Pre-Ph. D course work
Paper-2
Unit-5

SIMPLE REGRESSION ANALYSIS

Regression is the determination of a statistical relationship between two or more variables. In simple regression, we have only two variables, one variable (defined as independent) is the cause of the behaviour of another one (defined as dependent variable). Regression can only interpret what exists physically i.e., there must be a physical way in which independent variable *X* can affect dependent variable *Y*. The basic relationship between *X* and *Y* is given by

$$\hat{Y} = a + bX$$

where the symbol $\hat{Y}$ denotes the estimated value of *Y* for a given value of *X*. This equation is known as the regression equation of *Y* on *X* (also represents the regression line of *Y* on *X* when drawn on a graph) which means that each unit change in *X* produces a change of *b* in *Y*, which is positive for direct and negative for inverse relationships. Then generally used method to find the 'best' fit that a straight line of this kind can give is the least-square method. To use it efficiently, we first determine

$$\sum x_i^2 = \sum X_i^2 - n\overline{X}^2$$

$$\sum y_i^2 = \sum Y_i^2 - n\overline{Y}^2$$

$$\sum x_i y_i = \sum X_i Y_i - n\overline{X} \cdot \overline{Y}$$

$$b = \frac{\sum x_i y_i}{\sum x_i^2},\ a = \overline{Y} - b\overline{X}$$

These measures define *a* and *b* which will give the best possible fit through the original *X* and *Y* points and the value of *r* can then be worked out as under:

$$r = \frac{b\sqrt{\sum x_i^2}}{\sqrt{\sum y_i^2}}$$

Thus, the regression analysis is a statistical method to deal with the formulation of mathematical model depicting relationship amongst variables which can be used for the purpose of prediction of the values of dependent variable, given the values of the independent variable.

[Alternatively, for fitting a regression equation of the type $\hat{Y} = a + bX$ to the given values of *X* and *Y* variables, we can find the values of the two constants viz., *a* and *b* by using the following two normal equations:

$$\sum Y_i = na + b \sum X_i$$

$$\sum X_i Y_i = a \sum X_i + b \sum X_i^2$$

and then solving these equations for finding *a* and *b* values. Once these values are obtained and have been put in the equation

$$\hat{Y} = a + bX$$

we say that we have fitted the regression equation of *Y* on *X* to the given data. In a similar fashion, we can develop the regression equation of *X* and *Y* viz., $\hat{X} = a + bX,$ *a* + *bX*, presuming *Y* as an independent variable and *X* as dependent variable].

MULTIPLE CORRELATION AND REGRESSION

When there are two or more than two independent variables, the analysis concerning relationship is known as multiple correlation and the equation describing such relationship as the multiple regression equation. We here explain multiple correlation and regression taking only two independent variables and one dependent variable (Convenient computer programs exist for dealing with a great number of variables). In this situation the results are interpreted as shown below: Multiple regression equation assumes the form

$\hat{Y} = a + b_1 X_1 + b_2 X_2$ where *X*1 and *X*2 are two independent variables and *Y* being the dependent variable, and the constants *a*, *b*1 and *b*2 can be solved by solving the following three normal equations:

$$\sum Y_i = na + b_1 \sum X_{1i} + b_2 \sum X_{2i}$$

$$\sum X_{1i} Y_i = a \sum X_{1i} + b_1 \sum X_{1i}^2 + b_2 \sum X_{1i} X_{2i}$$

$$\sum X_{2i} Y_i = a \sum X_{2i} + b_1 \sum X_{1i} X_{2i} + b_2 \sum X_{2i}^2$$

(It may be noted that the number of normal equations would depend upon the number of independent variables. If there are 2 independent variables, then 3 equations, if there are 3 independent variables then 4 equations and so on, are used.) In multiple regression analysis, the regression coefficients (viz., *b*1 *b*2) become less reliable as the degree of correlation between the independent variables (viz., *X*1, *X*2) increases. If there is a high degree of correlation between independent variables, we have a problem of what is commonly described as the *problem of multicollinearity*. In such a situation we should use only one set of the independent variable to make our estimate. In fact, adding a second variable, say *X*2, that is correlated with the first variable, say *X*1, distorts the values of the regression coefficients. Nevertheless, the prediction for the dependent variable can be made even when multi collinearity is present, but in such a situation enough care should be taken in selecting the independent variables to estimate a dependent variable so as to ensure that multi-collinearity is reduced to the minimum. With more than one independent variable, we may make a difference between the collective effect of the two independent variables and the individual effect of each of them taken separately. The collective effect is given by the coefficient of multiple correlation,

$R_{y \cdot x_1 x_2}$ defined as under:

$$R_{y \cdot x_1 x_2} = \sqrt{\frac{b_1 \sum Y_i X_{1i} - n\overline{Y}\,\overline{X}_1 + b_2 \sum Y_i X_{2i} - n\overline{Y}\,\overline{X}_2}{\sum Y_i^2 - n\overline{Y}^2}}$$

Alternatively, we can write

$$R_{y \cdot x_1 x_2} = \sqrt{\frac{b_1 \sum x_{1i} y_i + b_2 \sum x_{2i} y_i}{\sum Y_i^2}}$$

where

$$x_{1i} = (X_{1i} - \overline{X}_1)$$

$$x_{2i} = (X_{2i} - \overline{X}_2)$$

$$y_i = (Y_i - \overline{Y})$$

and b_1 and b_2 are the regression coefficients.

PARTIAL CORRELATION

Partial correlation measures separately the relationship between two variables in such a way that the effects of other related variables are eliminated. In other words, in partial correlation analysis, we aim at measuring the relation between a dependent variable and a particular independent variable by holding all other variables constant. Thus, each partial coefficient of correlation measures the effect of its independent variable on the dependent variable. To obtain it, it is first necessary to compute the simple coefficients of correlation between each set of pairs of variables as stated earlier. In the case of two independent variables, we shall have two partial correlation coefficients denoted *ryx*1 ×*x*2 and *ryx x* 2 × 1 which are worked out as under:

$$r_{yx_1 \cdot x_2} = \frac{R^2{}_{y \cdot x_1x_2} - r^2_{yx_2}}{1 - r^2_{yx_2}}$$

This measures the effort of *X*1 on *Y*, more precisely, that proportion of the variation of *Y* not explained by *X*2 which is explained by *X*1. Also

$$r_{yx_2 \cdot x_1} = \frac{R^2_{y \cdot x_1x_2} - r^2_{yx_1}}{1 - r^2_{yx_1}}$$

in which *X*1 and *X*2 are simply interchanged, given the added effect of *X*2 on *Y*.

Alternatively, we can work out the partial correlation coefficients thus:

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{1 - r^2_{yx_2}}\ \sqrt{1 - r^2_{x_1x_2}}}$$

$$r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1x_2}}{\sqrt{1 - r^2_{yx_1}}\ \sqrt{1 - r^2_{x_1x_2}}}$$

These formulae of the alternative approach are based on simple coefficients of correlation (also known as zero order coefficients since no variable is held constant when simple correlation coefficients are worked out). The partial correlation coefficients are called first order coefficients when one variable is held constant as shown above; they are known as second order coefficients when two variables are held constant and so on.