UNIT 4 SIMPLE LINEAR REGRESSION MODEL: ESTIMATION*

Structure

4.0 OBJECTIVES

After going through this unit, you should be able to

- describe the classical linear regression model;
- differentiate between Population Regression Function (PRF) and Sample Regression Function (SRF);
- find out the Ordinary Least Squares (OLS) estimators;
- describe the properties of OLS estimators;
- explain the concept of goodness of fit of regression equation; and
- describe the coefficient of determination and its properties.

* Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi

4.1 INTRODUCTION

In Unit 5 of the course BECC 107: Statistical methods for Economics we discussed the topics correlation and regression. In that Unit we gave a brief idea about the concept of regression. You already know that there are two types of variables in regression analysis: i) dependent (or explained) variable, and ii) independent (or explanatory) variable. As the name (explained and explanatory) suggests the dependent variable is explained by the independent variable.

Usually we denote the dependent variable as Y and the independent variable as X. Suppose we took up a household survey and collected *n* pairs of observations in X and Y. The relationship between X and Y can take many forms. The general practice is to express the relationship in terms of some mathematical equation. The simplest of these equations is the linear equation. It means that the relationship between X and Y is in the form of a straight line, and therefore, it is called linear regression. When the equation represents curves (not a straight line) the regression is called non-linear or curvilinear.

Thus in general terms we can express the relationship between X and Y as follows in equation (4.1).

$$Y = f(X) \quad \ldots (4.1)$$

In this block (Units 4, 5 and 6) we will consider simple linear regression models with two variables only. The multiple regression model comprising more than one explanatory variable will be discussed in the next block.

Regression analysis may have the following objectives:

- To estimate the mean or average value of the dependent variable, given the values of the independent variables.
- To test the hypotheses regarding the underlying economic theory. For example, one may test the hypotheses that the price elasticity of demand is (–)1 that is, the demand is perfectly elastic, assuming other factors affecting the demand are held constant.
- To predict the mean value of the dependent variable given the values of the independent variable.

4.2 POPULATION REGRESSION FUNCTION

A population regression function hypothesizes a theoretical relationship between a dependent variable and a set of independent or explanatory variables. It is a linear function. The function defines how the conditional expectation of a variable Y responds to the changes in independent variable *X*.

$$Y_i = E(Y_i|X_i) + u_i \quad \ldots (4.2)$$

The function consists of a deterministic component $E(Y|X)$ and a non-deterministic or 'stochastic' component u, as depicted in equation (4.2).

We are concerned about examining the determinants of dependent variable (Y) conditional upon the given values of impendent variables (X).

4.2.1 Deterministic Component

The conditional expectation of Y constitutes the deterministic component of the regression model. It is obtained in the form of a deterministic line. It is also known as the Population Regression Line (PRL). The non-deterministic or stochastic component is represented by a random error term, denoted by u_i.

Let us take an example. Suppose we want to examine the impact of weekly personal disposable income (PDI) on the weekly expenditure for a set of population, then we consider weekly PDI as the independent variable (Y) and weekly expenditure as the dependent variable (X). For each given value of weekly PDI, the average value of weekly expenditure is plotted on the vertical axis. People with higher income are likely to spend more, therefore intuitively, the relationship between weekly PDI and weekly expenditure is positive. Thus the following Population Regression Line is obtained and plotted on a graph as explained below.

$$E(Y_i|X_i) = \beta_1 + \beta_2 X_i \quad \ldots (4.3)$$

Note that in equation (4.3), β_1 and β_2 are the parameters. Here β_1 is the intercept of the population regression function. It indicates the expected value of the dependent variable when the explanatory variable is zero. Further, β_2 is the slope of the population regression function. It indicates the magnitude by which the dependent variable will change if there is a one unit change in the independent variable. The population parameters describe the relationship between the dependent variable and the independent variable in the population.

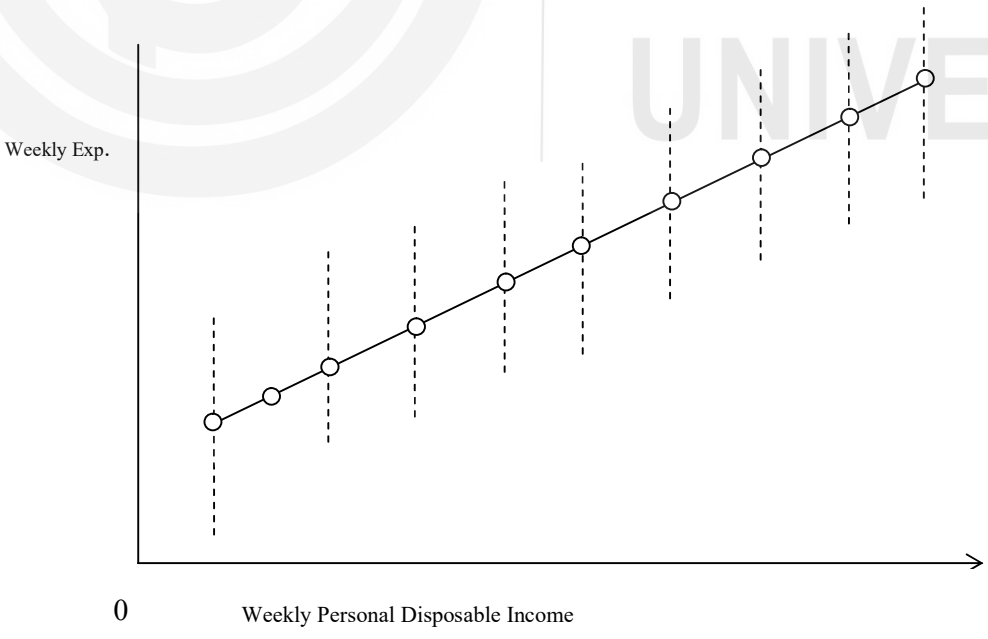


Fig. 4.1: Weekly Personal Disposable Income

Look into the circled points in Fig. 4.1. These points represent the mean or the average value of Y corresponding to various X_i. They are called the conditional means or conditional expectation values. If we connect the various expected values of Y, the resulting lines is called the Population Regression Line (PRL).

4.2.2 Stochastic Component

When we collect data from a sample, we do not a deterministic relationship between X and Y. For example, for the same level of income the expenditure of two persons could be different. Suppose there are two persons with monthly income of Rs. 20000 per month. While the monthly expenditure of one person is Rs. 15000, that of the other person could be Rs. 19000. The differences in monthly expenditure for the second person could be higher due to his health condition or living style. Such differences in the dependent variable are captured by the stochastic error term. In Fig. 4.1, for a particular value of X, the value of the Y variable is depicted by a vertical dotted line. The expected value of Y for a particular value of X is circled (see Fig. 4.1).

Thus, there is a need to specify the stochastic relationship between X and Y. The specification of the sample regression function (SRF) is

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad \ldots (4.5)$$

In equation (4.5) the term u_i is called stochastic error or random error.

The first component of equation (4.5) is the deterministic component ($\beta_1 + \beta 2Xi$, which we have already discussed. The deterministic component is the mean or average expenditure in the example under consideration. The deterministic component is also called the systematic or deterministic component.

The second component u_i is called the random component (determined non-systematically by factors other than income). The error term u_i is also known as the 'noise component'. The error term u_i is a random variable. The value of u_i cannot be controlled or known.

There are three reasons for including the error term u_i in a regression model: (i) The error term represents the influence of those variables that are not explicitly introduced in the regression model. For example, there are several variables that influence consumption expenditure of a household (such as number of family members, health status, neighbourhood, etc.). These variables affect the dependent variable, and there exists intrinsic randomness between X and Y. (ii) Human behaviour is not predictable. This sort of randomness is reflected and captured by the random error term. (iii) The errors in measuring data such as rounding off of annual family income, absence of many students from the school, etc.

Because of the randomness the actual value of the data would either remain above or below the expected value of the dependent variable. In other words, the actual value will deviate from the average, that is, the systematic component.

Having understood the elementary concept of Population Regression Function and Population Regression Line (PRL), the following section describes the estimation of PRL using the sample.

Check Your Progress 1

1) What are the objectives of estimating regression models?

.....

.....

.....

.....

.....

.....

2) Why does the average value of the dependent variable differ from the actual value?

.....

.....

.....

.....

.....

3) Why do we include an error term ($\boldsymbol{u_i}$) to the regression model?

.....

.....

.....

.....

.....

4.3 SAMPLE REGRESSION FUNCTION

We rarely have the data related to the entire population at our disposal. We only have a sample from the population. Thus, we need to use the sample to estimate the population parameters. We may not be able to find out the population regression line (PRL) because of sampling fluctuations or sampling error. Suppose we have two samples from the given population. Using the samples separately, we obtain Sample Regression Lines (SRLs). A sample represents the population. In Fig. 4.2 we have shown two sample regression lines, SRL_1 and SRL_2.

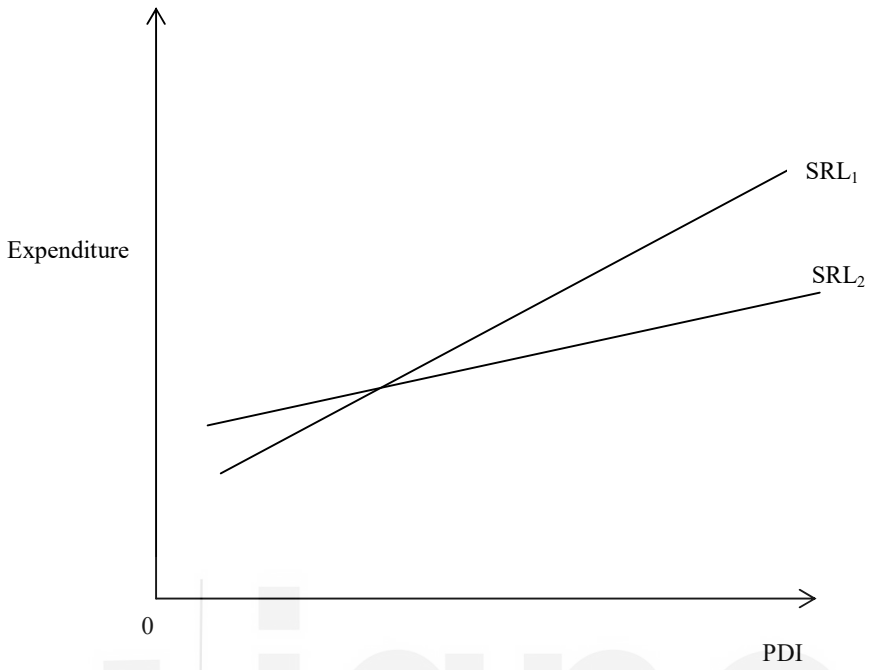


Fig: 4.2: Two Sample Regression Lines

Both the sample regression lines represent the population regression line. However, due to sampling fluctuation, the slope and intercept of both the SRLs are different. Analogous to population regression function (PRF) that underlies the PRL, we develop the concept of Sample Regression Function (SRF) comprising Sample Regression Line (SRL) and the error term u_i.

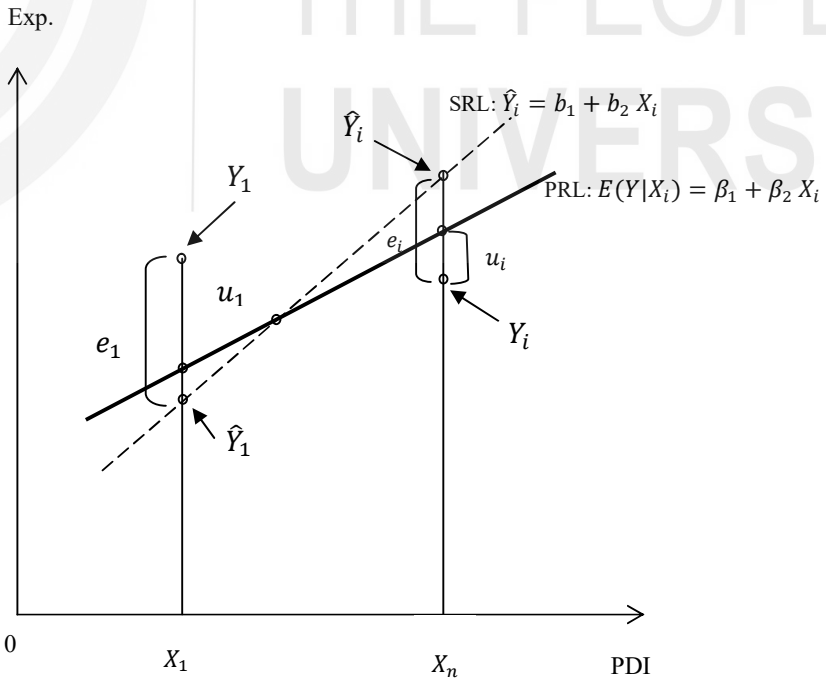


Fig 4.3: Population Regression Line and Sample Regression Line

In Fig. 4.3 we depict the population regression line (PRL) and the sample regression line (SRL). We observe that the slopes of both the lines are different. Thus, $b_1 \neq \beta_1$ and $b_2 \neq \beta_2$. Let us consider a particular value of the explanatory variable, X_1. The corresponding value of the explained variable is Y_1. On the basis of the sample regression line we obtain estimated value of the explained variable, $\hat{Y}_1$. Now let us find out the distinction between the error term (u) and the residual (e). The distance between the actual value Y_1 and the corresponding point on the population regression line is u_1. This error u_1 is not known to us, because we do not know the values of β_1 and β_2. What we know is $\hat{Y}_1$, which is estimated on the basis of b_1 and b_2. The distance between Y_1 and $\hat{Y}_1$ is the residual, e_1.

The population regression line as given in equation (4.2) is

$Y_i = E(Y_i|X_i) + u_i$

The sample regression line that we estimate is given by

$\hat{Y}_i = b_1 + b_2 X_i$ … (4.6)

In equation (4.6) the symbol (^) is read as 'hat' or 'cap'. Thus, $\hat{Y}_i$ is read as 'Y_i-hat'.

You should remember that what we observe are proxies b_1, b_2 and e in place of β_1, β_2 and u_i.

$Y_i = \hat{Y}_i + e_i = b_1 + b_2 X_i + e_i$ ... (4.7)

where $\hat{Y}_i$ = estimator of $E(Y|X_i)$, the estimator of the population conditional mean $\hat{Y}_i$ is an estimator (or a sample statistic) in equation (4.7). A particular value obtained by the estimator is considered an estimate.

The actual value of Y is obtained by adding the residual term to the estimated value of Y, also referred as the residual. The residual is the estimated value of random error term of the population regression function. The sample regression function in equation (4.7) is combination of sample regression line given by $\hat{Y}_i$ and the estimated residual term e_i. The dark straight line in Fig. 4.3 is the Population Regression Line (PRL) and it is given by the following equation:

$E(Y|X) = \beta_1 + \beta_2 X_i$. …(4.8)

Therefore, the Population Regression function (PRF) can be expressed as

$Y_i = E(Y_i|X_i) + u_i$

Or,

$Y_i = \beta_1 + \beta_2 X_i + u_i$ … (4.9)

Thus, the Population Regression Function in equation (4.9) is a combination of population regression line (PRL) $E(Y_i|X_i)$ and random error term u_i. The SRF is only an approximation of PRF. We attempt to find the most appropriate sample that yields estimators b_1 and b_2 which are as close as possible to population

parameters $\beta_1$and β_2. In other words, $b_1$is as close as possible to β_1, and b_2 is as close as possible to β_2.

4.4 ASSUMPTIONS OF CLASSICAL REGRESSION MODEL

A linear regression model is based on certain assumptions as specified below. If a regression model fulfils the following assumptions, it is called the classical linear regression model (CLRM). The assumptions of CLRM are as follows:

(i) The regression model is linear in parameters. It may or may not be linear in variables. For example, the equation given below is linear in parameters as well as variables as shown in equation (4.10)

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad \ldots(4.10)$$

(ii) The explanatory variable is not correlated with the disturbance term *u.* This assumption requires that $\sum u_i X_i = 0$. In other words, the covariance between error term and explanatory variable is zero. This assumption is automatically fulfilled if X is non-stochastic. It requires that the X_i values are kept fixed in repeated samples.

(iii) The expected value or mean value of the error term u is zero. In symbols, $E(u_i|X_i) = 0$. It does not mean that all error terms are zero. It implies that the error terms cancel out each other.

(iv) The variance of each u_i is constant. In symbols, $var(u_i) = \sigma^2$. The conditional distribution of the error term has been displayed in Fig. 4.4(a). The corresponding error variance for a specific value of the error term has been depicted in Fig. 4.4(b). From the figure you can make out that the error variance is constant at all levels of the X variable. It describes the case of 'homoscedasticity'.

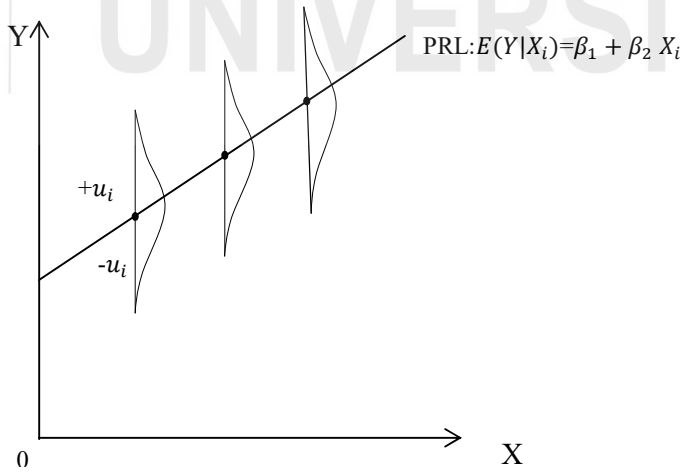


Fig 4.4 (a) Conditional Distribution of Error Term u_i

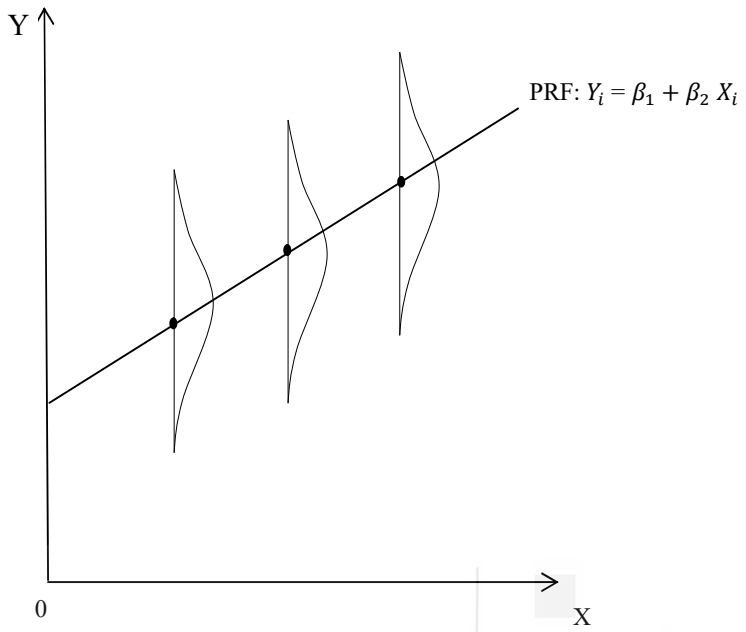


Fig 4.4 (b) Homoscedasticity (equal variance)

Fig. 4.5 depicts the case of unequal error variance, i.e., heteroscedasticity. Here the variance of the error terms varies across the values of X_i.

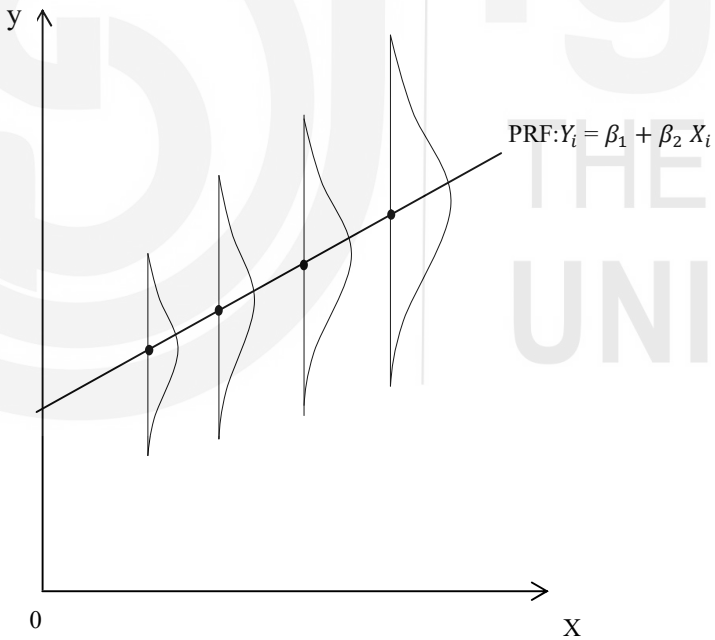


Fig. 4.5: Case of Heteroscedasticity (Unequal Variance)

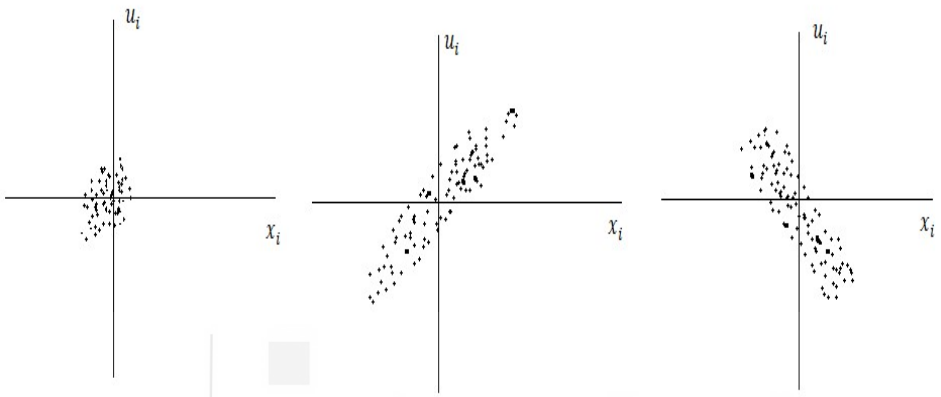
(v) There is no correlation between the two error terms. This is the assumption of no autocorrelation.

$$cov(u_i, u_j) = 0 \quad i \neq j$$

It implies that there is no systematic relationship between two error terms. This assumption implies that the error terms u_i are random.

Since two error terms are assumed to be uncorrelated, any two Y values will also be uncorrelated, i.e., $cov(Y_i, Y_j) = 0$.

Fig 4.6(i) depicts the case of no autocorrelation. Fig 4.6(ii) depicts positive autocorrelation, and Fig 4.7(iii) shows the case of negative autocorrelation.



(i) No Autocorrelation *(ii) Positive Autocorrelation* *(iii) Negative Autocorrelation*

Fig 4.6: Various Cases of Autocorrelation

(vi) The regression model is correctly specified, that is, there is no specification error in the model. If certain relevant variable is not included or certain irrelevant variable is included in the regression model then we commit model specification error. For instance, suppose we study the demand for automobiles. If we take the price of automobiles only and do not include the income of the consumer income then there is some specification error. Similarly, if we do not take into account costs of adverting, financing, gasoline prices, etc., we will be committing model specification error (we will discuss the issue of specification error in Unit 13).

4.5 ORDINARY LEAST SQUARES METHOD OF ESTIMATION

As mentioned in Unit 1 of this course, we need to estimate the parameters of the regression model. There are quite a few methods of estimation of the parameters. In this course will discuss about two such methods: (i) Least Squares, and (ii) Maximum Likelihood. We discuss about the Ordinary Least Squares (OLS) method below.

The Ordinary Least Squares (OLS) method estimates the parameters of a linear regression model by minimising the error sum of squares (ESS). In other words, it minimizes the sum of the squares of the differences between the observed dependent variable (Y_i) and the predicted or expected value of the dependent variable ($\hat{Y}_i$).

In symbols,

$e_i = (Y_i - \hat{Y}_i)$

$e_i^2 = (Y_i - \hat{Y}_i)^2$

$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ ... (4.11)

In OLS method we minimise $\sum_{i=1}^{n} e_i^2$.

We know that

$\hat{Y}_i = b_1 + b_2 X_i$

If we substitute the value of $\hat{Y}_i$ in equation (4.11) we obtain

$\sum_{i=1}^{n} e_i^2 = \Sigma(Y_i - b_1 - b_2 X_i)^2$

The first order condition of minimization requires that the partial derivatives are equal to zero. Note that we have to decide on the values of b_1 and b_2 such that ESS is the minimum. Thus, we have take partial derivates with respect to b_1 and b_2. This implies that

$\frac{\partial \Sigma e_i^2}{\partial b_1} = 0$ … (4.13)

and

$\frac{\partial \Sigma e_i^2}{\partial b_2} = 0$ … (4.14)

From equation (4.13) we have

$2\Sigma(Y_i - b_1 - b_2 X_i)(-1) = 0$

By re-arranging terms in the above equation we obtain

$\Sigma Y_i = nb_1 + b_2 \Sigma X_i$ … (4.15)

In equation (4.15), note that *n* is the sample size.

From equation (4.14) we have

$2\Sigma(Y_i - b_1 - b_2 X_i)(-X_i) = 0$

By re-arranging terms in the above equation we obtain

$\Sigma X_i Y_i = b_1 \Sigma X_1 + b_2 \Sigma X_i^2$ … (4.16)

Equations (4.15) and (4.16) are called normal equations. We have two equations with two unknowns (b_1 and b_2) .

Thus, by solving these two normal equations we can find out unique values of b_1 and b_2.

By solving the normal equations (4.15) and (4.16) we find that

$$b_1 = \bar{Y} - b_2\bar{X} \quad \ldots (4.17)$$

and

$$b_2 = \frac{n\Sigma XY - \Sigma X\Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - X)^2}$$

Let us take the variables X and Y in deviation forms such that

$$x_i = X_i - \bar{X} \qquad y_i = Y_i - \bar{Y}$$

Thus,

$$b_2 = \frac{\Sigma x_i y_i}{\Sigma x_i^2} \quad \ldots (4.18)$$

As you can see from the formula for b_2, it is simpler to write the estimator of the slope coefficient in deviation form. Expressing the values of a variable from its mean value does not change the ranking of the values, since we are subtracting the same constant from each value. It is crucial to note that b_1 and b_2 are expressed in terms of quantities computed from the sample, given by the formula in expressions in (4.17) and (4.18).

We mention below the formulae for variance and standard deviation of the estimators b_1 and b_2

$$Var(b_1) = \sigma_{b_1}^2 = \frac{\Sigma X_i^2}{n\Sigma x_i^2}\sigma^2 \quad \ldots (4.19)$$

$$SE(b_1) = \sqrt{Var(b_1)} \quad \ldots (4.20)$$

$$Var(b_2) = \sigma_{b_2}^2 = \frac{\sigma^2}{\Sigma x_i^2}$$

$$SE(b_2) = \sqrt{\text{var}}(b_2) \quad \ldots (4.21)$$

$$\hat{\sigma}^2 = \frac{\Sigma e_i^2}{n-2} = \frac{RSS}{n-2} = \frac{RSS}{d.f.} \quad \ldots (4.22)$$

$$\text{S.E. of the residual } (e_i) = \sqrt{\hat{\sigma}^2} \quad \ldots (4.23)$$

The formulae mentioned in equations (4.19), (4.20), (4.21), (4.22) and (4.23) are the variance and standard errors of estimated parameters b_1 and b_2.

Smaller the value of $\hat{\sigma}^2$, closer is the actual Y value to its estimated value. Recall that any linear function of a normally distributed variable to itself normally distributed. If b_1 and b_2 are linear functions of normally distributed variable u_i they themselves are normally distributed. Thus,

$$b_1 \sim N(\beta_1, \sigma_{b_1}^2) \quad \ldots (4.24)$$

$$b_2 \sim N(\beta_2, \sigma_{b_2}^2) \quad \ldots (4.25)$$

Check Your Progress 2

1) Distinguish between the error term and the residual by using appropriate diagram.

.....
.....
.....
.....
.....

2) Prove that the sample regression line passes through the mean values of X and Y.

.....
.....
.....
.....
.....

4.6 ALGEBRAIC PROPERTIES OF OLS ESTIMATORS

The OLS estimators b_1 and b_2 fulfil certain important properties.

a) SRF obtained by OLS method passes through sample mean values of X and Y. This mainly implies that the point $(\bar{X}, \bar{Y})$ passes through the Sample Regression Line.

$\bar{Y} = b_1 + b_2\bar{X}$ …(4.26)

Mean value of residuals $\bar{e}$ is always zero $\bar{e} = \frac{\Sigma e_i}{n} = 0$. This implies that on an average, the positive and negative residual terms cancel each other.

b) $\Sigma e_i X_i = 0$ …(4.27)

The sum of product of residuals e_i and the values of explanatory variable X is zero, i.e., the two variables are uncorrelated.

c) $\Sigma e_i \hat{Y}_i = 0$ …(4.28)

The sum of product of residuals $e_i$and estimated $\hat{Y}_i$ is zero, i.e., $e_i \hat{Y}_i = 0$.

4.7 COEFFICIENT OF DETERMINATION

Let us consider the regression model:

$Y_i = \beta_1 + \beta_2 X_i + u_i$

Recall from equation (4.7) that

$Y_i = \hat{Y}_i + e_i$

If we subtract $\bar{Y}$ from both sides of the above equation, we obtain

$$(Y_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i) \quad \ldots (4.29)$$

[Since $e_i = Y_i - \hat{Y}_i$]

In equation (4.20) there are three terms: (i) $(Y_i - \bar{Y})$ which is the variation in Y_i, (ii) $(\hat{Y}_i - \bar{Y})$ which is the explained variation, and (iii) $(Y_i - \hat{Y}_i)$ which is the unexplained or residual variation.

Now, let us use the lower case letters to indicate deviation from mean of a variable. Equation (4.30) can be written as

$$y_i = \hat{y}_i + e_i \quad \ldots (4.30)$$

Since $\sum e_i = 0$, we have $\bar{e} = 0$.

Therefore, we have $\bar{Y} = \bar{\hat{Y}}$, that is, the mean values of the actual Y and the estimated Y are the same.

Recall that

$$Y_i = b_1 + b_2 X_i + e_i \quad \ldots (4.7)$$

and

$$\bar{Y} = b_1 + b_2 \bar{X} \quad \ldots (4.26)$$

If we subtract equation (4.26) from equation (4.7), we get

$$y_i = b_2 x_i + e_i \quad \ldots (4.31)$$

If find OLS estimator of (4.31), we obtain

$$\hat{y}_i = b_2 x_i .$$

Therefore,

$$y_i = \hat{y}_i + e_i \quad \ldots (4.32)$$

Now let us takes squares of equation (4.32) on both sides and sum it over the sample. After re-arranging terms, we obtain

$$\Sigma y_i^2 = \Sigma \hat{y}_i^2 + \Sigma e_i^2 \quad \ldots (4.33)$$

Or, equivalently,

$$\Sigma y_i^2 = b_2^2 \Sigma x_i^2 + \Sigma e_i^2 \quad \ldots (4.34)$$

Equation (4.34) can be expressed in the following manner;

TSS = ESS + RSS … (4.35)

where TSS = Total Sum of Squares

ESS = Explained Sum of Squares

RSS = Residual Sum of Squares

Let us divide equation (4.35) by TSS. This gives us

$$1 = \frac{ESS}{TSS} + \frac{RSS}{TSS} \quad \ldots (4.36)$$

Now, let us define

$$R^2 = \frac{ESS}{TSS} \quad \ldots (4.37)$$

The R^2 is called the coefficient of determination. It is considered as measure of goodness of fit of a regression model. It is an overall 'goodness of fit' that tells us how well the estimated regression line fits the actual Y values.

4.7.1 Formula of Computing R^2

Using the definition of R^2 given at equation (4.37), we can write equation (4.36) as:

$$1 = R^2 + \frac{RSS}{TSS} = R^2 + \frac{\Sigma e_i^2}{\Sigma y_i^2}$$

Therefore,

$$R^2 = 1 - \frac{\Sigma e_i^2}{\Sigma y_i^2} \quad \ldots (4.38)$$

You should note that R^2 gives the percentage of TSS explained by ESS. Thus, if $R^2 = 0.75$, we can say that 75 per cent variation in the dependent variable is explained by explanatory variable in the regression model. The value of R^2 or coefficient of determination lies between 0 and 1. This is mainly because it represents the ratio of explained sum of squares to total sum of squares.

Now let us look into the algebraic properties of R^2 and interpret it. When $R^2 = 0$ we have ESS $= 1$. It indicates that no proportion of the variation in the dependent variable is explained by ESS. If $R^2 = 1$, the sample regression is a perfect fit. If $R^2 = 1$, all the observations lie on the estimated regression line. A higher value of the R^2 implies a better fit of a regression model.

4.7.2 F-Statistic for Goodness of Fit

The statistical significance of a regression model is tested by the F-statistic. By using the t-test we can test the statistical significance of a particular parameter of the regression model. For example, the null hypothesis $H_0: \beta_2 = 0$ implies that there is no relationship between Y and X in the population. By using F-statistic, we can test the null hypothesis that all the parameters in the model are zero. Therefore, we use F-statistics for goodness of fit.

F-statistics for goodness of fit is given by the following:

$$F = \frac{ESS/(k-1)}{RSS/(n-k)} \quad \ldots (4.39)$$

where k is the number of parameters in regression equation and n is the sample size.

4.7.3 Relationship between F and R^2

From equation (4.39) we know that $F = \frac{ESS/(k-1)}{RSS/(n-k)}$. If we divide the numerator and the denominator by TSS, we have

$$F = \frac{ESS/TSS/(k-1)}{RSS/TSS/(n-k)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \quad \ldots (4.40)$$

Note that the F-statistic is an increasing function of R^2. An increase in the value of R^2 means an increase in the numerator and a decrease in the denominator. Now let us explain the interpretation of F-static obtained in equation (4.41). The value obtained by applying equation (4.41) to a dataset is the calculated value of F or F-calculated. We compare this value with the tabulated value or critical value of F given at the end of the book. For comparison purpose the degrees of freedom are $((k-1),(n-k))$.

If F-calculated is greater than F-critical we reject the null hypothesis $H_0: \beta_2 = 0$. An implication of the above is that the independent variables explain the dependent variable. In other words, there exists a statistically significant relationship between Y and X.

If F-calculated is less than F-critical we do not reject the null hypothesis $H_0: \beta_2 = 0$. Thus there is no significant relationship between Y and X.

4.7.4 Relationship between F and t^2

There is relationship between the F-statistic and the t-statistic in a regression model. Suppose, the number of explanatory variables *k* = 2.

$$F = \frac{ESS/(k-1)}{RSS/(n-2)}$$

For the two-variable model,

$$F = \frac{ESS/(2-1)}{RSS/(n-2)} = \frac{ESS}{RSS/(n-2)} \quad \ldots(4.41)$$

We know that ESS $\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$ and $RSS = \sum_{i=1}^{n} e_i^2$

Therefore,

$$\mathrm{F} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n} e_i^2/(n-2)} = \frac{\sum_{i=1}^{n}([b_1 + b_2 X_i] - [b_1 + b_2 \bar{X}])^2}{\hat{\sigma}^2} \quad \ldots (4.42)$$

$$\text{Estimation of error variance} = \hat{\sigma}^2 = \frac{RSS}{n-k} = \frac{\Sigma e_i^2}{n-2} \quad \ldots(4.43)$$

$$F = \frac{1}{\hat{\sigma}^2} \cdot \sum_{i=1}^{n} b_2^2 (X_i - \bar{X})^2 \quad \ldots(4.44)$$

We know that

$$var(b_2) = \frac{\hat{\sigma}^2}{\Sigma x_i^2}$$

Substituting equation (4.43) in equation (4.44) we get,

$$F = \frac{b_2^2}{\hat{\sigma}^2} = \frac{b_2^2}{var(b_2)} = \frac{b_2^2}{[SE(b_2)]^2} = t^2 \quad \ldots (4.45)$$

Therefore, the F-statistic is equal to square of the t-statistic ($F = t^2$). The above result, however, is true for the two-variable model only. If the number of explanatory variable increases in a regression model, the above result may not hold.

Check Your Progress 3

1) Is it possible to carry out F-test on the basis of the coefficient of determination? Explain how.

.....
.....
.....
.....
.....

2) Can the coefficient of determination be greater than 1? Explain why.

.....
.....
.....
.....
.....

4.8 LET US SUM UP

In this unit we discussed about the classical linear regression model, which is based on certain assumptions. We distinguished between the population regression function and the sample regression function. We explained why a stochastic error term is added in a regression equation. We explained the meaning of each of the assumptions of the classical regression model. The procedure of obtaining OLS estimators of a regression model is given in the Unit. The unit further elaborated on the notion of goodness of fit and concept of R-squared.

4.9 ANSWERS/ HINTS TO CHECK YOUR PORGRESS EXERCISES

Check Your Progress 1

1) The objectives of carrying out a regression model could be as follows:

- To estimate the mean or the average value of the dependent variable, given the values of independent variables.
- To test the hypotheses regarding the underlying economic theory. For example, one may test the hypotheses that price elasticity of demand is (–)1.
- To predict or forecast the mean value of the dependent variable given the value of the independent variable.

2) The relationship between Y and X is stochastic in nature. There is an error term added to the regression equation. The inclusion of the random error term leads to a difference between the expected value and the actual value of the dependent variable.

3) There are three reasons for inclusion of the error term in the regression model. See Sub-Section 4.2.2 for details.

Check Your Progress 2

1) Go through Section 4.3. You should explain the difference between the error term and the residual by using Fig. 4.3.

2) In the OLS method we minimise Σe_i^2 by equating its partial derivates to zero. The condition $\frac{\partial \Sigma \ e_i^2}{\partial b_1} = 0$ gives us the first normal equation:
$Y_i = nb_1 + b_2 \Sigma X_i$. If we divide this equation by the sample size, n, we obtain $\bar{Y} = b_1 + b_2 \bar{X}$. Thus, the estimated regression passes through the point $\bar{X}, \bar{Y}$.

Check Your Progress 3

1) Yes, we can carry out F-test on the basis of the R^2 value. Go through equation (4.40).

2) The value of R^2 or the coefficient of determination lies between 0 and 1. This is mainly because it represents the ratio of ESS to TSS. It indicates the proportion of variation in Y that has been explained by the explanatory variables. The numerator ESS cannot be more than the TSS. Therefore, R^2 cannot be greater than 1.