Econometrics

1.1 Introduction:

Most economic theory is concerned with relationships among variables and specifies exact functional relationships among its variables. For example the quantity demanded of a commodity by an individual is related to his income and price of the commodity; national consumption expenditure is related to national income, mortality and air pollution etc. . Also there are many situations where one has to consider time dependent phenomena such as monthly sales of news prints, stock prices, daily rain fall, etc. In these type of examples there is a dependent variable which may be explained by means of one or more exogenous factors. But in such cases it is not possible to write a deterministic model that allows for exact calculations of future behaviour of the phenomena. This is due to the fact that in such cases, the system is not fully specified or the character of much human behaviour is highly unpredictable. So exact functional relations connecting the variables are inadequate descriptions of economic behaviour. In other words there is always an element of uncertainty in any prediction of the system. Therefore to allow inexact relationship between the variables the econometrician introduces a random variable into the model which has well-defined probabilistic properties. A model defined in terms of random variables with certain consistently specified probability distribution is called a stochastic model or probabilistic model. A general stochastic model may include lagged variables to explain the endogenous dependence, regressors to explain the exogenous variables and an error term to introduce the stochastic nature to the phenomena.

Economic models that have a well-specified functional form and statistically defined error term are called econometric models. Institutional information, economic theory and other information are relied upon to formulate an econometric model. Econometrics is an integration of economics, Statistics and mathematics.

According to Samuelson, Koopmans and Stone, Econometrics may be defined as the quantitative analysis of actual economic phenomena based on concurrent development of theory and observations related by appropriate method of inference.

Econometrics gives empirical content to economic theory.

An econometric model consists of system of simultaneous equations connecting ‘k’ input variable or regressors (X) to ‘m ‘ output variables(Y – the response variable or endogenous variable). Each of these output variables is presumed to be generated by so-called structural relationship that comprises of its inputs not only some of the k primary inputs(X) of the

system but also some of the (m-1) out put variables(y) that are generated by other structural relationships.

Econometric model containing only one equation is the regression model.

1.2 Some Basic Aspects :

Types of Data: The type of data normally we come across are Cross-section data and time series data.

If the data on the characteristic of a variable is collected at a point of time then it is called cross section data. For example in a socio-economic survey data on consumption expenditure on food is collected on a number of households at a given point in time.

If the data on characteristic of a variable is collected over a period of time, then it is called time series data. For example, annual Industrial production of a particular firm over the years.

Types of Variables : There are two types of variables namely endogenous and exogenous variables. Endogenous variables are those determined within the economic system and exogenous variables are those determined outside the economic system. For example quantity(Q) demanded in a market is a function of price(P). Here quantity is endogenous and price is exogenous variable as w e do not have any control over the price and it is determined outside the economic system.

Endogenous variables are also called response variable or dependent variable or regressand and we denote this by 'Y'.

Exogenous variables are also called independent variables or regressors or explanatory variables or predetermined variables and we denote these variables by X1, X_2, ...,X_k .

Types of Relationships:

Linear Rerelation: Here we can come across two variable linear relationship and multiple linear relationship.

Two variable linear Model: In this case there is one independent variable X effecting the dependent variable Y and they are related through a relation

$Y=\alpha+\beta X+€$

Linearity is assumed to be in parameter not in variable. We can have X^2 in the linear relationship but not β^{-1} .

Many variable Linear relationship: In this case there is one dependent variable(Y) and more than one independent variables say X_1, X_2, …, X_k and they are said to have linear relationship if

$Y=\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + €$

The major uses of regression methodology can be placed in three broad categories: Prediction, Model specification, and parameter estimation and testing.

Model Specification: Regression analysis requires correct model specification. Using this model we assess the relative merit of individual predictor variables on the prediction of the response. One must ensure that all relevant predictor variables that effects the phenomena under study are included in the regression model with correct functional form for all the predictor variables.

For example one conducts an investigation into possible wage inequalities. Constructing a prediction equation for annual salaries could include predictor variables such as sex, race, job description, number of years employed, and other variables that are believed to be related to wages. One predictor variable that is difficult to define and measure in many employment situation is the quality of the work performed. Yet this might be one of the most important variable that influences on wage differentials. Particularly if one wishes to study the sex or race discrimination, this variable should be included. Similarly entering number of years employed as a variable in a prediction equation for annual salary should be carefully considered. In some employment categories, salary increases slowly during the first few years of employment and more rapidly thereafter. A linear term for wage may not be sufficiently accurate.

There is no hard and strict rule for the choice of a particular functional form for the prediction model related to a particular situation. However the form selected should be such that it is simple , compatible as much as possible to the real situation and posses good predictive power.

Commonly used functional forms are Linear relationships, semi logarithmic relationships and Doubly logarithmic relationships.

1.3 Regression Model

The regression curve is defined as the locus of conditional expectation of the response variable Y given the predictor variable X and it is a function of X. The regression function has the representation

$E(Y/X=x)=m(x)$.

Or

$$Y = m(x) + \varepsilon \qquad (1)$$

Here m(x) is the regression function and ϵ is the error or disturbance term. The disturbance term ϵ represents the sum effect of all those factors which influence the response variable

Y, but they are not included in the model. This error term is an unobservable random variable having well defined probability distribution. This error term brings stochastic nature to the phenomena. The regression function m(x) may be linear regression function or non-linear function.

The model in (1) is referred to as a population regression function. Estimate of the regression model (1)or fitted regression model using the sample data is the sample regression function

In case of linear regression model we have the following important models:

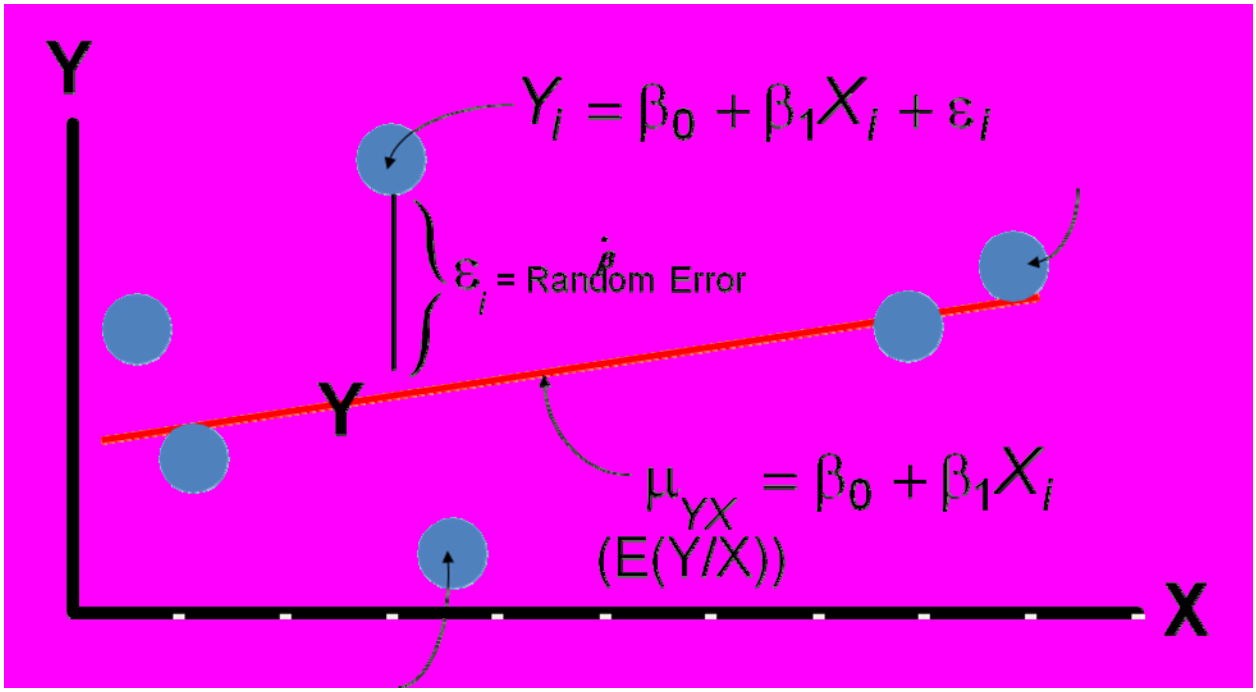
- Simple Linear regression model
- Multiple linear regression model

1.4 SIMPLE LINEAR REGRESSION MODEL:

In this case there is only one regressor (X) influencing the response variable Y and the mean response m(x) in model (1) becomes $m(x) = \alpha + \beta x$.

The simple linear regression model has the representation

$$Y_i = \alpha + \beta X_i + \varepsilon_i, i = 1,2,...n \qquad (2)$$



Where Y_i denotes the i-th observation on the dependent variable Y which could be for example consumption, investment or output, and X_i denotes the i^{th} observation on the independent variable or regressor which could be for example disposable income, the

interest rate or an input. n is the number of observations which could be number of households or firms in a cross-section or number of years if the observations are collected annually.

α and β are the intercept and slope of this simple linear relationship between Y and X. They are assumed to be unknown parameters to be estimated from the data. For example if Y is consumption and X is disposable income, now if we plot the data on a graph, not all the observations (Y_i, X_i) lie on the straight line $\alpha + \beta X$. The difference is due to the error ϵ_i.

This error may be due to-

- the omission of relevant factors that could influence consumption , other than disposable income like wealth, varying tastes, or unforeseen events that induce household to consume more or less,
- (ii) measurement error, which could be the result of households not reporting their consumption or income accurately,
- (iii) Wrong choice of the functional form i.e. linear relationship between consumption and income, when the true relationship may be non-lnear.
- Intrinsic randomness in economic data, which arises because human and organisational behaviour can by its very natur, never be captured fully in any mathematical model.

In the real life, α and β are not known and have to be estimated from the observed data $\{(Y_i, X_i), \quad i = 1,2,..., n\}$

1.5 Parameter Estimation: Regression analysis demands not only must the model be correctly specified and prediction be accurate but the data base must allow for good estimation . The regression model contains unknown parameters such as the regression coefficients and error variance. In order to have prediction with minimum prediction mean square error, the parameters to be estimated such that it possess desirable statistical properties such as, unbiased estimator, minimum variance unbiased linear estimator (BLUE), consistency, efficiency and asymptotic efficiency. Certain characteristics of a data base create problem (such as multicollinearity, autocorrelation, heteroscedasticity etc.) in the estimation of parameters. In such cases the regression model need to be corrected for such problems and then parameters are estimated using appropriate method of estimation.

Standard regression analysis commences by making the following ideal assumptions.

i) $E(\varepsilon_i) = 0, \quad \forall i \quad (no \quad specification \quad error);$ This insure that on the average we are on the true line.

ii) $E(\varepsilon_i \varepsilon_j) = 0, \quad i \neq j \qquad (absence \quad of \quad autocorrelation);$

iii) $V(\varepsilon_i) = \sigma^2, \quad \forall i \qquad (\text{hom}oscedasticity \quad);$ This insures that every observation is equally reliable.

iv) Either the regressor X is non stochastic i.e., X_i (i=1,2,..,n) are fixed in repeated samples or if the X_i are stochastic, then Cov(X_i, ϵ_i) =0, i=1,,2,...n.

v) ϵ_i is normally distributed for each i.

Estimatio by Least squares Method:

The genesis of regression lies in the attempt to obtain the truth from inconsistent observation by minimising some function of the errors.

Let us derive estimates of α and β in model (2) from n pairs of sample observations $\{(Y_i, X_i), \quad i = 1,2,...,n\}$. Assume that assumptions (i) to (v) hold. Plotting the observations in the X-Y plane, yields a 'scatter diagram'. We choose α and β such that the straight line Y=α +βX lies as close as possible to various points on the scatter diagram.

Least squares method and maximum likelihood(ML) method both of them concentrated on minimising the error sum of squares. Least squares and MLE of regression parameters coincide when the errors are normally distributed. The rapid development of the least squares method was largely due to its computational flexibility.

We use least squares principle to estimate α and β. According to this principle the sum of squares of the deviations of the observations from their expectation is minimum.

i.e $\sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2 = \sum_{i=1}^{n} \varepsilon_i^2 \; is \quad \min imum$ (3)

Thus we select the line (that is , the parameters α and β) such that the error sum of squares is minimum.

That is least squares estimator of the parameters α and β is obtained by minimising $ESS = \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2 \; with \quad respect \quad to \quad \alpha \; and \quad \beta.$

The first order condition for minimum are:

$$\frac{\partial (ESS)}{\partial \alpha} = \frac{\partial (ESS)}{\partial \beta} = 0$$

Which lead to the 'normal equations'

$$\sum_{i=1}^{n} Y_i = n\alpha + \beta \sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_i Y_i = \alpha \sum_{i=1}^{n} X_i + \beta \sum_{i=1}^{n} X_i^2$$

Their solutions leads to the following estimates of α and β.

$\hat{\alpha} = \overline{Y} - \hat{\beta}\,\overline{X}$ and

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \qquad (4)$$

Where $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \quad \overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}, \quad x_i = X_i - \overline{X} \quad and \quad y_i = Y_i - \overline{Y}$

$\hat{\alpha} \quad and \quad \hat{\beta}$ are called ordinary least squares(OLS) estimators.

The line of *best fit* is given by

$\hat{Y} = \hat{\alpha} + \hat{\beta} X$

For a given value of X_i, the best fit line yields the predicted value of Y, denoted as $\hat{Y}_i$

$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ (5)

1.6 Goodness of fit of the model:

The disturbance term in the regression model is unobservable. The residual acts as a proxy variable for the error term.

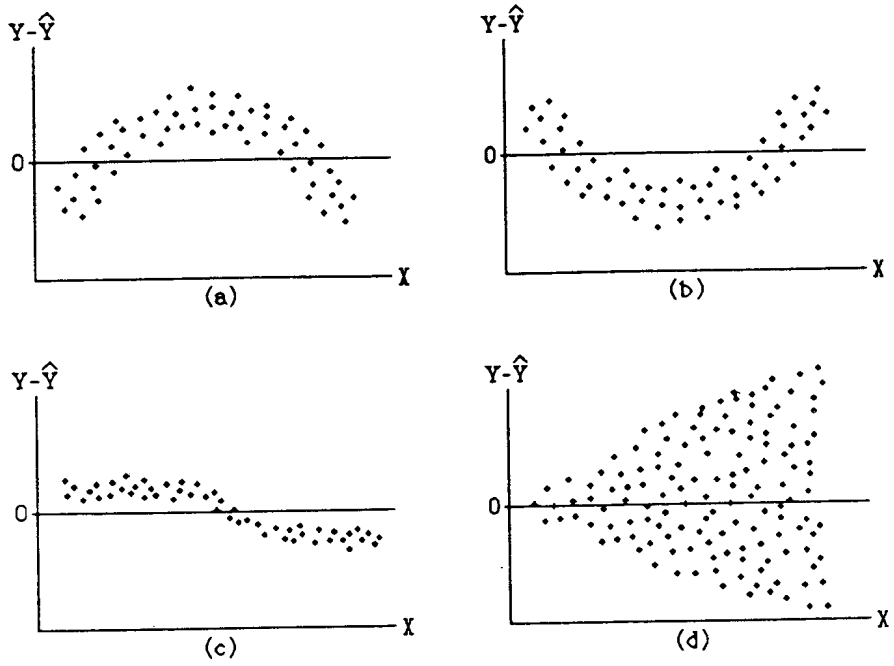
The OLS residual is defined as $e = Y - \hat{Y}$.

Actual observation Y_i depart from the value predicted by the line of best fit,

$$e_i = Y_i - \hat{Y}_i, \quad i = 1,2,...,n$$

Or $Y_i = \hat{\alpha} + \hat{\beta} X_i + e_i, \quad i = 1,2,\ldots,n.$ (6)

The OLS residual plays crucial role in examining the goodness of fit of the model ,overall significance of the model and also violation of the basic ideal assumptions



The plot of the residuals which shows a certain pattern like a set of positive residuals followed by a set of negative residuals as in (c) may be indicative of a violation of one of the basic ideal conditions.

If we fit a linear regression line to a true quadratic relation between Y and X then a scatter of residual like in (a) or(b) will be generated.

Figure (d) shows the residual variation growing with X. This indicates the violations of constant variance assumption.

Definition(Sample regression function): The relation(6) is termed the sample regression function(SRF).

Note: The PRF is the hypothesized relation between Y and X, whereas the SRF is an estimate of this relation, based on the sample.

$\sum_{i=1}^{n} e_i = 0 \Rightarrow \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$ if model contains an intercept term. Further

$$\sum_{i=1}^{n} e_i x_i = 0 \quad and \quad \sum e_i \hat{Y}_i = 0$$

This means OLS residuals are uncorrelated with the predicted values from the egression and also uncorrelated with the regressors.

We can write equation (6) as

$$e_i = Y_i - \hat{Y}_i, \quad i = 1,2,...,n$$

Or $e_i = Y_i - \hat{Y}_i, \quad i = 1,2,...,n$

$$= Y_i - \hat{\alpha} - \hat{\beta} X_i \quad , \quad i = 1,2,...,n.$$

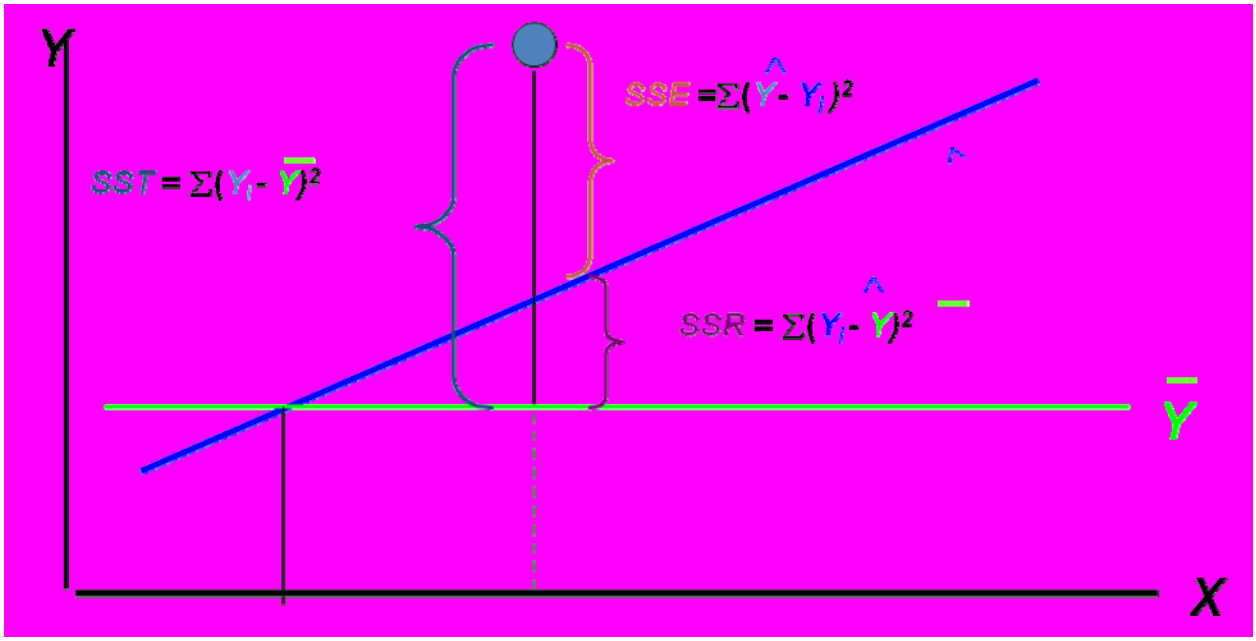
$$= (Y_i - \bar{Y}) - \hat{\beta}\left(X_i - \bar{X}\right)$$

$$= y_i - \hat{\beta} x_i \qquad (7)$$

On squaring and summing both sides of (7) we get

$$\sum_{i=1}^{n} y_i^2 = \hat{\beta}^2 \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} e_i^2 \qquad (8)$$

Equation (8) says that the total variation in the dependent variable Y_i around its sample mean can be decomposed into two parts.: the first is the regression sum of squares - the portion explained by the regressor given by the first term on the right hand side(RHS) and the second is variation explained by the error given by the residual sum of squares (second term).



Therefore equation (8) may be written as:

Total sum of squares(SST)= Explained sum of squares(SSR) + residual sum of squares (SSE)

The goodness of fit of the regression line may be judged by the quantity R^2 which is called coefficient of determination .

$$R^2 = \frac{SSR}{SST} = \hat{\beta}^2 \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2} = 1 - \frac{SSE}{SST} \qquad (9)$$

$0<R^2<1$.

R^2 represents the proportion of variation in the response variable Y explained by the regressors. Larger the value of R^2, model is a good fit.

Statistical Properties of OLSE of α and β

(i) OLS Estimators $\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$ and $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$ are unbiased and consistent estimators of β and α

Proof: we have $\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$ = β + $\frac{\sum_{i=1}^{n} x_i \varepsilon_i}{\sum_{i=1}^{n} x_i^2}$ (10)

$$E(\hat{\beta}) = \beta + \frac{\sum_{i=1}^{n} x_i E(\varepsilon_i)}{\sum_{i=1}^{n} x_i^2} = \beta .$$

and

$$E(\hat{\alpha}) = \overline{Y} - E(\hat{\beta})\overline{X}$$
$$= \overline{Y} - \beta\overline{X}$$
$$= \alpha$$

(ii) $Var(\hat{\beta}) = E(\hat{\beta} - \beta)^2$

$$= E\left(\frac{\sum_{i=1}^{n} x_i \varepsilon_i}{\sum_{i=1}^{n} x_i^2} \right)^2$$

$$= Var\left(\frac{\sum_{i=1}^{n} x_i \varepsilon_i}{\sum_{i=1}^{n} x_i^2} \right)$$

$$= \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$$

And $Var(\hat{\alpha}) = \frac{\sigma^2 \sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2}$ (11)

A sufficient condition for consistency is that the estimator is unbiased and its variance tends to zero as n tends to infinity.

We have shown that OLSE of α and β are unbiased. Consider

$$\lim_{n\to\infty} \operatorname{var}(\hat{\beta}) = \lim_{n\to\infty}\left(\frac{\sigma^2/n}{\sum_{i=1}^{n} x_i^2/n} \right)$$

$$= \left(\frac{\lim_{n\to\infty}(\sigma^2/n)}{\lim_{n\to\infty}\sum_{i=1}^{n} x_i^2/n} \right)$$

$$= \frac{0}{\lim_{n\to\infty}\sum_{i=1}^{n} x_i^2/n} = 0$$

Hence $\hat{\beta}_{OLS} = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$ is unbiased and consistent estimator of β.

Similarly it can be shown that OLSE of α is consistent.

(iii) OLS estimators are Best Linear Unbiased estimators.

Let $\tilde{\beta} = \sum_{i=1}^{n} a_i Y_i$ be any other linear unbiased estimator of β.

We can write

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2} = \sum_{i=1}^{n} w_i Y_i$$

$$\text{where} \quad w_i = \frac{x_i}{\sum_{i=}^{n} x_i^2}$$

Take $a_i = w_i + d_i$ for i= 1,2,..,n.

$$Var(\tilde{\beta}) = \sum a_i^2 \sigma^2$$

$$= \sigma^2 \sum_{i=1}^{n} w_i^2 + \sigma^2 \sum_{i=1}^{n} d_i^2$$

$$= Var(\hat{\beta}_{OLS}) + \sigma^2 + \sum_{i=1}^{n} d_i^2$$

$$\geq Var(\hat{\beta}_{OLS})$$

Hence OLSE of β is BLUE.

Similarly one can show that $\hat{\alpha} = \overline{Y} - \hat{\beta}\,\overline{X}$ is BLUE.

Estimating Error variance σ^2.

An unbiased estimator of error variance σ^2 is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{(n-2)}, \qquad (12)$$

where

$$e_i = Y_i - \hat{Y}_i, \quad i = 1,2,\ldots,n .$$

(i) $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{(n-2)}$, is unbiased and consistent for σ^2.

$$E(\hat{\sigma}^2) = \frac{E(\sum_{i=1}^{n} e_i^2)}{(n-2)},$$

$$E(\sum_{1}^{n} e_i^2) = \sum_{1}^{n} x_i^2 Var(\hat{\beta}_{OLS}) + (n-1)\sigma^2 - 2\frac{E(\sum_{i=1}^{n} x_i \varepsilon_i)^2}{\sum_{i=1}^{n} x_i^2}$$

$= \sigma^2 \_ (n-1)\sigma^2 - 2\sigma^2$

$= (n-2)\,\sigma^2$

Therefore $E(\hat{\sigma^2}) = \sigma^2$.

Conclusion: Regression analysis can be described as a study of the relationships between response variable and one or more other variables called predictor variables. Regression analysis specifically addresses how predictor variables influence, describe or control the response. The relationship is intended to be directional in that one ignores whether the response variable affects the predictors. Relationship in correlation is not directional. Regression analysis techniques can be applied to analyse the data related to a dependent-independent system. The analysis is based on a set of assumptions. To start with we stated a set of assumptions called basic ideal conditions. Under these assumptions ordinary least squares method yields unbiased, BLUE, consistent and efficient estimators for the regression coefficients. Using these OLS estimators, proposed an unbiased and consistent estimator for error variance . Role of residual analysis in assessing the goodness of fit of the model is discussed.