20

Partial and Multiple Correlation and Regression

INTRODUCTION

The simple correlation and regression analysis discussed earlier measure the degree and nature of the effect of one variable on another. While it is useful to know how one phenomenon is influenced by another, it is also important to know how one phenomenon is affected by several other variables. One variable is related to a number of other variables, many of which may be interrelated among themselves. For example, yield of rice is affected by the type of soil, temperature, amount of rainfall, etc. It is part of the statistician's task to determine the effect of one case when the effect of others is estimated. This is done with the help of multiple and partial correlation analysis. Thus, it shall be possible for us to compare the relative importance of television advertisement and newspaper advertisement on increasing sales.

The basic distinction between multiple and partial correlation analysis is that whereas in the former, we measure the degree of association between the variable Y and all the variables, $X_1, X_2, X_3, \ldots X_n$, *taken together,* in the latter we measure the degree of association between Y and one of the variables $X_1, X_2, X_3, \ldots X_n$, *with the effect of all the other variables removed.* It should be noted that when only two variables are included in a study, the dependent variable is usually designated by Y, and the independent variable by X. However, when more than one independent variable is used it becomes advantageous to distinguish between the variables by means of subscripts and use only the letter X. The dependent variable is generally denoted by X_1 and the independent variables by X_2, X_3, etc. This scheme of notation can be expanded to take care of any number of independent variables.

PARTIAL CORRELATION

It is often important to measure the correlation between a dependent variable and one particular independent variable when all other variables involved are kept constant, *i.e.,* when the effects of all other variables are removed (often indicated by the phrase "other things being equal"). This can be obtained by calculating coefficient of partial correlation. For example, if we have three variables : yield of wheat, amount of rainfall and temperature and if we limit our analysis of yield and rainfall to periods when a certain average daily temperature existed, or if we treat the problem mathematically in such a way that changes in temperature are allowed for, the problem becomes one of partial correlation. Thus, partial correlation analysis measures the strength of the relationship between Y and one independent variable in such a way that variations in the other independent variables are taken into account. A partial correlation coefficient is analogous to a partial regression coefficient in that all other factors are "*held constant*". Simple correlation, on the other hand, ignores the effect of all the other variables even through these variables might be quite closely related to the dependent variable, or to one another.

Partial Correlation Coefficients

Partial correlation coefficients provide a measure of the relationship between the dependent variable and other variables, with the effect of the rest of the variables eliminated.

If we denote by $r_{12.3}$, the coefficient of partial correlation between X_1 and X_2 keeping X_3 constant, we find that

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

Similarly,

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

where, $r_{13.2}$ is the coefficient of partial correlation between X_1 and X_3 keeping X_2 constant.

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{13}^2}}$$

where $r_{23.1}$ is the coefficient of partial correlation between X_2 and X_3 keeping X_1 constant. Thus, for three variables X_1, X_2 and X_3, there will be three coefficients of partial correlation each studying the relationship between two variables when the third is held constant. It should be noted that the squares of partial correlation coefficients are called *coefficients of partial determination.*

It will be clear from above that the partial correlation coefficients measure the degree of correlation between the dependent variable and each independent variable when the values of specified combinations of the other independent variables are held constant. These coefficients enable us to determine the direct relationship between any two variables independent of the indirect effect of the other variables : Partial correlation coefficient helps in deciding whether to include or not, an additional independent variable in regression analysis. Depending on the number of independent variables held constant, we often talk of *zero-order, first-order, second-order* correlation coefficients. The correlation coefficients obtained in case of two variables, *i.e.,* X and Y is called *zero-order correlation coefficient* since no restrictions are imposed on the values of all variables other than X and Y. The zero-order coefficients possess no secondary subscripts—that is subscripts after the point. If one independent variable is held constant in correlating two other variables, the resulting coefficient is known as the *first-order correlation* coefficient. Thus in a trivariate case, the partial correlation and regression coefficients such as $r_{12.3}$, $r_{13.2}$, $r_{23.1}$ are called first order coefficients. In a similar manner, a correlation between two variables while holding the values of the two other variables constant is known as a *second-order correlation coefficient.* Examples of second order coefficient are $r_{12.23}$, $r_{14.23}$, $r_{13.24}$, etc.

The notation of partial correlation coefficients always follows the same principle ; namely, the two variables being correlated are identified by the subscripts of r before the dot and the variables held constant are identified by the subscripts after the dot. It may be noted that so long as the particular subscripts of r are on the correct side of the period, the order in which they are placed is of no consequence. For example, $r_{12.34} = r_{12.43} = r_{21.43}$. However, the usual practice among statisticians is to place the subscripts in the ascending order.

Coefficients of a given order can generally be expressed in terms of the next lower order such as expressing partial correlations for the trivariate case in terms of simple correlations. This possibility simplifies greatly, the computational work involved in case of three or four independent variables.

A partial correlation coefficient measures the net co-variation between two of the variables under consideration. It is interpreted in terms of its squared values—coefficient of partial determination. For

instance, if $r_{12.3} = 0.8$ then $r^2_{12.3}$ would be 0.64 which means that the errors made in estimating X_1 from X_2 are reduced by 64 per cent when X_3 is employed as an additional explanatory variable.

Illustration 1. In a trivariate distribution it is found that

$$r_{12} = 0.7,\ r_{13} = 0.61, r_{23} = 0.4$$

Find the values of $r_{23.1}$, $r_{13.2}$ and $r_{12.3}$

Solution.
$$r_{23.1} = \frac{r_{23} - r_{12}\, r_{13}}{\sqrt{1-r_{12}^2}\ \sqrt{1-r_{13}^2}}$$

Substituting the given values

$$r_{23.1} = \frac{0.4 - 0.7 \times 0.61}{\sqrt{1-(0.7)^2}\ \sqrt{1-(0.61)^2}} = \frac{0.4 - 0.427}{\sqrt{0.51}\ \sqrt{1-0.3721}} = 0.048$$

$$r_{13.2} = \frac{r_{13} - r_{12}\, r_{23}}{\sqrt{1-r_{12}^2}\ \sqrt{1-r_{23}^2}}$$

$$= \frac{0.61 - 0.7 \times 0.4}{\sqrt{1-(0.7)^2}\ \sqrt{1-(0.4)^2}} = \frac{0.61 - 0.28}{\sqrt{1-0.49}\ \sqrt{1-0.16}} = 0.504$$

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{1-r_{13}^2}\ \sqrt{1-r_{23}^2}} = \frac{0.7 - 0.6 \times 0.4}{\sqrt{1-(0.61)^2}\ \sqrt{1-(0.4)^2}} = 0.633$$

Illustration 2. On the basis of observation made on agricultural production (X_1) the use of fertilizers (X_2) and the use of irrigation (X_3), the following zero order correlation coefficients were obtained :

$$r_{12} = 0.8, r_{13} = 0.65, r_{23} = 0.7$$

Compute the partial correlation between agricultural production and the use of fertilizers eliminating the effect of irrigation.

Solution. We have to calculate the value of $r_{12.3}$

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

Substituting the values $r_{12.3} = 0.8, r_{13} = 0.65$ and $r_{23} = 0.7$

$$r_{12.3} = \frac{0.8 - (0.65 \times 0.7)}{\sqrt{1-(0.65)^2}\ \sqrt{1-(0.7)^2}} = \frac{0.8 - 0.455}{\sqrt{1-0.4225}\ \sqrt{1-0.49}} = 0.636.$$

Illustration 3. Is it possible to get the following from a set of experimental data.

(*a*) $r_{23} = 0.8, r_{31} = -0.5, r_{12} = 0.6$ (*b*) $r_{23} = 0.7,\ r_{31} = -0.4, r_{12} = 0.6.$

Solution. In order to see whether there is any inconsistency, we should calculate $r_{12.3}$. If its value exceeds one, there is inconsistency otherwise not.

(*a*)
$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\ \sqrt{1-r_{23}^2}} = \frac{0.6 - (-0.5)\,(0.8)}{\sqrt{1-(-0.5)^2}\ \sqrt{1-(0.8)^2}}$$

$$= \frac{0.6 + 0.4}{\sqrt{0.75}\ \sqrt{0.36}} = \frac{1}{0.52} = 1.92$$

Since the value of $r_{12.3}$ is greater than one, therefore, there is some inconsistency in the given data.

(*b*)
$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\ \sqrt{1-r_{23}^2}}$$

$$= \frac{0.6 - (-0.4)\,(0.7)}{\sqrt{1-(0.4)^2}\ \sqrt{1-(0.7)^2}} = \frac{0.6 + 0.28}{\sqrt{0.84}\ \sqrt{0.51}} = \frac{0.88}{0.65} = 1.35$$

This again is greater than one, therefore, there is some inconsistency in the given data.

Partial Correlation Coefficients in more than three variables

When four variables are involved in a correlation problem, there are twelve possible first-order coefficients. Some of these are :

$$r_{14.2} = \frac{r_{14} - r_{12}\, r_{24}}{\sqrt{1 - r_{12}^2}\ \sqrt{1 - r_{24}^2}}$$

$$r_{14.3} = \frac{r_{14} - r_{13}\, r_{34}}{\sqrt{1 - r_{13}^2}\ \sqrt{1 - r_{34}^2}}$$

$$r_{13.4} = \frac{r_{13} - r_{14}\, r_{34}}{\sqrt{1 - r_{14}^2}\ \sqrt{1 - r_{34}^2}}$$

$$r_{12.4} = \frac{r_{12} - r_{14}\, r_{24}}{\sqrt{1 - r_{14}^2}\ \sqrt{1 - r_{24}^2}}$$

$$r_{24.3} = \frac{r_{24} - r_{23}\, r_{34}}{\sqrt{1 - r_{23}^2}\ \sqrt{1 - r_{34}^2}}$$

$$r_{34.2} = \frac{r_{34} - r_{23}\, r_{24}}{\sqrt{1 - r_{23}^2}\ \sqrt{1 - r_{24}^2}}$$

$$r_{23.4} = \frac{r_{23} - r_{24}\, r_{34}}{\sqrt{1 - r_{24}^2}\ \sqrt{1 - r_{34}^2}}.$$

Similarly, the formulae for other partial correlation coefficients, *i.e.*, $r_{12.3}$, $r_{13.2}$, $r_{23.1}$, $r_{24.1}$, $r_{34.1}$, can also be written.

Second-order Partial Correlation Coefficients

Second-order coefficients may be obtained from order coefficients. In case of four variables, if $r_{12.34}$ is the coefficient of partial correlation between X_1 and X_2 keeping X_3 and X_4 constant, then

$$r_{12.34} = \frac{r_{12.4} - r_{13.4}\, r_{23.4}}{\sqrt{1 - r_{13.4}^2}\ \sqrt{1 - r_{23.4}^2}}.$$

Similarly,

$$r_{13.24} = \frac{r_{13.4} - r_{12.4}\, r_{23.4}}{\sqrt{1 - r_{12.4}^2}\ \sqrt{1 - r_{23.4}^2}}$$

and

$$r_{14.23} = \frac{r_{14.3} - r_{12.3}\, r_{24.3}}{\sqrt{1 - r_{12.3}^2}\ \sqrt{1 - r_{24.3}^2}}.$$

Alternative formulae giving the same results are available for all three of the second-order coefficients. They are :

$$r_{12.34} = \frac{r_{12.3} - r_{14.3}\, r_{24.3}}{\sqrt{1 - r_{14.3}^2}\ \sqrt{1 - r_{24.3}^2}}$$

$$r_{13.24} = \frac{r_{13.2} - r_{14.2}\, r_{34.2}}{\sqrt{1 - r_{14.2}^2}\ \sqrt{1 - r_{34.2}^2}}$$

$$r_{14.23} = \frac{r_{14.2} - r_{13.2}\, r_{34.2}}{\sqrt{1 - r_{13.2}^2}\ \sqrt{1 - r_{34.2}^2}}.$$

The value of a partial correlation coefficient is usually interpreted via the corresponding coefficient of partial determination, which is merely the square of the former. Thus, if $r_{12.3} = 0.4$, $r^2_{12.3} = 0.16$.

The t-test employed to test the significance of a simple correlation can be employed to test the significance of a partial correlation when the number of degrees of freedom is reduced by the number of variables eliminated.

Characteristics and Uses of Partial Correlation Analysis. The purpose of partial correlation analysis is the measurement of relationship between two factors, with the effects of one or more other factors eliminated. If the assumptions of the method are true for a series of data, the power of partial analysis is great. The problem of holding certain variables constants while the relationship between the other is measured often presents itself in statistical analysis. Partial correlation is especially useful in the analysis of interrelated series. It is particularly pertinent to uncontrolled experiments of various kinds in which, such interrelationship usually exists. Most economic data fall in this category.

Partial correlation is of greatest value when used in conjunction with gross and multiple correlation in the analysis of factors affecting variations in many kinds of phenomena.

It has the advantage that the relationships are expressed concisely in a few well-defined coefficients. Also it is adaptable to small amounts of data and the reliability of the results can be rather easily tested.

Limitations of Partial Correlation Analysis. 1. The usefulness of the partial analysis is somewhat limited by the following basic assumptions of the method :

(*i*) The gross or zero-order correlation must have linear regressions.

(*ii*) The effects of the independent variable must be additively and not jointly related.

(*iii*) Because the reliability of partial coefficients decreases as its order increases. The number of observations in gross correlations should be fairly large. Often the student carries the analysis beyond the limits of the data. Thus, weakness to some extent can be guarded against by test of reliability.

2. When the above assumptions have been satisfied, partial correlation analysis still possess the disadvantages of laborious calculations and difficult interpretation even for statisticians.

The interpretation of the partial and multiple correlation results tends to assume that the independent variable have casual effects on dependent variable. The assumption is sometimes true, but more often untrue in varying degrees.

MULTIPLE CORRELATION

In problems of multiple correlation, we are dealing with situations that involve three or more variables. For example, we may consider the association between the yield of wheat per acre and both the amount of rainfall and the average daily temperature. We are trying to make estimates of the value of one of these variables based on the values of all the others. The variable whose value we are trying to estimate is called the dependent variable and the other variables on which our estimates are based are known as independent variables. The statistician himself chooses which variable is to be dependent and which variables are to be independent. It is merely a question of problem being studied. If we are trying to determine the most probable weight of men, we make weight, the dependent variable and height, age, etc., independent variables. If on the other hand, we are interested in estimating height, we will make height the dependent variable and weight, age, etc., the independent variables. Thus in problems of multiple correlation, we always have three or more variables (one dependent and the rest independent). In order that we may distinguish them easily, we follow the custom of representing them by the letter X with subscript. The dependent variable is always denoted by X_1 and the others by X_2, X_3, etc. Thus in the height, age and weight problem, if we are trying to estimate men's weight (that is, if weight be dependent variable), we might denote

$X_1 \rightarrow$ weight in lbs.
$X_2 \rightarrow$ height in inches.
$X_3 \rightarrow$ age in years.

The multiple correlation is of great practical significance—for rarely is it ever true that a variable is influenced solely or predominantly by one other factor. For example, the sales of a manufacturer are influenced, among other things, by his prices, his competitive position in the industry, his sales promotion campaign, industry sales, competitors' prices and national prosperity. In simple correlation, only one of the independent variables at a time could be correlated with the manufact ırer's sales and there is no direct way of determining the extent to which the observed correlation might have been caused by the interacting influence of other factors on the two variables under study. For instance, in times of prosperity a high level of national income may lead to increased industry sales, a share of which is captured by this manufacturer. But to what extent are the manufacturer's sales influenced by the universally buoyant affect of national prosperity and to what extent are his sales affected by the particular trend of the industry sales within the economy, *i.e.*, assuming that the nation's economy remain fairly stable ? To answer questions of this type which are of vital importance in framing suitable managerial policies, one has to depend on multiple correlation analysis.

Coefficient of Multiple Correlation

The coefficient of multiple linear correlation* is represented by R_1 and it is common to add subscript designating the variables involved. Thus $R_{1.234}$ would represent the coefficient of multiple linear correlation between X_1 on the one hand, and X_2, X_3 and X_4 on the other. The subscript of the dependent variable is always to the left of the point.

The coefficient of multiple correlation can be expressed in terms of r_{12}, r_{13} and r_{23} as follows :

$$R_{1.23} = \sqrt{\frac{r^2_{12} + r^2_{13} - 2r_{12}\, r_{13}\, r_{23}}{1 - r^2_{23}}}$$

$$R_{2.13} = \sqrt{\frac{r^2_{21} + r^2_{23} - 2r_{12}\, r_{13}\, r_{23}}{1 - r^2_{13}}}$$

and

$$R_{3.12} = \sqrt{\frac{r^2_{13} + r^2_{23} - 2r_{12}\, r_{13}\, r_{23}}{1 - r^2_{12}}}$$

A coefficient of multiple correlation such as $R_{1.23}$ lies between 0 and 1. The closer it is to 1, the better is the linear relationship between the variables. The closer it is to 0, the worse is the linear relationship. If the coefficient of multiple correlation is 1, the correlation is called *perfect*. Although a correlation coefficient of 0 indicates no linear relationship between the variables, it is possible that a nonlinear relationship may exist. It should be noted that whereas the simple correlation coefficients range from + 1.0 to 0 to – 1.0, the coefficients of multiple correlation are always positive in sign and range from + 1.0 to 0.

An alternative formula for calculating $R_{1.23}$ is :

$$R_{1.23} = \sqrt{r^2_{12} + r_{13.2}\,(1 - r^2_{12})}.$$

*When a linear regression equation is used, the coefficient of multiple correlation is called the coefficient of linear multiple correlation unless otherwise specified. Whenever we refer to multiple correlation, we shall imply linear multiple correlation.

similarly, $$R_{1.24} = \sqrt{\frac{r^2_{12} + r^2_{14} - 2r_{12}\, r_{14}\, r_{24}}{1 - r^2_{24}}}$$

or $$R_{1.24} = \sqrt{r^2_{12} + r^2_{14.2}\,(1 - r^2_{12})}$$

and $$R_{1.34} = \sqrt{\frac{r^2_{13} + r^2_{14} - 2r_{13}\, r_{14}\, r_{34}}{1 - r^2_{34}}}$$

or $$R_{1.34} = \sqrt{r^2_{13} + r^2_{14.3}\,(1 - r^2_{13})}.$$

Coefficient of Multiple Determination

In chapter on correlation, we talked of coefficient of determination r^2 which measures the fit of a straight line to the two-variable scatter. In exactly the same way, the coefficient of multiple determination denoted by $R^2_{1.23}$ is also defined. Thus, $R^2_{1.23}$ may be thought of as a measure of closeness of fit of the regression plane to the actual points relative to the point of the means of the variable. Or, just as does r^2, $R^2_{1.23}$ measures the percentage of total error that is accounted for by the regression. Obviously, the greater the value of $R^2_{1.23}$, the smaller is the scatter and the better is the fit. Thus, if coefficient of multiple determination between yield of rice (X_1) and fertilizers (X_2) and rain (X_3) is 0.953, it means that 95.3 per cent of the variations in yield have been explained by the variation in fertilizers and rain. There remains only 4.7 per cent of the variations in yield of rice that can be explained only by factors which have not been taken into consideration in our analysis.

Illustration. 4. The following zero-order, correlation coefficients are given

$$r_{12} = 0.98,\ r_{13} = 0.44 \text{ and } r_{23} = 0.54.$$

Calculate multiple correlation coefficient treating first variable as dependent and second and third variables as independent.

Solution. We have to calculate the multiple correlation coefficient treating first variable as dependent and second and third variables as independent, *i.e.*, we have to find $R_{1.23}$

$$R_{1.23} = \sqrt{\frac{r^2_{12} + r^2_{13} - 2r_{12}\, r_{13}\, r_{23}}{1 - r^2_{23}}}$$

Substituting the given values

$$R_{1.23} = \sqrt{\frac{(0.98)^2 + (0.44)^2 - 2(0.98)\,(0.44)\,(0.54)}{1 - (0.54)^2}}$$

$$= \sqrt{\frac{0.9604 + 0.1936 - 0.4657}{0.7084}} = +0.986.$$

Advantages of Multiple Correlation Analysis. The coefficient of multiple correlation serves the following purposes :

1. It serves as a measure of the degree of association between one variable taken as the dependent variable and a group of other variables taken as the independent variables.

2. Hence it also serves as a measure of goodness of fit of the calculated plane of regression and consequently, as a measure of the general degree of accuracy of estimates made by reference to equation for the plane or regression.

Limitations of Multiple Correlation Analysis. 1. Multiple correlation analysis is based on the assumption that the relationship between the variables is linear. In other words, the rate of change in one variable in terms of another is assumed to be constant for all values. In practice, most relationship are not linear but follow some other pattern. This limits somewhat the use of multiple correlation analysis. The linear regression coefficients are not accurately descriptive of curvilinear data.

2. A second important limitation is the assumption that effects of independent variables on the dependent variables are separate, distinct and additive. When the effects of variables are additive, a given change in one has the same effect on the dependent variable regardless of the sizes of the other two independent variables.

3. Linear multiple correlation involves a great deal of work relative to the results frequently obtained. When the results are obtained, only a few students well trained in the method are able to interest them. The misuse of correlation results has probably cast more doubt on the method than is justified. However, this lack of understanding and resulting misuses are due to the complexity of the method.

Multiple Regression

In the simple linear regression model discussed earlier, we talked of one dependent variable and one independent variable.

In multiple regression analysis which is a logical extension of two-variable regression analysis, instead of a single independent variable, two or more independent variables are used to estimate the values of a dependent variable. However, the fundamental concepts in the analysis remain the same. The multiple regression and correlation analysis serves highly useful purpose in practice. Its main objectives are :

(*a*) To derive an equation which provides estimates of the dependent variable from values of the two or more independent variables.

(*b*) To obtain a measure of the error involved in using this regression equation as a basis for estimation.

(*c*) To obtain a measure of the proportion of variance in the dependent variable accounted for or "explained by" the independent variables.

The first purpose is accomplished by deriving an appropriate regression equation by the method of least squares. The second purpose is achieved through the calculation of standard error of estimate and the third purpose is accomplished by computing the multiple coefficient of determination.

The multiple regression equation involving two independent variables shall take the form

$$Y_c = a + b_1X_1 + b_2X_2$$

The general form of the linear multiple regression function for k independent variables

$$X_1, X_2, \ldots\ldots, X_k, \text{ is}$$

$$Y_c = a + b_1X_1 + b_2X_2 + \ldots\ldots + b_kX_k$$

The linear function which is fitted to data for two variables is referred to as a *straight line,* for three variables a *plane,* for four or more variables a *hyperplane.*

If we have three variables X_1, X_2 and X_3, the multiple linear regression equation of X_1 on X_2 and X_3 shall have the following form :

$$X_1 = a_{1.23} + b_{12.3}X_2 + b_{13.2}X_3$$

Obviously X_1 is the dependent variable here and X_2 and X_3 are independent variables. The constant $a_{1.23}$ is the intercept made by the regression plane; it is zero when the regression line passes through the origin. The regression coefficients denoted by $b_{12.3}$ and $b_{13.2}$ represent the rate of change of the dependent variable per unit change in each of the independent variables when the other independent variables are held constant. The first subscript always represents the dependent variable and the second subscript denotes the particular independent variable being related to X_1. The subscripts after the period indicate the other independent variables, all of which are held constant while the effect of the particular independent variable on X_1 is measured. Thus, $b_{12.3}$ measures the amount by which a unit change in X_2 is expected to

affect X_1 when X_3 is held constant and $b_{12.3}$ measures the amount of change in X_1 when X_3 is held constant and $b_{13.2}$ measures the amount of change in X_1 per unit change in X_3 when X_2 is held constant. Similarly, $b_{13.24}$ would represent the change in X_1 per unit change in X_3 when the values of X_2 and X_4 are held constant.

The regression coefficients, *i.e.*, b's in multiple linear regression are termed *coefficients of net regression* : the regression is *net* in the sense that the regression of the dependent variable on the particular independent variable is measured while holding the values of the other independent variables constant. It contrast, the coefficients in simple regression are called *coefficients of gross regression* because no allowance is made for indirect influences on the regression.

The follwing are the usual assumptions made in a linear multiple regression analysis illustrated for the case of two independent variables :

1. The conditional distributions of Y for given X_1 and X_2 are assumed to be normal.
2. These conditional distributions are assumed to have equal standard deviations.
3. The $Y - Y_c$ deviation are assumed to be independent of one another.

Normal Equations for the Least Square Regression Plane

Just as there exist, least square regression line approximating a set of N data points (X, Y) in a two-dimensional scatter diagram, so also there exist *least square regression* planes fitting a set of N data points (X_1, X_2, X_3) in a three-dimensional by scatter diagram.

The least square regression plane of X_1 on X_2 and X_3 has the equation (i) where $b_{12.3}$ and $b_{13.2}$ are determined by solving simultaneously, the *normal equations.*

$$\Sigma X_1 = Na_{1.23} + b_{12.3}\,\Sigma X_2 + b_{13.2}\Sigma X_3$$
$$\Sigma X_1 X_2 = a_{1.23}\,\Sigma X_2 + b_{12.3}\,\Sigma X_2^2 + b_{13.2} X_2 X_3$$
$$\Sigma X_1 X_3 = a_{1.23}\,\Sigma X_3 + b_{12.3}\,\Sigma X_2 X_3 + b_{13.2} X_3^2.$$

These can be obtained formally by multiplying both sides of equation (i) by 1, X_2 and X_3 successively and summing on both sides.

When the number of variables is 4 or more, solving the above system of normal equation becomes a very tedious procedure. Efficient methods solving simultaneous equations require a knowledge of matrix algebra, which is not assumed for the reader of the text. Thus in our discussion that follows, we shall confine ourselves to the two independent variables cases, which of course, can be extended to cover cases with three or more independent variables.

The work involved in finding these regression equations can be reduced by proceeding in terms of deviations from the mean of the variables under consideration. The regression equation for these variables in this procedure is :

$$x_1 = b_{12.3}\,x_2 + b_{13.2}\,x_3$$

where $$x_1 = (X_1 - \bar{X}_1),\ x_2 = (X_2 - \bar{X}_2),\ x_3 = (X_3 - \bar{X}_3).$$

The value $b_{12.3}$ and $b_{13.2}$ can be obtained by solving simultaneously the following two normal equations :

$$\Sigma x_1 x_2 = b_{12.3}\,\Sigma x_2^2 + b_{13.2}\,\Sigma x_2 x_3$$
$$\Sigma x_1 x_3 = b_{12.3}\,\Sigma x_2 x_3 + b_{13.2}\Sigma x_3^2.$$

The value of $b_{12.3}$ and $b_{13.2}$ can also be obtained as follows :

$$b_{12.3} = r_{12.3}\,\frac{\sigma_{1.23}}{\sigma_{2.13}}$$

$$b_{13.2} = r_{13.2}\frac{\sigma_{13.2}}{\sigma_{3.12}}$$

The regression equation of X_1 on X_2 and X_3 can be expressed as follows :

$$X_1 - \bar{X}_1 = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r^2_{23}}\right)\left(\frac{S_1}{S_2}\right)(X_2 - \bar{X}_2) + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r^2_{23}}\right)\left(\frac{S_1}{S_3}\right)(X_3 - \bar{X}_3) \quad ...(i)$$

The regression equation of X_3 on X_2 and X_1 can be written as follows :

$$X_1 - \bar{X}_2 = \left(\frac{r_{23} - r_{13}r_{12}}{1 - r^2_{12}}\right)\left(\frac{S_1}{S_2}\right)(X_2 - \bar{X}_2) + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r^2_{12}}\right)\left(\frac{S_3}{S_1}\right)(X_1 - \bar{X}_1) \quad ...(ii)$$

From (*i*) and (*ii*), the coefficients of X_3 and X_1 are respectively

$$b_{13.2} = \left(\frac{r_{13} - r_{12}r_{23}}{1 - r^2_{23}}\right)\left(\frac{S_1}{S_2}\right) \text{ and } \left(\frac{r_{13} - r_{23}r_{12}}{1 - r^2_{12}}\right)\left(\frac{S_2}{S_1}\right)$$

$$b_{13.2}\, b_{31.2} = \frac{(r_{13} - r_{12}r_{23})^2}{(1 - r^2_{23})(1 - r^2_{12})} = r^2_{13.2}$$

This method of obtaining regression equations is much simpler compared to one where simultaneously several normal equations are to be solved. For calculating regression equation for three variables, when the above procedure is used we need the following :

X_1	X_2	X_3
S_1	S_2	S_3
r_{12}	r_{13}	r_{23}

Other Equations of Multiple linear Regression

In the case of two variables, there were two equations of regression—one of them indicating regression of Y on X, and the other, that of X on Y. When there are three variables, there will be three equations of regression indicating the regression of X_1 on X_2 and X_3, the other indicating the regression of X_2 on X_1 and X_3 and the third indicating the regression of X_3 on X_1 and X_2. The first of these has been given earlier. If X_2 and X_3 were to be treated as dependent variables the regression equation will respectively be :

$$X_2 = a_{2.13} + b_{21.3}X_1 + b_{23.1}X_3 \quad ...(ii)$$

$$X_3 = a_{3.12} + b_{31.2}X_1 + b_{32.1}X_2 \quad ...(iii)$$

The normal equations for fitting (*ii*) will be :

$$\Sigma X_2 = Na_{2.13} + b_{21.3}\Sigma X_1 + b_{23.1}\Sigma X_3$$

$$\Sigma X_1X_2 = a_{2.13}\Sigma X_1 + b_{21.3}\Sigma X_1^2 + b_{23.1}\Sigma X_1X_2$$

$$\Sigma X_2X_3 = a_{2.13}\Sigma X_3 + b_{21.3}\Sigma X_1X_3 + b_{23.1}\Sigma X_3^2$$

In case we want to fit equation (*iii*), the normal equations will be :

$$\Sigma X_3 = Na_{3.12} + b_{31.2}\Sigma X_1 + b_{32.1}\Sigma X_2$$

$$\Sigma X_1X_3 = a_{3.12}\Sigma X_1 + b_{31.2}\Sigma X_1^2 + b_{32.1}\Sigma X_1X_2$$

$$\Sigma X_2X_3 = a_{3.12}\Sigma X_2 + b_{31.2}\Sigma X_1X_2 + b_{32.1}\Sigma X_2^2$$

Generalization for More Than Three Variables

In case of four variables, the linear regression equation of X_1 on X_2, X_3 and X_4 can be written as

$$X_1 = a_{1.234} + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4$$

It represents a hyperplane in four-dimensional space. On formal multiplication of both sides of the above equation by X_1, X_2, X_3 and X_4 successively and then summing on both sides, we obtain the normal equations for determination of $a_{12.34}$, $b_{13.24}$, $b_{12.34}$, and $b_{14.23}$ which when substituted to above equation gives the least square regression equation of X_1 on X_2, X_3 and X_4.

The regression equation that uses all the factors which have an influence on the dependent variable can be an extremely useful device for estimating a variable. The chief difficulty, however, with this type of analysis has been the burden of making the calculations. As the number of variables increases, the number of equations to be solved simultaneously and the number of cross products to sum increase to the point that the burden of the arithmetic makes the analysis extremely difficult. The electronic computer is ideally adopted to do this type of analysis and with its help it is possible to use a large number of variables and a large sample and perform all the calculations in a matter of a few seconds.

Relationship between Partial and Multiple Correlation Coefficients

Interesting results connecting the multiple correlation coefficients and the various partial correlation coefficients can be found. For example, we find :

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

$$1 - R^2_{1.234} = (1 - r^2_{13})(1 - r^2_{13.2})(1 - r^2_{14.23})$$

Illustration 5. Find multiple linear regression equation of X_1 on X_2 and X_3 from the data relating to three variables given below :

X_1 :	4	6	7	9	13	15
X_2 :	15	12	8	6	4	8
X_3 :	30	24	20	14	10	4

Solution. The regression equation of X_1 on X_2 and X_3 is

$$X_1 = a_{1.23} + b_{12.3}\, X_2 + b_{13.2}\, X_3$$

The value of the constants $a_{1.23}$, $b_{12.3}$ and $b_{13.2}$ are obtained by solving the following three normal equations :

$$\Sigma X_1 = N a_{1.23} + b_{12.3}\, \Sigma X_2 + b_{13.2}\, \Sigma X_3$$

$$\Sigma X_1 X_2 = a_{1.23}\, \Sigma X_2 + b_{12.3}\, \Sigma X_2^2 + b_{13.2}\, \Sigma X_2 X_3$$

$$\Sigma X_1 X_3 = a_{1.23}\, \Sigma X_3 + b_{12.3}\, \Sigma X_2 X_3 + b_{13.2}\, \Sigma X_3^2$$

Calculating the required values :

X_1	X_2	X_3	X_1X_2	X_1X_3	X_2X_3	X_2^2	X_3^2	X_1^2
4	15	30	60	120	450	225	900	16
6	12	24	72	144	288	144	576	36
7	8	20	56	140	160	64	400	49
9	6	14	54	126	84	36	196	81
13	4	10	52	130	40	16	100	169
15	3	4	45	60	12	9	16	225
$\Sigma X_1 = 54$	$\Sigma X_2 = 48$	$\Sigma X_3 = 102$	$\Sigma X_1X_2 = 339$	$\Sigma X_1X_3 = 720$	$\Sigma X_2X_3 = 1{,}034$	$\Sigma X_2^2 = 494$	$\Sigma X_3^2 = 2{,}188$	$\Sigma X_1^2 = 576$

Substituting the values in the normal equations :

$$6a_{1.23} + 48b_{12.3} + 102b_{13.2} = 54 \quad \ldots (i)$$

$$48a_{1.23} + 49b_{12.3} + 1034b_{13.2} = 339 \quad \ldots (ii)$$

$$102a_{1.23} + 1034b_{12.3} + 2188b_{13.2} = 720 \quad \ldots (iii)$$

Multiplying Eqn. (*i*) by 8, we get

$$48a_{1.23} + 384b_{12.3} + 816b_{13.2} = 432 \quad \ldots (iv)$$

Subtracting Eqn. (*ii*) from (*iv*), we get

$$110b_{12.3} + 218b_{13.2} = -93 \quad \ldots (v)$$

Multiplying Eqn. (*i*) by 17, we get

$$102a_{1.23} + 816b_{12.3} + 1734b_{13.2} = 918 \quad \ldots (vi)$$

Subtracting Eqn. (*iii*) from Eqn. (*vi*) we get

$$218b_{12.3} + 454b_{13.2} = -198 \quad \ldots (vii)$$

Multiplying Eqn. (*v*) by 109, we obtain

$$11990b_{12.3} + 23762b_{13.2} = -10137 \quad \ldots (viii)$$

Multiplying Eqn, (*vii*) by 55, we get

$$11990b_{12.3} + 24970b_{13.2} = -10890 \quad \ldots (ix)$$

Subtracting Eqn. (*viii*) from Eqn. (*ix*), we get

$$1208b_{13.2} = -753$$

$$b_{13.2} = \frac{-753}{1208} = -0.623.$$

Substituting this value of $b_{13.2}$ in Eqn. (*v*), we get

$$110b_{12.3} + 218\,(-0.623) = -93$$

$$110b_{12.3} = 135.814 - 93$$

$$b_{12.3} = \frac{42.814}{110} = +0.389.$$

Substituting the value of $b_{12.3}$ and $b_{13.2}$ in Eqn. (*i*), we get

$$6a_{1.23} + 48\,(0.389) + 102\,(-0.623) = 54$$

$$6a_{1.23} = 54 + 63.546 - 18.672 = 98.874$$

$$a_{1.23} = 16.479.$$

Thus, the required regression equation is :

$$X_1 = 16.479 + 0.389\,X_2 - 0.623\,X_3.$$

Illustration. 6. Given the following, determine the regression equation of :

(*i*) x_1 on x_2 and x_3, and (*ii*) x_2 on x_1 and x_3

$r_{12} = 0.8$ $r_{13} = 0.6$ $r_{23} = 0.5$

$\sigma_1 = 10$ $\sigma_2 = 8$ $\sigma_3 = 5.$

Solution. Regression equation of X_1 on X_2 and X_3 is given by

$$X_1 = a + b_{12.3}\,X_2 + b_{13.2}\,X_3.$$

If the variates X_1, X_2 and X_3 are measured as deviations from their respective means, 'a' will be zero. The values of $b_{12.3}$ and $b_{13.2}$ can be calculated from the data given above but not for 'a'. So, let us assume x_1 and x_2 represent deviations from means. So the regression equation of x_1 on x_2 and x_3 is :

$$x_1 = b_{12.3}\,x_2 + b_{13.2}\,x_3$$

$$b_{12.3} = \frac{\sigma_1}{\sigma_2} \times \frac{r_{12} - r_{13}\,r_{23}}{1 - r^2_{23}}$$

$$= \frac{10}{8} \times \frac{0.8 - (0.6)\,(0.5)}{1 - (0.5)^2} = 0.833.$$

$$b_{13.2} = \frac{\sigma_1}{\sigma_2} \times \frac{r_{13} - r_{13}\,r_{23}}{1 - r^2_{23}}$$

$$= \frac{10}{5} \times \frac{0.6 - (0.8)\,(0.5)}{1 - (0.5)^2} = 0.533.$$

$\therefore$ Required regression equation is

$$x_1 = 0.833\,x_2 + 0.533\,x_3.$$

(*ii*) Regression equation of x_2 on x_1 and x_3

$$x_2 = b_{21.3}\,x_1 + b_{23.1}x_3$$

$$b_{21.3} = \frac{\sigma_1}{\sigma_2} \times \frac{r_{12} - r_{23}\,r_{13}}{1 - r^2_{13}}$$

$$= \frac{8}{10} \times \frac{(0.8) - (0.5)(0.6)}{1 - (0.6)^2}$$

$$= \frac{8}{10} \times \frac{0.8 - 0.3}{1 - 0.36} = \frac{8}{10} \times \frac{05}{0.64} = 0.625.$$

$$b_{23.1} = \frac{\sigma_1}{\sigma_2} \times \frac{r_{23} - r_{12}\, r_{13}}{1 - r^2_{13}}$$

$$= \frac{8}{5} \times \frac{0.5 - (0.8)(0.6)}{1 - (0.6)^2} = \frac{8}{5} \times \frac{0.02}{0.64} = 0.05.$$

Thus $x_2 = 0.625\, x_1 + 0.05\, x_3$ is the required regression equation.

Reliability of Estimates

The problem of determining the accuracy of estimates from the multiple regression is basically the same as for estimates from a simple regression equation. Since the correlation is seldom perfect, estimates made from regression equation will deviate from the correct value or the dependent variable. If an estimate is to be of maximum usefulness, it is necessary to have some indication of its precision. Just as with the simple regression equation, the measure of reliability is an average of the deviation of the actual value of non-dependent variable from the estimates from the regression equation or in other words, the standard error of estimate.

The standard error of estimate of X_1 on X_2 and X_3, is defined as

$$S_{1.23} = \sqrt{\frac{\Sigma (X_1 - X_{1\,est})^2}{N - 3}}.$$

$S_{1.23}$ represents standard error of estimate of X_1 on X_2 and X_3, $X_{1\,est}$ indicates the estimated value of X_1 as calculated from the regression equations.

It terms of the correlation coefficients r_{12}, r_{13} and r_{23}, the standard error of estimate can also be computed from the results :

$$S_{1.23} = S_1 \sqrt{\frac{1 - r^2_{12} - r^2_{13} - r^2_{23} + 2r_{12}r_{13}r_{23}}{1 - r^2_{23}}}.$$

MISCELLANEOUS ILLUSTRATIONS

Illustration 7. In a trivariate distribution $\sigma_1 = 3, \sigma_2 = \sigma_3 = 5, r_{12} = 0.6, r_{23} = r_{31} = 0.8$

Find (*i*) $r_{23.1}$, and (*ii*) $R_{1.23}$.

Solution .

(*i*) $$r_{23.1} = \frac{r_{23} - r_{12}\, r_{13}}{\sqrt{1 - r^2_{12}}\,\sqrt{1 - r^2_{13}}} = \frac{0.8 - 0.6 \times 0.8}{\sqrt{1 - (0.6)^2}\,\sqrt{1 - (0.8)^2}} = \frac{0.8 - 0.48}{\sqrt{0.64}\,\sqrt{0.36}} = 0.667.$$

(*ii*) $$R_{1.23} = \sqrt{\frac{r^2_{12} + r^2_{13} - 2r_{12}\, r_{13}\, r_{23}}{1 - r^2_{23}}}$$

$$= \sqrt{\frac{(0.6^2 + 0.8^2) - 2(0.6)(0.8)(0.8)}{1 - (0.8)^2}} = \sqrt{\frac{(0.36 + 0.64) - 0.768}{0.36}} = 0.803$$

Illustration 8. Calculate (*a*) $R_{1.23}$, (*b*) $R_{3.15}$ and (*c*) $R_{2.13}$ for the following data :

$\overline{X}_1 = 6.8$, $\overline{X}_2 = 7.0$, $\overline{X}_3 = 74$, $S_1 = 1.0$, $S_2 = 0.8$, $S_3 = 9$, $r_{12} = 0.6$, $r_{13} = 0.7$, $r_{23} = 65$.

Solution. $$R_{1.23} = \sqrt{\frac{r^2_{12} + r^2_{13} - 2\, r_{12}\, r_{13}\, r_{23}}{1 - r^2_{23}}} = \sqrt{\frac{(0.6)^2 + (0.7)^2 - (2 \times 0.6 \times 0.7 \times 0.65)}{1 - (0.65)^2}}$$

$$= \sqrt{\frac{0.36 + 0.49 - 0.546}{0.5775}} = \sqrt{0.527} = 0.726$$

$$R_{3.12} = \sqrt{\frac{r^2_{13} + r^2_{23} - 2\, r_{12}\, r_{13}\, r_{23}}{1 - r^2_{12}}} = \sqrt{\frac{(0.7)^2 + (0.65)^2 - (2 \times 0.6 \times 0.7 \times 0.65)}{1 - (0.6)^2}}$$

$$= \sqrt{\frac{0.49 + 0.4225 - 0.546}{1 - 0.36}} = \sqrt{0.573} = 0.757$$

$$R_{2.13} = \sqrt{\frac{r^2_{13} + r^2_{23} - 2\, r_{12}\, r_{13}\, r_{23}}{1 - r^2_{13}}} = \sqrt{\frac{(0.6)^2 + (0.65)^2 - (2 \times 0.6 \times 0.7 \times 0.65)}{1 - (0.7)^2}}$$

$$= \sqrt{\frac{0.36 + 0.4225 - 0.546}{0.51}} = \sqrt{0.464} = 0.681.$$

Illustration 9. The following constants are obtained from measurement on length in m.m. (X_1), volume in c.c (X_2) and weight in gm (X_3) of 300 eggs :

$\overline{X}_1 = 55.95$ $S_1 = 2.26$ $r_{12} = 0.578$
$\overline{X}_2 = 51.48$ $S_2 = 4.39$ $r_{13} = 0.581$
$\overline{X}_3 = 56.03$ $S_3 = 4.41$ $r_{23} = 0.974$

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 m.m. and volume is 52.5 c.c.

Solution. We have to obtain linear regression equation of egg weight on egg length and egg volume, *i.e.*, X_3 on X_1 and X_2. The regression equation of X_3 on X_1 and X_2 can be written as :

$$X_3 - \overline{X}_3 = \left(\frac{r_{23} - r_{12}\, r_{13}}{1 - r^2_{12}}\right)\left(\frac{S_3}{S_2}\right)(X_2 - \overline{X}_2) + \left(\frac{r_{13} - r_{12}\, r_{23}}{1 - r^2_{13}}\right)\left(\frac{S_3}{S_1}\right)(X_1 - \overline{X}_1).$$

Substituting the values

$$X_3 - 56.03 = \left(\frac{0.974 - (0.581) \times 0.578}{1 - (0.578)^2}\right)\left(\frac{4.41}{4.39}\right)(X_2 - 51.48) + \left(\frac{0.581 - (0.974 \times 0.578)}{1 - (0.581)^2}\right)\left(\frac{4.41}{2.26}\right)(X_1 - 55.95)$$

$$X_3 - 56.03 = \left(\frac{0.974 - 0.336)}{1 - 0.334}\right)\left(\frac{4.41}{4.39}\right)(X_2 - 51.48) + \left(\frac{0.581 - 0.563}{1 - 0.338}\right)\left(\frac{4.41}{2.26}\right)(X_1 - 55.95)$$

$$X_3 - 56.03 = 0.962\,(X_2 - 51.48) + 0.053\,(X_1 - 55.95)$$

$$X_3 - 56.03 = 0.962X_2 - 49.52 + 0.053X_1 - 2.97$$

$$X_3 = 3.54 + 0.053X_1 + 0.962X_2$$

When length, *i.e.*, X_1 is 58 and volume, *i.e.*, X_2 is 52.5, weight of the egg would be :

$$X_3 = 4.37 + 0.040\,(58) + 0.962\,(52.5) = 4.37 + 2.32 + 50.50 = 57.19 \text{ gm.}$$

Illustration 10. The table shows the corresponding values of three variables X_1, X_2, and X_3. Find the least square regression of X_3 on X_1 and X_2. Estimate X_3 when $X_1 = 10$ and $X_2 = 6$.

X_1	3	5	6	8	12	14
X_2	16	10	7	4	3	2
X_3	90	72	54	42	30	12

Solution. The regression equation of X_3 on X_2 and X_1 can be written as follows :

$$X_3 - \overline{X}_3 = \left(\frac{r_{23} - r_{13}\, r_{12}}{1 - r^2_{12}}\right)\left(\frac{S_3}{S_2}\right)(X_2 - \overline{X}_2) + \left(\frac{r_{13} - r_{23}\, r_{12}}{1 - r^2_{12}}\right)\left(\frac{S_3}{S_1}\right)(X_1 - \overline{X}_1)$$

Calculating $\overline{X}_1, \overline{X}_2, \overline{X}_3, S_1, S_2, S_3, r_{12}, r_{13}, r_{23}$.

X_1	$(X_1-\overline{X}_1)$ x_1	x_1^2	X_2	$(X_2-\overline{X}_2)$ x_2	x_2^2	X_3	$(X_3-\overline{X}_3)$ x_3	x_3^2	x_1x_2	x_1x_3	x_2x_3
3	–5	25	16	+9	81	90	+40	1600	–45	–200	+360
5	–3	9	10	+3	9	72	+22	484	–9	–66	+66
6	–2	4	7	0	0	54	+4	16	0	–8	0
8	0	0	4	–3	9	42	–8	64	0	0	+24
12	+4	16	3	–4	16	30	–20	400	–16	–80	+80
14	+6	36	2	–5	25	12	–38	1444	–30	–228	+190

$\Sigma X_1 = 48$ $\Sigma x_1 = 0$ $\Sigma x_1^2 = 90$ $\Sigma X_2 = 42$ $\Sigma x_2 = 0$ $\Sigma x_2^2 = 140$ $\Sigma X_3 = 300$ $\Sigma x_3 = 0$ $\Sigma x_3^2 = 4008$ $\Sigma x_1x_2 = -100$ $\Sigma x_1x_3 = -582$ $\Sigma x_2x_3 = 720$

$$\overline{X}_1 = \frac{48}{6} = 8, \qquad \overline{X}_2 = \frac{42}{6} = 7, \qquad \overline{X}_3 = \frac{300}{6} = 50$$

$$S_1 = \sqrt{\frac{\Sigma(X_1 - \overline{X}_1)^2}{N}} = \sqrt{\frac{90}{6}} = 3.87$$

$$S_2 = \sqrt{\frac{S(X_2 - \overline{X}_2)^2}{N}} = \sqrt{\frac{140}{6}} = 4.83$$

$$S_3 = \sqrt{\frac{S(X_3 - \overline{X}_3)^2}{N}} = \sqrt{\frac{4008}{6}} = 25.85$$

$$r_{12} = \frac{S\,x_1\,x_2}{\sqrt{S\,x_1^2 \times S\,x_2^2}} = \frac{-100}{\sqrt{90 \times 140}} = -0.891$$

$$r_{13} = \frac{S\,x_2\,x_3}{\sqrt{S\,x_1^2 \times S\,x_2^3}} = \frac{-582}{\sqrt{90 \times 4008}} = -0.969$$

$$r_{23} = \frac{S\,x_2\,x_3}{\sqrt{S\,x_2^2 \times S\,x_3^2}} = \frac{720}{\sqrt{140 \times 4008}} = 0.961$$

$$X_3 - 50 = \left[\frac{0.961 - (-0.969 \times -0.891)}{1 - (-0.891)^2}\right]\left(\frac{25.85}{4.83}\right)(X_2 - 7)$$
$$+ \left(\frac{-0.969 - (0.961 \times -0.891)}{1 - (-0.891)^2}\right)\left(\frac{25.85}{3.87}\right)(X_1 - 8)$$

$$X_3 - 50 = 2.546\,(X_2 - 7) - 3.664\,(X_1 - 8)$$
$$X_3 - 50 = 2.546\,X_2 - 17.822 - 3.664\,X_1 + 29.312$$
$$X_3 = 2.546\,X_2 - 3.664\,X_1 + 61.49.$$

When $X_1 = 10$ and $X_2 = 6$, X_3 will be

$$X_3 = 15.276 - 36.64 + 61.49 = 40.126 \text{ or } 40.$$

Illustration. 11. Suppose a computer has found, for a given set of values of X_1, X_2 and X_3

$r_{12} = 0.96$, $r_{13} = 0.36$ and $r_{23} = 0.78$.

Examine whether these computations may be said to be free from errors.

Solution. For determining whether the given computations are free from errors, or not, we compute the value of $r_{12.3}$. If $_{12.3}$ comes out to be greater than one, the computations cannot be regarded as free from errors.

$$r_{12.3} = \frac{r_{12} - r_{13}\,r_{23}}{\sqrt{1 - r_{13}^2}\,\sqrt{1 - r_{23}^2}} = \frac{0.96 - 0.36 \times 0.78}{\sqrt{1 - (0.36)^2}\,\sqrt{1 - (0.78)^2}}$$

$$= \frac{0.96 - 0.2808}{\sqrt{0.8704}\,\sqrt{0.3916}} = \frac{0.6792}{0.9329 \times 0.6258} = \frac{0.6792}{0.5838} = 1.163$$

Since $r_{12.3}$ is greater than one, the given computations about r_{12}, r_{13}, etc., do contain some error.

Illustration 12. If $r_{12} = 0.6$, $r_{13} = 0.5$ and $r_{23} = 0.2$, compute the values of $r_{12.3}$ and $R_{1.23}$.

Solution.

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

$r_{12} = 0.6$, $r_{13} = 5$, $r_{23} = 0.2$. Substituting the values, we get

$$r_{12.3} = \frac{0.6 - 0.5 \times 0.2}{\sqrt{1 - (0.5)^2}\ \sqrt{1 - (0.2)^2}} = \frac{0.6 - 0.1}{\sqrt{0.75 \times 0.96}}\ \frac{0.5}{0.8485} = 0.589$$

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}} = \sqrt{\frac{(0.6)^2 + (0.5)^2 - 2(0.6)\,(0.5)\,(0.2)}{1 - (0.2)^2}}$$

$$= \sqrt{\frac{0.36 + 0.25 - 0.12}{1 - 0.04}} = \sqrt{\frac{0.49}{0.96}} = 0.714.$$

Illustration 13. The simple correlation coefficients between variables X_1, X_2 and X_3 are respectively $r_{12} = 0.41$, $r_{13} = 0.71$ and $r_{23} = 0.50$. Calculate the partial correlation coefficients $r_{12.3}$, $r_{23.1}$ and $r_{31.2}$.

Solution. We are given $r_{12} = 0.41$; $r_{13} = 0.71$; $r_{23} = 0.5$. We have to find $r_{12.3}$, $r_{23.1}$ and $r_{31.2}$.

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{1 - r_{13}^2}\ \sqrt{1 - r_{23}^2}} = \frac{0.41 - (0.71 \times 0.5)}{\sqrt{1 - (0.7)^2}\ \sqrt{1 - (0.5)^2}}$$

$$= \frac{0.41 - 0.355}{\sqrt{0.51}\ \sqrt{0.7}} = \frac{0.055}{0.60} = 0.09$$

$$r_{23.1} = \frac{r_{23} - r_{12} r_{13}}{\sqrt{1 - r_{12}^2}\ \sqrt{1 - r_{13}^2}} = \frac{0.5 - (0.41 \times 0.71)}{\sqrt{1 - (0.41)^2}\ \sqrt{1 - (0.71)^2}}$$

$$= \frac{0.5 - 0.2911}{\sqrt{0.8319}\sqrt{0.4959}} = \frac{0.2089}{0.6423} = 0.325$$

$$r_{13.2} = \frac{r_{13} - r_{12} r_{23}}{\sqrt{1 - r_{12}^2}\ \sqrt{1 - r_{23}^2}} = \frac{0.71 - 0.41 \times 0.5}{\sqrt{1 - (0.41)^2}\ \sqrt{1 - (0.5)^2}}$$

$$= \frac{0.71 - 0.205}{\sqrt{0.8319}\sqrt{0.75}} = \frac{0.505}{0.7899} = 0.639$$

Illustration 14. In a trivariate distribution :

$$\bar{X}_1 = 28.20,\ \bar{X}_2 = 4.91,\ \bar{X}_3 = 594,\ S_1 = 4.4,\ S_2 = 1.1,\ S_3 = 80$$
$$r_{12} = 0.80,\ r_{23} = -0.56,\ r_{31} = -0.40.$$

(*a*) Find the correlation coefficient $r_{23.1}$ and $R_{1.23}$.

(*b*) Also estimate the value of X_1, when $X_2 = 6.0$ and $X_3 = 650$.

Solution. (*a*)

$$r_{23.1} = \frac{r_{23} - r_{12} r_{13}}{\sqrt{1 - r_{12}^2}\ \sqrt{1 - r_{13}^2}} = \frac{-0.56 - (0.8 \times -0.4)}{\sqrt{1 - (0.8)^2}\ \sqrt{1 - (0.4)^2}}$$

$$= \frac{-0.56 + 0.32}{\sqrt{1 - 0.64}\sqrt{1 - 0.16}} = \frac{-0.24}{\sqrt{0.36 \times 0.84}} = -0.436$$

$$R_{1.23} = \sqrt{\frac{r^2 12 + r^2 13 - 2 r_{12} r_{13} r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.8)^2 + (-0.4)^2 - 2(0.8)(-0.4)(-0.56)}{1 - (0.56)^2}}$$

$$= \sqrt{\frac{0.64 + 0.16 - 0.3584}{1 - 0.3136}} = \sqrt{\frac{0.4416}{0.6864}} = +0.802$$

(*b*) The regression equation of X_1 on X_2 and X_3 can be written as follows :

$$X_1 - \bar{X}_1 = \left(\frac{r_{12} - r_{12}\, r_{23}}{1 - r_{23}^2}\right)\left(\frac{S_1}{S_2}\right)(X_2 - \bar{X}_2) + \left(\frac{r_{13} - r_{12}\, r_{23}}{1 - r_{23}^2}\right)\left(\frac{S_1}{S_3}\right)(X_3 - \bar{X}_3)$$

$$X_1 - 28.02 = \left(\frac{0.8 - (-0.4) \times (0.56)}{1 - (-0.56)^2}\right)\left(\frac{4.4}{1.1}\right)(X_2 - 4.91) + \left(\frac{(-0.4) - (0.8)(-0.56)}{1 - (-0.56)^2}\right)\left(\frac{4.4}{80}\right)(X_3 - 0.594)$$

$$X_1 - 28.02 = \left(\frac{0.8 - 0.224}{0.6864}\right)\left(\frac{4.4}{1.1}\right)(X_2 - 4.91) + \left(\frac{-0.4 + 0.448}{0.6864}\right)\left(\frac{4.4}{80}\right)(X_3 - 0.594)$$

$$X_1 - 28.02 = 3.356\,(X_2 - 4.91) + 0.070\left(\frac{4.4}{80}\right)(X_3 - 0.594)$$

$$X_1 = 28.02 + 3.356\,X_2 - 16.478 + 0.0039\,(X_3 - 0.594)$$

$$= 11.539 + 3.356\,X_2 + 0.0039X_3$$ is the required regression equation of X_1 on X_2 and X_3.

When $X_2 = 6$ and $X_3 = 650$, estimated value of

$$X_1 = 11.539 + 3.356(6) + 0.0039(650)$$

$$= 11.539 + 20.136 + 2.535 = 34.2107.$$

Illustration 15. The simple correlation coefficients between temperature (X_1), corn yield (X_2) and rainfall (X_3) are : $r_{12} = 0.59$, $r_{13} = 0.46$ and $r_{23} = 0.77$

Calculate partial correlation coefficient $r_{12.3}$ and multiple correlation coefficient $R_{1.23}$.

Solution.

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

$$= \frac{0.59 - (0.46 \times 0.77)}{\sqrt{1 - (0.46)^2}\,\sqrt{1 - (0.77)^2}}$$

$$= \frac{0.59 - 0.354}{\sqrt{0.7884 \times 0.4071}} = \frac{0.236}{0.5665} = 0.417$$

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13}\, r_{32}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.59)^2 + (0.46)^2 - 2(0.59)(0.46)(0.77)}{1 - (0.77)^2}} = \sqrt{\frac{0.3481 + 0.2116 - 0.418}{1 - 0.5929}} = \sqrt{\frac{0.1417}{0.4071}} = 0.59.$$

Illustration 16. If $r_{12} = 0.8$, $r_{13} = -0.4$, and $r_{23} = -0.56$, find the value of $r_{12.3}$, $r_{13.2}$, and $r_{23.1}$.

Solution.

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{1 - r_{13}^2}\,\sqrt{1 - r_{23}^2}} = \frac{0.8 - (-0.4)(-0.56)}{\sqrt{1 - (0.4)^2}\,\sqrt{1 - (-0.56)^2}}$$

$$= \frac{0.8 - 0.224}{\sqrt{0.84 \times 0.6864}} = \frac{0.576}{0.759} = 0.759$$

$$r_{13.2} = \frac{r_{13} - r_{12}\, r_{23}}{\sqrt{1 - r_{12}^2}\,\sqrt{1 - r_{23}^2}} = \frac{-0.4 - (0.8)(-0.56)}{\sqrt{1 - (0.8)^2}\,\sqrt{1 - (-0.56)^2}}$$

$$= \frac{-0.4 + 0.448}{\sqrt{1 - 0.64}\,\sqrt{1 - 0.3136}} = \frac{0.048}{\sqrt{0.36 \times 0.6864}} = \frac{0.048}{0.4971} = 0.097$$

$$r_{23.1} = \frac{r_{23} - r_{12}\, r_{13}}{\sqrt{1 - r_{12}^2}\,\sqrt{1 - r_{13}^2}} = \frac{-0.56 - 0.8 \times -0.4}{\sqrt{1 - (0.8)^2}\,\sqrt{1 - (0.4)^2}}$$

$$= \frac{-0.56 + 0.32}{\sqrt{1 - 0.64}\,\sqrt{1 - 0.16}} = \frac{0.24}{0.55} = 0.436$$

Illustration 17. In a trivariate distribution

$$r_{12} = 0.863,\ r_{13} = 0.648 \text{ and } r_{23} = 0.709. \text{ Find } r_{12.3} \text{ and } R_{1.23}.$$

Solution.

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{1 - r_{13}^2}\ \sqrt{1 - r_{23}^2}} = \frac{0.863 - (0.648 \times 0.709)}{\sqrt{1 - (0.648)^2}\ \sqrt{1 - (0.709)^2}}$$

$$= \frac{0.863 - 0.4594}{\sqrt{0.58 \times 0.50}} = \frac{0.4036}{0.5385} = 0.749$$

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2 r_{12} r_{13} r_{23}}{1 - r_{23}^2}} = \sqrt{\frac{(0.863)^2 + (0.648)^2 - 2(0.863)(0.648)(0.709)}{1 - (0.709)^2}}$$

$$= \sqrt{\frac{0.7448 + 0.4199 - 0.7930}{1 - 0.5027}} = \sqrt{\frac{0.3717}{0.4973}} = 0.865.$$

Illustration 18. If $r_{12} = 0.80$; $r_{13} = -0.56$ and $r_{23} = -0.40$, then obtain, $r_{12.3}$ and $R_{1.23}$.

Solution.

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{1 - r_{13}^2}\ \sqrt{1 - r_{23}^2}} = \frac{0.8 - (-0.56)(-0.4)}{\sqrt{1 - (-0.56)^2}\ \sqrt{1 - (-0.4)^2}}$$

$$= \frac{0.8 - 0.224}{\sqrt{0.6864}\sqrt{0.84}} = \frac{0.576}{0.759} = 0.759$$

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2 r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.8)^2 + (-0.56)^2 - 2 \times 0.8 \times (-0.56)(-0.4)}{1 - (-0.4)^2}}$$

$$= \sqrt{\frac{0.64 + 0.3136 - 0.3584}{1 - 0.16}} = \sqrt{\frac{0.5952}{0.84}} = 0.842.$$

Illustration 19. Calculate (*a*) $R_{1.23}$, (*b*) $R_{3.12}$ and (*c*) $R_{2.13}$ for the following data :

$$r_{12} = 0.6 \qquad r_{13} = 0.7 \quad r_{23} = 0.65$$

Solution.

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2 r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.6)^2 + (0.7)^2 - (2 \times 0.6 \times 0.7 \times 0.65)}{1 - (0.65)^2}}$$

$$= \sqrt{\frac{0.36 + 0.49 - 0.546}{0.5775}} = \sqrt{0.526} = 0.725$$

$$R_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2 r_{12}\, r_{13}\, r_{23}}{1 - r_{12}^2}}$$

$$= \sqrt{\frac{(0.7)^2 + (0.65)^2 - 2(0.6 \times 0.7 \times 0.65)}{1 - (0.6)^2}}$$

$$= \sqrt{\frac{0.49 + 0.4225 - 0.546}{1 - 0.36}} = \sqrt{0.573} = 0.757$$

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2 r_{12}\, r_{13}\, r_{23}}{1 - r_{13}^2}}$$

$$= \sqrt{\frac{(0.6)^2 + (0.65)^2 - (2 \times 0.6 \times 0.7 \times 0.65)}{1 - (0.7)^2}}$$

$$= \sqrt{\frac{0.36 + 0.4225 - 0.546}{0.51}} = \sqrt{0.464} = 0.681.$$

PROBLEMS

1-A: Answer the following questions, each question carries **one** mark:

(*i*) What is the difference between $r_{12.3}$ and $r_{21.3}$?

(*ii*) Write the formula for partial correlation $r_{23.1}$. *(MBA, Madurai-Kamaraj Univ., 2003)*

(*iii*) Define partial and multiple correlation.

(*iv*) What purpose does partial correlation coefficient serves ?

(*v*) Write the formula for coefficient of multiple correlation $R_{1.23}$.

(*vi*) What is coefficient of determination ?

(*vii*) Write the formula for standard error of estimate $S_{1.23}$.

(*viii*) What is the difference between simple linear and multiple linear regression ?

(*ix*) How multiple correlation differs from partial correlation ?

(*x*) What do you understand by reliability of estimates ?

1-B : Answer the following questions, each question carries **four** marks:

(*i*) In a three variate multiple correlation analysis, the following results were obtained :
$r_{12} = 0.7$, $r_{13} = 0.6$, and $r_{23} = 0.4$
Find the multiple correlation coefficient $R_{1.23}$. *(M. Com., M.K. Univ., 2002)*

(*ii*) Describe the three steps in the process of multiple regression and correlation analysis. *(MBA, Madras Univ., 2003)*

(*iii*) What are zero-order, first-order and second-order coefficients ?

(*iv*) What are the uses and limitations of partial correlation analysis ?

(*v*) What are the advantages and limitations of multiple correlation analysis ?

(*vi*) What are 'normal equations' and how are they used in multiple regression analysis ?

2. Define partial and multiple correlation. With the help of an example distinguish clearly between partial and multiple correlation.
3. What is partial correlation ? Under what circumstances is it to be preferred to the total correlation?
4. (*a*) What is multiple linear regression? Explain clearly the difference between simple linear and multiple linear regression.
   (*b*) With the help of an example illustrate how does multiple linear regression help in the analysis of business problems.
5. Explain the concept of multiple regression and try to find out an example in practical field where multiple regression analysis is likely to be helpful.
6. Distinguish between partial and multiple correlation and point out their usefulness in statistical analysis.
7. Explain the terms : (*i*) Coefficient of determination. (*ii*) Regression coefficient, and (*iii*) Partial and multiple correlation.
8. How do we determine the reliability of estimates obtained from the multiple regression of X_1 on X_2 and X_3 ?
9. (*a*) In the multiple regression equation of X_1 on X_2 and X_3, what are the two regression coefficients and how do you interpret them ?
   (*b*) Explain the concepts of simple, partial and multiple correlation.
   (*c*) When is multiple regression needed? Explain with the help of an example. *(M.Com, DU, 2004)*
10. Within what limits the coefficient of multiple correlation $R_{1.23}$ lies? What inference would you draw if $R_{1.23} = 0$, $R_{1.23} = 1$, $R_{1.23} = 0.92$?
11. How do we distinguish between zero-order, first-order and second-order correlation coefficients ? Illustrate your answer with the help of some examples.
12. What precautions do you think must be observed while making use of partial and multiple correlation techniques ?
13. If $r_{12} = 0.6$, $r_{13} = 0.8$ and $r_{23} = -0.4$, find the values of $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$. Also calculate $R_{1.23}$ and $R_{3.12}$.
14. Calculate $R_{1.23}$, $R_{2.13}$ and $R_{3.12}$ for the following :
    $r_{12} = 0.6$, $r_{13} = 0.7$, $r_{23} = 0.65$
    and comment on these values.
15. In a certain investigation, the following values were obtained :
    $r_{12} = 0.8$, $r_{13} = 0.2$, $r_{23} = -0.5$.
    Do you think that the computations are free from error ?

16. The following information about a trivariate population is given to you :

$$\sigma_1 = 3.2,\ \sigma_2 = 4.5,\ \sigma_3 = 2.8,\ r_{12} = 0.3,\ r_{23} = 6 \text{ and } r_{13} = 0.8.$$

Do you think that the given data are consistent ? If so, calculate $r_{23.1}$ and $r_{1.23}$.

17. Given the following data, find the regression equation of X_1 on X_2 and X_3.

X_1 :	12	22	32	28
X_2 :	6	12	16	22
X_3 :	4	6	12	18

Also predict the value of X_1 when $X_2 = 5$ and $X_3 = 7$.

18. Given the following data :

Performance evaluation	(X_1) :	28	33	21	40	38	46
Aptitude Test Score	(X_2) :	74	87	69	69	81	97
Prior Experience	(X_3) :	5	11	4	9	7	10

(*i*) Develop the estimating equation, best describing these data.

(*ii*) If an employee scored 83 on the aptitude test and had a prior experience of 7 years, what performance evaluation would be expected ?

(M.Com., DU, 2007)

*****