Chapter 1
Introduction to Econometrics

Econometrics deals with the measurement of economic relationships. It is an integration of economics, mathematical economics and statistics with an objective to provide numerical values to the parameters of economic relationships. The relationships of economic theories are usually expressed in mathematical forms and combined with empirical economics. The econometrics methods are used to obtain the values of parameters which are essentially the coefficients of the mathematical form of the economic relationships. The statistical methods which help in explaining the economic phenomenon are adapted as econometric methods. The econometric relationships depict the random behaviour of economic relationships which are generally not considered in economics and mathematical formulations.

It may be pointed out that the econometric methods can be used in other areas like engineering sciences, biological sciences, medical sciences, geosciences, agricultural sciences etc. In simple words, whenever there is a need of finding the stochastic relationship in mathematical format, the econometric methods and tools help. The econometric tools are helpful in explaining the relationships among variables.

Econometric Models:

A model is a simplified representation of a real-world process. It should be representative in the sense that it should contain the salient features of the phenomena under study. In general, one of the objectives in modeling is to have a simple model to explain a complex phenomenon. Such an objective may sometimes lead to oversimplified model and sometimes the assumptions made are unrealistic. In practice, generally, all the variables which the experimenter thinks are relevant to explain the phenomenon are included in the model. Rest of the variables are dumped in a basket called “disturbances” where the disturbances are random variables. This is the main difference between economic modeling and econometric modeling. This is also the main difference between mathematical modeling and statistical modeling. The mathematical modeling is exact in nature, whereas the statistical modeling contains a stochastic term also.

An economic model is a set of assumptions that describes the behaviour of an economy, or more generally, a phenomenon.

An econometric model consists of

- a set of equations describing the behaviour. These equations are derived from the economic model and have two parts – observed variables and disturbances.
- a statement about the errors in the observed values of variables.
- a specification of the probability distribution of disturbances.

Aims of econometrics:

The three main aims econometrics are as follows:

1. Formulation and specification of econometric models:

The economic models are formulated in an empirically testable form. Several econometric models can be derived from an economic model. Such models differ due to different choice of functional form, specification of the stochastic structure of the variables etc.

2. Estimation and testing of models:

The models are estimated on the basis of the observed set of data and are tested for their suitability. This is the part of the statistical inference of the modelling. Various estimation procedures are used to know the numerical values of the unknown parameters of the model. Based on various formulations of statistical models, a suitable and appropriate model is selected.

3. Use of models:

The obtained models are used for forecasting and policy formulation, which is an essential part in any policy decision. Such forecasts help the policymakers to judge the goodness of the fitted model and take necessary measures in order to re-adjust the relevant economic variables.

Econometrics and statistics:

Econometrics differs both from mathematical statistics and economic statistics. In economic statistics, the empirical data is collected recorded, tabulated and used in describing the pattern in their development over time. The economic statistics is a descriptive aspect of economics. It does not provide either the explanations of the development of various variables or measurement of the parameters of the relationships.

Statistical methods describe the methods of measurement which are developed on the basis of controlled experiments. Such methods may not be suitable for the economic phenomenon as they don't fit in the framework of controlled experiments. For example, in real-world experiments, the variables usually change continuously and simultaneously, and so the set up of controlled experiments are not suitable.

Econometrics uses statistical methods after adapting them to the problems of economic life. These adopted statistical methods are usually termed as econometric methods. Such methods are adjusted so that they become appropriate for the measurement of stochastic relationships. These adjustments basically attempt to specify attempts to the stochastic element which operate in real-world data and enters into the determination of observed data. This enables the data to be called a random sample which is needed for the application of statistical tools.

The **theoretical econometrics** includes the development of appropriate methods for the measurement of economic relationships which are not meant for controlled experiments conducted inside the laboratories. The econometric methods are generally developed for the analysis of non-experimental data.

The **applied econometrics** includes the application of econometric methods to specific branches of econometric theory and problems like demand, supply, production, investment, consumption etc. The applied econometrics involves the application of the tools of econometric theory for the analysis of the economic phenomenon and forecasting economic behaviour.

Types of data

Various types of data is used in the estimation of the model.

1. Time series data

Time series data give information about the numerical values of variables from period to period and are collected over time. For example, the data during the years 1990-2010 for monthly income constitutes a time series of data.

2. Cross-section data

The cross-section data give information on the variables concerning individual agents (e.g., consumers or produces) at a given point of time. For example, a cross-section of a sample of consumers is a sample of family budgets showing expenditures on various commodities by each family, as well as information on family income, family composition and other demographic, social or financial characteristics.

3. Panel data:

The panel data are the data from a repeated survey of a single (cross-section) sample in different periods of time.

4. Dummy variable data

When the variables are qualitative in nature, then the data is recorded in the form of the indicator function. The values of the variables do not reflect the magnitude of the data. They reflect only the presence/absence of a characteristic. For example, variables like religion, sex, taste, etc. are qualitative variables. The variable `sex' takes two values – male or female, the variable `taste' takes values-like or dislike etc. Such values are denoted by the dummy variable. For example, these values can be represented as '1' represents male and '0' represents female. Similarly, '1' represents the liking of taste, and '0' represents the disliking of taste.

Aggregation problem:

The aggregation problems arise when aggregative variables are used in functions. Such aggregative variables may involve.

1**. Aggregation over individuals**:

For example, the total income may comprise the sum of individual incomes.

2. Aggregation over commodities:

The quantity of various commodities may be aggregated over, e.g., price or group of commodities. This is done by using suitable index.

3. **Aggregation over time periods**

Sometimes the data is available for shorter or longer time periods than required to be used in the functional form of the economic relationship. In such cases, the data needs to be aggregated over the time period. For example, the production of most of the manufacturing commodities is completed in a period shorter than a year. If annual figures are to be used in the model, then there may be some error in the production function.

4. Spatial aggregation:

Sometimes the aggregation is related to spatial issues. For example, the population of towns, countries, or the production in a city or region etc..

Such sources of aggregation introduce "aggregation bias" in the estimates of the coefficients. It is important to examine the possibility of such errors before estimating the model.

Econometrics and regression analysis:

One of the very important roles of econometrics is to provide the tools for modeling on the basis of given data. The regression modeling technique helps a lot in this task. The regression models can be either linear or non-linear based on which we have linear regression analysis and non-linear regression analysis. We will consider only the tools of linear regression analysis and our main interest will be the fitting of the linear regression model to a given set of data.

Linear regression model

Suppose the outcome of any process is denoted by a random variable y, called as dependent (or study) variable, depends on k independent (or explanatory) variables denoted by $X_1, X_2,..., X_k$. Suppose the behaviour of y can be explained by a relationship given by

$$y = f(X_1, X_2,..., X_k, \beta_1, \beta_2,..., \beta_k) + \varepsilon$$

where f is some well-defined function and $\beta_1, \beta_2,..., \beta_k$ are the parameters which characterize the role and contribution of $X_1, X_2,..., X_k$, respectively. The term ε reflects the stochastic nature of the relationship between y and $X_1, X_2,..., X_k$ and indicates that such a relationship is not exact in nature. When $\varepsilon = 0,$ then the relationship is called the mathematical model otherwise the statistical model. The term "**model**" is broadly used to represent any phenomenon in a mathematical framework.

A model or relationship is termed as linear if it is linear in parameters and non-linear, if it is not linear in parameters. In other words, if all the partial derivatives of y with respect to each of the parameters $\beta_1, \beta_2,..., \beta_k$ are independent of the parameters, then the model is called as a **linear model**. If any of the partial derivatives of y with respect to any of the $\beta_1, \beta_2,..., \beta_k$ is not independent of the parameters, the model is called non-linear. Note that the linearity or non-linearity of the model is not described by the linearity or non-linearity of explanatory variables in the model.

For example

$$y = \beta_1 X_1^2 + \beta_2 \sqrt{X_2} + \beta_3 \log X_3 + \varepsilon$$

is a linear model because $\partial y / \partial \beta_i, \ (i = 1,2,3)$ are independent of the parameters $\beta_i, (i = 1,2,3).$ On the other hand,

$$y = \beta_1^2 X_1 + \beta_2 X_2 + \beta_3 \log X + \varepsilon$$

is a non-linear model because $\partial y / \partial \beta_1 = 2\beta_1 X_1$ depends on β_1 although $\partial y / \partial \beta_2$ and $\partial y / \partial \beta_3$ are independent of any of the β_1, β_2 or β_3.

When the function f is linear in parameters, then $y = f(X_1, X_2, ..., X_k, \beta_1, \beta_2, ..., \beta_k) + \varepsilon$ is called a linear model and when the function f is non-linear in parameters, then it is called a non-linear model. In general, the function f is chosen as

$$f(X_1, X_2, ..., X_k, \beta_1, \beta_2 ..., \beta_k) = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

to describe a linear model. Since $X_1, X_2, ..., X_k$ are pre-determined variables and y is the outcome, so both are known. Thus the knowledge of the model depends on the knowledge of the parameters $\beta_1, \beta_2, ..., \beta_k$.

The statistical linear modeling essentially consists of developing approaches and tools to determine $\beta_1, \beta_2, ..., \beta_k$ in the linear model

$$y = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$$

given the observations on y and $X_1, X_2, ..., X_k$.

Different statistical estimation procedures, e.g., method of maximum likelihood, the principle of least squares, method of moments etc. can be employed to estimate the parameters of the model. The method of maximum likelihood needs further knowledge of the distribution of y whereas the method of moments and the principle of least squares do not need any knowledge about the distribution of y.

The regression analysis is a tool to determine the values of the parameters given the data on y and $X_1, X_2, ..., X_k$. The literal meaning of regression is “to move in the backward direction”. Before discussing and understanding the meaning of “backward direction”, let us find which of the following statements is correct:

$S1$: model generates data or

$S2$: data generates the model.

Obviously, $S1$ is correct. It can be broadly thought that the model exists in nature but is unknown to the experimenter. When some values to the explanatory variables are provided, then the values for the output or study variable are generated accordingly, depending on the form of the function f and the nature of the phenomenon. So ideally, the pre-existing model gives rise to the data. Our objective is to determine the functional form of this model. Now we move in the backward direction. We propose to first collect the data on study and explanatory variables. Then we employ some statistical techniques and use this data to know the form of function f. Equivalently, the data from the model is recorded first and then used to determine the parameters of the model. The regression analysis is a technique which helps in determining the statistical model by using the data on study and explanatory variables. The classification of linear and non-linear regression analysis is based on the determination of linear and non-linear models, respectively.

Consider a simple example to understand the meaning of "regression". Suppose the yield of the crop (y) depends linearly on two explanatory variables, viz., the quantity of fertilizer (X_1) and level of irrigation (X_2) as

$$y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon.$$

There exist the true values of β_1 and β_2 in nature but are unknown to the experimenter. Some values on y are recorded by providing different values to X_1 and X$_2$. There exists some relationship between y and X_1, X_2 which gives rise to a systematically behaved data on y, X_1 and X_2. Such a relationship is unknown to the experimenter. To determine the model, we move in the backward direction in the sense that the collected data is used to determine the unknown parameters β_1 and β_2 of the model. In this sense, such an approach is termed as regression analysis.

The theory and fundamentals of linear models lay the foundation for developing the tools for regression analysis that are based on valid statistical theory and concepts.

Steps in regression analysis

Regression analysis includes the following steps:

- Statement of the problem under consideration
- Choice of relevant variables
- Collection of data on relevant variables
- Specification of model
- Choice of method for fitting the data
- Fitting of model
- Model validation and criticism
- Using the chosen model(s) for the solution of the posed problem.

These steps are examined below.

1. Statement of the problem under consideration:

The first important step in conducting any regression analysis is to specify the problem and the objectives to be addressed by the regression analysis. The wrong formulation or the wrong understanding of the problem will give the wrong statistical inferences. The choice of variables depends upon the objectives of the study and understanding of the problem. For example, the height and weight of children are related. Now there can be two issues to be addressed.

(i) Determination of height for a given weight, or

(ii) determination of weight for a given height.

In case 1, the height is the response variable, whereas weight is the response variable in case 2. The role of explanatory variables is also interchanged in cases 1 and 2.

2. Choice of relevant variables:

Once the problem is carefully formulated and objectives have been decided, the next question is to choose the relevant variables. It has to be kept in mind that the correct choice of variables will determine the statistical inferences correctly. For example, in any agricultural experiment, the yield depends on explanatory variables like quantity of fertilizer, rainfall, irrigation, temperature etc. These variables are denoted by $X_1, X_2, ..., X_k$ as a set of k explanatory variables.

3. Collection of data on relevant variables:

Once the objective of the study is clearly stated, and the variables are chosen, the next question arises how to collect data on such relevant variables. The data is essentially the measurement of these variables. For example, suppose we want to collect the data on age. For this, it is important to know how to record the data on age. Then either the date of birth can be recorded which will provide the exact age on any specific date or the age in terms of completed years as on specific date can be recorded. Moreover, it is also important to decide whether the data has to be collected on variables as quantitative variables or qualitative variables. For example, if the ages (in years) are 15,17,19,21,23, then these are quantitative values. If the ages are defined by a variable that takes value 1 if ages are less than 18 years and 0 if the ages are more than 18 years, then the earlier recorded data is converted to 1,1,0,0,0. Note that there is a loss of information in converting the quantitative data into qualitative data. The methods and approaches for qualitative and quantitative data are also different. If the study variable is binary, then **logistic** and **probit regressions** etc. are used. If all explanatory variables are qualitative, then **analysis of variance** technique is used. If some explanatory variables are qualitative and others are quantitative, then **analysis of covariance** technique is used. The techniques of analysis of variance and analysis of covariance are the special cases of regression analysis.

Generally, the data is collected on n subjects, then y on data, then y denotes the response or study variable and $y_1, y_2, ..., y_n$ are the n values. If there are k explanatory variables $X_1, X_2, .., X_k$ then x_{ij} denotes the i^{th} value of the j^{th} variable $i = 1, 2, ..., n;\ j = 1, 2, ..., k$. The observation can be presented in the following table:

Notation for the data used in regression analysis

Observation number	Response y	Explanatory variables $X_1 \quad X_2 \ \cdots \ X_k$
1	y_1	$x_{11} \quad x_{12} \cdots x_{1k}$
2	y_2	$x_{21} \quad x_{22} \cdots x_{2k}$
$\vdots$	$\vdots$	$\vdots \quad \vdots \ \ddots \ \vdots$
n	y_n	$x_{n1} \quad x_{n2} \cdots x_{nk}$

4. Specification of model:

The experimenter or the person working in the subject usually help in determining the form of the model. Only the form of the tentative model can be ascertained, and it will depend on some unknown parameters. For example, a general form will be like

$$y = f(X_1, X_2, ..., X_k; \beta_1, \beta_2, ..., \beta_k) + \varepsilon$$

where ε is the random error reflecting mainly the difference in the observed value of y and the value of y obtained through the model. The form of $f(X_1, X_2, ..., X_k; \beta_1, \beta_2, ..., \beta_k)$ can be linear as well as non-linear depending on the form of parameters $\beta_1, \beta_2, ..., \beta_k$. A model is said to be linear if it is linear in parameters. For example,

$$y = \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \varepsilon$$
$$y = \beta_1 + \beta_2 \ln X_2 + \varepsilon$$

are linear models whereas

$$y = \beta_1 X_1 + \beta_2^2 X_2 + \beta_3 X_2 + \varepsilon$$
$$y = (\ln \beta_1) X_1 + \beta_2 X_2 + \varepsilon$$

are non-linear models. Many times, the non-linear models can be converted into linear models through some transformations. So the class of linear models is wider than what it appears initially.

If a model contains only one explanatory variable, then it is called a **simple regression model.** When there are more than one independent variables, then it is called a **multiple regression model.** When there is only one study variable, the regression is termed as **univariate regression.** When there are more than one study variables, the regression is termed as **multivariate regression.** Note that the simple and multiple regressions are not same as univariate and multivariate regressions. The simple and multiple regression are determined by the number of explanatory variables, whereas univariate and multivariate regressions are determined by the number of study variables.

5. Choice of method for fitting the data:

After the model has been defined, and the data have been collected, the next task is to estimate the parameters of the model based on the collected data. This is also referred to as **parameter estimation** or **model fitting**. The most commonly used method of estimation is the least-squares method. Under certain assumptions, the least-squares method produces estimators with desirable properties. The other estimation methods are the maximum likelihood method, ridge method, principal components method etc.

6. Fitting of model:

The estimation of unknown parameters using appropriate method provides the values of the parameter. Substituting these values in the equation gives us a usable model. This is termed as model fitting. The estimates of parameters $\beta_1, \beta_2, ..., \beta_k$ in the model

$$y = f(X_1, X_2, ..., X_k, \beta_1, \beta_2, ..., \beta_k) + \varepsilon$$

are denoted by $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$ which gives the fitted model as

$$y = f(X_1, X_2, ..., X_k, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k).$$

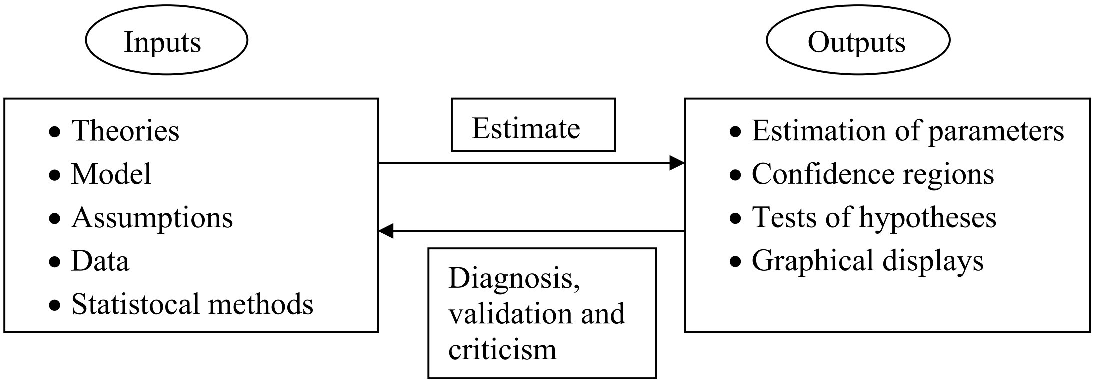
When the value of y is obtained for the given values of $X_1, X_2, ..., X_k,$ it is denoted as $\hat{y}$ and called as fitted value.

The fitted equation is used for prediction. In this case, $\hat{y}$ is termed as the **predicted value.** Note that the fitted value is where the values used for explanatory variables correspond to one of the n observations in the data, whereas predicted value is the one obtained for any set of values of explanatory variables. It is not generally recommended to predict the y-values for the set of those values of explanatory variables which lie outside the range of data. When the values of explanatory variables are the future values of explanatory variables, the predicted values are called forecasted values.

7. Model criticism and selection

The validity of the statistical method to be used for regression analysis depends on various assumptions. These assumptions become the assumptions for the model and the data essentially. The quality of statistical inferences heavily depends on whether these assumptions are satisfied or not. For making these assumptions to be valid and to be satisfied, care is needed from the beginning of the experiment. One has to be careful in choosing the required assumptions and to decide as well to determine if the assumptions are valid for the given experimental conditions or not? It is also important to decide that the situations is which the assumptions may not meet.

The validation of the assumptions must be made before drawing any statistical conclusion. Any departure from the validity of assumptions will be reflected in the statistical inferences. In fact, the regression analysis is an iterative process where the outputs are used to diagnose, validate, criticize and modify the inputs. The iterative process is illustrated in the following figure.



8. Objectives of regression analysis

The determination of the explicit form of the regression equation is the ultimate objective of regression analysis. It is finally a good and valid relationship between study variable and explanatory variables. The regression equation helps in understanding the interrelationships of variables among them. Such a regression equation can be used for several purposes. For example, to determine the role of any explanatory variable in the joint relationship in any policy formulation, to forecast the values of the response variable for a given set of values of explanatory variables.