Contents

CHAPTER 11

MULTIPLE REGRESSION AND CORRELATION

MULTIPLE REGRESSION AND CORRELATION

INTRODUCTION

The technique of simple regression which involves one dependent variable and one independent ble is often inadequate in most real-world situations where a variable depends upon two or more endent variables or regressors. For example, the yield of a crop depends upon the fertility of the fertilizer applied, rainfall, quality of seed, etc. Likewise, the systolic blood pressure of a person ds upon one's weight, age, etc. In such cases, the technique of simple regression may be expanded lude several independent variables. A regression which involves two or more independent variables ed a *multiple regression*. Thus, in case of multiple linear

regression where k independent variables influence the dependent variable Y, the general format of del is

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i, \ (i = 1, 2, ..., n)$$

re ε_i 's are the random errors,

α and β_i 's are the unknown population parameters, α is the intercept and β_1, β_2, ..., β_k are the regression coefficients for variables $X_1, X_2, ..., X_k$ respectively,

$X_{1i}, X_{2i}, ..., X_{ki}$ are the fixed values of k independent variables, the first of the two subscripts attached to each regressor denotes the variable and the second refers to the observation number,

We assume that

i) $E(\varepsilon_i) = 0$ for all i. This implies that for given values of X_i's,

$$E(Y_i) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki}.$$

ii) $\text{Var}(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2$ for all i, i.e. the variance of error terms is constant.

iii) $E(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$, *i.e.* error terms are independent of each other.

iv) $E(X, \varepsilon_i) = 0$ for all regressors, i.e. ε and each X variable are independent.

v) ε_i's are normally distributed with a mean of zero and a constant variance σ^2.

vi) We assume further in a multiple regression model that there exists no exact linear relationship between any two of the regressors.

The corresponding regression equation estimated from sample data then takes the following form

$$\hat{Y}_i = a + b_1 X_{1i} + b_2 X_{2i} + ... + b_k X_{ki},$$

re a and b_i's are the least-squares estimates of the population parameters α and β_i 's. The parameters or their estimates b_i's are called the *partial regression co-efficients* as β_i or its estimate 2,..., k) measures the change in the mean value of Y for a unit change in X_i, while all other ables remain unchanged.

MULTIPLE LINEAR REGRESSION WITH TWO REGRESSORS

For two independent variables X_1 and X_2, the predicting equation for an individual Y value is

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i,$$

Y on X_1 & X_2

and the *estimated* multiple linear regression based on sample data is

$$\hat{Y} = a + b_1 X_{1i} + b_2 X_{2i}$$

for a set of n observations, each of which is a number triple (X_{1i}, X_{2i}, Y_i). The error or residual in eac is given as

$$e_i = Y_i - \hat{Y}_i = Y_i - (a + b_1 X_{1i} + b_2 X_{2i})$$

Using the least-squares criterion, we determine those values of a, b_1 and b_2 which will minimize of squared residual, $\sum e_i^2$. To minimize $\sum e_i^2$, we find $\frac{\partial \sum e_i^2}{\partial a}$, $\frac{\partial \sum e_i^2}{\partial b_1}$ and $\frac{\partial \sum e_i^2}{\partial b_2}$ and set equal

Thus

$$\frac{\partial \sum e_i^2}{\partial a} = -2\sum[Y_i - (a + b_1 X_{1i} + b_2 X_{2i})] = 0,$$

$$\frac{\partial \sum e_i^2}{\partial b_1} = -2\sum X_{1i}[Y_i - (a + b_1 X_{1i} + b_2 X_{2i})] = 0,$$

$$\frac{\partial \sum e_i^2}{\partial b_2} = -2\sum X_{2i}[Y_i - (a + b_1 X_{1i} + b_2 X_{2i})] = 0.$$

Simplifying, we get the following three normal equations

$$\sum Y = na + b_1 \sum X_1 + b_2 \sum X_2,$$

$$\sum X_1 Y = a \sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1 X_2,$$

$$\sum X_2 Y = a \sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum X_2^2$$

The values of a, b_1 and b_2 are determined by solving these three normal equations sim and are substituted into

$$\hat{Y} = a + b_1 X_1 + b_2 X_2$$

to obtain the estimated multiple linear egression equation.

Example 11.1 A statistician wants to predict the incomes of restaurants, using two variables: the number of restaurant employees and restaurant floor area. He collected the follo

Income (000) Y	Floor area (000 sq. ft) X_1	Number of employees X_2
30	10	15
22	5	8
16	10	12
7	3	7
14	2	10

Calculate the estimated multiple linear regression equation (*i.e.* $\hat{Y} = a + b_1X_1 + b_2X_2$) for the above

The estimated multiple linear regression equation is

$$\hat{Y} = a + b_1X_1 + b_2X_2$$

a, b_1 and b_2 are the least squares estimates of α, β_1 and β_2. The three normal equations are:

$$\sum Y = na + b_1\sum X_1 + b_2\sum X_2$$
$$\sum X_1Y = a\sum X_1 + b_1\sum X_1^2 + b_2\sum X_1X_2,$$
$$\sum X_2Y = a\sum X_2 + b_1\sum X_1X_2 + b_2\sum X_2^2.$$

The calculations needed to find a, b_1 and b_2 are showing in the following table:

Y	X_1	X_2	X_1^2	X_2^2	X_1X_2	X_1Y	X_2Y
30	10	15	100	225	150	300	450
22	5	8	25	64	40	110	176
16	10	12	100	144	120	160	192
7	3	7	9	49	21	21	49
14	2	10	4	100	20	28	140
89	30	52	238	582	351	619	1007

Substituting the sums in the normal equations, we get

$$5a + 30b_1 + 52b_2 = 89$$
$$30a + 238b_1 + 351b_2 = 619$$
$$52a + 351b_1 + 582b_2 = 1007$$

Solving them simultaneously, we obtain

$$a = -1.33,\ b_1 = 0.38 \text{ and } b_2 = 1.62.$$

Hence the desired estimated multiple linear regression is

$$\hat{Y} = -1.33 + 0.38X_1 + 1.62X_2.$$

11.2.1 Expression of Multiple Linear Regression in Deviation Form. The computational cedure is considerably simplified by working with the deviations from the respective means of the ables. With two independent variables, the estimated multiple regression equation is

$$\hat{Y}_i = a + b_1X_{1i} + b_2X_{2i}, \qquad (i = 1, 2, \ldots, n)$$

As the regression equation goes through the point of means, we have

$$\hat{Y} = a + b_1\overline{X}_1 + b_2\overline{X}_2.$$

Subtracting, we get

$$\hat{y}_i = b_1 x_{1i} + b_2 x_{2i},$$

where $\hat{y}_i = \hat{Y}_i - \overline{Y}$, $x_{1i} = X_{1i} - \overline{X}_1$ and $x_{2i} = X_{2i} - \overline{X}_2$.

Then $e_i = y_i - \hat{y}_i = y_i - b_1 x_{1i} - b_2 x_{2i}$, and

$$\sum e_i^2 = \sum (y_i - b_1 x_{1i} - b_2 x_{2i})^2.$$

Differentiating $\sum e_i^2$ partially w.r.t b_1 and b_2, and equating to zero, we get

$$\frac{\partial \sum e_i^2}{\partial b_1} = -2\sum x_{1i}(y_i - b_1 x_{1i} - b_2 x_{2i}) = 0$$

$$\frac{\partial \sum e_i^2}{\partial b_2} = -2\sum x_{2i}(y_i - b_1 x_{1i} - b_2 x_{2i}) = 0$$

which yield, on simplification, the following two normal equations:

$$\sum x_1 y = b_1 \sum x_1^2 + b_2 \sum x_1 x_2,$$
$$\sum x_2 y = b_1 \sum x_1 x_2 + b_2 \sum x_2^2,$$

where the subscript i is dropped for convenience in printing.

Solving these two equations simultaneously, we get

$$b_1 = \frac{(\sum x_1 y)(\sum x_2^2) - (\sum x_2 y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}, \text{ and}$$

$$b_2 = \frac{(\sum x_2 y)(\sum x_1^2) - (\sum x_1 y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}.$$

Then a, the constant, is determined by

$$a = \overline{Y} - b_1 \overline{X}_1 + b_2 \overline{X}_2.$$

This is an alternative approach to solving the normal equations directly.

Example 11.2 Compute the estimated multiple linear regression $\hat{Y} = a + b_1 X_1 + b_2 X_2$ in Example 11.1, using the multiple regression in the deviation form.

In Example 11.1, we found that

$$\sum Y = 89, \sum X_1 = 30, \sum X_2 = 52, \sum X_1^2 = 238, \sum X_2^2 = 582,$$
$$\sum X_1 X_2 = 351, \sum X_1 Y = 619, \sum X_2 Y = 1007 \text{ and } n = 5.$$

Now we first calculate

$$\sum x_1^2 = \sum X_1^2 - \frac{(\sum X_1)^2}{n} = 238 - \frac{(30)^2}{5} = 58,$$

$$\sum x_2^2 = \sum X_2^2 - \frac{(\sum X_2)^2}{n} = 582 - \frac{(52)^2}{5} = 41.2,$$

$$\sum x_1 x_2 = \sum X_1 X_2 - \frac{(\sum X_1)(\sum X_2)}{n} = 351 - \frac{(30)(52)}{5} = 39,$$

$$\sum x_1 y = \sum X_1 Y - \frac{(\sum X_1)(\sum Y)}{n} = 619 - \frac{(30)(89)}{5} = 85,$$

$$\sum x_2 y = \sum X_2 Y - \frac{(\sum X_2)(\sum Y)}{n} = 1007 - \frac{(52)(89)}{5} = 81.4,$$

Next, we compute the regression co-efficients and constant as follows:

$$b_1 = \frac{(\sum x_1 y)(\sum x_2^2) - (\sum x_2 y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$= \frac{(85)(41.2) - (81.4)(39)}{(58)(41.2) - (39)^2} = \frac{327.4}{868.6} = 0.38,$$

$$b_2 = \frac{(\sum x_2 y)(\sum x_1^2) - (\sum x_1 y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$= \frac{(81.4)(58) - (85)(39)}{(58)(41.2) - (39)^2} = \frac{1406.2}{868.6} = 1.62,$$

and $$a = \overline{Y} - b_1 \overline{X}_1 + b_2 \overline{X}_2$$

$$= 17.8 - (0.38)(6) - (1.62)(10.4) = -1.33$$

Hence the desired multiple linear regression equation is

$$\hat{Y} = -1.33 + 0.38X_1 + 1.62X_2.$$

It is to be noted that we have exactly the same results as previously.

11.2.2 Standard Error of Estimate. The *standard error of estimate* is the standard deviation of ple regression. It measures the dispersion of Y values about the population multiple regression tion. For a multiple regression with two independent variables X_1 and X_2, it is denoted symbolically $_{Y.12}$ where the subscripts indicate that Y is regressed against two independent variables X_1 and X_2. lly, the value of $\sigma_{Y.12}$ is not known, it is therefore estimated from sample data.

mple standard error of estimate (unbiased estimate), denoted by $s_{Y.12}$ is given by

$$s_{Y.12} = \sqrt{\frac{\sum(Y - \hat{Y})^2}{n-3}},$$

but it can be computed more readily by using the following relations:

$$\Sigma(Y-\hat{Y})^2 = \Sigma[Y-(a+b_1X_1+b_2X_2)]^2$$
$$= \Sigma Y^2 - a\Sigma Y - b_1\Sigma X_1Y - b_2\Sigma X_2Y.$$

A larger value of $s_{Y.12}$ means that the multiple regression equation is of little use in estimatio prediction.

11.2.3 Co-efficient of Multiple Determination and Multiple Correlation. The *co-efficie* *multiple determination*, which measures as in the case of simple regression, the proportion of variabi the values of the dependent variable Y explained by its linear relation with the independent variabl defined by the ratio of the variation in Y explained by the regression equation to the total variation. F multiple regression with two regressors X_1 and X_2, the co-efficient of multiple determination is de symbolically by $R^2_{Y.12}$ and is computed by

$$R^2_{Y.12} = \frac{\Sigma(\hat{Y}-\overline{Y})^2}{\Sigma(Y-\overline{Y})^2},$$

where $\hat{Y} = a + b_1X_1 + b_2X_2$, but it can be readily computed by using the relation

$$\Sigma(\hat{Y}-\overline{Y})^2 = a\Sigma Y + b_1\Sigma X_1Y + b_2\Sigma X_2Y - (\Sigma Y)^2/n.$$

The co-efficient of multiple determination lies between 0 and 1, and has same meaning as in s linear regression.

The positive square root of the co-efficient of multiple determination, *i.e.* $\sqrt{R^2_{Y.12}}$ is call *co-efficient* of multiple *correlation*. $R_{Y.12}$ measures the degree of association between Y and bo regressors X_1 and X_2 combined, and is always taken to be positive.

Example 11.3 Compute the standard error of estimate, co-efficient of multiple determinati coefficient of multiple correlation for the data in Example 11.1.

For the data in Example 11.1, we found from the regression calculation, that

$$\Sigma Y = 89, \Sigma Y^2 = 1885, n = 5, a = -1.33.$$

$$\Sigma X_1Y = 619, \Sigma X_2Y = 1007, b_1 = 0.38, b_2 = 1.62$$

Therefore
$$s_{Y.12} = \sqrt{\frac{\Sigma Y^2 - a\Sigma Y - b_1\Sigma X_1Y - b_2\Sigma X_2Y}{n-3}}$$
$$= \sqrt{\frac{1885-(-1.33)(89)-(0.38)(619)-(1.62)(1007)}{5-3}}$$
$$= \sqrt{\frac{136.81}{2}} = \sqrt{68.405} = 8.27$$

which is the standard deviation of the multiple regression.

The coefficient of multiple determination is

$$R^2_{Y.12} = \frac{\Sigma(\hat{Y}-\overline{Y})^2}{\Sigma(Y-\overline{Y})^2}$$

$$= \frac{a\sum Y + b_1 \sum X_1 Y + b_2 \sum X_2 Y - (\sum Y)^2/n}{\sum Y^2 - (\sum Y)^2/n}$$

$$= \frac{(-1.33)(89) + (0.38)(619) + (1.62)(1007) - (89)^2/5}{1885 - (89)^2/5.}$$

$$= \frac{163.99}{300.80} = 0.55$$

This means that 55% of the variability in income is explained by its linear relationship with floor and the number of employees.

The co-efficient of multiple correlation, $R_{Y.12}$ is

$$R_{Y.12} = \sqrt{0.55} = 0.74.$$

11.2.4 Subscript Notation. For the purposes of generalization and change of variables, it is enient to adopt a notation due to *G*. Udny Yule (1871–1951). This notation involves subscripts. For ple, the individual *Y* value in case of the multiple linear regression with two independent variables, is en as

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

Using Yule's notation, this can be written as

$$X_{1i} = \beta_{1.23} + \beta_{12.3} X_{2i} + \beta_{13.2} X_{3i} + \varepsilon_i$$

the variables are numbered 1, 2 and 3 by the use of subscripts. The subscripted number 1 denotes ependent variable, 2 and 3 denote the independent variables X_2 and X_3 respectively, and $\beta_{1.23}$ is the of X_1 when X_2 and X_3 are both equal to zero.

There are three subscripts attached to each parameter. The subscripts preceding the point are called *ry subscripts* and those following the point are known as *secondary subscripts*. The dependent ble is always indicated by the first primary subscripts, while the second primary subscript indicates ariable to which the β co-efficient is attached. The secondary subscript(s) indicates which other ble(s) has been included in the regression equation. The secondary subscript, if more than one, may itten in any order.

The advantage of this notation is that it indicates the number of variables involved in the regression on and also shows which is the dependent variable and which are the independent variables.

The estimated multiple regression equation of X_1 on X_2 and X_3 is

$$\hat{X}_1 = b_{1.23} + b_{12.3} X_2 + b_{13.2} X_3.$$

It should be noted that in general $b_{12.3}$ is different from $b_{13.2}$.

Allowing a change of variables, the estimated regression equation of X_2 on X_1 and X_3 is given b

$$\hat{X}_2 = b_{2.13} + b_{23.1}X_3 + b_{21.3}X_1.$$

Similarly, the estimated regression equation of X_3 on X_1 and X_2 is

$$\hat{X}_3 = b_{3.12} + b_{31.2}X_1 + b_{32.1}X_2.$$

If the two variables, measured from their means, be x_1 and x_2 then the two simple regr equations of x_1 on x_2 and of x_2 and x_1 are

$$x_1 = b_{12}x_2 \text{ and } x_2 = b_{21}x_1$$

The residuals may be expressed as

$$x_{1.2} = x_1 - b_{12}x_2 \text{ and } x_{2.1} = x_2 - b_{21}x_1.$$

If x_1, x_2 and x_3 are three variables, measured from their respective means, then the m regression equation of x_1 on x_2 and x_3 is

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

and its residual is expressed by

$$x_{1.23} = x_1 - b_{12.3}x_2 - b_{13.2}x_3$$

The two normal equations may be written as

$$\sum x_2x_{1.23} = 0 \text{ and } \sum x_3x_{1.23} = 0.$$

11.2.5 Properties of Residuals. The residuals or errors have the following properties:

1. "The sum of the products of corresponding values of a variable and a residual is zero, p the subscript of the variable is included among the secondary subscripts of the residual."

Let the regression equation (in deviation form) of x_1 on x_2 and x_3 be

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

Then the two normal equations for determining the b's are

$$\sum x_2x_{1.23} = 0 = \sum x_3x_{1.23},$$

where $x_{1.23} = x_1 - b_{12.3}x_2 - b_{13.2}x_3$.

Similarly, the normal equations for the regression of x_2 on x_1 and x_3 and of x_3 on x_1 and x_2, a

$$\sum x_1x_{2.13} = 0 = \sum x_3x_{2.13},$$
$$\sum x_1x_{3.12} = 0 = \sum x_2x_{3.12}.$$

2. The sum of the products (or covariance) of two residuals remains unchanged by omit one residual any or all of secondary subscripts which are common to both".

Let the residual defined as $x_{1.2} = x_1 - b_{12}x_2$ be considered.

Then $\sum x_{1.2}x_{1.23} = \sum x_{1.23}(x_1 - b_{12}x_2)$.

$$= \sum x_1x_{1.23} - b_{12}\sum x_2x_{1.23}$$

The second term vanishes as $\sum x_2 x_{1.23} = 0$

Then $\sum x_{1.2} x_{1.23} = \sum x_1 x_{1.23}$

Again $$\sum x_{1.2} x_{1.23} = \sum x_{1.23}(x_1 - b_{12.3}x_2 - b_{13.2}x_3)$$
$$= \sum x_1 x_{1.23} - b_{12.3}\sum x_2 x_{1.23} - b_{13.2}\sum x_3 x_{1.23}$$

Here again the second and third terms vanish due to their being normal equations.

Hence $\sum x_{1.2} x_{1.23} = \sum x_1 x_{1.23}$.

3. "The sum of the products (or covariance) of two residuals is zero provided all the subscripts of one residual are included among the secondary subscripts of the second."

Let us consider the residuals defined by $x_{3.2}$ and $x_{1.23}$.

Then $\sum x_{3.2} x_{1.23} = \sum x_{3.2}(x_1 - b_{12.3}x_2 - b_{13.2}x_3)$

But this vanishes because of normal equation and property 1.

Similarly, $\sum x_{2.3} x_{1.23} = 0$.

11.2.6 Multiple Regression in terms of Linear Correlation Coefficients. The multiple ession equation of a variable, say X_1, on other variables, say X_2 and X_3, can be sometimes expressed in s of r_{12}, r_{13} and r_{23}, the linear correlation coefficients. The sample regression equation (in deviation) of x_1 on x_2 and x_3 is given by

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

The two normal equations are obtained as

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

$$\sum x_1 x_3 = b_{12.3}\sum x_2 x_3 + b_{13.2}\sum x_3^2$$

Let S_i^2 be the variance of x_i and let r_{ij} be the linear correlation co-efficient between x_i and x_j. Then ssing the normal equations in terms of variances and linear correlation co-efficient, we get

$$nr_{12}S_1S_2 = nb_{12.3}S_2^2 + nb_{13.2}r_{23}S_2S_3$$

$$nr_{13}S_1S_3 = nb_{12.3}r_{23}S_2S_3 + nb_{13.2}S_3^2$$

Simplification gives

$$r_{12}S_1 = b_{12.3}S_2 + b_{13.2}r_{23}S_3, \text{ and}$$

$$r_{13}S_1 = b_{12.3}r_{23}S_2 + b_{13.2}S_3$$

Solving these equations simultaneously for b's, we get

$$b_{12.3} = \frac{S_1}{S_2}\left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right), \text{ and}$$

$$b_{13.2} = \frac{S_1}{S_3}\left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right).$$

Substituting these values in the regression equation, we obtain

$$x_1 = \left(\frac{S_1}{S_2}\right)\left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right)x_2 + \left(\frac{S_1}{S_3}\right)\left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right)x_3.$$

Or dividing both sides of the equation by S_1, we get

$$\frac{x_1}{S_1} = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right)\left(\frac{x_2}{S_2}\right) + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right)\left(\frac{x_3}{S_3}\right)$$

as the multiple regression of x_1 on x_2 and x_3 in terms of standard deviations and the linear corre co-efficients of the variables involved. Similarly, the other two multiple regression equations of x_2 and x_3 and of x_3 on x_1 and x_2 are obtained as

$$\frac{x_2}{S_2} = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{13}^2}\right)\left(\frac{x_1}{S_1}\right) + \left(\frac{r_{23} - r_{12}r_{13}}{1 - r_{23}^2}\right)\left(\frac{x_3}{S_3}\right), \text{ and}$$

$$\frac{x_3}{S_3} = \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{12}^2}\right)\left(\frac{x_1}{S_1}\right) + \left(\frac{r_{23} - r_{12}r_{13}}{1 - r_{12}^2}\right)\left(\frac{x_2}{S_2}\right)$$

To obtain the regression equations in terms of original values, we replace x_1 by $X_1 - \overline{X}_1$, $X_2 - \overline{X}_2$, and x_3 by $X_3 - \overline{X}_3$ respectively.

11.3 MULTIPLE CORRELATION CO-EFFICIENT

The *co-efficient of multiple correlation* measures the degree of relationship between a variab its estimate from the regression equation. In other words, it is a product moment correlation betw variable, say x_1, and its value estimated by the regression equation $x_1 = b_{12.3}\, x_2 + b_{13.2}\, x_3$. The co-efficie multiple correlation between x_1 and the variables x_2 and x_3 combined, is denoted symbolically by $R_{1.2}$

Let us denote the *estimated value* of x_1 by $\hat{x}_1$. Then by definition,

$$R_{1.23} = \frac{Cov(x_1, \hat{x}_1)}{\sqrt{Var(x_1)\, Var(\hat{x}_1)}} = \frac{\sum x_1 \hat{x}_1}{\sqrt{\sum x_1^2 \sum(\hat{x}_1)^2}}$$

Now $\quad \sum x_1 \hat{x}_1 = \sum x_1(x_1 - x_{1.23}) \qquad (\because \hat{x}_1 = x_1 - x_{1.23})$

$$= \sum x_1^2 - \sum x_1 x_{1.23}$$

$$= \sum x_1^2 - \sum x_{1.23}x_{1.23} \qquad (\because \sum x_1 x_{1.23} = \sum x_{1.23}x_{1.23})$$

$$= n(S_1^2 - S_{1.23}^2),$$

where $S_{1.23}^2$ is the sample variance of residuals.

Also
$$\begin{aligned}\Sigma(\hat{x}_1)^2 &= \Sigma(x_1 - x_{1.23})^2\\ &= \sum x_1^2 + \sum x_{1.23}^2 - 2\sum x_1 x_{1.23}\\ &= \sum x_1^2 + \sum x_{1.23}^2 - 2\sum x_{1.23}^2 \qquad (\because \sum x_1 x_{1.23} = \sum x_{1.23} x_{1.23}\ )\\ &= \sum x_1^2 - \sum x_{1.23}^2 = n(S_1^2 - S_{1.23}^2),\end{aligned}$$

and $\sum x_1^2 = nS_1^2$

Substituting these values in the formula, we get

$$R_{1.23} = \frac{S_1^2 - S_{1.23}^2}{S_1\sqrt{S_1^2 - S_{1.23}^2}} = \left(1 - \frac{S_{1.23}^2}{S_1^2}\right)^{1/2}$$

Squaring, we get $R_{1.23}^2 = 1 - \dfrac{S_{1.23}^2}{S_1^2}$.

The quantity $S_{1.23}^2$ can be expressed in terms of the simple correlation co-efficients between the pairs of the variables as below:

$$S_{1.23}^2 = \frac{1}{n}\sum x_{1.23}^2 = \frac{1}{n}\Sigma(x_1 - b_{12.3}x_2 - b_{13.2}x_3)$$

$$= \frac{1}{n}\sum x_1(x_1 - b_{12.3}x_2 - b_{13.2}x_3)$$

(second property of residuals)

$$= \frac{1}{n}\sum x_1^2 - b_{12.3}\frac{1}{n}\sum x_1 x_2 - b_{13.2}\frac{1}{n}\sum x_1 x_3$$

$$= S_1^2 - b_{12.3}S_1 S_2 r_{12} - b_{13.2}S_1 S_3 r_{13}$$

Substituting the values of $b_{12.3}$ and $b_{13.2}$ in terms of simple correlation co-efficient and simplifying, we get

$$S_{1.23}^2 = S_1^2\left(\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{23}r_{13}}{1 - r_{23}^2}\right)$$

Hence
$$R_{1.23}^2 = 1 - \frac{S_1^2(1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{23}r_{13})}{S_1^2(1 - r_{23}^2)}$$

$$= \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{23}^2}, \text{ so that}$$

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

It should be noted that $R_{1.23}$ is necessarily positive or zero as the term $\sum x_1\hat{x}_1$ being equal to $\sum(\hat{x}$ cannot be negative. If $R_{1.23} = 1$, the $S^2_{1.23} = 0$, *i.e.* all the residuals $x_{1.23}$ are zero; the observed estimated values of x_1 coincide. The multiple correlation in this case, is called *perfect*, indicating a lin relationship between the variables.

Similarly, by the change of variables, we get

$$R_{2.31} = \sqrt{\frac{r_{23}^2 + r_{21}^2 - 2r_{12}r_{23}r_{13}}{1-r_{31}^2}}, \text{ and}$$

$$R_{3.12} = \sqrt{\frac{r_{31}^2 + r_{32}^2 - 2r_{12}r_{23}r_{31}}{1-r_{12}^2}},$$

Example 11.4 An instructor of mathematics wished to determine the relationship of grades final examination to grades on two quizzes given during the semester. Calling X_1, X_2 and X_3 the grad a student on the first quiz, second quiz and final examination respectively, he made the follo computations for a total of 120 students.

$$\overline{X}_1 = 6.8 \qquad S_1 = 1.0 \qquad r_{12} = 0.60$$
$$\overline{X}_2 = 7.0 \qquad S_2 = 0.8 \qquad r_{13} = 0.70$$
$$\overline{X}_3 = 74 \qquad S_3 = 9.0 \qquad r_{23} = 0.65$$

a) Find the least-squares regression equation of X_3 on X_1 and X_2.

b) Estimate the final grades of two students who scored respectively (1) 9 and 7, and (2) 4 on the two quizzes.

c) Compute $R_{3.12}$. (B.Sc. Eng

a) Since the standard deviations and linear correlation co-efficients are given, therefo estimated regression equation of X_3 on X_1 and X_2 is

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

Now $\dfrac{r_{13} - r_{12}r_{23}}{1-r_{12}^2} = \dfrac{0.70-(0.60)(0.65)}{1-(0.60)^2} = \dfrac{0.31}{0.64}$, and

$$\frac{r_{23} - r_{12}r_{13}}{1-r_{12}^2} = \frac{0.65-(0.60)(0.70)}{1-(0.60)^2} = \frac{0.23}{0.64}.$$

Substituting these values, we get

$$\frac{X_3 - 74}{9.0} = \left(\frac{0.31}{0.64}\right)\left(\frac{X_1 - 6.8}{1.0}\right) + \left(\frac{0.23}{0.64}\right)\left(\frac{X_2 - 7.0}{0.8}\right)$$

or $X_3 - 74 = 4.36\,(X_1 - 6.8) + 4.04\,(X_2 - 7.0)$

$$= 4.36\,X_1 - 29.648 + 4.04\,X_2 - 28.28$$

$$X_3 = 16.07 + 4.36\,X_1 + 4.04\,X_2$$

the desired least squares regression equation of X_3 on X_1 and X_2.

b) **Student 1:** When $X_1 = 9$ and $X_2 = 7$, we get

$$\hat{X}_3 = 16.07 + 4.36\,(9) + 4.04\,(7) = 83.59 = 84$$

Student 2: When $X_1 = 4$ and $X_2 = 8$, we get

$$\hat{X}_3 = 16.07 + 4.36\,(4) + 4.04\,(8) = 65.83 = 66.$$

c) The co-efficient of multiple correlation $R_{3.12}$ is

$$R_{3.12} = \sqrt{\frac{r_{31}^2 + r_{32}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{12}^2}},$$

$$= \sqrt{\frac{(0.70)^2 + (0.65)^2 - 2(0.60)(0.65)(0.70)}{1 - (0.60)^2}}$$

$$= \sqrt{\frac{0.3665}{0.64}} = \sqrt{0.5727} = 0.757.$$

4 PARTIAL CORRELATION

Relationship b/w any two variables by neglecting the effect of 3r variable.

A partial correlation measures the degree of linear relationship between any two variables in a ivariable problem under the condition that any common relationship or influence with all other ables (or some of them) has been removed. Stated differently, if there are three variables X_1, X_2 and hen the correlation between X_1 and X_2 after removing the linear effect of X_3 from X_1 and from X_2, is partial correlation. The sample *co-efficient of partial correlation* measuring the strength of the onship (correlation) between X_1 and X_2, when the influence of X_3 has been removed, is demoted olically by $r_{12.3}$. By removing the influence, we mean subtracting the fitted regression $\hat{X}_1$ from the ved values X_1 obtaining the residual – a part of X_1 not explained by X_3.

To derive the co-efficient of partial correlation $r_{12.3}$, we use the variables x_1, x_2 and x_3 which are ations from their means. The linear regression of x_1 on x_3 and of x_2 on x_3 are $x_1 = b_{13}x_3$ and $x_2 = b_{23}x_3$. oving the linear effect of x_3 from x_1 and from x_2 and denoting the residuals by $x_{1.3}$ and $x_{2.3}$, we get

$$x_{1.3} = x_1 - b_{13}x_3, \text{ and } x_{2.3} = x_2 - b_{23}x_3.$$

These residuals may be written as

$$x_{1.3} = x_1 - r_{13}\frac{S_1}{S_3}x_3, \text{ and } x_{2.3} = x_2 - r_{23}\frac{S_2}{S_3}x_3.$$

Now the co-efficient of partial correlation is the product moment correlation co-efficient between als $x_{1.3}$ and $x_{2.3}$. Thus by definition

$$r_{12.3} = \frac{\sum x_{1.3}x_{2.3}}{\sqrt{\sum x_{1.3}^2 \sum x_{2.3}^2}}.$$

Now $$\sum x_{1.3}x_{2.3} = \sum\left[x_1 - r_{13}\frac{S_1}{S_3}x_3\right]\left[x_2 - r_{23}\frac{S_2}{S_3}x_3\right]$$

$$= \sum\left[x_1x_2 - r_{23}\frac{S_2}{S_3}x_1x_3 - r_{13}\frac{S_1}{S_3}x_2x_3 + r_{13}r_{23}\frac{S_1S_2}{S_3^2}x_3^2\right]$$

$$= \sum x_1x_2 - r_{23}\frac{S_2}{S_3}\sum x_1x_3 - r_{13}\frac{S_1}{S_3}\sum x_2x_3 + r_{13}r_{23}\frac{S_1S_2}{S_3^2}\sum x_3^2$$

$$= n\,[r_{12}S_1S_2 - r_{23}r_{13}S_1S_2 - r_{13}r_{23}S_1S_2 + r_{13}r_{23}S_1S_2]$$

$$= n\,S_1S_2\,(r_{12} - r_{13}r_{23})$$

And $$\sum x_{1.3}^2 = \sum\left[x_1 - r_{13}\frac{S_1}{S_3}x_3\right]^2$$

$$= \sum x_1^2 + r_{13}^2\frac{S_1^2}{S_3^2}\sum x_3^2 - 2r_{13}\frac{S_1}{S_3}\sum x_1x_3$$

$$= n[S_1^2 + r_{13}^2S_1^2 - 2r_{13}^2S_1^2]$$

$$= nS_1^2(1 - r_{13}^2).$$

Similarly, $\sum x_{2.3}^2 = nS_2^2(1 - r_{23}^2)$.

Substituting these values in the formula, we obtain

$$r_{12.3} = \frac{S_1S_2(r_{12} - r_{13}r_{23})}{S_1S_2\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$ ✓

$$\boxed{r_{12\cdot3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}}$$ ✓

Alternatively. The partial correlation co-efficient between x_1 and x_2 when the influence been eliminated, is also defined as the geometric mean of the regression co-efficient $b_{12.3}$ and two partial regression lines of x_1 on x_2 and of x_2 on x_1 respectively, *i.e.*

$$r_{12.3} = \sqrt{b_{12.3} \times b_{21.3}}$$

$$= \sqrt{\frac{S_1}{S_2}\left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right)\cdot\frac{S_2}{S_1}\left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{13}^2}\right)}$$

$$= \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}} \qquad (r_{12.3} \text{ has the same sign as } b_{12.3} \text{ and } b_{21.3})$$

In a similar way, we can prove that

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{23}^2}}, \text{ and } r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{13}^2}}.$$

Example 11.5 From the following data, determine the linear regression equations of X_1 on X_3 and X_2 on X_3.

X_1	7	12	14	17	20
X_2	4	7	8	9	12
X_3	1	2	4	5	8

Find the deviations of observed values of X_1 from the regression, viz. $X_{1.3}$. Repeat the same for X_2, we obtain $X_{2.3}$. Determine the simple correlation co-efficient between the two sets of deviations $X_{1.3}$ and $X_{2.3}$.
(P.U., B.A./B.Sc. 1977)

The estimated simple regression equation of X_1 on X_3 is

$$\hat{X}_1 = b_{1.3} + b_{13}X_3,$$

where $b_{13} = \dfrac{n\sum X_1X_3 - (\sum X_1)(\sum X_3)}{n\sum X_3^2 - (\sum X_3)^2}$ and $b_{1.3} = \overline{X}_1 - b_{13}\overline{X}_3$.

The estimated simple regression equation of X_2 on X_3 is

$$\hat{X}_2 = b_{2.3} + b_{23}X_3,$$

where $b_{23} = \dfrac{n\sum X_2X_3 - (\sum X_2)(\sum X_3)}{n\sum X_3^2 - (\sum X_3)^2}$ and $b_{2.3} = \overline{X}_2 - b_{23}\overline{X}_3$.

The computations needed to find the b's are given in the table below:

X_1	X_2	X_3	X_1X_3	X_2X_3	X_3^2
7	4	1	7	4	1
12	7	2	24	14	4
14	8	4	56	32	16
17	9	5	85	45	25
20	12	8	160	96	64
70	40	20	332	191	110

$$\overline{X}_1 = \frac{\sum X_1}{n} = \frac{70}{5} = 14,\ \overline{X}_2 = \frac{\sum X_2}{n} = \frac{40}{5} = 8 \text{ and } \overline{X}_3 = 4.$$

And the regression co-efficient are obtained as

$$b_{13} = \frac{(5)(332)-(70)(20)}{(5)(110)-(20)^2} = \frac{260}{150} = 1.73,$$

$$b_{1.3} = 14-(1.73)(4) = 7.08,$$

$$b_{23} = \frac{(5)(191)-(40)(20)}{(5)(110)-(20)^2} = \frac{155}{150} = 1.03, \textit{ and}$$

$$b_{2.3} = 8-(1.03)(4) = 3.88.$$

Hence the desired regression equations are

$$\hat{X}_1 = 7.08 + 1.73X_3 \text{ and } \hat{X}_2 = 3.88 + 1.03X_3.$$

Next, we compute the residuals $X_{1.3} = X_1 - 7.08 - 1.73\ X_3$ and $X_{2.3} = X_2 - 3.88 - 1.03\ X_3$, and the s correlation between them. The necessary computations are given in the following table:

X_1	X_2	X_3	$X_{1.3}$	$X_{2.3}$	$X_{1.3}X_{2.3}$	$X_{1.3}^2$	$X_{2.3}^2$
7	4	1	–1.81	–0.91	1.6471	3.2761	0.8281
12	7	2	1.46	1.06	1.5476	2.1316	1.1236
14	8	4	0	0	0	0	0
17	9	5	1.27	–0.03	0.0381	1.6129	0.0009
20	12	8	–0.92	–0.12	0.1104	0.8464	0.0144
70	40	20	0	0	3.2670	7.8670	1.9670

Hence the co-efficient of correlation between $X_{1.3}$ and $X_{2.3}$, which is the co-efficient of correlation between X_1 and X_2 when the influence of X_3 has been removed, is obtained as

$$r_{12.3} = \frac{\Sigma X_{1.3}X_{2.3}}{\sqrt{\Sigma X_{1.3}^2 \, \Sigma X_{2.3}^2}} \qquad (\because \Sigma X_{1.3} = \Sigma X_{2.3} = 0)$$

$$= \frac{3.2670}{\sqrt{(7.8670)(1.9670)}} = \frac{3.2670}{3.9340} = 0.83$$

Example 11.6 Given r_{12}= 0.492, r_{13}= 0.927 and r_{23}= 0.758, find all the partial co co-efficients.

We have $r_{12.3} = \dfrac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}} = \dfrac{0.492-(0.927)(0.758)}{\sqrt{1-(0.927)^2}\sqrt{1-(0.758)^2}}$

$$= \frac{-0.2107}{\sqrt{0.1407 \times 0.4254}} = -0.86;$$

$$r_{23.1}=\frac{r_{23}-r_{21}r_{31}}{\sqrt{1-r_{21}^2}\sqrt{1-r_{31}^2}}=\frac{0.758-(0.492)(0.927)}{\sqrt{1-(0.492)^2}\sqrt{1-(0.927)^2}}$$

$$=\frac{0.302}{\sqrt{0.7579\times 0.1407}}=0.92;\text{ and}$$

$$r_{31.2}=\frac{r_{31}-r_{32}r_{12}}{\sqrt{1-r_{32}^2}\sqrt{1-r_{12}^2}}=\frac{0.927-(0.758)(0.492)}{\sqrt{1-(0.758)^2}\sqrt{1-(0.492)^2}}$$

$$=\frac{0.5541}{\sqrt{0.4254\times 0.7579}}=0.98.$$

xample 11.7 Show that if $x_3=ax_1+bx_2$, the three partial correlations are numerically equal to $_{2}$ having the sign of a, $r_{32.1}$, the sign of b and $r_{12.3}$, the opposite sign of a/b.

the multiple regression equation $x_3=ax_1+bx_2$, we treat x_3 as dependent and x_1 and x_2 as ent variables. Let the three variables be measured from their respective means.

quaring and summing over all values, we get

$$\sum x_3^2 \quad =a^2\sum x_1^2+b^2\sum x_2^2 \qquad \text{(the product vanishes as } x_1 \text{ and } x_2 \text{ are independent)}$$

$$=n(a^2S_1^2+b^2S_2^2)$$

ltiplying the given equation by x_1 and summing, we have

$$\sum x_1x_3 \quad =a\sum x_1^2 \qquad (\sum x_1x_2=0\text{, as } x_1 \text{ and } x_2 \text{ are independent)}$$

$$=naS_1^2$$

ow $\quad r_{31}=\dfrac{\sum x_1x_3}{\sqrt{\sum x_1^2\sum x_3^2}}=\dfrac{aS_1^2}{\sqrt{S_1^2(a^2S_1^2+b^2S_2^2)}}$

$$=\frac{aS_1}{\sqrt{a^2S_1^2+b^2S_2^2}}=\frac{aS_1}{w},\text{ where } w^2=a^2S_1^2+b^2S_2^2.$$

milarly, $r_{23}=\dfrac{bS_2}{w}$ and $r_{12}=0$

ence $r_{13.2}=\dfrac{r_{13}-r_{12}r_{32}}{\sqrt{(1-r_{12}^2)(1-r_{32}^2)}}=\dfrac{\frac{aS_1}{w}-0}{\sqrt{(1-0)\left(1-\frac{b^2S_2^2}{w^2}\right)}}$

$$=\frac{a}{+\sqrt{a^2}}=\pm 1,\text{ according as } a \text{ is +ve or –ve.}$$

In other words, $r_{13.2}$ has the sign of a.

Again $$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}} = \frac{0 - \frac{aS_1}{w}\cdot\frac{bS_2}{w}}{\sqrt{\left(1-\frac{a^2S_1^2}{w^2}\right)\left(1-\frac{b^2S_2^2}{w^2}\right)}}$$

$$= \frac{-abS_1S_2}{\sqrt{a^2b^2S_1^2S_2^2}} = \frac{-ab}{+\sqrt{a^2b^2}}.$$

$\because$ a^2b^2 is always positive, therefore, $\sqrt{a^2b^2}$ is always positive.

Now ab may be positive or negative.

Thus $r_{12.3}$ has the sign opposite to ab or $\frac{a}{b}$.

Similarly, $$r_{32.1} = \frac{r_{23} - r_{31}r_{21}}{\sqrt{(1-r_{31}^2)(1-r_{21}^2)}}$$

$$r_{32.1} = \frac{\frac{bS_2}{w} - 0}{\sqrt{\left(1-\frac{a^2S_1^2}{w^2}\right)(1-0)}}$$

$$= \frac{b}{+\sqrt{b^2}} = \pm 1,$$ according as b is +ve or –ve.

Hence the result.

11.4.1 Relationship between Multiple and Partial Correlation Co-efficients. correlation co-efficients can be connected with the various partial correlation co-efficients. we have shown earlier that

$$1 - R^2_{1.23} = \frac{S^2_{1.23}}{S_1^2},$$

where $$nS^2_{1.23} = \sum x^2_{1.23} = \sum x_{1.2}x_{1.23}$$ (second property of residuals)

$$= \sum x_{1.2}(x_1 - b_{12.3}x_2 - b_{13.2}x_3)$$

$$= \sum x^2_{1.2} - b_{13.2}\sum x_{1.2}x_{3.2}$$ because of the properties of residuals.

$$= \sum x^2_{1.2}\left[1 - b_{13.2}\frac{\sum x_{1.2}x_{3.2}}{\sum x^2_{1.2}}\right]$$

$$= nS^2_{1.2}[1 - b_{13.2}b_{31.2}] = nS^2_{1.2}(1 - r^2_{13.2})$$

$$S_{1.23}^2 = S_{1.2}^2(1 - r_{13.2}^2)$$
$$= S_1^2(1 - r_{12}^2)(1 - r_{13.2}^2)$$

$$1 - R_{1.23}^2 = (1 - r_{12}^2)(1 - r_{13.2}^2).$$

ding in the same way, we can find

$$1 - R_{1.234}^2 = (1 - r_{12}^2)(1 - r_{13.2}^2)(1 - r_{14.23}^2).$$

CURVILINEAR REGRESSION

Sometimes a scatter diagram indicates that the relationship between the two variables will be more ely described by a non-linear regression line. When this occurs, either we may transform one or f the variables so that the transformed data appear approximately linear or we may use a ial equation. In the former case, the estimating equation may be an exponential or a logarithmic . In the latter case, the estimating equation may be

$$\hat{Y} = a + bX + cX^2,$$

b and c are the least-squares estimates of the population parameters in

$$E(Y) = \alpha + \beta X + \gamma X^2.$$

ey are determined from the following set of normal equations:

$$\begin{aligned} \sum Y &= na + b\sum X + c\sum X^2 \\ \sum XY &= a\sum X + b\sum X^2 + c\sum X^3 \\ \sum X^2Y &= a\sum X^2 + b\sum X^3 + c\sum X^4. \end{aligned}$$

s quadratic equation may also be changed into a multiple linear form

$$\hat{X}_1 = a_{1.23} + b_{12.3}X_2 + b_{13.2}X_3,$$

$\hat{X}_1 = \hat{Y}$, $a_{1.23} = a$, $b_{12.3} = b_1$, $b_{13.2} = c$, $X_2 = X$ and $X_3 = X^2$. A number of other curvilinear are available. The co-efficient of determination and standard error of estimate can be obtained ame way as in the case of *linear* regressions.

EXERCISES

IVE

er 'True' or 'False'. If the statement is not true then replace the underlined words with words ake the statement true:

A partial correlation coefficient measures the degree of relationship between a variable and its stimate from the regression line.

ii) A <u>multiple correlation coefficient</u> measures the degree of linear relationship between variables in a multivariable problem when the influence with all other variables has removed.

iii) The multiple correlation coefficient is the <u>square</u> of the coefficient of multiple determin

iv) The multiple correlation coefficient R^2 will be <u>negative</u> in sign when all of the two correlation coefficients are negative in sign.

v) The <u>regression sum of squares</u> in case of multiple regression is the explained variation.

vi) For a multiple regression analysis, if $\sum(Y-\bar{Y})^2 = 50$ and $\sum(Y-\hat{Y})^2 = 20$, then the coefficient of determination R^2 is equal to <u>0.70</u>.

vii) The standard error of estimate in multiple regression has <u>n – k</u> degrees of freedom.

viii) The standard error of estimate is a measure of <u>scatter</u> of the observations about the line.

ix) The regression coefficients are the other name for <u>multiple regression coefficients</u>.

x) In a multiple regression the addition of new variables will <u>always reduce</u> the standa estimate.

b) MULTIPLE CHOICE QUESTIONS

i) The range of multiple correlation coefficient is

a) –1 to +1

b) 0 to ∞

[✓ c) 0 to 1]

d) none of above

ii) The range of partial correlation coefficient is

a) 0 to 1

[✓ b) –1 to +1]

c) 0 to +∞

d) –1 to 0

iii) If the multiple correlation coefficient $R_{3.12} = 1$, then it implies a

[✓ a) perfect relationship]

b) high relationship

c) weak linear relationship

d) perfect linear relationship

iv) In the regression analysis, the explained variation of the dependent variable Y is given by

a) $\Sigma(Y-\overline{Y})^2$

b) $\Sigma(Y-\hat{Y})^2$

[✓ c)] $\Sigma(\hat{Y}-\overline{Y})^2$

d) $\Sigma(Y-\hat{Y})$

v) Which of the following is not a standard deviation?

a) Standard error of the slope coefficient

[✓ b)] Mean square errors

c) Standard error of estimator

d) Standard deviation of the Y variable

vi) The coefficient of determination in multiple regression is given by

a) $R^2_{Y.13} = 1-(SST/SSE)$

b) $R^2_{Y.13} = 1-(SSR/SST)$

c) $R^2_{Y.13} = 1-(SSE/SSR)$

[✓ d)] $R^2_{Y.13} = 1-(SSE/SST)$

vii) The slope b_1 in the multiple regression equation $\hat{Y} = a + b_1X_1 + b_2X_2$ measures

a) the amount of variation in $\hat{Y}$ explained by X_1

b) the change in $\hat{Y}$ per unit change in X_1

[✓ c)] the change in $\hat{Y}$ per unit change in X_1, holding X_2 constant

d) the change in $\hat{Y}$ per unit change in X_2, holding X_1 constant

viii) The predicted value of Y for $X_1 = 1$, $X_2 = 5$, and $X_3 = 10$ by using the regression line

$$\hat{Y} = 30 - 10X_1 + 18X_2 - 7.5X_3 \text{ is}$$

a) 45

b) 15

[✓ c)] 35

d) 50

ix) Which of the following statements remains always true?

[✓ a)] The coefficient of multiple determination will increase when new variables are added

b) The coefficient of multiple determination will decrease when new variables are added

c) The adjusted coefficient of multiple determination will not decrease when new variables are added

d) Both a and c above

x) Which of the following relationship holds?

a) $r_{13.2} = \sqrt{b_{12.3} \times b_{21.3}}$

b) $r_{13.2} = \sqrt{b_{13.2} \times b_{31.2}}$

c) $r_{13.2} = \sqrt{b_{23.1} \times b_{32.1}}$

d) All of above

SUBJECTIVE

11.1 a) What is a multiple regression? Explain the basic differences between simple regression an multiple regression.

b) What is meant by the co-efficient of multiple determination and multiple correlation?

c) Explain the assumptions underlying a multiple linear regression model.

11.2 Carryout the necessary computations to obtain the least-squares estimates of the parameters in multiple regression model $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, given

Y	12	10	9	13	20
X_1	2	2	3	4	4
X_2	1	1	0	0	3

(B.Z.U., M.A. Econ. 19

11.3 Given the data

Y	2	5	7	8	5
X_1	8	8	6	5	3
X_2	0	1	1	3	4

a) Calculate the estimated regression equation, (*i.e* $\hat{Y} = a + b_1 X_1 + b_2 X_2$) for the above data

b) State the meaning of the partial regression co-efficients b_1 and b_2.

(B.Z.U., M.A. Econ. 19

11.4 Given the following data

X_1	1	4	1	3	2	4
X_2	1	8	3	5	6	10
X_3	2	8	1	7	4	6

a) Find the least-squares regression line where X_1 is the dependent variable and X_2 and independent variables

b) Calculate the standard error of estimate, $s_{1.23}$

c) Calculate the co-efficient of multiple determination and multiple correlation and interpr result.

The following table shows the corresponding values of three variables X_1, X_2 and X_3.

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

a) Find the regression equation of X_3 on X_1 and X_2.

b) Estimate X_3 when $X_1 = 10$and X_2=6.

c) Compute $R_{3.12}$ and $s_{3.12}$. (I.U., M.Sc. 1991)

The following data were collected to determine a suitable regression equation relating the length of an infant, Y(cm), to age, X_1 (days), and weight at birth, X_2 (kg):

Y	57.5	52.8	61.3	67.0	53.5	62.7	56.2	68.5	69.2
X_1	78	69	77	88	67	80	74	94	102
X_2	2.75	2.15	4.41	5.52	3.21	4.32	2.31	4.30	3.71

Fit a least-squares regression equation of the form

$$\hat{Y} = a + b_1 X_1 + b_2 X_2$$

Predict the average length of infants who are 75 days old and weighed 3.15 kg at birth.

Calculate the standard error of estimate $s_{Y.12}$.

Define the multiple correlation co-efficient and prove that

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{23}r_{31}}{1 - r^2_{23}}$$

Calculate the multiple correlation co-efficient $R_{1.23}$ of X_1 on X_2 and X_3 from the following data:

X_1	4	3	2	4	6	7	8
X_2	2	5	3	2	1	4	5
X_3	8	10	13	5	17	16	20

(P.U., B.A. (Hons.) Part-I, 1969)

The following data represent concomitant values of three variables.

X_1	32	18	52	16	42	48
X_2	3	2	5	1	4	6
X_3	2	4	2	5	3	9

Calculate all the multiple correlation coefficients, working out the usual simple correlation co-efficients. (B.Z.U., M.A. Econ. 1991)

b) Given $\overline{X}_1 = 20$, $S_1 = 1.0$, $r_{12} = -0.20$,

$$\overline{X}_2 = 36, \quad S_2 = 2.0, \quad r_{13} = 0.40,$$

$$\overline{X}_3 = 12, \quad S_3 = 1.5, \quad r_{23} = 0.50.$$

Find the regression equation of X_3 on X_1 and X_2. (P.U., B.A./B.Sc. 19

11.9 a) Distinguish between the simple and the multiple correlation co-efficients.

b) If b_{ij} is the regression co-efficient of X_i on X_j, then calculate the multiple correla co-efficient of X_2 with X_1 and X_3, where

$$b_{12} = 0.75, b_{13} = 0.58, b_{21} = 0.88,$$

$$b_{23} = 0.53, b_{31} = 1.68, \text{ and } b_{32} = 1.30$$ (P.U., B.A./B.Sc. 19

c) Three variables have in pairs simple correlation coefficients: $r_{12} = 0.60$; $r_{13} =$ $r_{23} = 0.65$. Find the multiple correlation coefficient $R_{2.13}$ of X_2 on X_1 and X_3. (P.U., B.A./B.Sc. 19

11.10 a) Three variable have in pairs simple correlation coefficients given by

$$r_{12} = 0.8, r_{13} = -0.7, r_{23} = -0.9.$$

Find the multiple correlation co-efficient $R_{1.23}$ of X_1 on X_2 and X_3. (P.U., B.A./B.Sc 19

b) Calculate the multiple correlation co-efficient $R_{2.13}$ and the partial correlation co-efficie from the values given below:

$$b_{12} = -0.1, b_{21} = -0.4, b_{13} = 0.27$$

$$b_{31} = 0.6, b_{23} = 0.67, b_{32} = 0.38$$ (P.U., B.A./B.Sc 19

11.11 a) Explain what is meant by partial correlation. Establish a formula for the co-efficient of correlation.

b) From the following data, determine the linear regression equations of X_1 on X_3 and X_2

X_1	5	9	7	10	12	8	6	10
X_2	10	12	8	9	11	7	5	8
X_3	2	6	4	5	7	6	4	6

Find the deviations of observed values of X_1 from the regression equation, viz. $X_{1.3}$ the same for X_2, *i.e.* obtain $X_{2.3}$. Determine the simple correlation co-efficient bet two sets of deviations $X_{1.3}$ and $X_{2.3}$.

11.12 The following means, standard deviations and correlations are found for

X_1 = Seed-hay crops in cwts. per acre,

X_2 = Spring rainfall in inches,

X_3 = Accumulated temperature above 42°F in spring in a certain district in England d years.

$\overline{X}_1 = 28.02, \quad S_1 = 4.42, \quad r_{12} = 0.80,$

$\overline{X}_2 = 4.91, \quad S_2 = 1.10, \quad r_{13} = -0.40,$

$\overline{X}_3 = 594, \quad S_3 = 85, \quad r_{23} = -0.56.$

Find the partial correlation and the regression equation for hay-crop on spring rainfall and accumulated temperature. (P.U., B.A./B.Sc. 1974)

11.13 The following values represent sample values of 450 college students in which the three variables represent marks obtained (X_1), general intelligence scores (X_2) and hours of study (X_3). Find the regression equation for estimating marks obtained. Find all three partial correlations and interpret them in the light of the corresponding simple correlations.

$\overline{X}_1 = 18.5, \quad S_1 = 11.2, \quad r_{12} = 0.60,$

$\overline{X}_2 = 100.6, \quad S_2 = 15.8, \quad r_{13} = 0.32,$

$\overline{X}_3 = 24, \quad S_3 = 6.0, \quad r_{23} = 0.35.$ (P.U., M.A. Stat., 1960)

11.14 a) Prove that a variable and a residual are uncorrelated if the subscript of the variable is included among the secondary subscripts of the residual.

b) Given the equations of the three regression planes as

$x_1 = 0.41\, x_2 + 0.23\, x_3,$

$x_2 = 0.96\, x_1 - 0.025\, x_3,$

$x_3 = 1.04\, x_1 - 0.05\, x_2.$

Calculate the partial correlation co-efficients. Do we have sufficient data to determine the correlation co-efficients r_{23}, r_{31} and r_{12}? (P.U., B.A. (Hons.) Part-I, 1970)

15 a) If $X_1 = a + b_{12.3}\, X_2 + b_{13.2}\, X_3$ and $X_3 = d + b_{32.1}\, X_2 + b_{31.2}\, X_1$ are the regression equations of X_1 on X_2 and X_3, and of X_3 on X_2 and X_1 respectively, prove that $r_{13.2}^2 = b_{13.2} \times b_{31.2}$.

b) Is it possible to obtain the following from a set of data?

(i) $r_{12} = 0.6, r_{23} = 0.8, r_{31} = -0.5.$

(ii) $r_{23} = 0.7, r_{13} = -0.4, r_{12} = 0.6.$

(iii) $r_{21} = 0.01, r_{13} = 0.66, r_{23} = -0.70.$

16 If X_1, X_2 and X_3 are three correlated variables, where S_1=1, S_2=1.3, S_3=1.9 and r_{12}=0.370, r_{13}= −0.641, and r_{23}= −0.736, find $r_{13.2}$. If X_4= X_1+ X_2, obtain r_{42}, r_{43} and $r_{43.2}$. Verify that the two partial correlation co-efficients are equal and explain this result. (M.Sc. Stat., P.U., 1972, I.U., 1990, 92, 94)

7 a) Differentiate between multiple correlation and partial correlation.

b) If $R_{1.23} = 1$, prove that (i) $R_{2.13} = 1$ and (ii) $R_{3.12} = 1$.

c) If $R_{1.23} = 0$, does it necessarily follow that $R_{2.13} = 0$?

d) If $r_{12} = r_{23} = r_{13} = r \neq 1$, then show that $R_{1.23} = R_{2.13} = R_{3.12} = \frac{r\sqrt{2}}{\sqrt{1+r}}$. Discuss the case when r=1. (B.Sc. Eng. 1976)

11.18 a) Show that if r_{12} is zero, $r_{12.3}$ will not be zero unless one at least of r_{13} and r_{23} is zero.

b) If the relation $aX_1 + bX_2 + cX_3 = 0$ holds true for all sets of values of X_1, X_2 and X_3, find out the three partial correlation co-efficients.

11.19 Show that the correlation co-efficient between the residuals $x_{1.23}$ and $x_{2.13}$ is equal and opposite to that between $x_{1.3}$ and $x_{2.3}$. (P.U., M.A. 1963)

Solution. The co-efficient of correlation between $x_{1.23}$ and $x_{2.13}$ is given by

$$\frac{Cov\ (x_{1.23},\ x_{2.13})}{\sqrt{Var\ (x_{1.23})\ Var(x_{2.13})}} = \frac{1}{n}.\frac{\sum x_{1.23}x_{2.13}}{S_{1.23}\,S_{2.13}}$$

$$= \frac{1}{n}.\frac{\sum x_{2.13}(x_1 - b_{12.3}x_2 - b_{13.2}x_3)}{S_{1.23}\,S_{2.13}}$$

$$= \frac{1}{n}.\frac{0 - b_{12.3}\sum x_{2.13}^2 - 0}{S_{1.23}S_{2.13}} = \frac{b_{12.3}S_{2.13}}{S_{1.23}}$$

Substituting the values of $S_{1.23}$ and $S_{2.13}$ and simplifying, we get

$$Corr. = -b_{12.3}\left\{\frac{S_2\sqrt{1-r_{23}^2}}{S_1\sqrt{1-r_{13}^2}}\right\} = -b_{12.3}\,\frac{S_{2.3}}{S_{1.3}}$$

Again the co-efficient of correlation between $x_{1.3}$ and $x_{2.3}$ is

$$\frac{Cov\ (x_{1.3},\ x_{2.3})}{\sqrt{Var\ (x_{1.3})\ Var(x_{2.3})}} = \frac{1}{n}.\frac{\sum x_{1.3}x_{2.3}}{S_{1.3}\,S_{2.3}} = b_{12.3}\,\frac{S_{2.3}}{S_{1.3}}$$

Hence the result.

11.20 Using the method of least-squares, fit a quadratic model $Y = \alpha + \beta_1 X + \beta_2 X^2 + \varepsilon$ to the follo data:

X	−2	−1	0	1	2
Y	0.4	1.3	2.2	2.5	3.0

Also calculate the standard error of estimate.

◆◆◆ ◆◆◆◆◆◆◆

Contents

CHAPTER 12

CURVE FITTING BY LEAST SQUARES

CURVE FITTING BY LEAST SQUARES

.1 INTRODUCTION

Let us suppose that we wish to *approximate* (describe) a certain type of function that best expresses association that exists between variables. A *scatter plot* of the set of values of the variables makes it ssible to visualize a smooth curve that effectively approximates the given data set. A more useful way represent this sort of approximating curve is by means of an equation or a formula. A term applied to process of determining the equation and/or estimating the parameters appearing in the equation of roximating curve, is commonly called *curve fitting*.

It is relevant to point out that the relationship between the variables may be *functional* or *ressional*. In functional relationship, a variable Y has a *true* value corresponding to each possible value nother variable X, *i.e* there is no question of random variation in the values of Y, and we make no abilistic assumptions in this respect. In this chapter, we shall limit our discussion to some functional tionships, *i.e.* problems of approximation and not of regression (already discussed earlier). Such tionships which are common in the natural sciences may be *linear* or *non-linear*.

2 APPROXIMATING CURVES AND THE PRINCIPLE OF LEAST SQUARES

The data sets encountered in practice greatly vary in nature. It is therefore necessary to decide h type of approximating curve and equation should be used. For this purpose, some of many mon types of approximating curves and their equations are given below:

ight line or linear curve,	$Y = a + bX$
abola of second degree or quadratic curve,	$Y = a + bX + cX^2$
abola of third degree or cubic curve,	$Y = a + bX + cX^2 + dX^3$
ponential curve,	$Y = ab^X$ or $Y = ae^{bX}$
metric or power curve,	$Y = aX^b$
erbola,	$\frac{1}{Y} = a + bX$.

o on.

these equations, Y is the *dependent* variable and X, the *independent* variable. In some situations, ver, the variables X and Y can be reversed.

We may approximate a given set of data by drawing a *free hand curve*, covering most of the points ed. But it is clear that different individuals would draw different curves according to their personal ent. Therefore this procedure of fitting a curve is not satisfactory.

The *principle of least squares* is applicable to curve fitting where the purpose is simply one of thing (or approximation) of a set of observations. Accordingly, we choose to determine the values of rameters in the equations of approximating curves so as to make the sum of squares of residuals a mum. A residual has been defined as the difference between the observed value and the ponding value of the approximating curve.

12.2.1 Fitting a Straight Line. A straight line is the simplest type of approximating curve and its n is written as

$$Y = a + bX$$

e the values of a and b are to be determined.

Given n pairs of observations $[(X_i, Y_i), i = 1, 2, \ldots, n]$ to which we wish to fit a straight line. We ne the values of a and b by the *principle of least squares*, which calls for the minimization of S, the sum of squares of the differences between the actual Y_i values and the corresponding values pro by $a + bX_i$. That is we minimize

$$S = \sum_{i=1}^{n} (Y_i - a - bX_i)^2$$

To do so, we need to solve the two equations $\frac{\partial S}{\partial a} = 0$ and $\frac{\partial S}{\partial b} = 0$.

That is $\frac{\partial S}{\partial a} = 2\sum(Y - a - bX)(-1) = 0$, and

$$\frac{\partial S}{\partial b} = 2\sum(Y - a - bX)(-X) = 0;$$

which on simplification become

$$\sum Y = na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2.$$

Solving these two normal equations simultaneously, we get

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} \text{ and } a = \overline{Y} - b\overline{X}.$$

The value of a indicates that the least squares line passes through the means of observation $(\overline{X}, \overline{Y}$

It should be noted that, when the origin is believed to lie on the curve, the straight line is simply $Y=bX$ and the sum of squared deviations to be minimized is

$$S = \sum(Y - bX)^2.$$

For a minimum value of S, $\frac{\partial S}{\partial b}$ must be zero, that is

$$\frac{\partial S}{\partial b} = 2\sum(Y - bX)(-X) = 0, \text{ which gives } \sum XY = b\sum X^2$$

as the *normal equation* and whence $b = \frac{\sum XY}{\sum X^2}$.

The sum of squares of residuals for a straight line is

$$S = \sum(Y - a - bX)^2$$

$$= \sum[Y(Y - a - bX)] = \sum Y^2 - a\sum Y - b\sum XY.$$

Example 12.1 Fit a straight line by the method of least squares to the following data:

X	1	2	3	4	5
Y	3	4	6	9	10

Also find the sum of squares of residuals. (P.U., B.A.

Let the equation of the straight line to be fitted to the data, be $Y=a+bX$, where a and b are to be valuated.

The normal equations for determining a and b are

$$\sum Y = na + b\sum X,$$

$$\sum XY = a\sum X + b\sum X^2$$

We now calculate $\sum X, \sum X^2, \sum Y$ and $\sum XY$ as below:

	X	Y	XY	X^2	Y^2
	1	3	3	1	9
	2	4	8	4	16
	3	6	18	9	36
	4	9	36	16	81
	5	10	50	25	100
Σ	15	32	115	55	242

Thus the normal equations become

$$5a + 15b = 32$$

$$15a + 55b = 115$$

Solving these two equations simultaneously, we obtain

$$a = 0.7 \text{ and } b = 1.9$$

Hence the equation of the required straight line is

$$Y = 0.7 + 1.9X$$

The sum of squares of residuals is given by

$$S = \sum(Y_i - a - bX_i)^2$$

$$= \sum[Y(Y - a - bX)] = \sum Y^2 - a\sum Y - b\sum XY$$

nce

$$S = 242 - 0.7\,(32) - 1.9\,(115)$$

$$= 242 - 240.9 = 1.1.$$

12.2.2 Fitting a Second Degree Parabola. The simplest type of a *non-linear* approximating curve second degree parabola that has the equation

$$Y = a + bX + cX^2$$

the values of a, b and c are to be determined.

Let us suppose that we wish to fit this parabolic curve to n pairs of observations $[(X_i, Y_i), i = 1, 2,$ Then we need to find those values of a, b and c which will minimize the sum of squares of ces between actual Y values and corresponding values obtained by $a+bX+cX^2$. (the principle of squares). That is we minimize

$$S = \sum(Y_i - a - bX_i - cX_i^2)^2$$

Minimizing S, we need to set its partial derivatives *w.r.t a*, b and c equal to zero. Thus

$$\frac{\partial S}{\partial a} = 2\sum(Y_i - a - bX_i - cX_i^2)(-1) = 0,$$

$$\frac{\partial S}{\partial b} = 2\sum(Y_i - a - bX_i - cX_i^2)(-X_i) = 0, \text{ and}$$

$$\frac{\partial S}{\partial c} = 2\sum(Y_i - a - bX_i - cX_i^2)(-X_i^2) = 0.$$

Simplifying, we get the following three normal equations

$$\sum Y = na + b\sum X + c\sum X^2$$

$$\sum XY = a\sum X + b\sum X^2 + c\sum X^3$$

$$\sum X^2Y = a\sum X^2 + b\sum X^3 + c\sum X^4$$

These equations are solved simultaneously to determine the values of a, b and c.

The sum of squares of residuals in case of second degree parabola is given by

$$S = \sum(Y - a - bX - cX^2)^2 = \sum[Y(Y - a - bX - cX^2)]$$

$$= \sum Y^2 - a\sum Y - b\sum XY - c\sum X^2Y$$

Example 12.2 Fit a second degree parabola to the following data, taking X as indepe variable.

X	0	1	2	3	4
Y	1	1.8	1.3	2.5	6.3

(P.U., B.A./B.Sc. 1961,

Let the equation of the second degree parabola be

$$Y = a + bX + cX^2$$

The normal equations are

$$\sum Y = na + b\sum X + c\sum X^2$$

$$\sum XY = a\sum X + b\sum X^2 + c\sum X^3$$

$$\sum X^2Y = a\sum X^2 + b\sum X^3 + c\sum X^4$$

The computations involved are shown in the following table:

	X	Y	XY	X^2	X^2Y	X^3	X^4
	0	1.0	0	0	0	0	0
	1	1.8	1.8	1	1.8	1	1
	2	1.3	2.6	4	5.2	8	16
	3	2.5	7.5	9	22.5	27	81
	4	6.3	25.2	16	100.8	64	256
Total	10	12.9	37.1	30	130.3	100	354

Putting these values in the normal equations, we get

$$12.9 = 5a + 10b + 30c$$

$$37.1 = 10a + 30b + 100c$$

$$130.3 = 30a + 100b + 354c$$

lving them as simultaneous equations in a, b and c, we obtain

$$a = 1.42,\ b = -1.07,\text{ and } c = 0.55.$$

nce the equation of the required second degree parabola is

$$Y = 1.42 - 1.07X + 0.55X^2$$

Example 12.3 Fit an equation of the form $Y = aX^2 + bX$ to the following data:

X	0	1	2	3	4	5
Y	1	5	12	20	25	36

Also find the sum of squares of residuals. (P.U., B.A./B.Sc. 1993)

The curve to be fitted is $Y = aX^2 + bX$.

The normal equations are

$$\sum X^2Y = a\sum X^4 + b\sum X^3 \text{ and } \sum XY = a\sum X^3 + b\sum X^2$$

The arithmetic is arranged in the table below:

X	Y	X^2	X^3	X^4	XY	X^2Y	Y^2
0	1	0	0	0	0	0	1
1	5	1	1	1	5	5	25
2	12	4	8	16	24	48	144
3	20	9	27	81	60	180	400
4	25	16	64	256	100	400	625
5	36	25	125	625	180	900	1296
15	99	55	225	979	369	1533	2491

ubstitution gives

$$979a + 225b = 1533$$
$$225a + 55b = 369$$

lving them simultaneously, we get

$$a = 0.4006 \text{ and } b = 5.0703.$$

nce the desired equation is $Y = 0.40X^2 + 5.07X$.

e sum of squares of residuals is given by

$$S = \sum(Y - aX^2 - bX)^2 = \sum[Y(Y - aX^2 - bX)]$$

$$= \sum Y^2 - a\sum X^2Y - b\sum XY$$
$$= 2491 - (0.40)(1533) - (5.07)(369)$$
$$= 2491 - 613.2 - 1870.83 = 6.97$$

12.2.3 Fitting of Higher Degree Parabolic Curves. A parabolic curve of degree p, approxim a set of observations $[(X_i, Y_i), i=1, 2, \ldots, n]$ has the equation

$$Y_i = a + bX_i + cX_i^2 + \ldots + kX_i^p$$

where $a, b, c, \ldots, k$ are the unknown quantities and where $k \neq 0$, and $n > p+1$..

The problem is to determine the $(p + 1)$ unknown quantities $a, b, c, \ldots, k$ in such a way th resulting values of Y_i should be as close as possible to the observed values. We, therefore, take the su squares of the residuals, *i.e.*

$$S = \sum_{i=1}^{n}(Y_i - a - bX_i - cX_i^2 - \ldots - kX_i^p)^2$$

which is a function of $a, b, c, \ldots, k$ as (X_i, Y_i) are certain numbers. The *principle of least-squares* ca the selection of that parabolic curve that minimizes S, the sum of squares of differences between th values of Y and the corresponding values calculated from the curve. To minimize S, we ca $\frac{\partial S}{\partial a}, \frac{\partial S}{\partial b}, \frac{\partial S}{\partial c}, \ldots, \frac{\partial S}{\partial k}$ and set them equal to zero. Simplification leads to the following $(p+1)$ equ

$$\sum Y = na + b\sum X + c\sum X^2 + \ldots + k\sum X^p$$
$$\sum XY = a\sum X + b\sum X^2 + c\sum X^3 + \ldots + k\sum X^{p+1}$$
$$\sum X^2Y = a\sum X^2 + b\sum X^3 + c\sum X^4 + \ldots + k\sum X^{p+2}$$
$$\vdots$$
$$\sum X^pY = a\sum X^p + b\sum X^{p+1} + c\sum X^{p+2} + \ldots + k\sum X^{2p}$$

These are the *normal equations* for fitting the parabolic curve of degree p. So simultaneously, we determine $a, b, c, \ldots, k$.

For the particular case, $p = 3$, the normal equations for fitting the cub $Y_i = a + bX_i + cX_i^2 + dX_i^3$ become

$$\sum Y = na + b\sum X + c\sum X^2 + d\sum X^3$$
$$\sum XY = a\sum X + b\sum X^2 + c\sum X^3 + d\sum X^4$$
$$\sum X^2Y = a\sum X^2 + b\sum X^3 + c\sum X^4 + d\sum X^5$$
$$\sum X^3Y = a\sum X^3 + b\sum X^4 + c\sum X^5 + d\sum X^6$$

Similarly, parabolic curves of higher degree may be fitted.

The sum of squares of residuals in case of cubic parabola is given by

$$S = \sum(Y - a - bX - cX^2 - dX^3)^2$$
$$= \sum Y^2 - a\sum Y - b\sum XY - c\sum X^2Y - d\sum X^3Y$$

A better fit. It is important to note that the sum of squares of residuals enables us to make some of comparison. A simple way of judging whether a straight line, a quadratic parabola or a cubic abola is likely to give the better fit, is to calculate the sum of squares of residuals in each case. The aller the sum of squares, the better is the fit.

12.2.4 Change of Origin and Unit. The computational labour may be reduced by a suitable ce of origin and unit. If the given values of X_i (i=1, 2, …, n) are equally spaced with a common val h and n is an *odd* number of values, say, n=2k+1, the normal equations are simplified by taking mid value of X as the origin and the common interval h as unit of measurement. That is, if X_0 be the alue, then $u_i=(X_i - X_0)/h$ takes the values $-k$, $-(k-1)$, …, -2, -1; 0, 1, 2, …, $(k-1)$, k. Hence we get $= 0 = \sum u^3 = \ldots$ If instead, n is an *even* number, say n=2k, we take the origin at the mean of the two values of X and h/2 as the new unit. The values of u_i then become $-(2k-1)$, $-(2k-3)$, …, -3, -1, 1, 3, 2k–3), (2k–1), so that $\sum u = 0 = \sum u^3 = \ldots$ (Also see chapter 13).

Example 12.4 The profits, £Y, of a certain company in the Xth year of its life are given by

X	1	2	3	4	5
Y	2500	2800	3300	3900	4600

Taking $u = X-3$ and $v = (Y-3300)/100$, find the parabolic curve of v on u in the form $v=a+bu+cu^2$ educe the curve of Y on X. (P.U., B.A./B.Sc. (Hons) 1964)

Since u=X–3 (given), so we find that the sums of odd powers of u are zero, *i.e.* $\sum u = 0 = \sum u^3$.

The normal equations are thus reduced to

$$\sum v = na + c\sum u^2,$$
$$\sum uv = b\sum u^2,$$
$$\sum u^2v = a\sum u^2 + c\sum u^4.$$

tions are computed in the following table.

X	u	Y	v	u^2	u^4	uv	u^2v
1	–2	2500	–8	4	16	16	–32
2	–1	2800	–5	1	1	5	–5
3	0	3300	0	0	0	0	0
4	1	3900	6	1	1	6	6
5	2	4600	13	4	16	26	52
Σ	0	---	6	10	34	53	21

Substituting these values in the normal equations, we get

$$5a + 10c = 9,$$
$$10b = 53,$$
$$10a + 34c = 21.$$

Solving them, we find $a = -0.086$, $b = 5.3$ and $c = 0.643$.

The equation of the required parabolic curve is therefore

$$v = -0.086 + 5.3u + 0.643u^2.$$

In order to deduce the parabolic curve of Y and X we replace u by $X-3$ and v by $\frac{Y-3300}{100}$ in the above relation. Thus we obtain

$$\frac{Y-3300}{100} = -0.086 + 5.3(X-3) + 0.643(X-3)^2.$$

Simplifying, we get

$$Y = 2280 + 144.2X + 64.3X^2.$$

as the required parabolic curve of Y on X.

12.3 EXPONENTIAL CURVES

Equations in which one of the variable quantities occurs as an exponent such as $Y=ac^{bX}$, are called *exponential equations* and graphs showing these equations as *exponential curves*. Exponential curves are used to describe a relation in which one variable forms approximately a geometric progression, while the other forms an arithmetic progression. Data of this hybrid type frequently occurs in the fields of business, banking and economics. In the equation $Y=ac^{bX}$, the letter c is a fixed constant, usually either 10 or e and a and b are determined from the data. If a and b are estimated by method of least-squares, we have to minimize S, where

$$S = \sum[Y_i - ae^{bX_i}]^2$$

Finding the partial derivatives with respect to a and b, and equating them to zero, we get

$$\frac{\partial S}{\partial a} = 2\sum[Y_i - ae^{bXi}][-e^{bXi}] = 0, \text{ and}$$

$$\frac{\partial S}{\partial b} = 2\sum[Y_i - ae^{bXi}][-ae^{bXi} \,.\, X_i] = 0.$$

Simplifying, we get

$$\sum Y_i e^{bXi} = a\sum e^{2bXi}, \text{ and}$$

$$\sum X_i Y_i e^{bXi} = a\sum X_i e^{2bXi}.$$

It is difficult to solve these equations as the solution requires tedious numerical methods. The solution simplifies if the *non-linear* curve may be reduced to the *linear* form by some suitable transformation of one or both the variables. The equation $Y=ae^{bX}$ can be *linearized* by taking the logarithms to the base 10, of both sides. Thus the exponential curve becomes.

$$\log Y = \log a + (b \log e)\, X$$

which may be written as

$$Y' = A + BX$$

Where Y' =log Y, A=log a and B=b log e. But this is the equation of a straight line in log Y and X. Hence the method of fitting an exponential curve to the observed set of data is to fit a straight line to the logarithms of the Y_is. It should be noted that it is the deviations of log Y, and not of Y, which are being minimized. It is relevant to point out that log form is better for calculating the values from the fitted curve.

We give some of the more common non-linear curves with suitable transformations to convert them into linear form Y=a+bX.

Non-linear Form	**Transformation**	**Linearized Form**
$Y = aX^b$	$Y' = \log Y, A = \log a,$ $X' = \log X$	$Y' = A + bX'$
$Y = ab^X$	$Y' = \log Y, A = \log a,$ $B = \log b$	$Y' = A + BX$
$Y = \frac{1}{a+bX}$ or $\frac{1}{Y} = a + bX$	$Y' = \frac{1}{Y}$	$Y' = a + bX$
$\frac{1}{Y} = a + \frac{b}{1+X}$	$Y' = \frac{1}{Y}, X' = \frac{1}{1+X}$	$Y' = a + bX'$
$Y = a + b\sqrt{X}$	$X' = \sqrt{X}$	$Y = a + bX'$
$Y = aX^2 + bX$	$Y' = \frac{Y}{X}$	$Y' = aX + b$

It is worth remaking that, if the variable Y incorporates an element of random variation, we introduce a random error term e and the equations become

$$Y = a + bX + e$$
$$Y = a + bX + cX^2 + e \quad \text{etc.}$$

which will be very similar to the regression models discussed in an earlier chapter.

Example 12.5 Fit an exponential curve $Y = ae^{bX}$ to the following data:

X	1	2	3	4	5	6
Y	1.6	4.5	13.8	40.2	125.0	363.0

(P.U., B.A./B.Sc. (Hons.), 1962; B.Z.U., 1976)

We can write the given equation as

$$\log Y = \log a + (b \log e) X$$

or $\quad Y' = A + BX.$ (From of a st. line)

where $\quad Y' = \log_{10} Y, A = \log_{10} a$ and $B = b \log_{10} e$.

As the equation is linear in $Y' = \log Y$ and X, therefore the two normal equations are

$$\sum Y' = nA + B\sum X$$

$$\sum XY' = A\sum X + B\sum X^2,$$

The necessary calculations are shown in the following table:

	X	Y	X^2	Y' (=log Y)	XY'
	1	1.6	1	0.2041	0.2041
	2	4.5	4	0.6532	1.3064
	3	13.8	9	1.1399	3.4197
	4	40.2	16	1.6042	6.4168
	5	125.0	25	2.0969	10.4845
	6	363.0	36	2.5599	15.3594
Total	21	----	91	8.2582	37.1909

Substituting these values, the normal equations become

$$6A + 21B = 8.2582$$

$$21A + 91B = 37.1909$$

Solving these equations simultaneously, we get

$A = -0.2805$, and $B = 0.4734$.

$\therefore$ $a = \text{anti-log } A = \text{anti-log } (-0.2805)$

$= \text{anti-log } \bar{1}.7195 = 0.52$

and $0.4343\ b = 0.4734$ or $b = 1.09$ $(\because \log_{10} e = 0.4343)$

Hence the equation of the curve fitted to the data is

$$Y = 0.52\,(e)^{1.09X}$$

Example 12.6 Fit an equation of the form $Y = aX^b$ to the following data:

X	1	2	3	4	5	6
Y	2.98	4.26	5.21	6.10	6.80	7.50

We may reduce the given equation to a linear form by taking logs to the base 10. Thus

$\log Y = \log a + b \log X$

or $Y' = A + bX'$

where $Y' = \log Y$, $A = \log a$ and $X' = \log X$.

As the equation is linear in $Y' = \log Y$ and $X' = \log X$, therefore the two normal equations are

$$\sum Y' = nA + b\sum X'$$

$$\sum X'Y' = A\sum X' + b\sum X'^2.$$

The following table contains the necessary calculations:

X	X' (=log X)	Y	Y' (=log Y)	$X'Y'$	X'^2
1	0	2.98	0.4742	0	0
2	0.3010	4.26	0.6294	0.189449	0.0906
3	0.4771	5.21	0.7168	0.341986	0.2276
4	0.6021	6.10	0.7853	0.472829	0.3625
5	0.6990	6.80	0.8325	0.581918	0.4886
6	.0.7782	7.50	0.8751	0.681003	0.6056
$\sum$	2.8574	---	4.3133	2.267185	1.7749

Substituting these summations, we get

$$6A + 2.8574b = 4.3133,$$

$$2.8574A + 1.7749b = 2.2672.$$

Solving them simultaneously and taking antilog of A, we get

$$a = 2.978 \text{ and } b = 0.5144$$

Hence the required equation is

$$Y = 2.978\,(X)^{0.5144}$$

$$= 3X^{1/2} \text{ approximately.}$$

OTHER TYPES OF CURVES

Some other types of curves frequently encountered in applied statistics are the following:

4.1 Modified Exponential Curve. A *modified exponential* curve, which is obtained by adding k to an exponential curve, is defined by the relation

$$Y = k + ab^X.$$

describes a set of data, the absolute growth of which decreases by a constant proportion when gative and "b" is less than one.

e first method to fit this curve is to transform it into a linear form by taking logarithms of both then to use the least-squares method. But this method is difficult for practical use. In the second e need three equations, because there are three constants k, a and b which are to be determined. ved data are therefore divided into three equal parts, leaving one or two values at the beginning, if necessary, to obtain the three equations, the criterion of fit being that the three partial totals of the trend values must equal those of the original data.

Let n denote the number of values in each third of the data.

Then the first equation is

$$\begin{aligned}\Sigma_1 Y &= nk + a + ab + ab^2 + ab^3 + \ldots + ab^{n-1}\\ &= nk + a[1 + b + b^2 + b^3 + \ldots + b^{n-1}]\\ &= nk + a\left[\frac{b^n - 1}{b-1}\right] \qquad (\because \frac{b^n-1}{b-1} = 1 + b + b^2 + \ldots + b^{n-1})\end{aligned}$$

In a similar way. the other two equations are obtained as

$$\Sigma_2 Y = nk + ab^n\left(\frac{b^n-1}{b-1}\right), \text{ and}$$

$$\Sigma_3 Y = nk + ab^{2n}\left(\frac{b^n-1}{b-1}\right).$$

Now we find the constant k, a and b.

Subtracting the first equation from the second one, we get

$$\Sigma_2 Y - \Sigma_1 Y = a\left(\frac{b^n-1}{b-1}\right)(b^n - 1) = a.\frac{(b^n-1)^2}{b-1}$$

Again, subtracting the second equation from the third one, we get

$$\Sigma_3 Y - \Sigma_2 Y = ab^n\frac{(b^n-1)^2}{b-1}.$$

Dividing, we have

$$\frac{\Sigma_3 Y - \Sigma_2 Y}{\Sigma_2 Y - \Sigma_1 Y} = \left[ab^n.\frac{(b^n-1)^2}{b-1}\right] \div \left[a\frac{(b^n-1)^2}{b-1}\right] = b^n$$

which gives $$b = \sqrt[n]{\frac{\Sigma_3 Y - \Sigma_2 Y}{\Sigma_2 Y - \Sigma_1 Y}}.$$

Finally, $$a = (\Sigma_2 Y - \Sigma_1 Y)\frac{b-1}{(b^n-1)^2}, \text{ and}$$

$$k = \frac{1}{n}\left[\Sigma_1 Y - \left(\frac{b^n-1}{b-1}\right)a\right]$$

12.4.2 The Compertz Curve, named after Benjamin Gompertz, is given by the equation

$$Y = ka^{b^X},$$

where k, a and b are constants. The equation is changed to *modified exponential equation* by taking logarithms of both sides. Thus

$$\log Y = \log k + b^X \log a$$

or $$Y' = k' + a' b^X$$

where $Y' = \log Y$, $k' = \log k$ and $a' = \log a$.

The Compertz curve, which increases first at an increasing rate, then increases at a decreasing rate until it reaches a maximum level, is frequently used in business and actuarial work.

12.4.3 The Logistic Curve, which is widely used to represent growth, is defined by the relation

$$Y = \frac{k}{1+bc^X},$$

r inverting, we get $$\frac{1}{Y} = \frac{1}{k} + \frac{b}{k}c^X$$

$$= k' + ac^X,$$

here $k' = \frac{1}{k}$ and $a = \frac{b}{k}$. This is similar in form to the *modified* exponential curve if $\frac{1}{Y}$ is expressed as function of X and the same method of fitting may therefore be applied with the reciprocals $\frac{1}{Y_i}$ instead f Y_i. The use of this curve to analyse population and biological growth was advocated by Raymond Pearl nd L.J. Reed. It should be noted that the *logistic curve* has four different stages, viz., (i) a period of latively slow growth, (ii) then a period of accelerated growth (iii) then a period of decelerated growth nd (iv) finally a period of stability, when the curve does not go up at all. The growth of human pulation and that of economic variables are appropriately described by the curve as they conform to ese stages.

12.4.4 The Makeham Curve is defined as

$$Y = ks^X b^{c^X}.$$

r in the logarithmic form;

$$\log Y = \log k + X \log s + c^X \log b$$

$$= A + CX + Bc^X$$

ere $A = \log k$, $C = \log s$ and $B = \log b$.

This type of curve, which is actually a combination of a straight line with a Compertz curve, is ed in actuarial and insurance work.

12.5 CRITERIA FOR A SUITABLE CURVE

Frequently, we are required to choose a suitable form of curve to obtain a reasonable fit to the observed sets of data in two variables. The suitability of several curves may be determined by examining the *differences* in the values of the dependent variable *Y*. The first difference, denoted by ΔY read as *delta Y*) is defined by $\Delta Y_i = Y_{i+1} - Y_i$, the second difference is defined by $\Delta^2 Y_i = \Delta Y_{i+1} - \Delta Y_i$, and so on. A straight line has the property that its first difference is equal to b (a constant), a second degree parabola has the property that its second difference is equal to $2c$ (a constant) and, in general, a parabolic curve of the *nth* degree has the property that its *nth* differences are constant. Thus we fit

- i) a straight line, if the first differences between successive values are approximately constant;
- ii) a second degree parabola, if the second differences are approximately constant;
- iii) a third degree parabola, if the third differences prove to be constant;
- iv) an exponential curve, if the first differences of the logarithms are approximately constant;
- v) a log parabola $Y = ab^X c^{X^2}$, if the second differences of the logarithms of the *Y*-values tend to be constant;
- vi) a modified exponential curve, if each first difference is a constant percentage of the preceding first difference;
- vii) a Gompertz curve, if the first differences of logarithms are changing by a constant percentage.
- viii) a logistic curve, if the first differences of the reciprocals are changing by a constant percentage; and
- ix) a reciprocal line $\frac{1}{Y} = a + bX$, if the reciprocals of the data show a straight line when plotted on a graph.

12.6 FINDING PLAUSIBLE VALUES BY THE PRINCIPLE OF LEAST-SQUARES

The *principle of least squares* can also be applied to find the most satisfactory values of the unknown quantities from a number of independent liner equations in the unknowns when the number of equations is greater than the number of unknowns.

Suppose there are *k* unknown quantities $X_1, X_2, \ldots, X_k$ and let the *n* observed relations where $n > k$ be

$$\begin{array}{ccc} a_1 X_1 + b_1 X_2 + \ldots + f_1 X_k = l_1 \\ a_2 X_1 + b_2 X_2 + \ldots + f_2 X_k = l_2 \\ \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \\ a_n X_1 + b_n X_2 + \ldots + f_n X_k = l_n \end{array}$$

where *a's*, *b's*, …, *l's* are constants.

When $n>k$, *i.e.* the number of equations is greater than the number of unknowns, there may not exist a unique solution. In such cases, we therefore try to find those values of $X_1, X_2, \ldots, X_k$ which will simultaneously satisfy the given set of independent linear equations as nearly as possible. Such values are obtained by the least-squares method and are called the *best* or *most plausible* values.

The least-squares criterion calls for the selection of those values of $X_1, X_2, \ldots, X_k$ which make the sum of squares of the discrepancies D_i's, also called *errors* or *residuals*, a minimum, where

$$D_i = a_iX_1 + b_iX_2 + \ldots + f_iX_k - l_i, \qquad i = 1, 2, \ldots, n$$

In other words, we have to select those values of $X_1, X_2, \ldots, X_k$ which minimize

$$S = \sum_{i=1}^{n} D_i^2 = \sum_{i=1}^{n} (a_iX_1 + b_iX_2 + \ldots + f_iX_k - l_i)^2$$

It is obvious that $S = f(X_1, X_2, \ldots, X_k)$, that is, the sum of squares of residuals is some function of $X_1, X_2, \ldots, X_k$. If S is to have a minimum value, it is necessary that its partial derivatives with respect to $X_1, X_2, \ldots, X_k$, if they exist, vanish there; hence $X_1, X_2, \ldots, X_k$ must satisfy the equations

$$\begin{aligned}
\frac{\partial S}{\partial X_1} &= 2\sum a_i(a_iX_1 + b_iX_2 + \ldots + f_iX_k - l_i) = 0 \\
\frac{\partial S}{\partial X_2} &= 2\sum b_i(a_iX_1 + b_iX_2 + \ldots + f_iX_k - l_i) = 0 \\
&\vdots \\
\frac{\partial S}{\partial X_k} &= 2\sum f_i(a_iX_1 + b_iX_2 + \ldots + f_iX_k - l_i) = 0
\end{aligned}$$

The equations given above may be written in the standard form as

$$\begin{aligned}
X_1\sum a_i^2 + X_2\sum a_ib_i + \ldots + X_k\sum a_if_i &= \sum a_il_i \\
X_1\sum a_ib_i + X_2\sum b_i^2 + \ldots + X_k\sum b_if_i &= \sum b_il_i \\
&\vdots \\
X_1\sum a_if_i + X_2\sum b_if_i + \ldots + X_k\sum f_i^2 &= \sum f_il_i
\end{aligned}$$

ese simultaneous equations obtained by minimizing process, are the normal equations which are ultaneously solved to obtain the best or the *most plausible values* of $X_1, X_2, \ldots, X_k$.

It should be noted that the normal equations for a set of variables are obtained by multiplying each ation by the co-efficient of the respective variable in the equations and adding them together. This is a venient way for remembering the normal equations.

Example 12.7 Apply the principle of least-squares to solve

$$2X + Y = 0,\ 3X - 2Y = 0,\ -X + Y = -2. \qquad \text{(P.U., B.A./B.Sc. 1971, 75)}$$

There are 3 linear equations and 2 unknown variables X and Y, therefore we apply the least-squares hod to get the most plausible values of X and Y.

Now $\quad S = (2X + Y - 0)^2 + (3X - 2Y - 0)^2 + (-X + Y + 2)^2$

The normal equations are $\frac{\partial S}{\partial X} = 0$ and $\frac{\partial S}{\partial Y} = 0$,

i.e. $2(2X + Y) + 3(3X - 2Y) - (-X + Y + 2) = 0$

and $(2X + Y) - 2(3X - 2Y) + (-X + Y + 2) = 0$

or $14X - 5Y = 2$ and $-5X + 6Y = -2$.

Solving these equations simultaneously, we get

$X = 0.034$, and $Y = -0.305$.

Example 12.8 Find the most plausible values of X and Y from the following equations:

$$X - Y - 3 = 0$$
$$3X + 2Y - 4 = 0$$
$$2X - 3Y + 1 = 0$$

(P.U., B.A. (Hons.) Part-I, 1963, B.A./B.Sc. 19

We first find the normal equation for X. Multiplying each equation by the co-efficient of X in it, we have

$$X - Y = 3$$
$$9X + 6Y = 12$$
$$4X - 6Y = -2$$

Adding, we get $14X - Y = 13$, which is the normal equation for X.

We then find the normal equation for Y. Again multiplying each equation by the co-efficient of it, we get

$$-X + Y = -3$$
$$6X + 4Y = 8$$
$$-6X + 9Y = 3$$

Adding them together, we have $-X + 14Y = 8$ as the normal equation for Y.

Thus the two normal equations are

$$14X - Y = 13$$
$$-X + 14Y = 8$$

Solving them simultaneously, we obtain

$X = 0.97$ and $Y = 0.64$

which is the required solution.

EXERCISES

12.1 a) What is meant by Curve Fitting? (P.U., B.A./B.Sc.

b) Explain the principle of Least Squares with particular reference to a straight line fit. sense, does it give the "best" solution? (P.U., B.A./B.Sc.

c) Fit a straight line to the following data:

X	1	2	3	4	5	6
Y	2	6	7	8	10	11

Calculate the values of Y for each value of X, obtain the values of residuals e_i's and check that $\sum e_i = 0$.

2.2 a) By means of Least Squares, show how a straight line can be fitted to a set of given observations, and obtain the normal equations.

b) Prove that a least squares line always passes through the point ($\overline{X}$, $\overline{Y}$).

(P.U., B.A./B.Sc. 1978)

c) Fit a straight line to the following data and plot on the graph paper the actual and calculated values.

X	0	1	2	3	4	5	6	7	8
Y	5	11	8	14	10	16	2	20	15

3 a) Write down the equation of a straight line through the origin and derive an expression for finding its slope by the principle of least squares. (P.U., B.A./B.Sc. 1991)

b) Fit a least-squares line to the following data:

Year (X)	1	2	3	4	5	6	7	8	9
Output (Y)	1	3	2	4	3	5	4	6	5

Measure the deviations from the fitted line and find the sum of squared deviations.

4 a) Find the normal equations which determine the values of a and b in least squares line $Y = a + bX$; and show that the sum of squares of residuals from the least squares line is given by

$$S = \sum Y^2 - a\sum Y - b\sum XY$$

b) Fit a straight line to the following data:

X	0	5	10	15	20	25
Y	12	15	17	22	24	30

(P.U., B.A./B.Sc. 1962; 80)

5 a) Fit the least squares line for 20 pairs of observations having $\overline{X} = 2$, $\overline{Y} = 8$, $\sum X^2 = 180$ and $\sum XY = 404$. (P.U., B.A./B.Sc. 1986)

b) Given

X	1	2	3	4	5
Y	8	9	13	18	27

Fit $Y=a+bX$ by least-squares and estimate Y for $X=6$. Also fit $X=c+dY$ and use this equation to estimate Y for $X=6$. Account for the difference in two estimates.

(P.U., B.A. (Hons.) Part-II, 1963-S)

12.6 a) Find the normal equations for *a*, *b* and *c* that will minimize

$$S = \sum[Y - (a + bX + cX^2)]^2$$

b) Show that the sum of squares of residuals for a second degree parabola is

$$S = \sum Y^2 - a\sum Y - b\sum XY - c\sum X^2 Y$$

c) Fit a parabola of the form $Y=a+bX+cX^2$ to the data:

X	–2	–1	0	1	2
Y	–5	–2	1	2	1

12.7 a) By means of the principle of least squares, show how a parabola of second order can fitted to a set of *n* observation (X_i, Y_i) and obtain the normal equations.

b) For 5 pairs of observations, it is given that A.M. of *X* series is 2 and A.M. of *Y* series is is also known that

$$\sum X^2 = 30, \sum X^3 = 100, \sum X^4 = 354, \sum XY = 242, \sum X^2Y = 850$$

Fit a second degree parabola, taking X as the independent variable.

(P.U., B.A./B.Sc.

c) Fit a second degree parabola to the following data:

X	0	1	2	3	4
Y	1	5	10	22	38

(P.U., B.A. (Part-I),

12.8 Fit a second degree parabola to the following seven pairs of values:

X	1.0	1.5	2.0	2.5	3.0	3.5	4.0
Y	1.1	1.3	1.6	2.0	2.7	3.4	4.1

(P.U., B.A./B.Sc.

12.9 Fit a second degree equation to the following data:

X	8	12	16	20	24	28	32	36	40
Y	2.4	4.8	8.3	9.5	11.2	24.3	22.2	21.2	25.4

(P.U., B.A. (Hons.) Part-

12.10 The profits, £*Y*, of a certain company in the *X*th year of its life are given by:

X	1	2	3	4	5
Y	1250	1400	1650	1950	2300

Taking $u = X - 3$, $v = (Y - 1650)/50$, show that the parabolic curve of v on u is

$$v + 0.086 = 5.30\,u + 0.643\,u^2,$$

and deduce that the parabolic curve of *Y* on *X* is

$$Y = 1140 + 72.14X + 32.14X^2.$$

(P.U., B.A./B.S

2.11 Fit, by the method of least-squares,

i) the straight line of best fit,

ii) the 2nd degree parabola of best fit, to the following data:

X	20	25	30	35	40	45	50	55
Y	240	315	403	450	488	520	525	532

Also calculate the sum of squares of residuals in the two cases.

12 Fit a straight line and parabolas of the second and third degrees to the following data, taking X to be the independent variable;

X	0	1	2	3	4
Y	1	1.8	1.3	2.5	6.3

and calculate the sum of squares of residuals in the three cases.

3 a) You are given data in two variables X and Y and you have to take a decision about fitting a suitable trend. How will you proceed?

(P.U., B.A./B.Sc. 1987)

b) Given the following pairs of values of X and Y.

X	0	1	2	3	4
Y	10	17	28	43	62

Fit a suitable curve. (P.U., B.A./B.Sc. 1976)

4 a) Explain the principle of least squares and use it to obtain the normal equations when a cubic parabola is fitted to n pairs of observations.

b) Fit a curve of the form $Y=ab^X$ to the following data in which Y represents the number of bacteria per unit volume existing in a culture at the end of X hours:

X	0	1	2	3	4
Y	73	91	112	131	162

Estimate the value of Y when X=5 and 6.

(P.C.S. 1972; P.U., B.A./B.Sc. 1978)

The number (Y) of bacteria per unit volume present in a culture after X hours is given in the following table:

No. of hours (X)	0	1	2	3	4	5	6
No. of bacteria per unit volume (Y)	32	47	65	92	132	190	275

Fit a least-squares curve having the form $Y=ab^X$ to the data. Estimate the value of Y when X=7.

(P.U., B.A./B.Sc. 1969, 79, 80)

12.16 Fit a simple exponential to the following data for a growing plant by taking the logarithms of exponential equation.

Day	0	1	2	3	4	5	6	7	8
Height	0.75	1.20	1.75	2.50	3.45	4.70	6.20	8.25	11.50

12.17 Fit a curve of the type $Y=ab^X$ to the following data:

X	1	2	3	4	5	6	7
Y	10	12.2	14.5	17.3	21.0	25.0	29.0

12.18 The following data represent the enrolments at a small liberal arts college during the past years:

X (years)	1	2	3	4	5	6	7
Y (enrolments)	304	341	393	457	548	670	882

Use the method of least-squares to estimate a curve of the form $Y=ab^X$ and predict the enrolm years from now. (P.U., B.A./B.Sc.

12.19 a) Given n=8, $\sum X=16$, $\sum X^2=204$, $\sum X^3=582$, $\sum \log Y$=23, $\sum X \log Y$=104. Fit a curve. (P.U., B.A./B.Sc. Hons.

b) Fit a curve of form $Y=a+b\sqrt{X}$ to the following data:

X	1.20	2.50	3.40	4.70	5.30
Y	6.33	8.03	8.95	10.09	10.56

12.20 Fit a curve $Y=aX^b$ to the following data:

X	1	2	3	4	5	6
Y	1200	900	600	200	110	50

12.21 Fit a curve of the form $Y=aX^b$ to the following data on the unit cost in dollars of producing electronic components and the number of units produced.

Lot size (X)	50	100	250	500	1000
Unit cost (Y)	108	53	24	9	5

Use the result to estimate the unit cost for a lot of 400 components.

(P.U., B.A./B.Sc.

12:22 It is thought that two physical quantities X and Y should be connected by a relation of the $Y=aX^n$. The experimental values are:

X	0.5	1.5	2.5	5.0	10.0
Y	3.4	7.0	12.8	29.8	68.2

Find the best values of a and n. (P.U., B.A./B.Sc.

2.23 The discharge of a capacitor through a resistance gave he following results:

t (seconds)	0.5	0.8	1.4	2.0	2.5
v (volts)	9.1	8.5	7.5	6.7	6.1

Fit a curve of the type $v = ae^{bt}$ to these data.

2.24 a) Fit a curve of the type $Y = ae^{bX}$ to the following data:

X	1	2	3	4	5
Y	27	73	200	545	1484

where $e = \lim\limits_{n\to\infty}\left(1+\dfrac{1}{n}\right)^n$.

b) Obtain the values of Y from the approximating line for various values of X. Do the deviations of the observed values of Y from the corresponding calculated values add to zero? Explain your result.

(P.U., B.A./B.Sc. 1977)

25 Estimate the constant of Pareto Curve, $n = AX^{-a}$, which fits the data below:

Income (£X)	Number (n)
150	14,000,000
500	825,000
1,000	173,000
2,000	35,500

26 The pressure (p) of a gas and its volume (v) are known to be related by an equation of the form pv^{γ} = constant. From the following data, find the value of γ by fitting a straight line to the logarithms of p and v, taking p to be the independent variable.

p (kg. per sq. cm)	0.5	1.0	1.5	2.0	2.5	3.0
v (litres)	1.62	1.00	0.75	0.62	0.52	0.46

a) Derive the *least-squares* equations for fitting a curve of the type $\dfrac{1}{Y} = a + bX$ to a set of n observations. Also find the values of a and b.

b) Fit a reciprocal curve $\dfrac{1}{Y} = a + bX$ to the following data:

X	0	1	4	6	12	16
Y	10	8	5	4	2.5	2

a) Find the normal equations for determining a, b and c from the linear equation $Y = a + bX_1 + cX_2$.

b) Find the least-squares fit $Y = a + bX_1 + cX_2$, given

Y	2	5	7	8	5
X_1	8	8	6	5	3
X_2	0	1	1	3	4

12.29 a) What is the modified exponential curve? Describe the method of fitting it.

(P.U., B.A. (Hons.) Part-II, 196

b) Derive the least-squares equations for fitting a modified exponential, $Y = c + ae^{bX}$ to a set
n observations, and indicate why these equations would be difficult to solve.

12.30 Write a critical note on the law of growth as portrayed by the logistic curve and the Gomp
curve. (P.U., B.A./B.Sc. 196

12.31 Use the "principle of least-squares" to find the normal equations when the number of equations
greater than the number of unknown quantities.

(P.U., B.A./B.Sc. 1981, 84, 86,

12.32 a) Explain the method of *least-squares*. Apply it to solve the equations

$X + 7Y = 17, 2X - Y = 0, 3X - 2Y = -1$ (P.U., B.A./B.Sc. 19

b) Find the most plausible values of X and Y from the following equations. Also compute
sum of squares of residuals.

$2X + Y = 4.8,$ $-X + 3Y = 6.3$

$3X - 2Y = -2.1,$ $3X + 2Y = 8.0,$ (P.U., B.A./B.Sc. 1981,

12.33 a) Find the most plausible values of X and Y from the following equations:

$X + Y = 3.01,$ $2X - Y = 0.03$

$X + 3Y = 7.02,$ $3X + Y = 4.97$

b) Obtain the best possible values of X and Y from

$2X + Y = 4,$ $3X - Y = 10.02,$

$X + 2Y = 5.02,$ $3X + 2Y = 0.97.$ (P.U., B.A./B.Sc.

12.34 Form normal equations and solve

$X + 2Y + Z = 1,$ $2X + Y + Z = 4,$

$-X + Y + 2Z = 3,$ $4X + 2Y - 5Z = -7.$ (P.U., B.A./B.Sc. 1962,

12.35 Find the most plausible values of X, Y and Z from the following equations:

$X - Y + 2Z = 3,$ $3X + 2Y - 5Z = 5,$

$4X + Y + 4Z = 21,$ $-X + 3Y + 3Z = 14.$

(P.U., B.A./B.Sc

◆ ◆ ◆