Unit II : The Two Variable Linear Model

Economic theory is concerned with the relations between various parts of the economy. Most of the economic phenomena may be formulated as a relationship between economic variables. Demand function, cost function, production function are examples of some of such relations.

A relation between x and y is defined as a set of all values of x and y which are characterized by a given equation. All relations can be classified as either deterministic or stochastic. A relation between x and y which is characterized as $y = f(x)$ will be deterministic if for each value of x there is only one corresponding value of y. On the other hand, a relation between x and y is said to be stochastic if for each value of x there is a whole probability distribution of values of y. This means that, for any given value of x the variable y may assume some specific value or fall within some specific interval. In economic theory all relations are, as a rule, stated in a deterministic form.

Suppose we have a statement that consumption depends upon income. This proposition can be expressed in mathematical form as $y = f(x)$ where y indicates the dependent (explained) variable and x the independent (explanatory) variable.

It shows that a change in x causes a change in y and a change in y is due to a change in x or there is a cause and effect relationship between the two.

Several forms of the function are common in use in economics. The simplest one is that the linear relationship between the two variable. The linear form of the relationship will be as follows.

$$Y = \alpha + \beta X \quad \text{---} \quad (1)$$

$$Y = \alpha X^{\beta}$$

$$\Rightarrow \log Y = \log \alpha + \beta \log X \quad \text{---} \quad (2)$$

$$Y = \exp\left\{\alpha + \beta \frac{1}{X}\right\}$$

$$\log Y = \alpha + \beta \frac{1}{X} \quad \text{---} \quad (3)$$

These functions are linear because its graph will always be a straight line. EConometrics is concerned with testing the above form of relations and with estimating parameters 'α' and 'β'.

Linear Regression Model

In econometrics we deal exclusively with stochastic relations. The most simple type of stochastic relation between two variables x and Y is a linear one. This model is called a simple linear regression model and is described as

$$Y_i = \alpha + \beta X_i + U_i$$

where Y is the dependent variable; x is the independent variable; u is the 'stochastic disturbance' or 'error term' and α and β are the parameters of the model. The subscript 'i' refers to the i^{th} observation.

Error term

Our economic theory formulates exact functional relationship among the variables. But dealing with common data even an ordinary investigator will feel that all observations do not fall exactly on a straight line. The best we can expect is that the observed quantities will be closer to the line, that is, why our regression model requires extension and introduction of stochastic disturbance term. The introduced term is known as disturbance or error term, because it represents the effect of all those factors which are not suspected by the investigator.

For example, we are studying Consumption function. Let Consumption and income be denoted as Y and X respectively. Further suppose that we have 100 samples which are divided into 10 sub-groups on the basis of income levels. we have $X_1, X_2, \cdots X_{10}$ income levels whose Consumption levels are $Y_1, Y_2, \cdots Y_{10}$.

In our daily life, we do not expect that all families within one sub-group having X_1 income level will display an identical Consumption Y_1. Some will spend more while some others will spend less. But all values will vary around a central figure Y_1. Hence disturbance term is introduced which will show the deviations of actual value from central value. Here, in our example, Consumption of different families of sub-group one will be $Y_1 + u_1$, $Y_1 + u_2$, $Y_1 + u_3$ and so on. Error term may have positive or negative values and is drawn at random.

Reasons for introducing error term in econometric models

There are several justifications for the introduction of error term in our model. Some of them are as follows:

1. We may have error of specification. As we know the economic world is very complex in its nature. At any point of time a single economic variable is affected by number of variables. Economic theory includes only important variables because it is neither possible nor desirable to include all variables in our functional relationship. In such situation, omitted variables may influence the dependent variable which can lead to error. Even the same variables have different behaviour in different environments. Consumption of each household of the same income level group is bound to be different from central value of consumption for that level of income. If we want to introduce all variables, our efforts of estimation will be in vain. No statistical method is yet discovered for estimation which could measure the effects of unlimited variables. The net effect of all these factors is represented by error term in regression model.

2. Another type of error which arises in our model is sampling error. If chosen samples are not representative and unbiased the sampling error may cause in our model. As in our example of Consumption function, if observations of Consumption expenditure are predominated by poor families, the actual Consumption expenditure is bound to differ from estimated value.

3. Measurement also causes error in the model. If investigator has collected the wrong data, the observed outcome will contain two ingredients; the theoritical prediction and experimental error.

Due to all above reasons we introduce a stochastic disturbance term in regression model.

Statistical Assumptions in Linear Model

The regression model includes not only the regression equation but also a specification of the probability distn. of the distubance and a statement indicating how the values of the explanatory variables are determined. This information is provided by the basic assumptions. These are as follows:

i) Normality

First assumption is that distubance term is distributed normally.

ii) Zero mean

The error term has zero mean

ie., $E(u_i) = 0$ for $i = 1, 2, 3, \ldots n$

This means that for each value of x, some values of u_i are greater than zero while some are smaller than zero, but the average value of u_i would always be equal to zero.

iii) Homoscedasticity

Every disturbance has the same variance whose value is unknown.

ie., $E(u_i^2) = \sigma^2$ for $i = 1, 2, \ldots, n$.

ie., for all values of x_i, the u's will show the same dispersion around their mean.

iv) Non- auto regression

The various disturbance terms are uncorrelated

ie, $E(u_i \cdot u_j) = 0$ for $i \neq j$; i and $j = 1, 2, 3, \cdots n$

ie. all the co-variances of u_i, u_j are equal to zero. The value which the random term assume in one period does not depend upon the value which it assumes in any other period

v) Non- Stochastic

$X - X_i$ is a non-stochastic variable with fixed values in repeated samples and such that for any sample size $\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$ is a finite value different from Zero.

The specific model is given below. For.

For a household with income X_i the average or expected expenditure is given by $\alpha + \beta X_i$. The actual expenditure will be $\alpha + \beta X_i + u_i$, where u_i is a random drawing from $N(0, \sigma_u^2)$. The complete mathematical specification of the model is

$$Y_i = \alpha + \beta X_i + u_i \quad , \quad i = 1, 2, \cdots, n \qquad \text{—— (1)}$$

$$E(u_i) = 0 \quad \text{for all } i \qquad \text{—— (2)}$$

$$E(u_i u_j) = \begin{cases} 0 & i \neq j, \text{ all } i, j \\ \sigma_u^2 & i = j, \text{ all } i, j \end{cases} \qquad \text{—— (3)}$$

$$P(u_i) = N(0, \sigma_u^2) \quad \text{for all } i \qquad \text{—— (4)}$$

Assumptions (2), (3) & (4) are a more extensive way of stating $u \sim NID(0, \sigma_u^2)$

The three unknown parameters of the model are α, β, and σ_u^2 and these may be estimated by using Least squares method.

Least Square Estimation

The main object of constructing statistical relationship is to predict or explain the effects on one dependent variable resulting from the changes in one or more explanatory variables. The first method of estimating the relationship is termed as least square estimation method. Under the least square criterion the line of best fit is said to be that which minimizes the sum of the squared residuals between the points of the graph and the points of straight line.

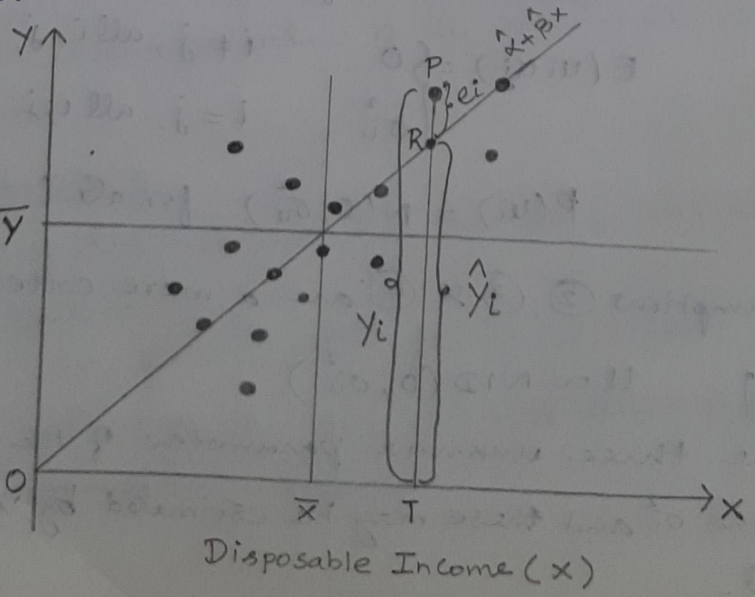
Let 'n' observations on Y and X are denoted as $Y_1, Y_2, Y_3, \ldots, Y_n$ and $X_1, X_2, X_3, \ldots, X_n$ and their arithmetic mean will be

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \quad \text{and} \quad \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

The estimated line is denoted as

$$\hat{Y} = \hat{\alpha} + \hat{\beta} X \qquad \text{——— (1)}$$

where $\hat{\alpha}, \hat{\beta}$ are the estimated parameters and $\hat{Y}$ is the estimated value of Y. As we know all observations do not fall exactly on straight line, the scatter diagram will be as follows:



Take any point 'P' on scatter diagram, here $OT = X_i$; $PT = Y_i$; $RT = \hat{Y}_i$. The difference between actual and estimated value of Y is the error.

$$e_i = Y_i - \hat{Y}_i = PR \quad \text{——} (2)$$

These deviations of actual values from estimated line will be positive or negative. Squares of these residuals will be positive. Estimated line changes with a change in $\hat{\alpha}$ and $\hat{\beta}$. So we can derive relationship between residual and parameters. The relationship will be

$$\sum_{i=1}^{n} e_i^2 = f(\hat{\alpha}, \hat{\beta})$$

The principle of least squares is to choose such values of $\hat{\alpha}$ and $\hat{\beta}$ that will minimize the sum of squared deviations. Now we have

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y})^2 \qquad \text{from } (2)$$

$$\text{or} \quad \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2 \qquad \text{from } (1)$$

Differentiating w.r.t. $\hat{\alpha}$ and $\hat{\beta}$ respectively and equate to zero, we get

$$\frac{\partial}{\partial \hat{\alpha}} \left[\sum_{i=1}^{n} e_i^2 \right] = 2 \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i)(-1) = 0$$

$$\Rightarrow -2 \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i) = 0 \quad \text{——} (3)$$

$$\text{and} \quad \frac{\partial}{\partial \hat{\beta}} \left[\sum_{i=1}^{n} e_i^2 \right] = 2 \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i)(-X_i) = 0$$

$$\Rightarrow -2 \sum X_i (Y_i - \hat{\alpha} - \hat{\beta} X_i) = 0 \quad \text{——} (4)$$

From (3), we get $$\sum_{i=1}^{n} Y_i = n\hat{\alpha} + \hat{\beta} \sum_{i=1}^{n} X_i \quad \text{——} (5)$$

From (4), we get $$\sum_{i=1}^{n} X_i Y_i = \hat{\alpha} \sum_{i=1}^{n} X_i + \hat{\beta} \sum_{i=1}^{n} X_i^2 \quad \text{——} (6)$$

The equations (5) & (6) are known as the least squares normal equations. With the help of these two normal equns/. the values of α and β can be determined and the linear equ... can be obtained.

① Estimate the demand curve from the following data

Price	5	6	7	8	9	11
Sales	12	17	19	24	30	29

Soln:

Let price be X and Sales Y, then the actual relationship to be estimated will be

$$Y_i = \alpha + \beta X_i + u_i$$

∴ estimated relationship will be

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} x_i$$

Price x	Sales y	x^2	xy
5	12	25	60
6	17	36	102
7	19	49	133
8	24	64	192
9	30	81	270
11	29	121	319
46	131	376	1076

The two normal equations are

$$\Sigma y = n\hat{\alpha} + \hat{\beta}\Sigma x \quad \text{—— (1)}$$

$$\Sigma xy = \hat{\alpha}\Sigma x + \hat{\beta}\Sigma x^2 \quad \text{—— (2)}$$

On putting the computed values in the above equations

$$131 = 6\hat{\alpha} + \hat{\beta}\cdot 46 \quad \text{—— (3)}$$

$$1076 = \hat{\alpha}\cdot 46 + \hat{\beta}\cdot 376 \quad \text{—— (4)}$$

on Solving (3) & (4), we get

(3) × 46, $276\hat{\alpha} + 2116\hat{\beta} = 6026$

(4) × 6, $276\hat{\alpha} + 2256\hat{\beta} = 6456$

$$140\hat{\beta} = 430$$

$$\Rightarrow \hat{\beta} = 3.07$$

Put $\hat{\beta} = 3.07$ in (3), we get

$$6\hat{\alpha} + (46 \times 3.07) = 131$$

$$6\hat{\alpha} = 131 - 141.22$$

$$\hat{\alpha} = \frac{-10.22}{6} = -1.7$$

∴ The estimated regression equation is

$$\hat{y} = -1.7 + 3.07x$$

⊙ From a sample of 200 pairs of observations, the following quantities were calculated:

$\Sigma x = 11.34$, $\Sigma y = 20.72$, $\Sigma x^2 = 12.16$, $\Sigma y^2 = 84.96$, $\Sigma xy = 22.13$.

Estimate the two regression lines and the variance of estimated reg. Co-eff of Y on X.

Soln:

The estimated regression line of Y on X is

$$\hat{y} = \hat{\alpha}_1 + \hat{\beta}_1 x \quad \text{where} \quad \hat{\beta}_1 = \frac{\Sigma xy - \frac{\Sigma x \cdot \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}$$

$$\hat{\beta}_1 = \frac{22.13 - \frac{11.34 \times 20.72}{200}}{12.16 - \frac{(11.34)^2}{200}} = \frac{20.96}{11.52} = 1.82$$

On passing the equation through means

$$\bar{x} = \frac{\Sigma x}{N} = \frac{11.34}{200} = 0.0567$$

and $$\bar{y} = \frac{\Sigma y}{N} = \frac{20.72}{200} = 0.1036$$

$$\bar{y} = \hat{\alpha}_1 + \hat{\beta}_1 \bar{x} \quad (\text{or}) \quad \hat{\alpha}_1 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\therefore \hat{\alpha}_1 = 0.1036 - 1.82(0.0567) = 0.0004$$

Thus regression line of Y on X is

$$\hat{y} = 0.0004 + 1.82x$$

$\hat{y} = \hat{\alpha}_1 + \hat{\beta}_1 x$

Taking Σ & $\div n$ on both sides,

$$\frac{\Sigma y}{n} = \frac{\Sigma \hat{\alpha}_1}{n} + \hat{\beta}_1 \frac{\Sigma x}{n}$$

$$\bar{y} = \hat{\alpha}_1 + \hat{\beta}_1 \bar{x}.$$

The other regression line of x on y is

$$\hat{x} = \hat{\alpha}_2 + \hat{\beta}_2 y \quad \text{where} \quad \hat{\beta}_2 = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma y^2 - \frac{(\Sigma y)^2}{n}}$$

On putting the values $\hat{\beta}_2 = \dfrac{22.13 - \frac{11.34 \times 20.72}{200}}{84.96 - \frac{(20.72)^2}{200}}$

$$= \frac{20.96}{82.81} = 0.253$$

On passing the line through means

$$\bar{x} = \hat{\alpha}_2 + \hat{\beta}_2 \bar{y} \quad (\text{or}) \quad \hat{\alpha}_2 = \bar{x} - \hat{\beta}_2 \bar{y}$$

$$= 0.0567 - 0.253 (0.1036)$$

$$= 0.0305$$

Thus the reg. line of x on y is

$$\hat{x} = 0.0305 + 0.253 y.$$

$$\text{Var}(\hat{\beta}_1) = \frac{s^2}{\Sigma x_i^2} \quad \text{where} \quad s^2 = \frac{\Sigma e^2}{n-2}$$

$$= \frac{1}{n-2} \left[\Sigma y^2 - \beta_1 \Sigma xy \right]$$

$$= \frac{1}{200-2} \left[82.81 - (1.82 \times 20.96) \right]$$

$$= \frac{1}{198} \left[82.81 - 38.14 \right]$$

$$= 0.225$$

$$\Sigma x^2 = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 11.52$$

$$\Sigma y^2 = \Sigma y^2 - \frac{(\Sigma y)^2}{n} = 82.81$$

$$\Sigma xy = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 20.96$$

$$\therefore V(\hat{\beta}_1) = \frac{s^2}{\Sigma x_i^2} = \frac{0.225}{11.52} = 0.019$$

Properties of Least Square Estimator

(i) The least square estimators are unbiased

(ii) The O.L.S. are linear estimators

(iii) The O.L.S. estimators are best estimators. In short the O.L.S. estimators are Best Linear Unbiased Estimators (BLUE)

Property 1

Statement:

The least square estimators are unbiased

ie., $E(\hat{\beta}) = \beta$.

Proof:

The mathematical specification of the model is

$$Y_i = \alpha + \beta X_i + u_i \quad \& \quad \bar{Y} = \alpha + \beta \bar{X} + \bar{u}$$

Taking deviations from actual means, we get

$$Y_i - \bar{Y} = \beta(X_i - \bar{X}) + (u_i - \bar{u})$$

$$y_i = \beta x_i + (u_i - \bar{u}) \quad \text{where } y_i = Y_i - \bar{Y} \ \& \ x_i = X_i - \bar{X}$$

We know that, if the deviations are taken from actual mean, the regression co-efft of Y on x is given by

$$\hat{\beta}_{yx} = \frac{\Sigma x_i y_i}{\Sigma x_i^2} \quad \text{where } x_i = X_i - \bar{X} \ \& \ y_i = Y_i - \bar{Y}$$

By putting the value of y_i to this equation, we get

$$\hat{\beta}_{yx} = \frac{\Sigma x_i [\beta x_i + (u_i - \bar{u})]}{\Sigma x_i^2}$$

$$(\text{or}) \quad \hat{\beta} = \beta \frac{\Sigma x_i^2}{\Sigma x_i^2} + \frac{\Sigma x_i u_i}{\Sigma x_i^2} - \bar{u} \frac{\Sigma x_i}{\Sigma x_i^2}$$

Since error term has zero mean, we have

$$\hat{\beta} = \beta + \frac{\Sigma x_i u_i}{\Sigma x_i^2} \quad \text{——— (1)}$$

On taking expectation of both sides, we get

$$E(\hat{\beta}) = \beta + E\left(\frac{\sum x_i u_i}{\sum x_i^2}\right)$$

$$= \beta + E(u_i)\frac{\sum x_i}{\sum x_i^2}$$

$$= \beta \qquad \because E(u_i) = 0$$

Thus $E(\hat{\beta}) = \beta$

$\therefore$ The estimator $\hat{\beta}$ is an unbiased estimator of β.

Property 2

Statement

The O.L.S. are Linear estimators

Proof:

We know that $\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$ where $y_i = Y_i - \bar{Y}$ & $x_i = X_i - \bar{X}$

$$\therefore \hat{\beta} = \frac{\sum x_i (Y_i - \bar{Y})}{\sum x_i^2}$$

$$= \frac{\sum x_i Y_i}{\sum x_i^2} - \frac{\bar{Y}\sum x_i}{\sum x_i^2}$$

$$= \frac{\sum x_i Y_i}{\sum x_i^2} \quad \text{since } \sum x_i = \sum (X_i - \bar{X}) = 0$$

$$= \frac{\sum x_i}{\sum x_i^2} Y_i \quad (\text{Since } x_i \text{ indicates the fixed set})$$

Let $\frac{x_i}{\sum x_i^2} = k = \text{Constant}$

Now $\hat{\beta} = \sum k Y_i$

Thus $\hat{\beta}$ is a linear one.

Property 3

Statement

The O.L.S estimators are best estimators

Proof:

We know that $\hat{\beta} = \beta + \frac{\Sigma x_i u_i}{\Sigma x_i^2}$ from equn. ① of Property 1.

$$\hat{\beta} - \beta = \frac{\Sigma x_i u_i}{\Sigma x_i^2}$$

We know that $V(\hat{\beta}) = E(\hat{\beta} - \beta)^2$

$$= E\left(\frac{\Sigma x_i u_i}{\Sigma x_i^2}\right)^2$$

$$V(\hat{\beta}) = E\left[\frac{1}{(\Sigma x_i^2)^2}\left(x_1^2 u_1^2 + x_2^2 u_2^2 + \cdots + x_n^2 u_n^2 + 2x_1 x_2 u_1 u_2 + \cdots + 2x_{n-1} x_n u_{n-1} u_n\right)\right]$$

$$= \frac{1}{(\Sigma x_i^2)^2}\left[x_1^2 E(u_1^2) + x_2^2 E(u_2^2) + \cdots + x_n^2 E(u_n^2) + 2x_1 x_2 E(u_1 u_2) + \cdots + 2x_{n-1} x_n E(u_{n-1} u_n)\right]$$

Since $E(u_i^2) = \sigma_u^2$ and $E(u_i u_j) = 0$, hence

$$V(\hat{\beta}) = \frac{1}{(\Sigma x_i^2)^2}\left(x_1^2 \sigma_u^2 + x_2^2 \sigma_u^2 + \cdots + x_n^2 \sigma_u^2\right)$$

$$= \frac{\sigma_u^2 \Sigma x_i^2}{(\Sigma x_i^2)^2}$$

$$= \frac{\sigma_u^2}{\Sigma x_i^2}$$

$\therefore V(\hat{\beta}) \leq V(u)$

ie. $\frac{\sigma_u^2}{\Sigma x_i^2} < \sigma_u^2$

Hence $\hat{\beta}$ is the best estimator.

Theorem:

P.T. the OLS estimators are Best Linear Unbiased Estimators (BLUE)

Proof:

Consider any arbitrary estimator of β

$$\hat{\beta}^* = \sum c_i Y_i$$

where $c_i = \frac{x_i}{\sum x_i^2} + d_i$ and d_i is any arbitrary constant

(or) $\hat{\beta}^* = \sum c_i(\alpha + \beta x_i + u_i)$

$= \alpha \sum c_i + \beta \sum c_i x_i + \sum c_i u_i$

Taking expectation of both sides,

$E(\hat{\beta}^*) = E[\alpha \sum c_i + \beta \sum c_i x_i]$

If $\sum c_i = 0$ and $\sum c_i x_i = 1$, then and then only

$$E(\hat{\beta}^*) = \beta$$

These conditions can be fulfilled if

$\sum d_i = 0$ and $\sum d_i x_i = \sum d_i x_i = 1$

Variance of $\hat{\beta}^*$ in that case will be

$Var(\hat{\beta}^*) = E(\hat{\beta}^* - \beta)^2 = E\left[(\sum c_i u_i)^2\right]$ using eqn ① & Property 1.

Solving this as in previous case, we get

$Var(\hat{\beta})^* = \sigma_u^2 \sum c_i^2$ where $c_i = \frac{x_i}{\sum x_i^2} + d_i$

$\therefore Var(\hat{\beta})^* = \sigma_u^2 \left[\frac{\sum x_i^2}{(\sum x_i^2)^2} + \sum d_i^2 + 2 d_i \frac{\sum x_i}{\sum x_i^2}\right] = \frac{x_i^2}{(\sum x_i^2)^2} + d_i^2 + 2\frac{x_i}{\sum x_i^2} d_i$

$= \sigma_u^2 \left[\frac{1}{\sum x_i^2} + \sum d_i^2\right]$ since $\frac{\sum x_i}{\sum x_i^2} = 0$

$= \frac{\sigma_u^2}{\sum x_i^2} + \sigma_u^2 \sum d_i^2$

$= V(\hat{\beta}) + \sigma_u^2 \sum d_i^2$

If $d_i = 0$, then only $V(\hat{\beta})^* = V(\hat{\beta})$, otherwise it would be greater than $V(\hat{\beta})$. But d_i is any arbitrary constant

$$\therefore V(\hat{\beta})^* \geq V(\hat{\beta})$$

Hence the OLS estimator has least variance than any other estimator. So OLS estimators are Best Linear Unbiased estimators. (BLUE).

Variance of the Random variable 'u'

Theorem:

$\hat{\sigma}_u^2 = \frac{\sum e_i^2}{n-2}$ is an unbiased estimate of the true Variance of u. ie. $E(\hat{\sigma}_u^2) = \sigma_u^2$.

Proof:

The variance of the residuals ($e_i = y_i - \hat{y}$) is defined as the expected value of the squared differences of e_is from their mean.

ie. $\text{Var}(e) = E[e_i - E(e)]^2$

$= E(e_i^2)$ since $E(e) = 0$

We know that

$e_i = Y_i - \hat{Y}_i$, $y_i = Y_i - \bar{Y}$ & $\hat{y}_i = \hat{Y}_i - \bar{Y}$

$\Rightarrow Y_i = y_i + \bar{Y}$ $\Rightarrow \hat{Y}_i = \hat{y}_i + \bar{Y}$

Substituting these values in e_i, we get

$$e_i = (y_i + \bar{Y}) - (\hat{y}_i + \bar{Y})$$

$$e_i = y_i - \hat{y}_i \quad \text{——— (1)}$$

But we know that

$$Y_i = \alpha + \beta x_i + u_i$$

$$\& \quad \bar{Y} = \alpha + \beta \bar{x} + \bar{u}$$

and $y_i = Y_i - \bar{Y} = \beta(x_i - \bar{x}) + (u_i - \bar{u})$ (2)

$= \beta x_i + (u_i - \bar{u})$ ——— (2) where $x_i = x_i - \bar{x}$

III^ly we know that $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ and

$\bar{y} = \hat{\alpha} + \hat{\beta} \bar{x}$

Hence $\hat{y}_i = \hat{y}_i - \bar{y}$

$= \hat{\beta}(x_i - \bar{x}) = \hat{\beta} x_i$ ——— (3)

Put (2) & (3) in (1), we get

$e_i = \beta x_i + (u_i - \bar{u}) - \hat{\beta} x_i$

$= (u_i - \bar{u}) - (\hat{\beta} - \beta) x_i$

$e_i^2 = [(u_i - \bar{u}) - (\hat{\beta} - \beta) x_i]^2$

$= (u_i - \bar{u})^2 + [(\hat{\beta} - \beta)^2 x_i^2] - 2(\hat{\beta} - \beta)(u_i - \bar{u}) x_i$

Taking sum of both sides,

$\Sigma e_i^2 = \Sigma (u_i - \bar{u})^2 + (\hat{\beta} - \beta)^2 \Sigma x_i^2 - 2(\hat{\beta} - \beta) \Sigma x_i (u_i - \bar{u})$

Taking expectation on both sides,

$E[\Sigma e_i^2] = E[\Sigma (u_i - \bar{u})^2 + (\hat{\beta} - \beta)^2 \Sigma x_i^2 - 2(\hat{\beta} - \beta) \Sigma x_i (u_i - \bar{u})]$ ——— (4)

The RHS terms may be rearranged as follows

a) $E[\Sigma (u_i - \bar{u})^2] = E[\Sigma u_i^2 - 2\bar{u} \Sigma u_i + n\bar{u}^2]$

$= E\left[\Sigma u_i^2 - 2 \frac{\Sigma u_i}{n} \times \Sigma u_i + n\left(\frac{\Sigma u_i}{n}\right)^2\right]$

$= E\left[\Sigma u_i^2 - 2 \frac{(\Sigma u_i)^2}{n} + \frac{(\Sigma u_i)^2}{n}\right]$

$= E\left[\Sigma u_i^2 - \frac{(\Sigma u_i)^2}{n}\right]$

$= \Sigma E(u_i^2) - \frac{1}{n} E[\Sigma u_i]^2$

$= n\sigma_u^2 - \frac{1}{n} E(u_1 + u_2 + \cdots + u_n)^2 \quad \because E(u_i^2) = \sigma_u^2$

$= n\sigma_u^2 - \frac{1}{n} E\left[\Sigma u_i^2 + 2 \sum_{i \neq j} u_i u_j\right]$

$= n\sigma_u^2 - \frac{1}{n}\left[\Sigma E(u_i^2) + 2 \sum_{i \neq j} E(u_i u_j)\right]$

$= n\sigma_u^2 - \frac{1}{n} n\sigma_u^2 - \frac{2}{n} \Sigma E(u_i u_j)$

$= n\sigma_u^2 - \sigma_u^2 \quad [\text{Given } E(u_i u_j) = 0]$

$= \sigma_u^2 (n-1)$ ——— (5)

b) $E\left[(\hat{\beta}-\beta)^2 \Sigma x_i^2\right] = \Sigma x_i^2 E(\hat{\beta}-\beta)^2$

Given that the x's are fixed in all samples. But we know that

$E[\hat{\beta}-\beta]^2 = Var(\hat{\beta}) = \sigma_u^2 \frac{1}{\Sigma x^2}$ from Property 3.

$\therefore E\left[(\hat{\beta}-\beta)^2 \Sigma x_i^2\right] = \Sigma x_i^2 \sigma_u^2 \frac{1}{\Sigma x_i^2} = \sigma_u^2$ — (6)

c) $E\left[(\hat{\beta}-\beta)\Sigma x_i (u_i - \bar{u})\right]$

$= E\left[(\hat{\beta}-\beta)(\Sigma x_i u_i - \bar{u}\Sigma x_i)\right]$

But $\hat{\beta} = \beta + \frac{\Sigma x_i u_i}{\Sigma x_i^2}$ from (1) of Property 1

$\Rightarrow \hat{\beta}-\beta = \frac{\Sigma x_i u_i}{\Sigma x_i^2}$

$\therefore E\left[(\hat{\beta}-\beta)\Sigma x_i (u_i-\bar{u})\right] = E\left[\frac{\Sigma x_i u_i}{\Sigma x_i^2}(\Sigma x_i u_i - \bar{u}\Sigma x_i)\right]$

$= E\left[\left(\frac{\Sigma x_i u_i}{\Sigma x_i^2}\right)(\Sigma x_i u_i) - 0\right]$ since $\Sigma x_i = \Sigma(x_i - \bar{x}) = 0$

$= E\left[\frac{(\Sigma x_i u_i)^2}{\Sigma x_i^2}\right]$

$= E\left[\frac{\Sigma x_i^2 u_i^2 + 2\sum_{i \neq j}(x_i x_j)(u_i u_j)}{\Sigma x_i^2}\right]$

$= E\left[\frac{\Sigma x_i^2 E(u_i^2)}{\Sigma x_i^2} + \frac{2\sum_{i \neq j}(x_i x_j) E(u_i u_j)}{\Sigma x_i^2}\right]$

$= \frac{\Sigma x_i^2 E(u_i^2)}{\Sigma x_i^2} + 0 \quad (\because E(u_i u_j) = 0$

$= E(u_i^2)$

$= \sigma_u^2$ ——— (7)

Put ⑤, ⑥ & ⑦ in ④, we get

$$E[\Sigma e_i^2] = \sigma_u^2(n-1) + \sigma_u^2 - 2\sigma_u^2$$

$$= n\sigma_u^2 - \sigma_u^2 + \sigma_u^2 - 2\sigma_u^2$$

$$= (n-2)\sigma_u^2$$

$$\Rightarrow E\left[\frac{\Sigma e_i^2}{n-2}\right] = \sigma_u^2$$

Defining $\hat{\sigma}_u^2 = \frac{\Sigma e_i^2}{n-2}$, we may write

$$E(\hat{\sigma}_u^2) = \sigma_u^2$$

Thus $\frac{\Sigma e_i^2}{n-2}$ is an unbiased estimate of the true variance of 'u'.

Note: As we Know that

$$\text{Var}(\hat{\beta}) = \sigma_{\hat{\beta}}^2 = \frac{\sigma_u^2}{\Sigma x_i^2} \quad \text{(Proof of property 3)}$$

$$S_{\hat{\beta}}^2 = \frac{S^2}{\Sigma x_i^2} = \frac{\Sigma e_i^2}{(n-2)\Sigma x_i^2}$$

$$\text{(or)} \quad S_{\hat{\beta}} = \sqrt{\frac{\Sigma e_i^2}{(n-2)\Sigma x_i^2}}$$

Hypothesis Testing

One of the prime objective of econometrics is to analyse data in a manner that allows us to test and evaluate the constructed model. Hypothesis testing constitutes an important part of the analysis and will result either in rejection or acceptance of the model. If data are consistent with the model, then model is implicitly accepted.

t - test

Null Hypothesis

Ho: ~~There is no significant difference between the variables.~~
ie., the relationship between x and y is Non-linear.
ie., Ho: $\beta = 0$ Vs $H_1: \beta \neq 0$

Level of Significance

$\alpha = 5\%$ or 1%

Critical region

At $\alpha = 5\%$ level for $\gamma = n-2$ d.f $t_{0.05}$
if $|t| \leq$ table value, accept Ho
if $|t| > t_{0.05}$ (table value), reject Ho.

Test statistic

$$t = \frac{\hat{\beta} - \beta}{s/\sqrt{\Sigma x_i^2}} = \frac{\hat{\beta} - \beta}{\sqrt{\frac{\Sigma e_i^2}{n-2}}} \times \sqrt{\Sigma x_i^2}$$

and since null hypothesis states $\beta = 0$,

$$t = \frac{\hat{\beta}}{\sqrt{\Sigma e_i^2/n-2}} \times \sqrt{\Sigma x_i^2}$$

with $\gamma = n-2$ d.f.

Conclusion

If $|t| \leq t_{0.05}$, accept Ho
$|t| > t_{0.05}$, reject Ho.

Confidence Interval

Confidence interval provides a range of values which are likely to contain the true regression co-efficients. At a given level of significance the confidence interval is constructed so that the probability that the interval contains the true coefficient is $1-\alpha$.

The estimation of confidence for β, say at a 95% level, indicates that in a repeated samples there would be a tendency to include the true value of β in the interval 95% of the time.

Confidence interval will be constructed according to the following formula

$$\hat{\beta} \pm t_{\alpha/2} S_{\hat{\beta}}$$

where α = (100 - level of confidence)

Similarly the confidence interval for a values of $\hat{\alpha}$ is given by

$$\hat{\alpha} \pm t_{(n-\alpha)/2} S_{\hat{\alpha}}$$

where $S_{\hat{\alpha}} = \frac{S_{yx}}{n} \sqrt{\frac{\Sigma x^2}{\Sigma (x-\bar{x})^2}}$ and $S_{yx} = \sqrt{\frac{\Sigma (y-\hat{y})^2}{n}}$

(or) $S_{yx} = \sigma_y \sqrt{1-r^2}$

① Using the following data, estimate the regression line $Y = \alpha + \beta x$, test the hypothesis that $\beta = 0$ Vs $\beta < 0$ at 5% level of significance, also construct 95% Confidence interval for β.

Investment (Y):	65	57	57	54	66
Change in Output (x):	26	13	16	−7	27

Soln:

The estimated line is $y = \hat{\alpha} + \hat{\beta} x$

On passing the equation through means,

$$\bar{y} = \hat{\alpha} + \hat{\beta}\bar{x} \quad \text{where} \quad \hat{\beta} = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}$$

x	y	x^2	xy	$\hat{y} = \hat{\alpha} + \hat{\beta} x$ $= 54.55 + 0.35x$	$e = y - \hat{y}$	e^2
26	65	676	1690	63.65	1.35	1.82
13	57	169	741	59.10	−2.10	4.41
16	57	256	912	60.15	−3.15	9.92
−7	54	49	−378	52.10	1.90	3.61
27	66	729	1782	64.00	2.00	4.00
75	299	1879	4747			23.76

$$\hat{\beta} = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}} = \frac{4747 - \frac{75 \times 299}{5}}{1879 - \frac{(75)^2}{5}} = \frac{4747 - 4485}{1879 - 1125} = \frac{262}{754} = 0.35$$

$$\bar{x} = \frac{75}{5} = 15 \quad \& \quad \bar{y} = \frac{299}{5} = 59.8$$

$$\bar{y} = \hat{\alpha} + \hat{\beta}\bar{x}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 59.8 - 0.35(15) = 54.55.$$

∴ Regression equation will be $\hat{y} = 54.55 + 0.35x$

Now we will test the Hypothesis.

Null Hypothesis

$H_0 : \beta = 0$ Vs $H_1 : \beta < 0$

Level of significance

$\alpha = 5\%$

Critical regim

At $\alpha = 5\%$, for $\gamma = n-2 = 5-2 = 3$ d.f.

if $|t| \leq 2.353$, accept H_0

$|t| > 2.353$, reject H_0.

Test Statistic

$$t = \frac{\hat{\beta}}{\sqrt{\frac{\Sigma e_i^2}{n-2}}} \sqrt{\Sigma x_i^2}$$

where $\Sigma x_i^2 = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$.

$$= \frac{0.35}{\sqrt{\frac{23.76}{3}}} \times \sqrt{754}$$

$$= 3.416.$$

Conclusion

Since $|t| = 3.416 > 2.353$, we reject H_0. ie, $\beta < 0$ is accepted.

Confidence interval at 95% level is

$$\hat{\beta} \pm t_{\alpha/2} S_{\hat{\beta}} \quad \text{where } S_{\hat{\beta}} = \sqrt{\frac{\Sigma e^2}{(n-2)\Sigma x^2}}$$

$$= 0.35 \pm (3.182 \times 0.102) \qquad = \sqrt{\frac{23.76}{3 \times 754}} = 0.102$$

$$= 0.35 \pm 0.325$$

$$= (0.025, 0.675)$$

Estimation of Elasticities from an Estimated Regression Line

Let us suppose that the estimated function

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

is the equation of a line whose intercept is $\hat{\alpha}$ and its slope is $\hat{\beta}$. The co-efficient $\hat{\beta}$ is the derivatives of $\hat{Y}$ with respect to X.

$$\text{ie., } \hat{\beta} = \frac{d\hat{Y}}{dx}$$

$\hat{\beta}$ is the change in $\hat{Y}$ as x changes with a very small amount. In the estimated line which is a linear consumption function the co-efficient $\hat{\beta}$ is not the income elasticity, but the component of the elasticity.

The income elasticity may be defined by

$$\eta_{iy} = \frac{dY/Y}{dx/x} = \frac{dy}{dx} \cdot \frac{x}{y}$$

Where η_{iy} = income elasticity

Y = quantity Consumed

X = income.

Clearly $\hat{\beta}$ is the component $\frac{dy}{dx}$. From an estimated linear line we obtain an average elasticity

$$\eta_{iy} = \hat{\beta} \cdot \frac{\bar{x}}{\bar{\hat{y}}} = \hat{\beta} \cdot \frac{\bar{x}}{\bar{y}}$$

Where $\bar{x}$ = the average income in the sample

$\bar{y}$ = average value of the quantity Consumed in the sample.

Estimation of a function whose intercept is zero

Let us suppose we have a regression line

$$Y = \alpha + \beta X + u$$

if the intercept is zero ie, $\alpha = 0$, then the formula for estimation of $\hat{\beta}$ becomes

$$\hat{\beta} = \frac{\Sigma XY}{\Sigma X^2}$$

~~① From the following data compute~~

Co-efficient of Determination (r^2) (Goodness of fit)

We are interested to know to what extent the line is good fit to observe the true relationship. The measure of the goodness of fit is the square of Correlation Co-efficient or r^2. This shows the percentage of total variation in the dependent variable which can be explained by the independent variable.

For example, if $r^2 = 0.80$, it means that the estimated regression line is able to explain 80 % of the total variation of dependent variable values around mean and remaining 20% of the total variation in Y is not accounted by regression line which can be attributed to the factors included in the error term.

ie.. $$r^2 = \frac{\Sigma \hat{y}_i^2}{\Sigma y_i^2}$$

(or) $$r^2 = 1 - \frac{\Sigma e^2}{\Sigma y^2}$$

Relationship between r^2 and the slope $\hat{\beta}$

The relationship which exists between r^2 and $\hat{\beta}$ is denoted as

$$r^2 = \hat{\beta} \frac{\Sigma xiyi}{\Sigma y_i^2}$$

① From the following data, Compute Co-efficient of determination (r^2) and price elasticity of demand.

Quantity:	69	76	52	56	57	77	58	55	67	53	72	64
Price :	9	12	6	10	9	10	7	8	12	6	11	8

Solm:

Let x be the price and y be the Quantity. The regression line of y on x is assumed as $y = \alpha + \beta x$

y	x	$x = x - \bar{x}$	$y = y - \bar{y}$	x^2	xy	y^2	$\hat{y}$	e	e^2
69	9	0	6	0	0	36	63	6	36
76	12	3	13	9	39	169	72.75	3.25	10.56
52	6	-3	-11	9	33	121	53.25	-1.25	1.56
56	10	1	-7	1	-7	49	66.25	-10.25	105.06
57	9	0	-6	0	0	36	63.00	-6	36.00
77	10	1	14	1	14	196	66.25	10.75	115.56
58	7	-2	-5	4	10	25	56.50	1.50	2.25
55	8	-1	-8	1	8	64	59.75	-4.75	22.56
67	12	3	4	9	12	16	72.75	-5.75	33.06
53	6	-3	-10	9	30	100	53.25	-0.25	0.063
72	11	2	9	4	18	81	69.50	2.5	6.25
64	8	-1	1	1	-1	1	59.75	4.25	18.06
756	108			48	156	894			

$$\bar{x} = \frac{\Sigma x}{N} = \frac{108}{12} = 9 \quad ; \quad \bar{y} = \frac{\Sigma y}{N} = \frac{756}{12} = 63$$

We know that $\hat{\beta} = \frac{\Sigma xy}{\Sigma x^2}$ where $x = X - \bar{X}$ & $y = Y - \bar{Y}$

$$= \frac{156}{48} = 3.25$$

$$\bar{y} = \hat{\alpha} + \hat{\beta}\bar{x} \Rightarrow \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

$$= 63 - (3.25)(9)$$

$$= 33.75$$

∴ The regression line of Y on X is

$$\hat{Y} = 33.75 + 3.25X$$

(i) Determination of r^2

We know that

$$r^2 = 1 - \frac{\Sigma e^2}{\Sigma y^2} = 1 - \frac{387.500}{894} = 1 - 0.432 = 0.568$$

e^2	y^2
6	36
3.25	10.56
1.25	1.56
10.25	105.06
6	36.00
10.75	115.56
2.25	5.06
4.75	22.56
5.75	33.06
0.25	0.063
2.5	6.25
4.25	18.06
387.500	

(ii) Price elasticity

$$\text{Price elasticity} = \frac{\bar{X}}{\bar{Y}} \cdot \frac{\Delta Y}{\Delta X}$$

$$= \hat{\beta} \cdot \frac{\bar{X}}{\bar{Y}}$$

$$= 3.25 \times \frac{9}{63}$$

$$= \frac{3.25}{7}$$

$$= 0.464$$