Multiple regression:

In many statistical investigations based on multivariate data there is just one variable, say x_1, which is of primary interest. x_1 is considered together with the other variables say $x_2, x_3, \ldots, x_p$ in order to study the nature of the dependence of x_1 on these variables.

Let us assume that the relationship between x_1 and $x_2, x_3, \ldots, x_p$ is linear and given by an equation of the form:

$$x_{1.23\ldots p} = a + b_2 x_2 + b_3 x_3 + \ldots + b_p x_p \quad \text{——— (1)}$$

Let us assume that each of the p variables has n values. The values of the variables for the α^{th} individual may be denoted by $x_{1\alpha}, x_{2\alpha}, \ldots, x_{p\alpha}$ for $\alpha = 1, 2, \ldots, n$.

The constants $a, b_2, b_3, \ldots, b_p$ are determined on the basis of the given data by the principle of least-square.

If we denote by $x_{1.23\ldots p}$ the difference $x_1 - x_{1.23\ldots p}$ then the error of estimate corresponding to the α^{th} individual is $x_{1.23\ldots p,\alpha}$. The least-square method means that the constants $a, b_2, b_3, \ldots, b_p$ are to be so determined that

$$\sum_{\alpha} x^2_{1.23\ldots p,\alpha} = \sum_{\alpha} (x_{1\alpha} - a - b_2 x_{2\alpha} - \ldots - b_p x_{p\alpha})^2$$

is minimum.

The normal equations to determining the constant are given by:

$$\sum_\alpha x_{1\alpha} = na + b_2 \sum_\alpha x_{2\alpha} + b_3 \sum_\alpha x_{3\alpha} + \cdots + b_p \sum_\alpha x_{p\alpha}$$

$$\sum_\alpha x_{1\alpha} \cdot x_{2\alpha} = a \sum_\alpha x_{2\alpha} + b_2 \sum_\alpha x_{2\alpha}^2 + b_3 \sum_\alpha x_{3\alpha} x_{2\alpha} + \cdots + b_p \sum_\alpha x_{p\alpha}$$

$$\vdots$$

$$\sum_\alpha x_{p\alpha} x_{p\alpha} = a \sum_\alpha x_{p\alpha} + b_2 \sum_\alpha x_{2\alpha} x_{p\alpha} + \cdots + b_p \sum_\alpha x_{p\alpha} \quad \text{——— (2)}$$

or,

$$\sum x_{1.23\cdots p,\alpha} = 0$$

$$\sum x_{2\alpha} \cdot x_{1.23\cdots p,\alpha} = 0 \quad \text{——— (3)}$$

$$\vdots$$

$$\sum x_{p\alpha} \cdot x_{1.23\cdots p,\alpha} = 0$$

Dividing the first equation of (2) by n we get

$$\bar{x}_1 = a + b_2 \bar{x}_2 + b_3 \bar{x}_3 + \cdots + b_p \bar{x}_p \quad \text{——— (4)}$$

Multiplying (4) by $n\bar{x}_2, n\bar{x}_3, \cdots, n\bar{x}_p$ and subtracting from the second, third, $\cdots$ p^{th} eq respectively of the system (2) we have $(p-1)$ equations determining the b's. viz.

$$S_{21} = b_2 S_{22} + b_3 S_{23} + \cdots + b_p S_{2p}$$

$$S_{31} = b_2 S_{32} + b_3 S_{33} + \cdots + b_p S_{3p} \quad \text{——— (5)}$$

$$\vdots$$

$$S_{p1} = b_2 S_{p2} + b_3 S_{p3} + \cdots + b_p S_{pp}$$

where $S_{ij} = \sum x_{i\alpha} x_{j\alpha} - n\bar{x}_i \cdot \bar{x}_j = \sum (x_{i\alpha} - \bar{x}_i)(x_{j\alpha}$

Now if we write $s_{ij} = \frac{1}{n} S_{ij}$ then the matrix $(s_{ij})_{p \times p}$ is called the dispersion matrix of $x_1, x_2, \ldots, x_p$.

Therefore the system of equation in (5) can be put into the form.

$$\begin{pmatrix} s_{21} \\ s_{31} \\ \vdots \\ s_{p1} \end{pmatrix}_{p-1 \times 1} = \begin{pmatrix} s_{22} & s_{23} & \cdots & s_{2p} \\ s_{32} & s_{33} & \cdots & s_{3p} \\ \vdots & \vdots & & \vdots \\ s_{p2} & s_{p3} & \cdots & s_{pp} \end{pmatrix}_{p-1 \times p-1} \begin{pmatrix} b_2 \\ b_3 \\ \vdots \\ b_p \end{pmatrix}_{p-1 \times 1}$$

$$\Rightarrow \begin{pmatrix} b_2 \\ b_3 \\ \vdots \\ b_p \end{pmatrix} = \begin{pmatrix} s_{22} & s_{23} & \cdots & s_{2p} \\ s_{32} & s_{33} & \cdots & s_{3p} \\ \vdots & \vdots & & \vdots \\ s_{p2} & s_{p3} & \cdots & s_{pp} \end{pmatrix}^{-1} \begin{pmatrix} s_{21} \\ s_{31} \\ \vdots \\ s_{p1} \end{pmatrix}$$

$$\therefore b_j = \frac{\begin{vmatrix} s_{22} & \cdots & s_{2j-1} & s_{21} & s_{2j+1} & \cdots & s_{2p} \\ s_{32} & \cdots & \cdots & \cdots & \cdots & \cdots & s_{3p} \\ \vdots & & & & & & \\ s_{p2} & & s_{pj-1} & s_{p1} & s_{pj+1} & \cdots & s_{pp} \end{vmatrix}}{\begin{vmatrix} s_{22} & s_{23} & \cdots & s_{2p} \\ s_{32} & s_{33} & \cdots & s_{3p} \\ \vdots & & & \\ s_{p2} & s_{p3} & \cdots & s_{pp} \end{vmatrix}}$$

For $j = 2, 3, \ldots, p$

$$\Rightarrow b_j = (-1)^{j-2} \times \frac{s_1}{s_j} \times \frac{\begin{vmatrix} r_{21} & r_{22} & \cdots & r_{2,j-1} & r_{2,j+1} & \cdots & r_{2p} \\ r_{31} & r_{32} & \cdots & \cdots & \cdots & \cdots & r_{3p} \\ \vdots & \vdots & & & & & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{p,j-1} & r_{p,j+1} & \cdots & r_{pp} \end{vmatrix}}{\begin{vmatrix} r_{22} & r_{23} & \cdots & \cdots & r_{2p} \\ r_{32} & r_{33} & \cdots & \cdots & r_{3p} \\ \vdots & \vdots & & & \vdots \\ r_{p2} & r_{p3} & \cdots & \cdots & r_{pp} \end{vmatrix}}$$

Now if we consider the correlation matrix

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{pmatrix}_{p \times p}$$

and R_{ij} is the cofactor of r_{ij} in R, then the numerator of b_j is the minor of r_{ij} $(i=2,3,\cdots)$ in R and is $(-1)^{1+j}$ times the cofactor of r_{ij}, while the denominator is also the cofactor of r_{11} in R.

Hence $b_j = (-1)^{2j-1} \frac{s_1}{s_j} \frac{R_{1j}}{R_{11}}$ for $j = 2,3,\cdots,p$

$$= -\frac{s_1}{s_j} \frac{R_{1j}}{R_{11}}$$

b_j is called the partial regression coefficient of X_1 on X_j

Thus the multiple regression equation becomes

$$X_{1.23\cdots p} = \bar{x}_1 - \frac{s_1}{s_2} \frac{R_{12}}{R_{11}} (x_2 - \bar{x}_2) - \frac{s_1}{s_3} \frac{R_{13}}{R_{11}} (x_3 - \bar{x}_3) - \cdots\cdots - \frac{s_1}{s_p} \frac{R_{1p}}{R_{11}} (x_p - \bar{x}_p)$$