Chapter 1

Introduction to Econometrics

Literally interpreted, *econometrics* means "economic measurement." Although measurement is an important part of econometrics, the scope of econometrics is much broader, as can be seen from the following quotations:

- Econometrics, the result of a certain outlook on the role of economics, consists of the application of mathematical statistics to economic data to lend empirical support to the models constructed by mathematical economics and to obtain numerical results.[1]
- Econometrics may be defined as the quantitative analysis of actual economic phenomena based on the concurrent development of theory and observation, related by appropriate methods of inference.
- Econometrics may be defined as the social science in which the tools of economic theory, mathematics, and statistical inference are applied to the analysis of economic phenomena.

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 behavior 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 behavior of an economy, or more generally, 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.

Methodology(Application) of Econometrics

How do econometricians proceed in their analysis of an economic problem? That is, what is their methodology? Although there are several schools of thought on econometric methodology, we present here the **traditional** or **classical** methodology, which still dominates empirical research in economics and other social and behavioral sciences.

Broadly speaking, traditional econometric methodology proceeds along the following lines:

1. Statement of theory or hypothesis.
2. Specification of the mathematical model of the theory.

3. Specification of the statistical, or econometric, model.
4. Obtaining the data.
5. Estimation of the parameters of the econometric model.
6. Hypothesis testing.
7. Forecasting or prediction.
8. Using the model for control or policy purposes.

To illustrate the preceding steps, let us consider the well-known Keynesian theory of consumption

1. Statement of Theory or Hypothesis

Keynes stated:

The fundamental psychological law . . . is that men [women] are disposed, as a rule and on average, to increase their consumption as their income increases, but not as much as the increase in their income.[10]

In short, Keynes postulated that the **marginal propensity to consume (MPC),** the rate of change of consumption for a unit (say, a dollar) change in income, is greater than zero but less than 1.

2. **Specification of the Mathematical Model of Consumption** Although Keynes postulated a positive relationship between consumption and income, he did not specify the precise form of the functional relationship between the two. For simplicity, a mathematical economist might suggest the following form of the Keynesian consumption function:

$$Y = \beta_1 + \beta_2 X \qquad 0 < \beta_2 < 1$$

where $Y =$ consumption expenditure and $X =$ income, and where β_1 and β_2 , known as the

parameters of the model, are, respectively, the **intercept** and **slope** coefficients.

The slope coefficient β_2 measures the MPC. Geometrically, Equation I.3.1 is as shown in Figure I.1. This equation, which states that consumption is linearly related to income, is an example of a mathematical model of the relationship between consumption and income that is called

the **consumption function** in economics. A model is simply a set of mathe- matical equations. If the model has only one equation, as in the preceding example, it is called a **single-equation model,** whereas if it has more than one equation, it is known as a **multiple-equation model** (the latter will be considered later in the book).

In Eq. (I.3.1) the variable appearing on the left side of the equality sign is called the **dependent variable** and the variable(s) on the right side is called the **independent,** or **explanatory,** variable(s). Thus, in the Keynesian consumption function, Eq. (I.3.1), con- sumption (expenditure) is the dependent variable and income is the explanatory variable.

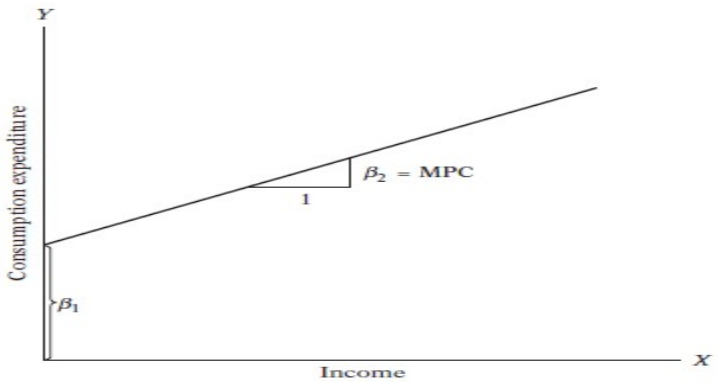


FIGURE I.1 Keynesian consumption function.

3. Specification of the Econometric Model of Consumption

The purely mathematical model of the consumption function given in Eq. (I.3.1) is of limited interest to the econometrician, for it assumes that there is an *exact* or *deterministic* relationship between consumption and income. But relationships between economic vari- ables are generally inexact. Thus, if we were to obtain data on consumption expenditure and disposable (i.e., aftertax) income of a sample of, say, 500 American families and plot these data on a graph paper with consumption expenditure on the vertical axis and disposable in- come on the horizontal axis, we would not expect all 500 observations to lie exactly on the straight line of Eq. (I.3.1) because, in addition to income, other variables affect consump- tion expenditure. For example, size of family, ages of the members in the family, family religion,

etc., are likely to exert some influence on consumption.

To allow for the inexact relationships between economic variables, the econometrician would modify the deterministic consumption function in Eq. (I.3.1) as follows:

$$Y = \beta_1 + \beta_2 X + u \qquad \text{(I.3.2)}$$

where u, known as the **disturbance,** or **error, term,** is a **random (stochastic) variable** that has well-defined probabilistic properties. The disturbance term u may well represent all those factors that affect consumption but are not taken into account explicitly.

Equation I.3.2 is an example of an **econometric model.** More technically, it is an example of a **linear regression model,** which is the major concern of this book. The econometric consumption function hypothesizes that the dependent variable Y (consumption) is linearly related to the explanatory variable X (income) but that the relationship between the two is not exact; it is subject to individual variation. The econometric model of the consumption function can be depicted as shown in Figure I.2.

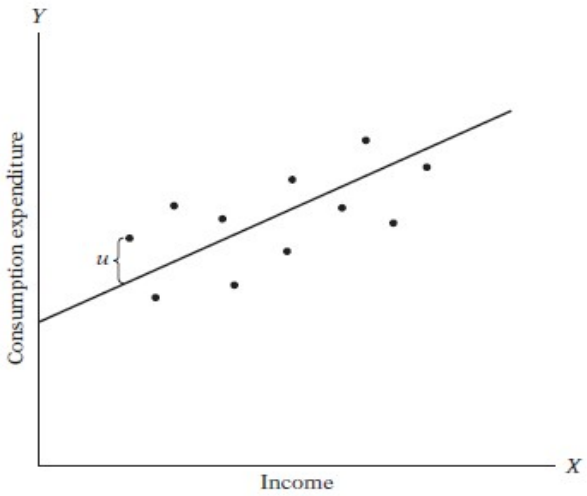


FIGURE I.2

Econometric model of the Keynesian consumption fun

4. Obtaining Data

To estimate the econometric model given in Eq. (I.3.2), that is, to obtain the numerical values of β_1 and β_2 , we need data. Although we will have more to say about the crucial importance of data for economic analysis in the next chapter, for now let us look at the data given in Table I.1, which relate to the U.S. economy for the period 1960–2005. The Y variable in this table is the *aggregate* (for the economy as a whole) personal consumption expenditure (PCE) and the X variable is gross domestic product (GDP), a measure of aggregate income, both measured in billions of 2000 dollars. Therefore, the data are in "real" terms; that is, they are measured in constant (2000) prices. The data are plotted in Figure I.3 (cf. Figure I.2). For the time being neglect the line drawn in the figure.

5. Estimation of the Econometric Model

Now that we have the data, our next task is to estimate the parameters of the consumption function. The numerical estimates of the parameters give empirical content to the con- sumption function. The actual mechanics of estimating the parameters will be discussed in Chapter 3. For now, note that the statistical technique of **regression analysis** is the main tool used to obtain the estimates. Using this technique and the data given in Table I.1, we obtain the following estimates of β_1 and β_2 , namely, -299.5913 and 0.7218. Thus, the estimated consumption function is:

$$\hat{Y}_t = -299.5913 + 0.7218 X_t$$

The hat on the Y indicates that it is an estimate. The estimated consumption function (i.e., regression line) is shown in Figure I.3.

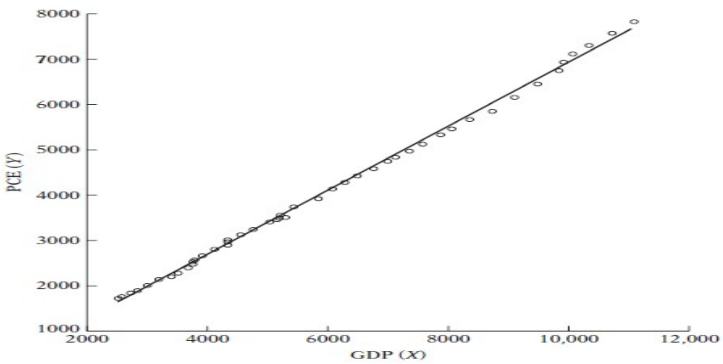


TABLE I.1
Data on *Y* (Personal Consumption Expenditure) and *X* (Gross Domestic Product, 1960–2005), both in 2000 Billions of Dollars

Year	PCE(*Y*)	GDP(*X*)
1960		2501.8
1961	1630.	2560.0
1962	1711.	2715.2
1963	1781.	2834.0
1964	1888.	2998.6
1965	2007.	3191.1
1966	2121.	3399.1
1967	2185.	3484.6
1968	2310.	3652.7
1969	2396.	3765.4
1970	2451.	3771.9
1971	2545.	3898.6
1972	2701.	4105.0
1973	2833.	4341.5
1974	2812.	4319.6
1975	2876.	4311.2
1976	3035.	4540.9
1977	3164.	4750.5
1978	3303.	5015.0
1979	3383.	5173.4
1980	3374.	5161.7
1981	3422.	5291.7
1982	3470.	5189.3
1983	3668.	5423.8
1984	3863.	5813.6
1985	4064.	6053.7
1986	4228.	6263.6
1987	4369.	6475.1
1988	4546.	6742.7
1989	4675.	6981.4
1990	4770.	7112.5
1991	4778.	7100.5
1992	4934.	7336.6
1993	5099.	7532.7
1994	5290.	7835.5
1995	5433.	8031.7
1996	5619.	8328.9
1997	5831.	8703.5
1998	6125.	9066.9
1999	6438.	9470.3
2000	6739.	9817.0
2001	6910.	9890.7
2002	7099.	10048.8
2003	7295.	10301.0
2004	7577.	10703.5
2005	7841.	11048.6

Source: *Economic Report of*

As Figure I.3 shows, the regression line fits the data quite well in that the data points are very close to the regression line. From this figure we see that for the period 1960–2005 the slope coefficient (i.e., the **MPC**) was about 0.72, suggesting that for the sample period an increase in real income of one dollar led, *on average,* to an increase of about 72 cents in real consumption expenditure.[12] We say *on average* because the relationship between con- sumption and income is inexact; as is clear from Figure I.3, not all the data points lie exactly on the regression line. In simple terms we can say that, according to our data, the *average,* or *mean,* consumption expenditure went up by about 72 cents for a dollar's increase in real income.

6. Hypothesis Testing

Assuming that the fitted model is a reasonably good approximation of reality, we have to develop suitable criteria to find out whether the estimates obtained in, say, Equation I.3.3 are in accord with the expectations of the theory that is being tested. According to "posi- tive" economists like Milton Friedman, a theory or hypothesis that is not verifiable by appeal to empirical evidence may not be admissible as a part of scientific enquiry.[13]

As noted earlier, Keynes expected the MPC to be positive but less than 1. In our exam- ple we found the MPC to be about 0.72. But before we accept this finding as confirmation of Keynesian consumption theory, we must enquire whether this estimate is sufficiently below unity to convince us that this is not a chance occurrence or peculiarity of the partic- ular data we have used. In other words, is 0.72 *statistically less than 1?* If it is, it may sup- port Keynes's theory.

Such confirmation or refutation of economic theories on the basis of sample evidence is based on a branch of statistical theory known as **statistical inference (hypothesis testing).** Throughout this book we shall see how this inference process is actually conducted.

7. Forecasting or Prediction

If the chosen model does not refute the hypothesis or theory under consideration, we may use it to predict the future value(s) of the dependent, or **forecast, variable** *Y* on the basis of the known or expected

future value(s) of the explanatory, or **predictor, variable** X.

To illustrate, suppose we want to predict the mean consumption expenditure for 2006. The GDP value for 2006 was 11319.4 billion dollars. Putting this GDP figure on the right-hand side of Eq. (I.3.3), we obtain:

$$\hat{Y}_{2006} = -299.5913 + 0.7218\,(11319.4)$$
$$= 7870.7516$$

or about 7870 billion dollars. Thus, given the value of the GDP, the mean, or average, fore- cast consumption expenditure is about 7870 billion dollars. The actual value of the consumption expenditure reported in 2006 was 8044 billion dollars. The estimated model Eq. (I.3.3) thus **under predicted** the actual consumption expenditure by about 174 billion dollars. We could say the **forecast error** is about 174 billion dollars, which is about

1.5 percent of the actual GDP value for 2006. When we fully discuss the linear regression model in subsequent chapters, we will try to find out if such an error is "small" or "large." But what is important for now is to note that such forecast errors are inevitable given the statistical nature of our analysis.

8. Use of the Model for Control or Policy Purposes

As these calculations suggest, an estimated model may be used for control, or policy, purposes. By appropriate fiscal and monetary policy mix, the government can manipulate the **control variable *X*** to produce the desired level of the **target variable *Y*.**

Statistical Data types

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.

Time Series Data

Data collected at a number of specific points in time is called time series

data. Such examples include stock prices, interest rates, exchange rates as well as product prices, GDP, etc. Time series data can be observed at many different *frequencies* (*hourly*, *daily*, *weekly*, *monthly*, *quarterly*, *annually*, etc.).

Unlike cross-sectional data, the ordering of the data is important in time-series data. Each point represents the values at specific points in time. As such, time series data are typically presented in chronological order. Changing the order of the data ignores the time- dimensionality of the data.

- **Time series data**:

No.	Bank	Year	Branch
1	Commercial Bank of Ethiopia	2014	977
2	Commercial Bank of Ethiopia	2015	1150
3	Commercial Bank of Ethiopia	2016	1310
4	Commercial Bank of Ethiopia	2017	1375
5	Commercial Bank of Ethiopia	2018	1578
6	Commercial Bank of Ethiopia	2019	1825

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.

With cross-sectional data the ordering of the data does not matter. In other words, we can order the data by ascending, descending or even randomized order and this will not affect out modeling results.

• Cross-sectional data:

CID	Bank	Year	Branch
1	AIB	2014	207
2	DB	2014	164
3	CBO	2014	141
4	BOA	2014	136
5	UB	2014	128
6	WB	2014	119
7	NIB	2014	115
8	OIB	2014	152
9	LIB	2014	88
10	BIB	2014	71
11	Bunna	2014	82
12	ZB	2014	7
13	AB	2014	89
14	AIB	2014	32
15	DGB	2014	22
16	EB	2014	11
17	CBE	2014	977

3. Panel (or Longitudinal) data

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

Panel data combines **cross-sectional** and **time series** data: the *same* individuals (persons, firms, cities, etc.) are observed at several points in time (days, years, before and after treatment etc

.

Longitudinal/panel data

cid	Bank	Year	Branch
1	NIB	2017	228
1	NIB	2018	280
1	NIB	2019	327
2	OIB	2017	260
2	OIB	2018	277
2	OIB	2019	314
3	UB	2017	233
3	UB	2018	294
3	UB	2019	351
4	WB	2017	292
4	WB	2018	355
4	WB	2019	399
5	ZB	2017	25
5	ZB	2018	44
5	ZB	2019	52

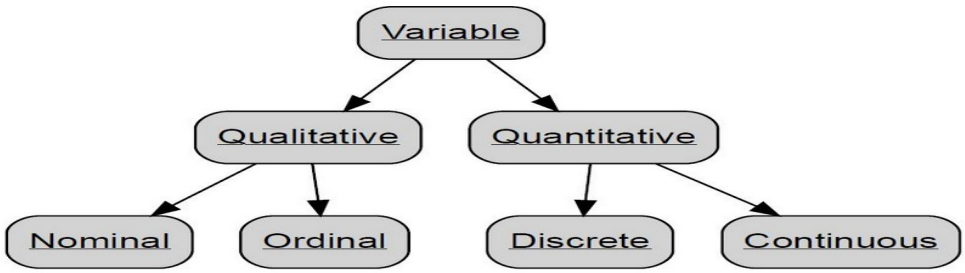
Pool data

Cross-sectional observations collected over time but the unit doesn't have to be the same

Pooled data

cid	Bank	Year	Branch
1	AIB	2014	207
2	DB	2014	164
3	CBO	2014	141
4	BOA	2014	136
.	.	.	.
5	UB	2015	144
6	WB	2015	161
7	NIB	2015	155
8	OIB	2015	210
.		.	.
9	LIB	2016	158
10	BIB	2016	177
11	Bunna	2016	143
12	ZB	2016	22

Types of variables



1. **Quantitative**

A **quantitative** variable is a variable that reflects a notion of **magnitude**, that is, if the values it can take are **numbers**. A quantitative variable represents thus a measure and is numerical.

Quantitative variables are divided into two types: **discrete** and **continuous**. The difference is explained in the following two sections.

Discrete

Quantitative discrete variables are variables for which the values it can take are **countable** and have a**finite number of possibilities**. The values are often (but not always) integers. Here are some examples ofdiscrete variables:

- Number of children per family
- Number of students in a class
- Number of citizens of a country

Even if it would take a long time to count the citizens of a large country, it is still technically doable. Moreover, for all examples, the number of possibilities is **finite**. Whatever the number of children in a family, it will never be 3.58 or 7.912 so the number of possibilities is a finite number and thus countable.

Continuous

On the other hand, **quantitative continuous** variables are variables for which the values are **not**
countable and have an **infinite number of possibilities**. For example:

- Age
- Weight
- Height

For simplicity, we usually referred to years, kilograms (or pounds) and centimeters (or feet and inches) forage, weight and height respectively.

However, a 28-year-old man could actually be 28 years, 7 months, 16 days, 3 hours, 4 minutes, 5 seconds, 31 milliseconds, 9 nanoseconds old.

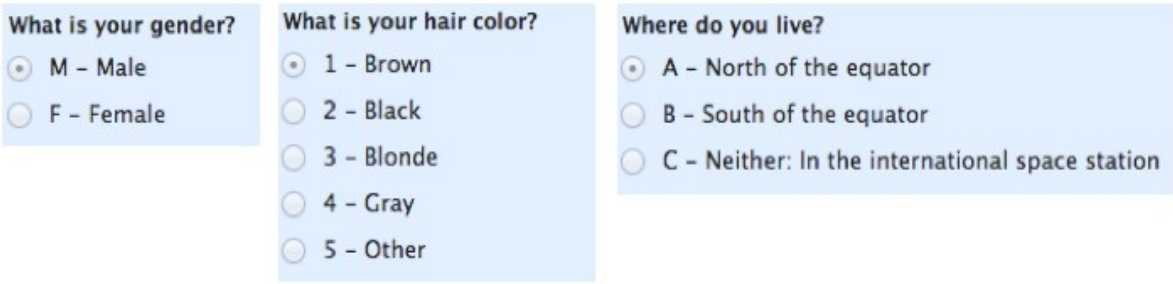
For all measurements, we usually stop at a standard level of granularity, but nothing (except our
measurement tools) prevents us from going deeper, leading to an **infinite number of potential values**.The fact that the values can take an infinite number of possibilities makes it uncountable.

2. Qualitative

In opposition to quantitative variables, **qualitative** variables (also referred as categorical variables or factors in R) are variables that are **not numerical** and which **values fits into categories**.
In other words, a **qualitative** variable is a variable which takes as its values modalities, **categories** or even levels, in contrast to **quantitative** variables which measure a **quantity** on each individual.
Qualitative variables are divided into two types: **nominal** and **ordinal**.

Nominal

A **qualitative nominal** variable is a qualitative variable where **no ordering** is possible or implied in the levels. For example, the variable gender is nominal because there is no order in the levels female/male. Eye color is another example of a nominal variable because there is no order among blue, brown or green eyes.
A nominal variable can have between two levels (e.g., do you smoke? Yes/No or what is your gender? Female/Male) and a large number of levels (what is your college major? Each major is a level in that case).
Note that a qualitative variable with exactly 2 levels is also referred as a **binary** or **dichotomous** variable.



Ordinal

On the other hand, a **qualitative ordinal** variable is a qualitative variable with an **order implied in the levels**. For instance, if the severity of road accidents has been measured on a scale such as light, moderate and fatal accidents, this variable is a qualitative ordinal variable because there is a clear order in the levels.
Another good example is health, which can take values such as poor, reasonable, good, or excellent. Again, there is clear order in these levels so health is in this case a qualitative ordinal variable.

