# 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).

$$Y = f(X) \qquad \dots (4.1)$$

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*.

$$Y_i = E(Y_i|X_i) + u_i \qquad \dots (4.2)$$

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.

$$E(Y_i|X_i) = \beta_1 + \beta_2 X_i \qquad \dots (4.3)$$

Note that in equation (4.3), $\beta_1$ and $\beta_2$ are the parameters. Here $\beta_1$ is the intercept of the population regression function. It indicates the expected value of the dependent variable when the explanatory variable is zero. Further, $\beta_2$ 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.

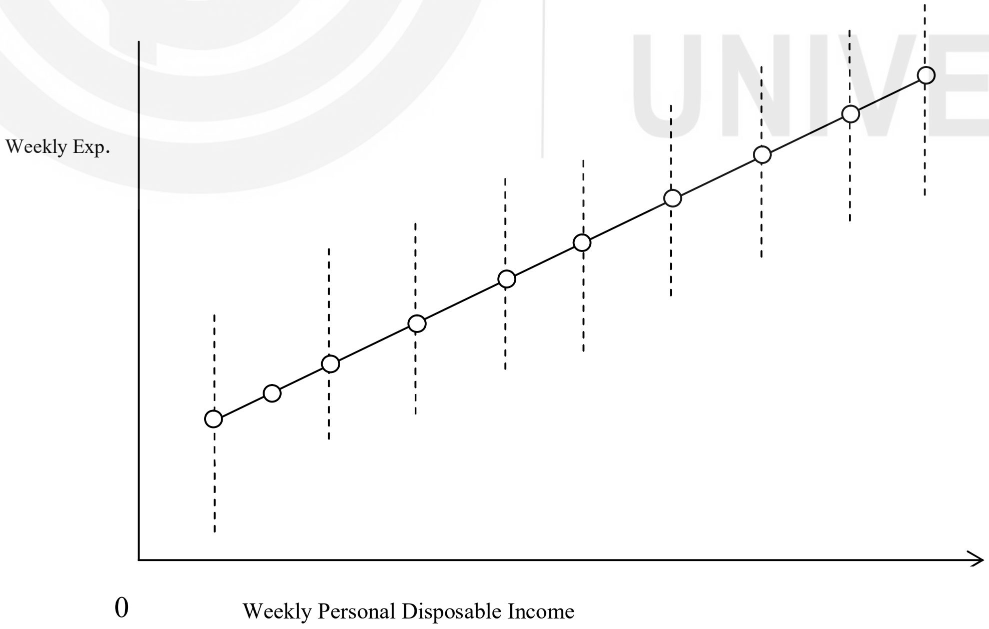


**Fig. 4.1: Weekly Personal Disposable Income**

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

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

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, $\text{SRL}_1$ and $\text{SRL}_2$.

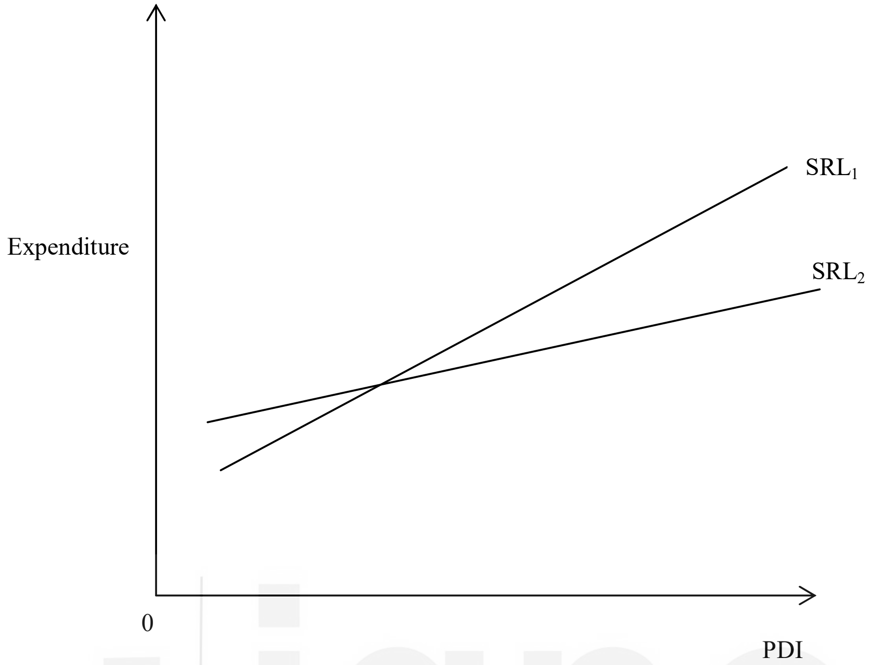


**Fig: 4.2: Two Sample Regression Lines**

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 $u_i$.

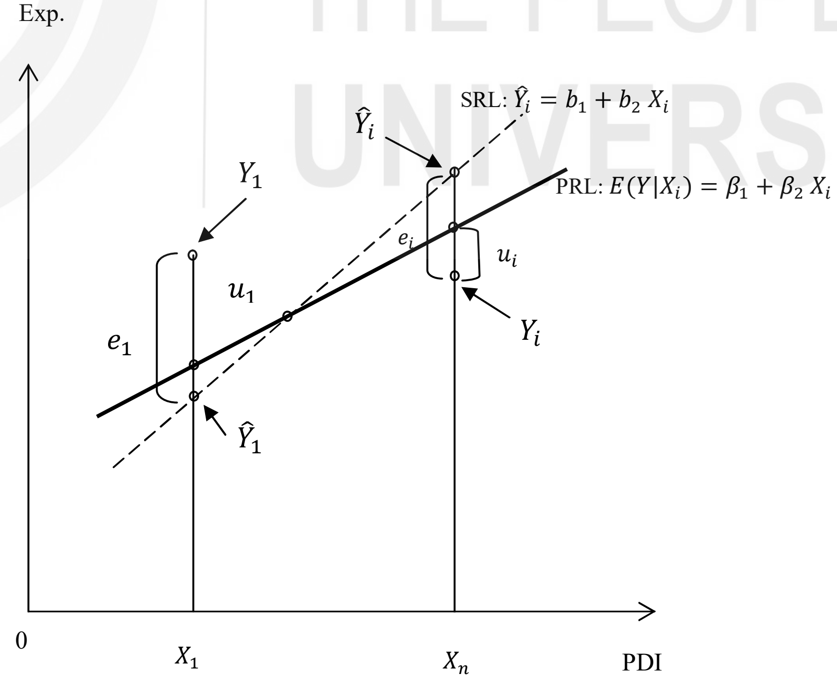

Exp.
SRL: $\hat{Y}_i = b_1 + b_2 X_i$
$\hat{Y}_i$
PRL: $E(Y|X_i) = \beta_1 + \beta_2 X_i$
$Y_1$
$e_i$
$u_i$
$u_1$
$Y_i$
$e_1$
$\hat{Y}_1$
0
$X_1$
$X_n$
PDI

**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 $\beta_1$ and $\beta_2$. 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

$$\hat{Y}_i = b_1 + b_2 X_i \qquad \dots (4.6)$$

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 $\beta_1$, $\beta_2$ and $u_i$.

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

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:

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

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

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

Or,

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

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 $\beta_2$. In other words, $b_1$is as close as possible to $\beta_1$, and $b_2$ is as close as possible to $\beta_2$.

## 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)

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

(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'.

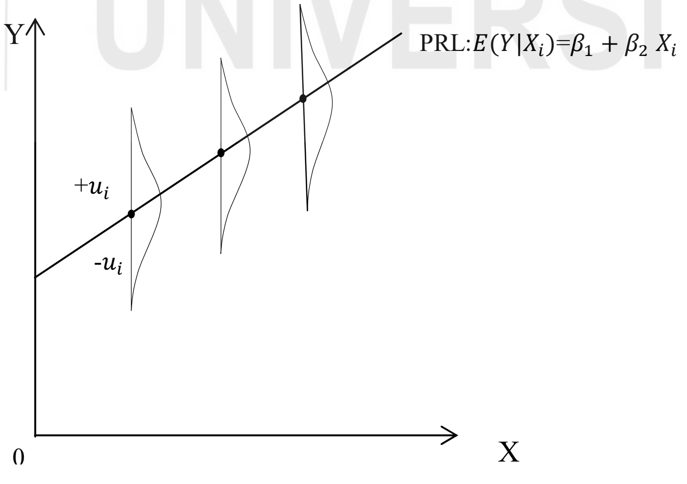


**Fig 4.4 (a) Conditional Distribution of Error Term $u_i$**

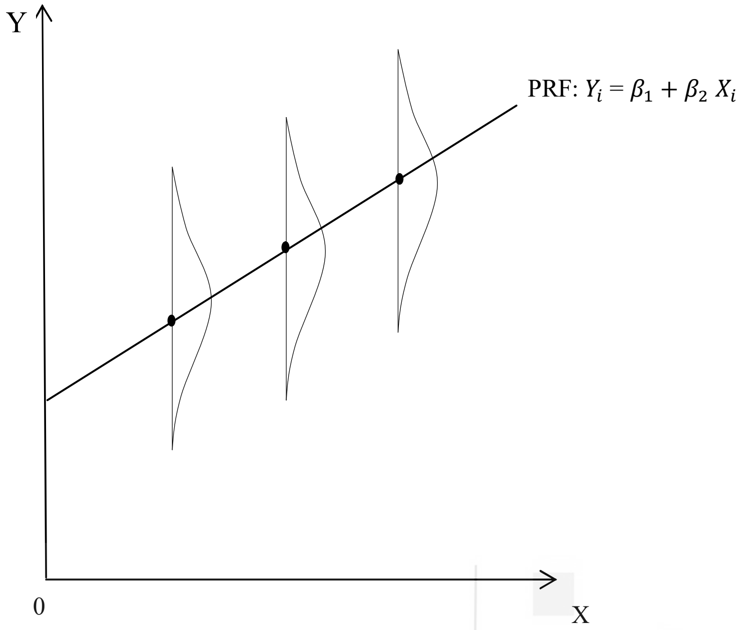


**Fig 4.4 (b) Homoscedasticity (equal variance)**

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 $X_i$.

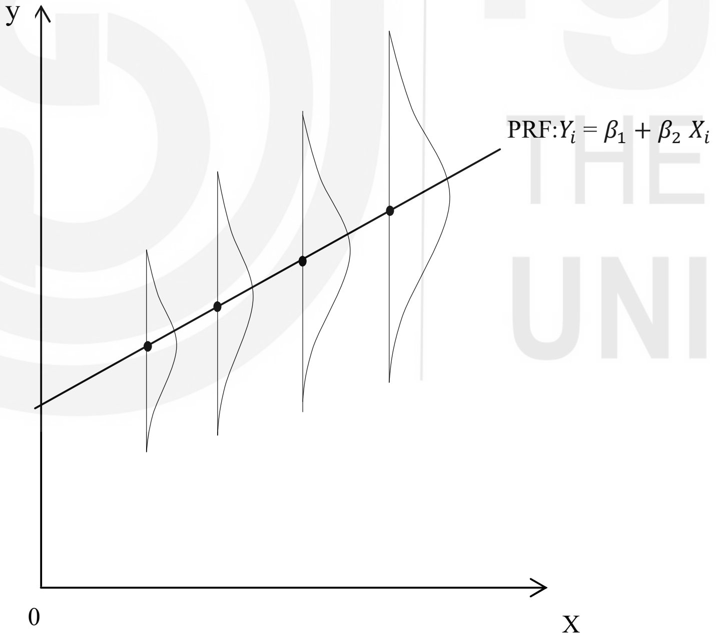


**Fig. 4.5: Case of Heteroscedasticity (Unequal Variance)**

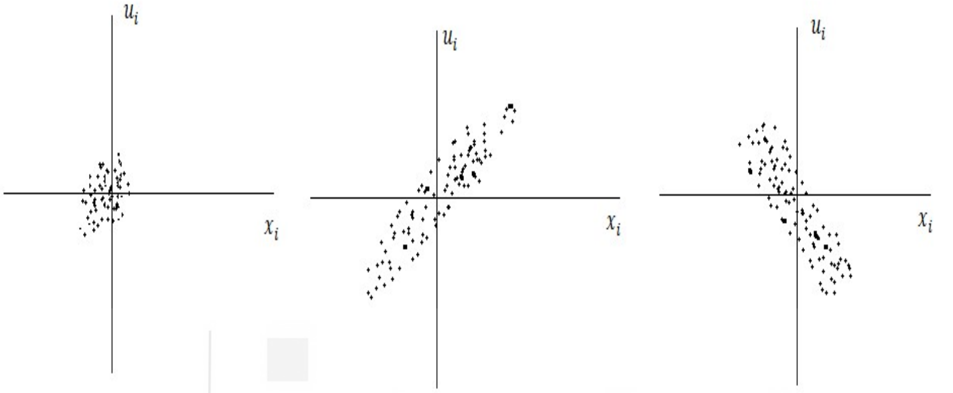
(v) There is no correlation between the two error terms. This is the assumption of no autocorrelation.

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

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,

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

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

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

$\frac{\partial \Sigma e_i^2}{\partial \mathrm{b}_2} = 0$ … (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

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

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

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

Thus,

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

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 $b_2$

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

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

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

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

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

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

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,

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

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

**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.

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

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

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

[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

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

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

$$Y_i = b_1 + b_2 X_i + e_i \qquad ... (4.7)$$

and

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

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

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

If find OLS estimator of (4.31), we obtain

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

Therefore,

$$y_i = \hat{y}_i + e_i \qquad ... (4.32)$$

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

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

Or, equivalently,

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

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

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

Now, let us define

$R^2 = \frac{ESS}{TSS}$ … (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,

$R^2 = 1 - \frac{\Sigma e_i^2}{\Sigma y_i^2}$ … (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:

$F = \frac{ESS/(k-1)}{RSS/(n-k)}$ … (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

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

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,

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

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,

$$\mathrm{F} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n} e_i^2/(n-2)} = \frac{\sum_{i=1}^{n}([b_1+b_2X_i]-[b_1+b_2\bar{X}])^2}{\hat{\sigma}^2} \qquad \ldots (4.42)$$

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

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

We know that

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

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

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

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 $\Sigma e_i^2$ by equating its partial derivates to zero. The condition $\frac{\partial\Sigma\ _i^2}{\partial b_1} = 0$ 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.

# UNIT 5 SIMPLE REGRESSION MODEL: INFERENCE*

**Structure**

## 5.0 OBJECTIVES

After reading this unit, you will be able to:

- explain the concept of Testing of Hypothesis;
- derive the confidence interval for the slope coefficient in a simple linear regression model;
- explain the approach of 'test of significance' for testing the hypothesis on the estimated slope coefficient;
- describe the concept of Analysis of Variance (ANOVA);
- state the Gauss Markov Theorem with its properties; and
- derive the confidence interval for the predicted value of $Y$ in a simple regression model.

## 5.1 INTRODUCTION

In Unit 4 we discussed the procedure of estimation of the values of the parameters. In this unit, we focus upon how to make inferences based on the estimates of parameters obtained. We consider a simple linear regression model with only one independent variable. This means we have one slope coefficient associated with the independent variable and one intercept term. We begin by recapitulating the basics of 'hypothesis testing'.

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

## 5.2 TESTING OF HYPOTHESIS

Testing of hypothesis refers to assessing whether the observation or findings are compatible with the stated hypothesis or not. The word compatibility implies "sufficiently close" to the hypothesized value. It further indicates that we do not reject the stated hypothesis. The stated hypothesis is also referred to as 'Null Hypothesis' and it is denoted by $H_0$. The null hypothesis is usually tested against the 'alternative hypothesis', also known as maintained hypothesis. The alternative hypothesis is denoted by $H_{1.}$ For instance, suppose the given population regression function is given by the equation:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad \text{... (5.1)}$$

where $X_i$ is personal disposable income (PDI) and $Y_i$ is expenditure. Now, the null hypothesis is:

$$H_0 : \beta_2 = 0 \qquad \text{... (5.2)}$$

while the alternative hypothesis is:

$$H_1 : \beta_2 \neq 0 \qquad \text{... (5.3)}$$

We deliberately set the null hypothesis to 'zero' in order to find out whether *Y* is related to *X* at all. If *X* really belongs to the model, we would fully expect to reject the zero-null hypothesis $H_0$ in favour of the alternatives hypothesis $H_1$. The alternative hypothesis implies that the slope coefficient is different from zero. It could be positive or it could be negative. Similarly, the true population intercept can be tested by setting up the null hypothesis:

$$H_1 : \beta_1 = 0 \qquad \text{… (5.4)}$$

while the alternative hypothesis is:

$$H_1 : \beta_1 \neq 0 \qquad \text{… (5.5)}$$

The null hypothesis states that the true population intercept is equal to zero, while the alternative hypothesis states that it is not equal to zero. In case of both the null hypotheses,, i.e., for true population parameter or slope and the intercept, the null hypothesis as stated is a 'simple hypothesis'. The alternative hypothesis is composite. It is also known as a **two-sided hypothesis.** Such a two-sided alternative hypothesis reflects the fact that we do not have a strong apriori or theoretical expectation about the direction in which the alternative hypothesis must move from the null hypothesis. However, when we have a strong apriori or theoretical expectations, based on some previous research or empirical work, then the alternative hypothesis can be one-sided or unidirectional rather than two-sided. For instance, if we are sure that the true population value of slope coefficient is positive then the best way to express the two hypotheses is

$$H_0 : \beta_2 = 0$$

$$H_1: \beta_2 > 0$$

Let us take an example from macroeconomics. The prevailing economic theory suggests that marginal propensity to consume is positive. This means that the slope coefficient is positive. Now, suppose that the given population regression function is estimated by using a sample regression by adopting Ordinary Least Squares estimate. Let us also suppose that the results of sample regression yield the value of estimated slope coefficient as $b_2 = 0.0814$. This numerical value will change from sample to sample. We know that $\beta_2$ follows normal distribution, i.e., $b_2 \sim N\left(\beta_2, \frac{\sigma^2}{\Sigma x_i^2}\right)$ . There are two methods of testing the null hypothesis that the true population slope coefficient is equal to zero. The next two sections of this unit describe the two methods of testing of hypothesis of regression parameters.

## 5.3 CONFIDENCE INTERVAL

In this section, we shall derive the confidence interval for the slope parameter in equation (5.1) above. Note that the confidence interval approach is a method of testing of hypothesis. This is because it refers to the probability that a population parameter falls within the set of critical values from the Table. We make two assumptions, viz. (i) $\alpha$, the level of significance on probability of committing type I error, is fixed at 5% level and (ii) the alternative hypothesis is two sided. From the *t*-table (given at the end of the book) we find the critical value of *t* at (n – k) degrees of freedom (d.f.) at $\alpha = 5\%$ is:

$$P(-2.306 \le t \le 2.306) = 0.95 \qquad \text{... (5.6)}$$

Substituting for 't', equation (5.6) can be re-written as:

$$P\left(2.306 \le \frac{b_2 - \beta_2}{\hat{\sigma} / \sqrt{\Sigma x_i^2}} \le 2.306\right) = 0.95 \qquad \text{... (5.7)}$$

Hence, the probability that *t* value lies between the limits –2.306, +2.306) is 0.95 or 95%. These are the critical *t* values. Substituting the value of *t* into equation (5.6) and rearranging the terms in (5.7) we get:

$$\mathrm{P}\left[\left(\mathrm{b}_2 - 2.306\,.\mathrm{SE}\left(\mathrm{b}_2\right) \le \beta_2 \le \mathrm{b}_2 + 2.306\,\mathrm{SE}\left(\mathrm{b}_2\right)\right)\right] = 0.95$$

The above equation provides a 95% confidence interval for $\beta_2$. Such a confidence interval is known as the region of acceptance (for $H_0$) and the area outside the confidence interval is known as the region of rejection [for $(H_0)$]. If this interval includes the value of $\beta_2$ we do not reject the hypothesis; but if it lies outside the confidence interval, we reject the null hypothesis.

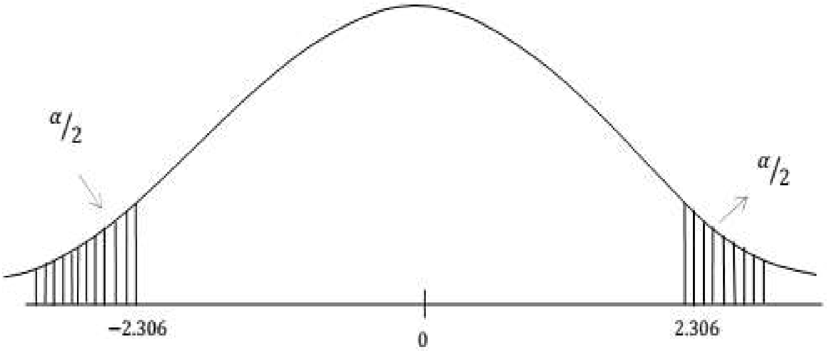


Fig 5.1: t-Distribution

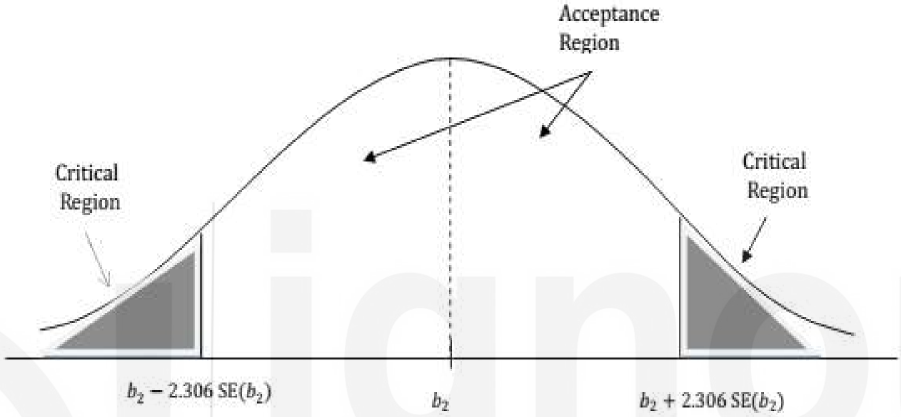


Fig 5.2: Confidence Interval for $\beta_2$

**Check Your Progress 1** [answer questions in about 50-100 words]

1) State the difference between a simple and a composite hypothesis.

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

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

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

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

2) Null hypothesis is the indicator of simple and composite hypothesis. Is this statement true? Justify.

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

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

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

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

3) What is meant by a 'confidence interval'?

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

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

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

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

4) Why do we say that the interval contains the hypothesized value of true population parameter?

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

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

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

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

## 5.4 TEST OF SIGNIFICANCE

Test of significance approach is another method of testing of hypothesis. The decision to accept or reject H$_0$ is made on the basis of the value of *t*- test. It is computed by the statistic from the sample data as:

$$t = \frac{b_2 - \beta_2}{SE(b_2)} \qquad ...(5.8)$$

Equation (5.8) follows *t*-distribution with (n – k) degrees of freedom. The null hypothesis that we are testing here is:

$$H_0 : \beta_2 = \beta_2^* \qquad \ldots (5.9)$$

Note that $\beta_2^*$ is some specific numerical value of $\beta_2$. Thus, the computed value of the test-statistic '*t*' will be like:

$$t = \frac{b_2 - \beta_2^*}{SE(b_2)} \qquad ...(5.10)$$

= [(estimated value) – (hypothesized value)] ÷ (standard error of estimator)

This can be computed from sample data as all values are available. The *t* value computed from (5.10) follows t distribution with (*n* – *k*) degrees of freedom (d.f.). This testing procedure is called the *t*-test. Fig. 5.3 depicts the region of rejection and the region of acceptance. One method of deciding on the result of the testing is to compare the computed value with the tabulated value (also called the 'critical value'). If the computed value of *t* is greater than the critical value of *t* then we reject the null hypothesis. This means we are rejecting the hypothesis that the true population parameter, or the slope coefficient, is zero. It implies that the explanatory variable plays a significant role in determining dependent variable. On the other hand, if the computed *t* value is less than critical value of *t*, then we do not reject the null hypothesis that the true value of the population parameter (or the slope coefficient) is zero. Not rejecting the null hypothesis implies that the value of slope coefficient is zero and that the explanatory variable does not play any significant role in determining the dependent variable.

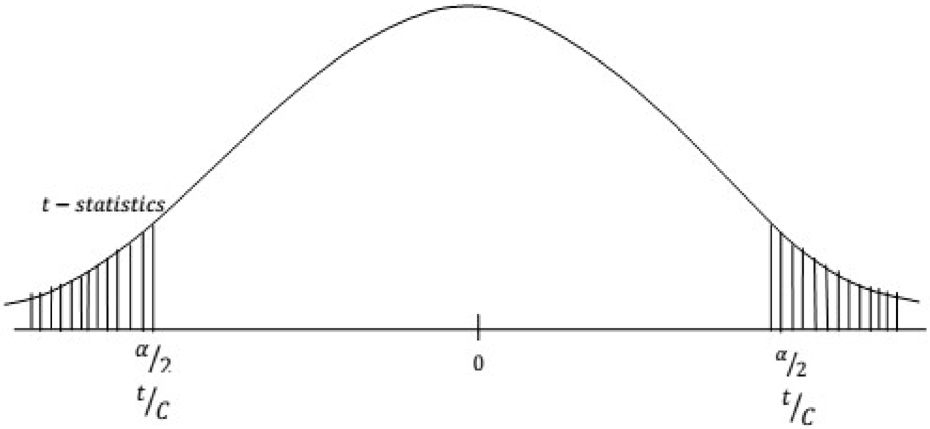


**Fig 5.3: Test of Significance**

In present times, when the results of the regression are obtained by computer, we usually get the *p*-value for the computed statistic. The p-value indicates the probability that the null hypothesis is true. If $p < 0.05$, we reject the null hypothesis and accept the alternative hypothesis. If $p > 0.05$, then we accept the null hypothesis. This means we base our test result at 5 percent level of significance. This also means that in 95 out of 100 independent samples, our result of the test will be similar. In other words, in 5 out of 100 cases, we could be coming to a wrong conclusion.

## 5.5 ANALYSIS OF VARIANCE (ANOVA)

Analysis of Variance (ANOVA) is a statistical tool used to analyse the given data for variations caused by several factors. These factors are divided into two parts: one is called the deterministic (or the systematic) part and the other is called the random part. This method of analysing the variance was developed by Ronald Fisher in 1918. Hence, this is also known as Fisher's analysis of variance. The ANOVA method separates the observed variance in the data into different components. It is used to determine the influence that the independent variables have on the dependent variable in a regression analysis. In a regression analysis ANOVA identifies the variability within a regression. Note that the total variability of dependent variable can be expressed in two parts as follows:

$$(Y_i - \bar{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i) \qquad \text{... (5.11)}$$

Equation (5.11) distributes the total variation in the dependent variable *Y* into two parts, i.e., the variation in mean and the residual value. Squaring each of the terms in equation (5.11) and adding over all the *n* observations, we get the following equation.

$$\sum(Y_i - \bar{Y})^2 = \sum(\hat{Y}_i - \bar{Y})^2 + \sum(Y_i - \hat{Y}_i)^2 \qquad \text{... (5.12)}$$

The above equation can be written as: TSS = ESS + RSS where TSS is the Total Sum of Squares, ESS is the Explained Sum of Squares, and RSS is the Residual

Sum of Squares. The RSS is also called the ‘Sum of Squares due to Error (SSE)’. The ratio ESS / TSS is defined as the coefficient of determination $R^2$. The $R^2$ indicates the proportion of total sum of squares explained by the regression model. An ANOVA analysis is carried out with the help of a table (Table 5.1). From such a table of analysis of variance, the *F*-statistic can be computed as: ESS/RSS. This *F*-statistic is used to test the overall level of significance of the model. The null hypothesis and the alternative hypothesis for testing the overall significance using ANOVA are given by:

$H_0$: Slope coefficient is zero

$H_1$: Slope coefficient is not equal to zero.

**Table 5.1: Format of a Typical ANOVA Table**

| **Sources** | **Degrees of Freedom (df)** | **Sum of Squares** | **Mean Square** | ***F* Statistics =**<br>**ESS /RSS** |
|---|---|---|---|---|
| Model | 1 | $\sum(\hat{Y}_i - \bar{Y})^2$ | ESS / df | |
| Error | *n-2* | $\sum(Y_i - \hat{Y}_i)^2$ | RSS / df | |
| Total | *n-1* | $\sum(Y_i - \bar{Y})^2$ | TSS / df | |

F = $\frac{ESS/(k-1)}{RSS/(n-k)}$ gives the observed value. The *F*-critical value at (*k-1*) and (*n-k*) degrees of freedom can be located from the statistical table. When the computed *F* is > than *F*-critical, the null hypothesis is rejected. Since the alternative hypothesis is accepted, the inference is that the explanatory variable plays a crucial role in determining the dependent variable. Similarly, when the *F* computed is < than the *F*-critical, the null hypothesis is not rejected. In this case, the hypothesis that the explanatory variable plays no role in determining the dependent variable is accepted. Again, here also, we can base our inference based on the *p*-value. This means if $p < 0.05$, we reject the null hypothesis.

**Check Your Progress 2** [answer questions in about 50-100 words]

1) What is ment by the ‘test of significance approach’ to hypothesis testing?

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

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

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

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

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

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

2) What does the 'level of significance' indicate?

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

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

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

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

3) What purpose does an ANOVA serve?

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

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

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

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

4) Distinguish between *t*-test in a regression model.

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

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

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

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

## 5.6 GAUSS-MARKOV THEOREM

This is an important theorem which gives us the condition under which the least squares estimator is the best estimator. When the assumptions of the classical linear regression model are not violated, the least-squares estimator fulfils certain optimum properties. These properties are summarised in the Gauss-Markov theorem which is stated as follows:

**Gauss-Markov Theorem:** Given the assumptions of classical linear regression model, the least-squares estimators, have minimum variance, in the class of all unbiased linear estimators, i.e., they are BLUE [best linear unbiased estimator(s)]. The characteristic of BLUE implies that the estimator obtained by the OLS method has the following properties.

a) It is *linear*, i.e., the estimator is a linear function of a random variable (such as the dependent variable $Y$ in the regression model).
b) It is *unbiased*, i.e., its average or expected value is equal to true value [E $(b_2) = \beta_2$].
c) It has *minimum variance* in the class of all such linear unbiased estimators. In other words, such an estimator with the least variance is an efficient estimator.

Thus, in the context of regression, the OLS estimators are BLUE. This is the essence of Gauss-Markov Theorem.

## 5.7 PREDICTION

So far we have spoken about estimation of population parameters. In the two variable model, we derived the OLS estimators of the intercept ($\beta_1$) and slope ($\beta_2$) parameters. Prediction refers to estimation of the dependent value at a particular value of the independent variable. In other words, we use the estimated regression model to predict the value of $Y$ corresponding to a given value of *X*.

Prediction is important to us for two reasons: First, it helps us in policy formulation. On the basis of the econometric model, we can find out the impact of changes in the explanatory variable on the dependent variable. Second, we can find out the robustness of our estimated model. If our econometric model is correct, the error between forecast value and actual value of the dependent variable should be small. Prediction could be of two types, as mentioned below.

### 5.7.1 Individual Prediction

If we predict an individual value of the dependent variable corresponding to a particular value of the explanatory variable, we obtain the individual prediction. Let us take a particular value of X, say X = $X_{0.}$ Individual prediction of Y at X = $X_0$ in obtained by:

$$Y_0 = \beta_1 + \beta_0 X_0 + u_0 \qquad \dots (5.13)$$

We know that $b_1$ and $b_2$ are unbiased estimators of $\beta_1$ and $\beta_2$. Hence, $\hat{Y}_0$ is an unbiased predictor of E (Y | $X_0$).

Therefore,

$$\hat{Y}_0 = b_1 + b_2 X_0 \qquad \dots (5.14)$$

Since $\hat{Y}_0$ is an estimator, the actual value $Y_0$ will be different from $\hat{Y}_0$, and there will be certain 'prediction error'.

The prediction error in $[\hat{Y}_0 - Y_0]$ is given by

$$\hat{Y}_0 - Y_0 = (b_1 + b_2 X_0) - (\beta_1 + \beta_2 X_0 + u_0) \qquad \dots (5.15)$$

We can re-arrange the terms in equation (5.15) to obtain

$$\hat{Y}_0 - Y_0 = (b_1 - \beta_1) + (b_2 - \beta_2) X_0 - u_0$$

Let us take expected value of (5.15).

$$\mathrm{E}(\hat{Y}_0 - Y_0) = E(b_1 - \beta_1) + E(b_2 - \beta_2) X_0 - E(u_0) \qquad \dots (5.16)$$

We know that $E(b_1) = \beta_1$, $E(b_2) = \beta_2$ and $E(u_0) = 0$.

*Thus, we find that expected value of prediction error is zero.*

Now let us find out the variance of the prediction error.

The variance of the prediction error,

$$\mathrm{V}(\hat{Y}_0 - Y_0) = V(b_1 - \beta_1) + V(b_2 - \beta_2)X_0$$

$$+2\, X_0\, cov(b_1 - \beta_1, b_2 - \beta_2) + V(u_0) \qquad \text{… (5.17)}$$

We know that

$$V(b_1) = \sigma^2 \frac{\Sigma X_i^2}{n\Sigma x_i^2} \qquad \text{… (5.18)}$$

$$V(b_2) = \frac{\sigma^2}{\Sigma x_i^2} \qquad \text{… (5.19)}$$

$$Cov(b_1, b_2) = -\bar{X}\left(\frac{\sigma^2}{\Sigma x_i^2}\right) \qquad \text{… (5.20)}$$

By combining the above three equations and re-arranging terms, we obtain

$$V\left(\hat{Y}_0 - Y_0\right) = \sigma^2\left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})}{\Sigma x_i^2}\right] \qquad \text{... (5.21)}$$

Thus, $Y_0$ follows normal distribution with mean $\beta_1 + \beta_0 X_0$ and variance $\sigma^2\left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma x_i^2}\right]$.

If we take estimator for $\sigma^2$, then we have

$$t = \frac{\hat{Y}_0 - (\beta_1 + \beta_0 X_0)}{SE(\hat{Y}_0)} \qquad \text{... (5.22)}$$

On the basis of (5.22) we can construct confidence interval for $\hat{Y}_0$. Since

$t = \frac{\hat{Y}_0 - E(Y/\alpha_0)}{SE(\hat{Y}_0)}$, we have

$$P\left[-t_{\alpha/2} \le t \le t_{\alpha/2}\right] = 1 - \alpha$$

Thus, the confidence interval of $\hat{Y}_0$ is

$$P\left[(b_1 + b_2 X_0) - t_{\alpha/2} SE\left(\hat{Y}_0\right) \le (\beta_1 + \beta_2 X_0) \le (b_1 + b_2 X_0) + t_{\alpha/2} SE\left(\hat{Y}_0\right)\right] = 1 - \alpha \qquad \text{…(5.23)}$$

Let us lok into equation (5.21) again. We see that the variance of $\hat{Y}_0$ increases with $(X_0 - \bar{X})^2$. Thus, there is an increase in variance if $X_0$ is farther away from $\bar{X}$, the mean of the sample on the basis of which $b_1$and $b_2$are computed. In Fig. 5.4 we depict the confidence interval for $\hat{Y}_0$ (see the dotted line)

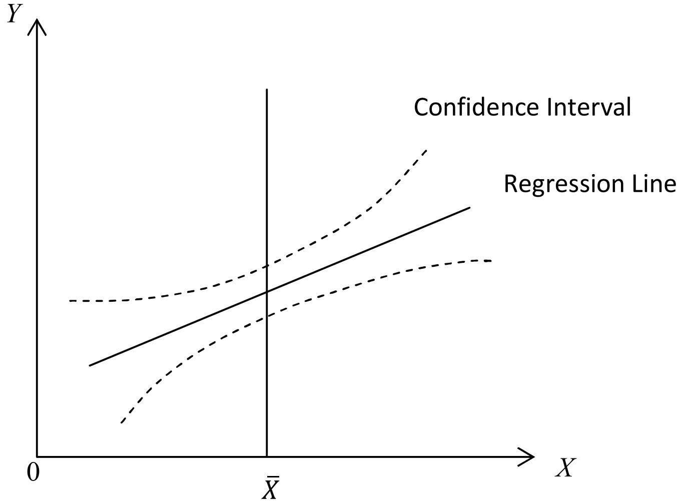


**Fig. 5.4:Confidence Interval for Individual Prediction**

### 5.7.2 Mean Prediction

It refers to prediction of expected values of $Y_0$, not the individual value. In other words, we are predicting the following:

$$\hat{Y}_0 = b_1 + b_2X_0$$

Thus the error term $u_0$ is not added.

In the case of mean prediction, the prediction error in $\left[\hat{Y}_0 - Y_0\right]$ is given by

$$\hat{Y}_0 - Y_0 = (b_1 + b_2X_0) - (\beta_1 + \beta_2X_0) \qquad \ldots (5.24)$$

We can re-arrange the terms in equation (5.24) to obtain

$$\hat{Y}_0 - Y_0 = (b_1 - \beta_1) + (b_2 - \beta_2)X_0$$

If we take the expected value of (5.24)

$$\mathrm{E}(\hat{Y}_0 - Y_0) = E(b_1 - \beta_1) + E(b_2 - \beta_2)X_0 \qquad \ldots (5.25)$$

Thus, we find that expected value of prediction error is zero.

Now let us find out the variance of the prediction error in thecase of mean prediction.

The variance of the prediction error,

$$\mathrm{V}(\hat{Y}_0 - Y_0) = V(b_1 - \beta_1) + V(b_2 - \beta_2)X_0$$
$$+2\,X_0\,cov(b_1 - \beta_1, b_2 - \beta_2) \qquad \ldots (5.26)$$

If we compare equations (5.17) and (5.26) we notice an important change – the term $V(u_0)$ is not there in (5.26). Thus the variance of the prediction error in the case of mean prediction is less compared to individual prediction. There is a change in the variance of $\hat{Y}_0$ in the case of mean prediction, however. Variance of the prediction error, in the case of mean prediction is given by

$$V\left(\hat{Y}_0 - Y_0\right) = \sigma^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})}{\Sigma x_i^2}\right] \qquad \text{... (5.27)}$$

Again, there is an increase in the variance of prediction error if $X_0$ is farther away from $\bar{X}$, the mean of the sample on the basis of which $b_1$and $b_2$are computed. It will look somewhat like the confidence interval we showed in Fig. 5.4, but the width of the confidence interval will be smaller.

An inference we draw from the above is that we can predict or forecast the value of the dependent variable, on the basis of the estimated regression equation, for a particular value of the explanatory variable ($X_0$). The reliability of our forecast, however, will be lesser if the particular value of X is away from $\bar{X}$.

**Check Your Progress 3** [answer questions within the given space in about 50-100 words]

1) State Gauss-Markov Theorem.

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

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

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

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

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

2) Differentiate between the two types of prediction possibilities in forecasting.

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

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

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

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

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

## 5.8 LET US SUM UP

This unit explains how to make inference on the estimated results of a simple regression model. After presenting an account of hypothesis testing to recapitulate the basics, it explains the two approaches for deciding on the validation of estimated results. The two methods are: confidence interval approach and test of significance approach. The testing of overall significance of the model is explained by the technique of ANOVA. Here, the application of $F$ – statistic is explained. The assumptions of classical linear regression model leads to the estimated parameters enjoying some unique properties. In light of this, the estimates are called BLUE (best linear unbiased estimates). This fact is stated in a result called the Gauss Markov theorem. The unit concludes with a detailed account of the concept of forecasting. This is once again a technique in which we have presented a confidence interval wherein the predicted or forecasted value of the dependent variable is shown to lie.

## 5.9 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

### Check Your Progress 1

1) In case of both the null hypothesis (, i.e., for true population parameters of slope and intercept), the null hypothesis are a simple hypothesis, whereas the alternative hypothesis are composite. The former is usually equated to zero (unless equated to a known value) and the latter in stated in inequality terms. The latter is also known as **two-sided hypothesis** when stated in 'not equal to' terms**.** It is considered one sided if stated in > or < terms.

2) False. It is the alternative hypothesis that decides whether it is composite or one-sided hypothesis. If the alternative hypothesis is stated as not equal to zero then it is composite or two-tailed test. Otherwise, i.e., if the alternative hypothesis is stated in positive or negative terms, then it will be a one-sided test.

3) The confidence interval approach is a method of testing of hypothesis. It refers to the probability that a population parameter falls within the set of critical values drawn from the Table.

4) We say that the hypothesised value is contained in the interval because the value of the interval depends upon the sample or the data used for estimation. The true population parameter value is fixed but the interval changes depending on the sample.

### Check Your Progress 2

1) The test of significance approach is another method of testing of hypothesis. The decision to accept or reject $H_0$ is made on the basis of the value of test statistic obtained from the sample data. This test statistic is given by: $t = \frac{b_2 - \beta_2}{SE(b_2)}$ and it follows $t-$distribution with (n −1 d.f.)

2) It is a measure of the strength of evidence when the null hypothesis is rejected It concludes that the effect is statistically significant. It is the probability of rejecting the null hypothesis when it is true. This is a grave error to commit and hence is chosen in a small measure like 1% or 5%.

3) Analysis of Variance (ANOVA) is a technique or a tool used to analyse the given data in two ways or direction. One is attributed to the deterministic factors, also called the explained part or the systematic part. The other is called the random or the unexplained part. This method of analysis of variance method was developed by Ronald Fisher in 1918.

4) The *t*-test is used to test the significance of estimated individual coefficients. It is distributed as *t* with ($k$ – 1) degrees of freedom (d.f.). where $k$ is the number of parameters estimated including the intercept term. Thus, for a simple linear regression, it is $[n – (2 – 1)] = (n –1)$. The *F*-distribution is used for testing the significance of the whole model. It has two parameters. The d.f. for a *F* test, in general is ($k$ – 1) and ($n$ – $k$). *K* includes the intercept term. Hence, in a simple linear regression, the d.f. for F is: (*2* – 1) and ($n$ – 2) or 1

and $(n - 2)$ Note that in a simple linear regression, the *t* test and the *F* test are equivalent because the number of independent variable is only one.

### Check Your Progress 3

1) The Gauss-Markov theorem states that the Ordinary Least Squares (OLS) estimators are also the best linear unbiased estimator (BLUE). The presence of BLUE property implies that the estimator obtained by the OLS method retains the following properties: (i) it is linear, i.e., the estimator is a linear function of a random variable such as the dependent variable Y in the regression model; (ii) it is unbiased, i.e., its average or expected value is equal to the true value in the sense that $E\ (b_2) = \beta_2$; (iii) it has minimum variance in the class of all such linear unbiased estimators. Such an estimator with the least variance is also known as an efficient estimator.

2) Prediction implies predicting two types of values: prediction of conditional mean, i.e., $E\ (Y \mid X_0) \rightarrow$ a point on the population regression line. This is called as the Mean Prediction. Prediction of individual Y value, corresponding $f\ (X_0)$ is called the Individual Prediction.

# UNIT 6 EXTENSION OF TWO VARIABLE REGRESSION MODELS[*]

**Structure**

## 6.0 OBJECTIVES

After going through this Unit, you should be in a position to

- interpret regression models passing through the origin;
- explain the impact of changes in the unit of measurement of dependent and independent variables on the estimates;
- interpret parameters in semi-log and log-linear regression models; and
- identify the correct functional form of a regression model.

## 6.1 INTRODUCTION

In the previous two Units we have discussed how a two variable regression model can be estimated and how inferences can be drawn on the basis of the estimated regression equation. In this context we discussed about the ordinary least squares (OLS) method of estimation. Recall that the OLS estimators are the best linear unbiased estimators (BLUE) in the sense that they are the best in the class of linear regression models.

The two variable regression model has the function as follows:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad \ldots (6.1)$$

[*] Prof. Kaustuva Barik, Indira Gandhi National Open University, New Delhi

where $Y$ is the dependent variable and $X$ is the independent variable. We added a stochastic error term $(u_i)$ to the regression model. We cited three reasons for inclusion of the error term in the regression model: (i) it takes care of the excluded variables in the model, (ii) it incorporates the unpredictable human nature into the model, and (iii) it absorbs the effects measurement error, incorrect function form, etc.

We assumed that the regression model is correctly specified. All relevant variables are included in the model. No irrelevant variable is included in the regression model. In this Unit we will continue with the two variables case as in the previous two units. We also continue with the same assumptions, as mentioned in Unit 4.

Let us look into the regression model given at equation (6.1). We observe that the regression model is linear in parameters. We do not have complex forms of the parameters such as $\beta_2^2$ or $\beta_1\beta_2$ as parameters. Further, the regression model is linear variables. We do not have $X^2$ or log $X$ as explanatory variable. Can we have these sorts of variables in a regression model? How do we interpret the regression model if such variables are there? We will extend the simple regression model given in equation (6.1) and explain how the interpretation of the model changes with the modifications.

## 6.2 REGRESSION THROUGH THE ORIGIN

Let us look into the simple regression model given at equation (6.1). There are two parameters in the regression model: $\beta_1$ and $\beta_2$. The intercept parameter is $\beta_1$ and the slope parameter is $\beta_2$. The intercept $\beta_1$ indicates the value of the dependent variable when the explanatory variable takes the value zero, i.e., $E(Y_0|X_0) = \beta_1$.

Suppose regression model takes the following form:

$$Y_i = \beta_2 X_i + u_i \qquad \ldots (6.2)$$

In equation (6.2) there is only one slope parameter, $\beta_2$. There is no intercept. The implication is that the regression line passes through the origin. The population regression function is $Y = \beta_2 X_i + u_i$ and the sample regression function is $Y_i = b_2 X_i + e_i$.

Now let us apply OLS method and find out the OLS estimator $b_2$. As you know from Unit 4, in OLS method we minimise the error sum of squares (ESS). Thus we minimise

$$ESS = \sum e_i^2 = \sum (Y_i - b_2 X_i)^2 \qquad \ldots (6.3)$$

We take derivative of the ESS and equate it to zero.

$$\frac{d\sum e_i^2}{db_2} = 0 \qquad \ldots (6.4)$$

$$\frac{d\sum e_i^2}{db_2} = 2\sum (Y_i - b_2 X_i)(-X_i) = 0 \qquad \ldots (6.5)$$

This implies

$-2\sum e_i(X_iY_i - b_2X_i^2) = 0$

$\sum X_iY_i - b_2\sum X_i^2 = 0$

$b_2 = \frac{\sum X_iY_i}{\sum X_i^2}$ … (6.6)

The estimator given at (6.6) is unbiased. The variance of the estimator is given by

$var(b_2) = \frac{\sigma^2}{\Sigma X_i^2}$ … (6.7)

Let us compare the above estimator with the estimator for the regression model $Y_i = \beta_1 + \beta_2X_i + u_i$ (see equation (4.18) in Unit 4)

$b_2 = \frac{\Sigma x_iy_i}{\Sigma x_i^2}$ … (6.8)

and

$var(b_2) = \frac{\sigma^2}{\Sigma x_i^2}$ … (6.9)

Note that in equation (6.6) the variables are not in deviation form. Thus when we do not have an intercept in the regression model, the estimator of the slope parameter is different from that of a regression model with intercept. Both the estimators will be the same if and only if $\bar{X} = 0$.

We present a comparison between the regression model with intercept and without intercept in Table 6.1.

**Table 6.1: Features of Regression Model without Intercept**

| Regression Model with Intercept | Regression Model without Intercept |
|---|---|
| $b_2 = \frac{\sum x_iy_i}{\sum x_i^2}$ | $b_2 = \frac{\sum X_iY_i}{\sum X_i^2}$ |
| $\text{var}(b_2) = \frac{\sigma^2}{\sum x_i^2}$ | $\text{var}(b_2) = \frac{\sigma^2}{\sum X_i^2}$ |
| $\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$ | $\hat{\sigma}^2 = \frac{\sum e_i^2}{n-1}$ |
| $R^2$ is non-negative | $R^2$ can be negative |

The estimated regression model is given as

$\hat{Y}_i = b_2X_i$ … (6.10)

Note that the coefficient of determination $R^2$ is not appropriate for regression models without the intercept. If the intercept in a regression model is not statistically significant, then we can have regression through the origin. Otherwise, it leads to specification error. There is omission of a relevant variable.

## 6.3 CHANGES IN MEASUREMENT UNITS

Suppose you are given time series data on GDP and total consumption expenditure of India for 30 years. You are asked to run a regression model with consumption expenditure as dependent variable and income as the independent variable. The objective is to estimate the aggregate consumption function of India. Suppose you took GDP and Consumption Expenditure in Rs. Crore. The estimated regression equation you found is

$$Y_i = 237 + 0.65X_i \qquad \dots (6.11)$$

When you presented the results before your seniors, they pointed out that the measure of GDP and consumption expenditure should have been in Rs. Million, so that it is comprehensible outside India also. If you re-estimate the results by converting the variables, will estimates be the same? Or, do you expect some changes in the estimates? Let us discuss the issue in details.

Suppose we transform both the dependent and independent variables as follows:

$$Y_i^* = w_1 Y_i \text{ and } X_i^* = w_1 X_i \qquad \dots (6.12)$$

The regression model (6.1) can be transformed as follows:

$$Y_i^* = \beta_1 + \beta_2 X_i^* + u_i \qquad \dots (6.13)$$

Estimation of equation (6.13) by OLS method gives us the following estimators

$$b_1^* = \bar{Y}^* - b_2^* \bar{X}^* \qquad \dots (6.14)$$

$$b_2^* = \frac{\sum x_i^* y_i^*}{\sum x_i^{*2}} \qquad \dots$$

(6.15)

In a similar manner you can find out the variance of $b_1^*$ and $b_2^*$, and the estimator of the error variance.

From equation (6.15) we can find out that

$$b_2^* = \frac{w_1}{w_2} b_2 \qquad \dots (6.16)$$

and

$$b_1^* = w_1 b_1 \qquad \dots (6.17)$$

Now let us look into the implications of the above.

(i) Let us begin with the dependent variable, $Y_i$ . Suppose $Y_i$ is doubled ($w_1 = 2$) and $X_i$ is unchanged ($w_2 = 1$). What will happen to $b_1$ and $b_2$? Substitute the values of $w_1$ and $w_2$ in equations (6.16) and (6.17). We find that both the estimates are doubled. Thus, if the dependent variable is multiplied by a constant *c*, then all OLS coefficients will be multipled by *c*.

(ii) Now let us take the case of the independent variable. Suppose $X_i$ is doubled ($w_2 = 2$) and $Y_i$ is unchanged ($w_1 = 1$). On substitution of the values of $w_1$ and $w_2$ in equations (6.16) and (6.17) we find that the slope coefficient ($b_2$) is halved, but the intercept ($b_1$) remains unchanged.

(iii) If we double both the variables $X_i$ and $Y_i$, then the slope coefficient ($b_2$)will remain unchanged, but the intercept will change. Remember that the intercept is changed by a change in the scale of measurement of the dependent variable.

Now the question arises: Will there be a change in the t-ratio and F-value of the model? No, the *t* and *F* statistics are not affected by a change in the scale of measurement of any variable.

**Check Your Progress 1**

1) Under what condition should we run a regression through the origin?

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

2) What are the implications of a regression model through origin?

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

3) What are the implications on the estimates if there is a change in the measurement scale of the explanatory variable?

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

4) What are the implications on the estimates if there is a change in the measurement scale of the dependent variable?

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

## 6.4 SEMI-LOG MODELS

In some of the cases the regression model is non-linear, but by taking logarithm on both sides of the regression equation, we get a linear model. If a model is non-linear , but becomes linear after transformation of its variables, then the model is said to be intrinsically linear . Thus, semi-log and log-linear models are intrinsically linear models. We discuss about the semi-log model in this section. We will discuss about the log-linear model in the next section.

Let us begin with a functional form as follows:

$$Y_t = e^{\beta_1+\beta_2 X_t+u_t} \qquad \ldots (6.18)$$

This regression model, in its present form, is non-linear . Therefore, it cannot be estimated by OLS method. However, if we take natural logs of both the sides, we obtain

$$\ln Y_t = \beta_1 + \beta_2 t + u_t \qquad \ldots (6.19)$$

It transforms into a semi-log equation. It is called a semi-log model as one of the variables is in log form.

If we take $\ln Y_t = Y_t^*$, then equation (6.19) can be written as

$$Y_t^* = \beta_1 + \beta_2 t + u_t \qquad \ldots (6.20)$$

Estimation of equation (6.20) is simple. The equation is linear in parameters and in variables. Thus, we can apply OLS method to estimate the parameters. The implication of the regression model (6.20), however, is much different from the regression model (6.1).

If we take the differentiation of the regression model $Y_t = \beta_1 + \beta_2 X_t + u_t$, we obtain

$$\frac{dY}{dt} = \beta_2 \qquad \ldots (6.21)$$

Equation (6.21) shows that the slope of the regression equation is constant. An implication of the above is that the absolute change in the dependent variable for unit increase in the independent variable is constant throughout the sample. If there is an increase in X by one unit, Y increases by $\beta_2$ unit.

Now let us consider the regression model $\ln Y_t = \beta_1 + \beta_2 t + u_t$ . If we take differentiation of equation (6.19) we find that

$$\frac{d\ln Y_t}{dt} = \beta_2$$

which means

$$\frac{1}{Y_t}\frac{dY_t}{dt} = \beta_2 \qquad \ldots (6.22)$$

An implication of equation (6.22) is that the slope of the regression model is variable. Thus its interpretation is different from that of the regression model $Y_t = \beta_1 + \beta_2 X_t + u_t$.

For equation (6.19), we interpret the slope coefficient ($\beta_2$) as follows: For every unit increase in *X*, there is $\beta_2$ per cent increase *Y*. Thus, for a semi-log model the change in the dependent variable in terms of percentages. The semi-log model is useful is estimating growth rates.

## 6.5 LOG-LINEAR MODELS

Let us consider the following regression equation:

Let us take the case of the following non-linear model

$$Y = \beta_1 X^{\beta_2} \qquad \dots (6.23)$$

This model will be intrinsically linear if it can be transformed into

$$Y^* = \beta_1 + \beta_2 X^* + u \qquad \dots (6.24)$$

Using the logarithm of each of the variable in equation (6.23), we get the following transformed equation:

$$\ln Y_i = \beta_1 + \beta_2 \ln X_i + u_i \qquad ... (6.25)$$

The regression model given at (6.25) is called log-linear model (because it is linear in logs of the variables) or double-log model (because both variables are in log form).

Let us take differentiation of equation (6.25) with respect to $X_i$

$$\frac{d(\ln Y_i)}{dX_i} = \frac{1}{Y_i} . \frac{\mathrm{d}Y_i}{dX_i} \qquad ... (6.26)$$

$$\frac{\mathrm{d}Y_i}{dX_i} = \frac{\beta_2}{X_i} \qquad ... (6.27)$$

By combining equations (6.26) and (6.27) we find that

$$\frac{\mathrm{d}Y_i}{dX_i} = \frac{Y_i}{X_i}\beta_2$$

Or,

$$\frac{\mathrm{d}Y_i}{dX_i}\frac{X_i}{Y_i} = \beta_2 \qquad ... (6.28)$$

A closer look at equation (6.28) shows that the slope parameter represents the elasticity between Y and X.

This attractive feature of the log-linear model has made it popular in applied work. The slope coefficient $\beta_2$ measures the elasticity of *Y* with respect to *X*, that is, the percentage change in *Y* for one per cent change in *X*. Thus, if *Y* represents the quantity of a commodity demanded and *X* its unit price, then $\beta_2$ measures the price elasticity of demand.

## 6.6 CHOICE OF FUNCTIONAL FORM

By you would have observed that the two variable regression model could have three functional forms as given below.

(I) $Y_i = \beta_1 + \beta_2 X_i + u_i$

(II) $lnY_i = \beta_1 + \beta_2 X_i + u_i$

(III) $\ln Y_i = \beta_1 + \beta_2 \ln x_i + u_i$

A question arises: which one is the best model? The choice of functional form depends on our objective. We should choose the model that gives us relevant answer to our queries. Suppose our objective is to estimate the impact of change in the independent variable on the dependent variable. In this case we can use model-I. On the other hand, if our objective is to estimate growth rate in the dependent variable as a result of the change in the independent variable, we should opt for semi-log model (model II). If our objective is to estimate elasticity between two variables, we choose the log-linear model.

The three regression models (Models –I, II, III) will give different estimates of the parameters. The standard error of the estimators will also be different. Further, the coefficient of determination, $R^2$, will be different for all three models. Can we compare the $R^2$ of the models and say that the model with the highest $R^2$ is the best fit? We cannot compare the value of $R^2$ obtained from regression models with different dependent variables. However, we can compare $R^2$ of regression models with the same dependent variable and the same estimation method. Thus the $R^2$ value of Model-I and Model-II cannot be compared. We can compare Model-II and Model-III in terms of their best fit.

If two regression models are almost similar in terms of their coefficient of determination, statistical significance of estimators and diagnostic checking (to be discussed in Units 13 and 14), we prefer the simpler model. The simpler model is easier to comprehend and usually accepted by others.

The log-linear regression model has certain advantages: (i) the parameters are invariant to change of scale since they measure percentage changes, (ii) the model gives elasticity figures directly, and (iii) the model moderates the problem of heteroscedasticity to some extent (see Unit 11 for the problem of heteroscedasticity).

**Check Your Progress 2**

1) In a semi-log model how do you interpret the slope coefficient?

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

2) Describe how the slope parameter of a log-linear regression model is estimated.

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

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

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

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

2) What are the advantages of the log-linear model?

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

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

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

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

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

3) What is meant by intrinsically linear model? Can you compare the results of an intrinsically linear model with that of a linear model? Why or why not?

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

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

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

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

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

## 6.7 LET US SUM UP

In this Unit we discussed about the functional forms that can be accommodated in a two variable regression model. We began with the regression model passing through the origin (there is no intercept). We pointed out the impact of changes in the scale of measurement of variables. Subsequently we considered three functional forms: the original model, the semi-log model and the log-linear model. The interpretations of the parameters in all three functional forms have been discussed in the Unit.

## 6.8 ANSWERS TO CHECK YOUR PROGRESS EXERCISES

**Check Your Progress 1**

1) The exclusion of intercept term from a regression model has serious implication. It should be omitted only when the intercept term in the unrestricted model is statistically not significant.

2) We have listed the implications of the omission of the intercept term in table 6.1. Go through it and answer.

3) When there is a change in the measurement scale of the explanatory variable the concerned estimate is affected. If X is multiplied by *c*, the parameter is divided by *c*.

4) If Y is multiplied by c, all parameters in the model are multiplied by c.

## Check Your Progress 2

1) In a semi-log model the slope parameter indicates growth rate. If there is 1 unit increase in the value of X, the expected value of Y increases by β per cent.

2) The estimation of the log-linear model is the same as the simple regression model, except that the variables are transformed. Write down the steps followed in estimation of a regression model.

3) We have mentioned three advantages in the text: (i) the parameters are invariant to change of scale since they measure percentage changes, (ii) the model gives elasticity figures directly, and (iii) the model lessens the problem of heteroscedasticity to some extent.

4) You cannot compare the results of two regression models unless the dependent variable is the same.