat{\beta}$ and $\hat{\hat{\beta}}$ are any two solutions of (3.51), then

$$(Y - A\hat{\beta})'(Y - A\hat{\beta}) = (Y - A\hat{\hat{\beta}})'(Y - A\hat{\hat{\beta}}).$$

This along with (3.52) shows that every solution of the normal equations is a set of LS estimators.

3.6 Unit Summary

In this unit, the concept of multiple regression and multiple correlation along with partial correlation are introduced. The multiple regression equation for $(p-1)$ independent variables is deduced. An example is also given. The expression for multiple and partial correlation coefficients are deduced. Some relations among simple correlation coefficients, multiple correlation coefficients and partial correlation coefficients are presented. An exercise is also given at the end of this unit.

3.7 Self Assessment Questions

1. Let $(x_i, y_i, z_i),\ i = 1, 2, \ldots, n$ be a sample of size n drawn from a papulation. Find the regression equation of z on x and y.

2. Let $(2, 5, 3), (8, 3, 9), (5, 3, 6), (5, 0, 1), (3, -1, 2)$ be a sample of the variables x_1, x_2, x_3 drawn from a random papulation. Find the regression line of x_2 on x_i and x_3.

3. Obtain the multiple regression equation of x_i on $x_2, x_3, \ldots, x_p$ in terms of the means, the standard deviations and the inter correlations of the variables.

4. Define multiple correlation and partial correlation, and indicate how they differ from simple correlation. Deduce the formulae for a multiple and partial correlation coefficient in terms of total correlation coefficients.

5. Prove that $1 - r^2_{1.23} = (1 - r^2_{12})(1 - r^2_{13.2})$. Use this relation to show that the multiple correlation coefficient is numerically greater than nay of the total or partial correlation coefficients of x_1 with the other variables.

6. Show that $r_{12},\ r_{13}$ and r_{23} must satisfy the inequality $r^2_{12} + r^2_{13} + r^2_{23} - 2r_{12}r_{13}r_{23} \leq 1$.

7. The following constants 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.
   $\bar{x}_1 = 55.95 \qquad s_1 = 2.26 \qquad r_{12} = 0.578$
   $\bar{x}_2 = 51.48 \qquad s_2 = 4.39 \qquad r_{13} = 0.581$
   $\bar{x}_3 = 56.03 \qquad s_3 = 4.41 \qquad r_{23} = 0.974$
   (a) Obtain the liner 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 cc.
   (b) Compute the partial correlation coefficient of weight and volume, estimating the effect of length.

3.8 References

1. Goon, A.M., Gupta, M.K. and Dasgupta, B., Fundamentals of Statistics, Vol-I, The World Press Pvt. Ltd.

2. Goon, A.M., Gupta, M.K. and Dasgupta, B., An outline of Statistical Theory, Vol-II, The World Press Pvt. Ltd.

3. Montgomery, B.C., Peck E.A. and Vining, G.G., Introduction to Linear Regression Analysis, John Wiley and Sons, Inc.

4. Stark, H., Woods, J. W., Probability, Statistic and Random Processes for Engineers, Pearson.

5. Mukhopadhyay, P., Mathematical Statistics, New Central Book agency.

6. Klimov, G., Probability Theory and Mathematical Statistics, Mir Publishers, Moscow.