ECONOMETRIC THEORY

MODULE – I

Lecture - 1

Introduction to Econometrics

Dr. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur
Econometrics deals with the measurement of economic relationships.

It is an integration of economic theories, 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 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 the economic modeling and econometric modeling. This is also the main difference between the 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 behavior of an economy, or more general, a phenomenon.

An econometric model consists of

- a set of equations describing the behavior. 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 of 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 stochastic structure of the variables etc.

2. Estimation and testing of models

The models are estimated on the basis of observed set of data and are tested for their suitability. This is the part of statistical inference of the modeling. 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 policy makers to judge the goodness of fitted model and take necessary measures 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 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 the stochastic element which operate in real world data and enters into the determination of observed data. This enables the data to be called as 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 economic phenomenon and forecasting the 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 data.

2. Cross section data

The cross section data give information on the variables concerning individual agents (e.g., consumers or producers) at a given point of time. For example, a cross section of 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 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 indicator function. The values of the variables do not reflect the magnitude of data. They reflect only the presence/absence of a characteristic. For example, the 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 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 economic relationship. In such cases, the data needs to be aggregated over 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 introduces “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 role 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 linear regression model to a given set of data.

Linear models and regression analysis

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 $X_1, X_2, \ldots, X_k$. Suppose the behaviour of $y$ can be explained by a relationship given by

$$y = f(X_1, X_2, \ldots, X_k, \beta_1, \beta_2, \ldots, \beta_k) + \epsilon$$

where $f$ is some well defined function and $\beta_1, \beta_2, \ldots, \beta_k$ are the parameters which characterize the role and contribution of $X_1, X_2, \ldots, X_k$, respectively. The term $\epsilon$ reflects the stochastic nature of the relationship between $y$ and $X_1, X_2, \ldots, X_k$ and indicates that such a relationship is not exact in nature. When $\epsilon = 0$, 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 frame work.

A model or relationship is termed as linear if it is linear in parameters and nonlinear, if it is not linear in parameters. In other words, if all the partial derivatives of $y$ with respect to each of the parameters $\beta_1, \beta_2, \ldots, \beta_k$ 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 $\beta_1, \beta_2, \ldots, \beta_k$ is not independent of the parameters, the model is called as nonlinear. Note that the linearity or non-linearity of the model is not described by the linearity or nonlinearity 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 nonlinear model because \( \partial y / \partial \beta_1 = 2\beta_1 X_1 \) depends on \( \beta_1 \) although \( \partial y / \partial \beta_2 \) and \( \partial y / \partial \beta_3 \) are independent of any of the \( \beta_1, \beta_2 \) or \( \beta_3 \).

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

\[ f(X_1, X_2, \ldots, X_k, \beta_1, \beta_2, \ldots, \beta_k) = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \]

to describe a linear model. Since \( X_1, X_2, \ldots, X_k \) 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 \( \beta_1, \beta_2, \ldots, \beta_k \).

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

\[ y = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon \]

given the observations on \( y \) and \( X_1, X_2, \ldots, X_k \).

Different statistical estimation procedures, e.g., method of maximum likelihood, 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 $X_1, X_2, \ldots, X_k$. 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 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 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 nonlinear regression analysis is based on the determination of linear and nonlinear models, respectively.

Consider a simple example to understand the meaning of “regression”. Suppose the yield of crop ($y$) depends linearly on two explanatory variables, viz., the quantity of a 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 $\beta_1$ and $\beta_2$ 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 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 $\beta_1$ and $\beta_2$ of the model. In this sense such an approach is termed as regression analysis.