

Chapter 1

Introduction to Econometrics

Econometrics deals with the measurement of economic relationships. It integrates economics, mathematical economics, and statistics to provide numerical values for the parameters of economic relationships. The relationships of economic theories are usually expressed in mathematical forms and combined with empirical economics. Econometric methods are used to estimate parameter values, which are essentially the coefficients in the mathematical form of economic relationships. The statistical methods used to explain economic phenomena are applied as econometric methods. The econometric relationships depict the stochastic behaviour of economic relationships, which are generally not accounted for in economic and mathematical formulations.

It may be noted that econometric methods can be used in other areas, such as engineering, biological, medical, geosciences, and agricultural sciences. In simple terms, whenever there is a need to model a stochastic relationship in mathematical format, econometric methods and tools help. Econometric tools are helpful for explaining 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 objective of modelling is to develop a simple model that explains a complex phenomenon. Such an objective may sometimes lead to an oversimplified model and, at other times, to unrealistic assumptions. In practice, the experimenter generally includes all variables they consider relevant to explaining the phenomenon in the model. The rest of the variables are dumped into a basket called “disturbances”, where they are random variables. This is the main difference between economic modelling and econometric modelling. This is also the main difference between mathematical modelling and statistical modelling. The mathematical modelling is exact, whereas the statistical modelling includes a stochastic term.

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 comprise 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 a form that is empirically testable. Several econometric models can be derived from an economic model. Such models differ due to different choices of functional forms, specifications of the stochastic structures of the variables, etc.

2. Estimation and testing of models

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

3. Use of models

The models obtained are used for forecasting and policy formulation, both essential parts of any policy decision. Such forecasts help policymakers assess the quality of the fitted model and take the necessary measures to adjust the relevant economic variables.

Econometrics and statistics

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

Statistical methods describe measurement procedures developed through controlled experiments. Such methods may not be suitable for economic phenomena, as they don't fit within the framework of controlled experiments. For example, in real-world experiments, the variables usually change continuously and simultaneously, and so the setup of controlled experiments is not suitable.

Econometrics uses statistical methods adapted to the problems of economic life. These adopted statistical methods are usually termed as econometric methods. Such methods are designed to measure stochastic

relationships. These adjustments aim to account for the stochastic element that operates in real-world data and influences the determination of observed data. This enables the data to be considered a random sample, which is necessary for applying statistical tools.

Theoretical econometrics includes the development of appropriate methods for measuring economic relationships that are not intended for controlled laboratory experiments. Econometric methods are generally developed for analysing non-experimental data.

Applied econometrics involves applying econometric methods to specific branches of econometric theory and to problems such as demand, supply, production, investment, and consumption. Applied econometrics involves applying the tools of econometric theory to analyse economic phenomena and forecast economic behaviour.

Types of data

Various types of data are used to estimate the model.

1. Time series data

Time series data provide information about the numerical values of variables over time and are collected as they change. For example, the data from 1990 to 2010 on monthly income constitute a time series.

2. Cross-section data

Cross-sectional data provide information on variables for individual agents (e.g., consumers or producers) at a given point in time. For example, a cross-section of consumers is a sample of family budgets that show expenditures on various commodities for each family, along with information on family income, family composition, and other demographic, social, or financial characteristics.

3. Panel data

Panel data are from a repeated survey of a single (cross-sectional) sample over different periods of time.

4. Dummy variable data

When the variables are qualitative, the data is recorded as indicator functions. 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 such as religion, sex, and taste are qualitative variables. The variable 'sex' takes two values – male or female, the variable 'taste' takes values like or dislikes, 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 quantities of various commodities may be aggregated, e.g., by price or by group of commodities. This is done by using a suitable index.

3. Aggregation over time periods

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

4. Spatial aggregation

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

Such sources of aggregation introduce “aggregation bias” into coefficient estimates. 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 tools for modelling based on given data. Regression modelling is very helpful for this task. The regression models can be either linear or non-linear, depending on whether we perform linear or non-linear regression analysis. We will consider only linear regression tools, and our main interest will be in fitting the linear regression model to a given dataset.

Linear regression model

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

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

where f is some well-defined function and $\beta_1, \beta_2, \dots, \beta_k$ are the parameters which characterize the role and contribution of $X_1, X_2, \dots, X_k$, respectively. The term ε reflects the stochastic nature of the relationship between y and $X_1, X_2, \dots, X_k$ 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 used broadly to refer to any phenomenon described within 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 $\beta_1, \beta_2, \dots, \beta_k$ 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 $\beta_1, \beta_2, \dots, \beta_k$ 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 β_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 $y = f(X_1, X_2, \dots, X_k, \beta_1, \beta_2, \dots, \beta_k) + \varepsilon$ 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

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

to describe a linear model. Since $X_1, X_2, \dots, X_k$ 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 $\beta_1, \beta_2, \dots, \beta_k$.

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

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

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

Various statistical estimation procedures, e.g., the method of maximum likelihood, the method of least squares, and the method of moments, can be used to estimate the model's parameters. 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 $X_1, X_2, \dots, X_k$. 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 is generally thought that the model exists in nature but is unknown to the experimenter. When some values for 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 generates the data. Our objective is to determine the functional form of this model. Now we move backwards. We propose first collecting data on the study and explanatory variables. Then we employ some statistical techniques and use this data to determine the form of the function f . Equivalently, the model's data is recorded first and then used to determine the model's parameters. Regression analysis is a technique that helps determine a statistical model using data from the 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 the 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 a relationship between y and X_1, X_2 which gives rise to 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 the model
- Choice of method for fitting the data
- Fitting of the 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. The wrong formulation or misunderstanding of the problem will yield incorrect statistical inferences. The choice of variables depends upon the objectives of the study and the understanding of the problem. For example, children's height and weight are related. Now, there are 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 in case 2, the weight is the response variable. The roles of the explanatory variables are 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 is important to keep in mind that the choice of variables determines the validity of statistical inferences. For example, in any agricultural experiment, yield depends on explanatory variables such as the quantity of fertiliser, rainfall, irrigation, and temperature. These variables are denoted by $X_1, X_2, \dots, X_k$ as a set of k explanatory variables.

3. Collection of data on relevant variables

Once the study's objective is clearly stated and the variables are chosen, the next question arises: how to collect data on these relevant variables. The data are essentially measurements of these variables. For example, suppose we want to collect the data on age. For this, it is important to know how to record age data. Then either the date of birth can be recorded, which provides the exact age on any specific date, or the age in completed years as of a specific date can be recorded. Moreover, it is important to decide whether the data should be collected as quantitative or qualitative variables. For example, if the ages (in years) are 15, 17, 19, 21, and 23, then these are quantitative values. If the ages are defined by a variable that takes the value 1 if the age is less than 18 years and 0 if the age is greater than 18 years, then the earlier recorded data is converted to 1, 1, 0, 0, 0. Note that information is lost when quantitative data are converted to qualitative data. The methods and approaches for qualitative and quantitative data are also different. If the study variable is binary, logistic and probit regressions, among others, are used. If all explanatory variables are qualitative, then the **analysis of variance** technique is used. If some explanatory variables are qualitative and others are quantitative, then the **analysis of covariance** technique is used. The techniques of analysis of variance and analysis of covariance are 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 $y_1, y_2, \dots, y_n$ are the n values. If there are k explanatory variables $X_1, X_2, \dots, X_k$ then x_{ij} denotes the i^{th} value of the j^{th} variable $i = 1, 2, \dots, n; j = 1, 2, \dots, k$. The observation can be presented in the following table:

Notation for the data used in regression analysis

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

4. Specification of model

The experimenter, or the person working with the subject, usually helps determine the form of the model. Only the form of the tentative model can be determined; it will depend on unknown parameters. For example, a general form will be like

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

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 $f(X_1, X_2, \dots, X_k; \beta_1, \beta_2, \dots, \beta_k)$ can be linear as well as non-linear, depending on the form of the parameters $\beta_1, \beta_2, \dots, \beta_k$. 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. Often, non-linear models can be converted to linear models through transformations. So the class of linear models is wider than it initially appears.

A model containing only one explanatory variable is called a **simple regression model**. When there is more than one independent variable, it is called a **multiple regression model**. When there is only one study variable,

the regression is termed as **univariate regression**. When there is more than one study variable, the regression is termed as **multivariate regression**. Note that the simple and multiple regressions are not the same as univariate and multivariate regressions. The simple and multiple regressions 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 step is to estimate the model's parameters using 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. Other estimation methods include the maximum likelihood method, the ridge method, and the principal components method.

6. Fitting of the model

Estimating unknown parameters using an appropriate method yields parameter values. Substituting these values into the equation yields a usable model. This is termed as model fitting. The estimates of parameters $\beta_1, \beta_2, \dots, \beta_k$ in the model

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

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

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

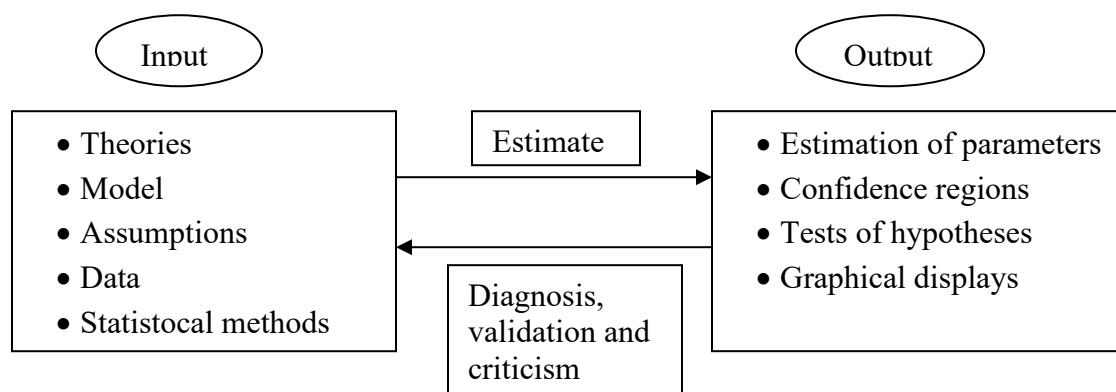
When the value of y is obtained for the given values of $X_1, X_2, \dots, X_k$, it is denoted as $\hat{y}$ and called the fitted value.

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 the 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 future values, the predicted values are called forecasted values.

7. Model criticism and selection

The validity of the statistical method used for regression analysis depends on several assumptions. These assumptions essentially serve as the basis for the model and the data. The quality of statistical inferences depends heavily on whether these assumptions are satisfied. For these assumptions to be valid, care must be taken from the beginning of the experiment. One has to be careful when choosing the required assumptions and decide whether they are valid for the given experimental conditions. It is also important to identify which situations the assumptions may not apply to.

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



8. Objectives of regression analysis

Determining the explicit form of the regression equation is the ultimate objective of regression analysis. It is finally a good, valid relationship between the study variable and the explanatory variables. The regression equation helps in understanding the interrelationships among the variables. 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, or to forecast the values of the response variable for a given set of values of explanatory variables.