MULTIPLE REGRESSION ANALYSIS

Multiple Regression and Correlation Analysis

Multiple regression analysis represents 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 concept in the analysis remains the same. The following are the three main objectives of multiple regression and correlation analysis:

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

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

(3) 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 a standard error of estimate. The third purpose is accomplished by computing the multiple coefficient of determination.

Multiple regression equation. The multiple regression equation describes the average relationship between these variables and this relationship is used to predict or control the dependent variable.

A *regression equation* is an equation for estimating a dependent variable, say, X_1 from the independent variables X_2, X_3, ... and is called a regression equation of X_1 on X_2, X_3, ... In functional notation, this is sometimes written briefly as $X_1 = F(X_2, X_3 .....)$ read as "X_1 is a function of X_2, X_3 and so on".

In case of three variables, the regression equation of X_1 on X_2 and X_3 has the form

$$X_{1\cdot 23} = a_{1\cdot 23} + b_{12\cdot 3} X_2 + b_{13\cdot 2} X_3 \quad ...(i)$$

$X_{1\cdot 23}$ is the computed or estimated value of the dependent variable and X_2, X_3 are the independent variables.

The constant $a_{1\cdot 23}$ is the intercept made by the regression plane. It gives the value of the dependent variable when all the independent variables assume a value equal to zero. $b_{12\cdot 3}$ and $b_{13\cdot 2}$ are called partial regression coefficients or the net regression coefficients. $b_{12\cdot 3}$ measures the amount by which a unit change in X_2 is expected to affect X_1 when X_3 is held constant and $b_{13\cdot 2}$ measures the amount of change in X_1 per unit change in X_3 when X_2 is held constant.

Due to the fact the X_1 varies partially because of variation in X_2 and partially because of variation in X_3, we call $b_{12\cdot 2}$ and $b_{13\cdot 2}$ the *partial regression coefficients* of X_1 on X_2 keeping X_3 constant and of X_1 on X_3 keeping X_2 constant.

Normal Equations for the Least Square Regression Plane

Just as there exist least square regression lines 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 scatter diagram.

The least square regression plane of X_1 on X_2 has the equation (i), where $b_{12\cdot 3}$ and $b_{13\cdot 2}$ are determined by solving simultaneously the *normal equations*

$$\begin{cases} \Sigma X_1 = N a_{1\cdot 23} + b_{12\cdot 3} \Sigma X_2 + b_{13\cdot 2} \Sigma X_3 \\ \Sigma X_1 X_2 = a_{1\cdot 23} \Sigma X_2 + b_{12\cdot 3} \Sigma X_2^2 + b_{13\cdot 2} \Sigma X_2 X_3 \\ \Sigma X_1 X_3 = a_{1\cdot 23} \Sigma X_3 + b_{12\cdot 3} \Sigma X_2 X_3 + b_{13\cdot 2} \Sigma X_2^2 \end{cases}$$

These equations can be obtained 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 of solving simultaneous equations require a knowledge of matrix algebra, which is not assumed for the reader of this text. Computer programmes are widely available for determining the variables of the constants in a multiple regression equation. In our discussion that follows, we shall confine ourselves to the two independent variables case which, of course, can be extended to cover case with three or more independent variables.

Assumptions of Linear Multiple Regression Analysis

For point estimation, the principle assumptions* of linear multiple regression analysis are:

(1) The dependent variable is a random variable whereas the independent variable need not be random variable.

(2) The relationship between the several independent variables and the one dependent variable is linear, and

(3) The variances of the conditional distributions of the dependent variable, given various combinations of values of the independent variables, are all equal. For internal estimation, an additional assumption is that the conditional distributions for the dependent variable follow the normal probability distri-bution.

Deviations taken from Actual Means: The work involved in finding these regression equations is reduced by taking deviations from the means of the variables under consideration. The regression equation for three variables then becomes:

$$x_1 = b_{12\cdot3}\, x_2 + b_{13\cdot2}\, x_3$$

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

The value of $b_{12\cdot3}$ and $b_{13\cdot2}$ can be obtained by solving simultaneously the following two normal equations:

$$\Sigma\, x_1\, x_2 = b_{12\cdot3}\, \Sigma\, x_2^2 + b_{13\cdot2}\, \Sigma\, x_2\, x_3$$

$$\Sigma\, x_1\, x_2 = b_{12\cdot3}\, \Sigma\, x_2\, x_3 + b_{13\cdot2}\, \Sigma\, x_3$$

The value of $b_{12\cdot3}$ and $b_{13\cdot2}$ can also be obtained as follows:

$$b_{12\cdot3} = r_{12\cdot3} \times \frac{\sigma_{1\cdot23}}{\sigma_{2\cdot13}}$$

$$b_{13\cdot2} = r_{13\cdot2} \times \frac{\sigma_{13\cdot2}}{\sigma_{3\cdot12}}$$

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

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

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

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

This method of obtaining regression equations is much simpler as 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:

---

* Leonard J. Kazmier, *Business Statistics*, p. 313.

$\bar{X}_1$ $\bar{X}_2$ $\bar{X}_3$

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, one 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:

$$X_2 = a_{2\cdot13} + b_{21\cdot3}\, X_1 + b_{23\cdot1}\, X_3 \quad \text{...(ii)}$$

$$X_3 = a_{3\cdot12} + b_{31\cdot2}\, X_1 + b_{32\cdot1}\, X_2 \quad \text{...(iii)}$$

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

$$\Sigma X_2 = N a_{2\cdot13} + b_{21\cdot3} \Sigma X_1 + b_{23\cdot1} \Sigma X_3$$

$$\Sigma X_1 X_2 = a_{2\cdot13} \Sigma X_1 + b_{21\cdot3} \Sigma X_1^2 + b_{23\cdot1} \Sigma X_1 X_2$$

$$\Sigma X_2 X_3 = a_{2\cdot13} \Sigma X_3 + b_{21\cdot3} \Sigma X_1 X_3 + b_{23\cdot1} \Sigma X_3^2$$

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

$$\Sigma X_3 = N a_{3\cdot12} + b_{31\cdot2} \Sigma X_1 + b_{32\cdot1} \Sigma X_2$$

$$\Sigma X_1 X_3 = a_{3\cdot12} \Sigma X_1 + b_{31\cdot2} \Sigma X_1^2 + b_{32\cdot1} \Sigma X_1 X_2$$

$$\Sigma X_2 X_3 = a_{3\cdot12} \Sigma X_3 + b_{31\cdot3} \Sigma X_1 X_2 + b_{32\cdot1} \Sigma X_2^2$$

Generalizations 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

$$X_1 = a_{1\cdot234} + b_{12\cdot34}\, X_2 + b_{13\cdot24}\, X_3 + b_{14\cdot23}\, X_4$$

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 $a_{1\cdot234}$, $b_{13\cdot24}$, $b_{12\cdot34}$ and $b_{14\cdot23}$ which when substituted to above equation, gives the least square regression equation of X_1 on X_2, X_3 and X_4.

While drawing statistical inference in multiple regression, it should be noted that the regression coefficients for highly intercorrelated independent variables tend to be unreliable. This is for the reason that when independent variables are highly intercorrelated, it is extremely difficult to separate out the individual influences of each variable. There is a great deal of concern in fields such as econometrics and applied statistics with this problem of intercorrelation among dependent variables, often referred to as *multicollinerarity*. As suggested by Morris Hamburg, one of the simplest solutions to the problem of two highly correlated independent variables is merely to discard one of the variables.

Use of Computers in Multiple Regression and Correlation Analysis

The application of multiple regression and correlation analysis requires

extensive and highly precise computations. As the number of variables increases, the computations become more and more difficult and time consuming. The computers are being used widely in the application of these techniques. A large number of computer installations have one or more multiple regression and correlation programmes in the programme library that are available to users. In fact, it is becoming increasingly unnecessary nowadays to carry out regression analysis by hand.

The availability of these programmes enables many analysts to obtain the desired regression and correlation result without the analyst having to spend time in writing a computer programme. The suitability of a given library programme for use in a particular problem depends upon the input requirements, operating procedure and results computed by the programme. Many library programmes are sufficiently general and comprehensive to fulfil the requirements of a wide variety of users.

It may be pointed out that though with the use of computers it is possible to test and include large number of independent variables in a regression analysis, good judgment and knowledge of the logical relationships involved must always be used as a guide to deciding which variables to include in the construction of a regression equation.

Illustration 8. Find the 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	3
X_3	30	24	20	14	10	4

(MBA, S.V. Univ., 1995; M.Com., D.U., 1997)

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 the regression equation will deviate from the correct value of 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 these deviations of the actual value of non-dependent variable from the estimate 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

$$S_{1\cdot 23} = \sqrt{\frac{\Sigma (X_1 - X_{last})^2}{N - 3}}$$

$S_{1\cdot 23}$ represents standard error of estimate of X_1 on X_2 and X_3. X_{last} indicates the estimated value of X_1 as calculated from the regression equations.

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

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

The standard error measures the closeness of estimates derived from the regression equation to actual observed values.

Coefficient of Multiple Determination

The coefficient of multiple determination is analogous to the coefficient of determination in the two-variable case. As explained earlier, the fit of a straight line to the two-variable scatter was measured by the simple coefficient of determination r^2 which was defined as the ratio of the explained sum of squares to the total sum of squares. In the same fashion, we can define coefficient of multiple determination which is denoted by R^2. Symbolically:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Similar to the case of r^2 and r in two-variable analysis, R^2 is easier to interpret since R^2 is a percentage figure whereas R is not.

As a ratio of explained variation to the total variation in X_1, R^2 can be interpreted as the proportion of the total variation in the dependent variable that is associated with or explained by the regression of X_1 on X_2 and X_3. We may also think of R^2 as a measure of closeness of fit of the regression plane to the actual points. The closer the value of R^2 to 1, the smaller is the scatter of the points about the regression plane and the better is the fit.

The square root of the coefficient of multiple determination is called the coefficient of multiple correlation, denoted as R. This measure is seldom used in practice.