Chapter 1 Introduction to Econometrics

1. What is Econometrics?

1.1 Definition

Econometrics is a mixture of economics, mathematics and statistics as illustrated in Figure 1.1.1.

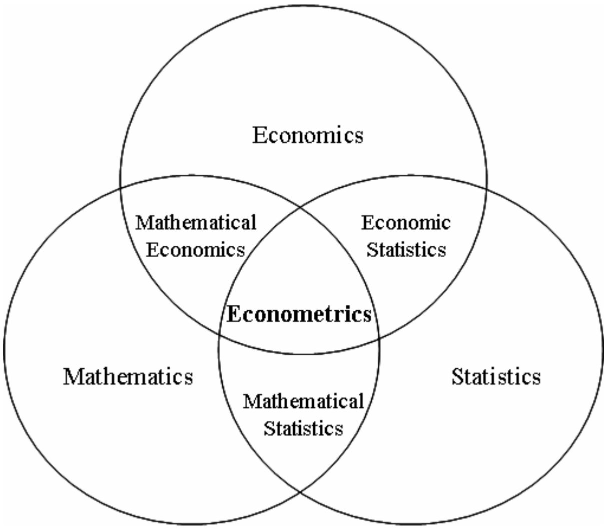


Figure 1.1.1

Economics, or economic theory, makes statements or hypotheses that are mostly qualitative in nature. Only econometrics gives empirical or numerical content to most economic theory.

The main concern of mathematical economics is to express economic theory in mathematical form (equations) without regard to measurability or empirical verification of the theory. Econometrics, as noted previously, is mainly interested in the empirical verification of economic theory.

Economic statistics is mainly concerned with collecting, processing, and presenting economic data in the form of charts and tables. The data thus collected constitute the raw data for econometric work. But the economic statistician does not go any further, not being concerned with using the collected data to test economic theories.

Although mathematical statistics provides many tools to analyze the data, the econometrician often needs special methods in view of the unique nature of most economic data, namely, that the data are not generated as the result of a controlled experiment. The econometrician generally depends on data that cannot be controlled directly.

In econometrics the modeler is often faced with observational as opposed to experimental data.

That is, in the social sciences, the data that one generally encounters are non-experimental in nature, that is, not subject to the control of the researcher. This lack of control often creates special problems for the researcher in pinning down the exact causes affecting a particular situation.

1.2 Categories

General- and narrow-defined econometrics

The former is a general designation of all econometrics methods such as regression analysis, input-output analysis, time series analysis, semi- and non-parametric analysis and so on.
The latter is just the usual-said classical econometric by using the regression analysis.

Theoretical and applied econometrics

Theoretical econometrics is concerned with the development of appropriate methods for measuring economic relationships specified by econometric models. It must spell out the assumptions of this method, its properties, and what happens to these properties when one or more of the assumptions of the method are not fulfilled.
In applied econometrics we use the tools of theoretical econometrics to study some special field(s) of economics and business, such as the production function, investment function, demand and supply functions, portfolio theory, etc.

1.3 Application

The application of econometrics includes four parts: structure analysis, economic forecasting or predicting, policy comments, verifying and developing economic theory.

Structure analysis

It studies the relations between economic variables and different from the usual-said industrial structure, consumption structure, invest structure and so on. That is, it studies the influence of the change of one or some variables on the others or the whole economic system by using the elastic analysis, multiplier analysis and comparative static analysis.

Economic forecasting

Econometrics can be used to either analyze the historic data or forecasting the future. Econometric model is developed from the application in economic forecasting.

Case: GHI Financial Services Limited
It is late February 1998, GHI Financial Services limited, a British company, has an offer to buy a call option on the US dollar expiring on 6 March. The following information is available:

Current exchange rate (GBP/USD)	0.61125
Exercise exchange rate	0.61140
Days to expiry	10
Types of option	American Call
Amount	USD50,000,000
Total cost	GBP5,000

The decision whether or not to take the offer depends on one thing only: the behavior of the exchange rate between the present time and the expiry date. The chief economist who embarks on the task chooses to prepare the forecasts from an analysis of the daily exchange rate movements over the past 90 days. He also chooses the following techniques to prepare the forecasts:

1. Simple average
2. 50-day moving average
3. 25-day moving average
4. Double moving average of order 20
5. ARIMA (2, 1,2) model

By applying these methods to the sample data, the forecast values reported in the following table are produced.

Table 1.1.1 Daily Exchange rate forecasts

Day	Daily Exchange rate forecasts					Actual exchange rates
	Simple	50-day	25-day	Double	ARIMA	
1	0.60465	0.60904	0.61007	0.60904	0.61111	0.61087
2	0.60471	0.60915	0.61008	0.60994	0.61110	0.61087
3	0.60478	0.60915	0.60099	0.60981	0.60820	0.60617
4	0.60480	0.60926	0.60962	0.60975	0.60960	0.60846
5	0.60484	0.60930	0.60951	0.60954	0.60974	0.60861
6	0.60488	0.60914	0.60941	0.60876	0.60669	0.60350
7	0.60486	0.60909	0.60959	0.60885	0.60870	0.60676
8	0.60488	0.60909	0.60951	0.60871	0.60820	0.60588
9	0.60489	0.60925	0.60940	0.60917	0.60894	0.60705
10	0.60491	0.60950	0.60941	0.60983	0.61185	0.61196

The economist seems to believe that ARIMA models produce more accurate forecasts than moving average methods, and accordingly he decides to go ahead with the decision to buy the option.

On no day does the exchange rate reach the 0.61100 level. On the expiry date, however, the exchange rate is 0.61196. This is good news: the option can be exercised to make a net profit of GBP28,000. The right decision has been taken.

The following measures of forecasting accuracy are employed:

1. Mean Absolute Error;
2. Mean Squared Error;
3. Coefficient of determination;
4. Correlation Coefficient;
5. Direction Accuracy Rate.

Table 1.1.2 Measures of forecasting accuracy

Measure	Simple	50-day	25-day	Double	ARIMA
MAE	0.569	0.393	0.430	0.379	0.235
MSE	0.0045	0.0020	0.0020	0.0019	0.0008
R^2	0.18	0.18	0.08	0.36	0.98
r	0.43	0.42	0.29	0.6	0.99
DA	0.56	0.44	0.33	0.44	0.44

On this occasion, therefore, the model generating the most accurate forecasts has led to the right decision.

Policy comments

Policy comment is to choose the better policy among many different policies to implement, or study the influence of different policies on economic targets. Because the economic policy cannot be experimented in advance, the econometric models that reveal the relations of variables in economic system could be used as an economic policy lab to comment the influence of different policies on economic targets.

There are three approaches to do so, in which the economic target is used as the dependent variable of the economic model and the economic policy as the explanatory variable.
1. Instrument-target method. Given the expected value of economic target, the value of policy variable could be obtained by resolving the model.
2. Policy Simulation method. Calculate respective value of target by inputting different policy variables and choose the better policy to implement.
3. Optimizing Control method. Choose the policy which leads to the optimized target by combing the economic model and optimizing methods.

Verifying and developing economic theory

There are two functions of econometric model. One is to verify the economic theory; that is, constructing the model using certain economic theory and then applied the historic sample data to fit the model; if the fit is good, the theory is confirmed. Another is to discover and develop economic theory. You can use the economic data to fit many kinds of econometric model you can imagine. The relations revealed by the econometric model with the best fit are just the law that the economic behavior follows.

2. Methodology of Econometrics

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

1. Statement of economic theory;
2. Specification of the mathematical model of the theory;
3. Specification of the econometric model of the theory;
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

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.

2. Specification of the Mathematical Model

A mathematical economist might suggest the following form of the Keynesian consumption function:

$$y = \beta_1 + \beta_2 x \quad 0 < \beta_2 < 1 \qquad 1.2.1$$

where y = consumption expenditure and x = income, and where β1 and β2, known as the parameters of the model, are, respectively, the intercept and slope coefficients.

3. Specification of the Econometric Model

To allow for the inexact relationships between economic variables, the econometrician would modify the deterministic consumption function (1.2.1) as follows:

$$y = \beta_1 + \beta_2 x + \varepsilon \qquad 1.2.2$$

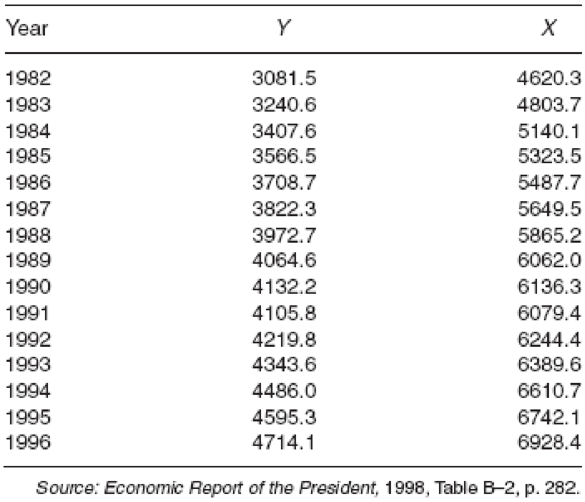
where ε, known as the disturbance, or error, term, is a random (stochastic) variable that has well-defined probabilistic properties. The disturbance term ε may well represent all those factors that affect consumption but are not taken into account explicitly.

4. Obtaining Data

To estimate the econometric model given in (1.2.2), that is, to obtain the numerical values of β 1 and β 2, we need data.

For example we have the data relate to the U.S. economies for the period 1981 – 1996 reported in Table 1.2.1. The data are plotted in the following Figure.

Table 1.2.1



Year	Y	X
1982	3081.5	4620.3
1983	3240.6	4803.7
1984	3407.6	5140.1
1985	3566.5	5323.5
1986	3708.7	5487.7
1987	3822.3	5649.5
1988	3972.7	5865.2
1989	4064.6	6062.0
1990	4132.2	6136.3
1991	4105.8	6079.4
1992	4219.8	6244.4
1993	4343.6	6389.6
1994	4486.0	6610.7
1995	4595.3	6742.1
1996	4714.1	6928.4

Source: Economic Report of the President, 1998, Table B–2, p. 282.

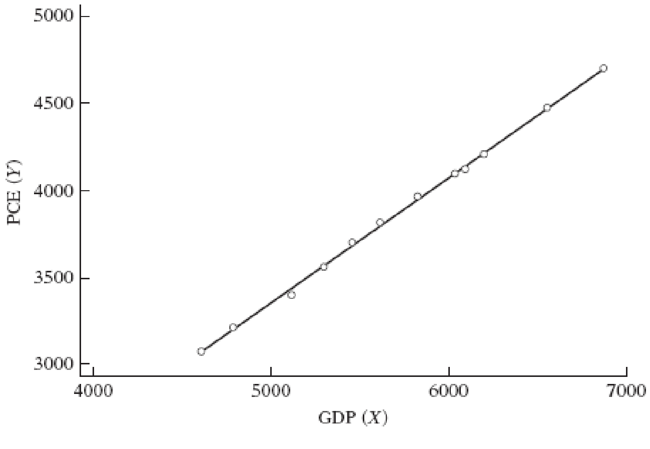


Figure 1.2.3

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 consumption function. The actual mechanics of estimating the parameters will be discussed in Chapter 3 and 4.

Using this technique and the given data, we obtain the following estimates of β1 and β2, namely, −184.08 and 0.7064. Thus, the estimated consumption function is:

$$\hat{y} = -184.08 + 0.7064x \qquad 1.2.3$$

The estimated consumption function (i.e., regression line) is shown in above Figure.

6. Hypothesis Testing

Four tests before forecast:
1. Economic meaning test
Investigate the sign, magnitude of the estimator of parameter of the model.

As noted earlier, Keynes expected the MPC to be positive but less than 1. In our example we found the MPC to be about 0.70.

2. Statistical test
Include the goodness of fit, significant test of the parameters and the function.

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 particular data we have used. In other words, is 0.70 statistically less than 1? If it is, it may support Keynes' theory.
We have to develop suitable criteria to find out whether the estimates are in accord with the expectations of the theory that is being tested. 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. Throughout this course we shall see how this inference process is actually conducted.

3. Econometric test
multicolinearity, heteroscedasticity, and autocorrelation

4. Forecast test
Cross validation

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 variable y on the basis of 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 1997. The GDP value for 1997 was 7269.8 billion dollars. Putting this GDP figure on the right-hand side of (1.2.3), we obtain:

$$\hat{y}_{1997} = -184.08 + 0.7064 \times 7269.8 = 4951.3167 \qquad 1.2.4$$

8. Use of the Model for Policy Purposes

If the regression results given in (1.2.3) seem reasonable, simple arithmetic will show that

$$4900 = -184.08 + 0.7064x \qquad 1.2.5$$

which gives x = 7197, approximately. That is, an income level of about 7197 (billion) dollars, given an MPC of about 0.70, will produce an expenditure of about 4900 billion dollars.

3. Regression Analysis

3.1 Why is called Regression?

The term regression was introduced by Francis Galton. In a famous paper, Galton found that, although there was a tendency for tall parents to have tall children and for short parents to have short children, the average height of children born of parents of a given height tended to move or "regress" toward the average height in the population as a whole.

Definition:

The modern interpretation of regression is, however, quite different. Broadly speaking, we may say Regression analysis is concerned with the study of the dependence of one variable, the dependent variable, on one or more other variables, the explanatory variables, with a view to estimating and/or predicting the (population) mean or average value of the former in terms of the known or fixed (in repeated sampling) values of the latter.

Examples

1. Reconsider Galton's law of universal regression.

Galton was interested in finding out why there was a stability in the distribution of heights in a population. But in the modern view our concern is not with this explanation but rather with finding out how the average height of sons changes, given the fathers' height.

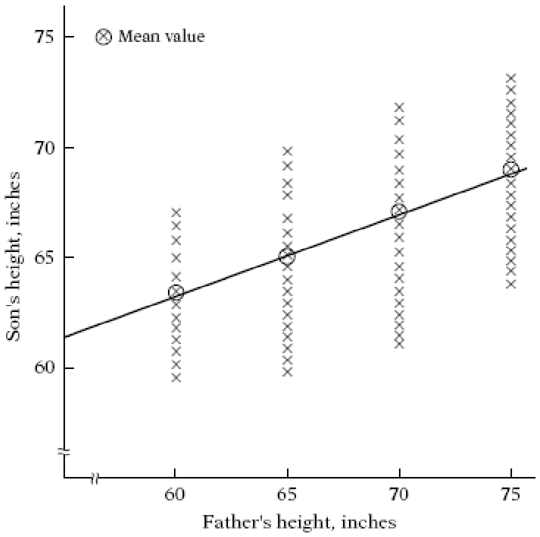


Figure 1.3.1

2. Consider the scatter-gram in Figure 1.3.2, which gives the distribution in a hypothetical population of heights of boys measured at fixed ages.

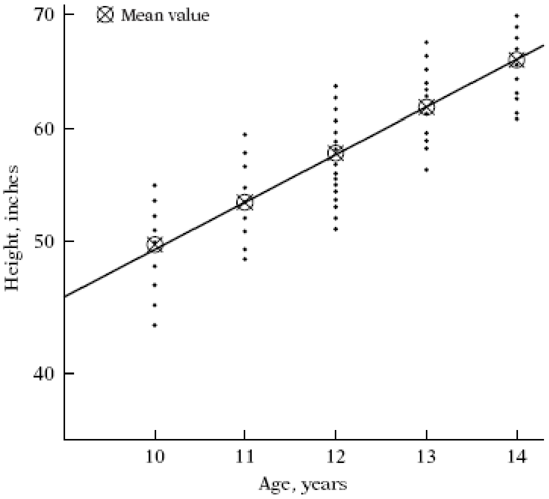


Figure 1.3.2

3. Turning to economic examples, an economist may be interested in studying the dependence of personal consumption expenditure on aftertax or disposable real personal income.

4. A monopolist who can fix the price or output (but not both) may want to find out the response of the demand for a product to changes in price. Such an experiment may enable the estimation of the price elasticity (i.e., price responsiveness) of the demand for the product and may help determine the most profitable price.

5. A labor economist may want to study the rate of change of money wages in relation to the unemployment rate.

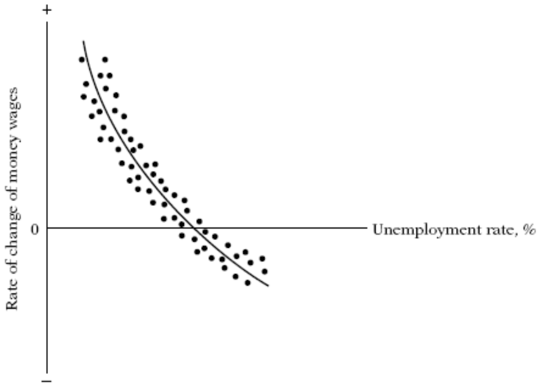


Figure 1.3.3

6. From monetary economics it is known that, other things remaining the same, the higher the rate of inflation π, the lower the proportion k of their income that people would want to hold in the form of money.

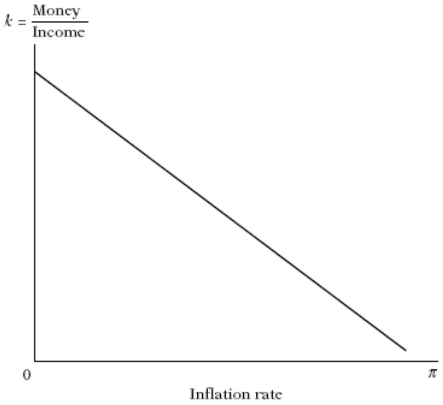


Figure 1.3.4

7. The marketing director of a company may want to know how the demand for the company's product is related to, say, advertising expenditure.

8. Finally, an agronomist may be interested in studying the dependence of crop yield, say, of wheat, on temperature, rainfall, amount of sunshine, and fertilizer. Such a dependence analysis may enable the prediction or forecasting of the average crop yield, given information about the explanatory variables.

The techniques of regression analysis discussed in this course are specially designed to study such dependence among variables.

3.2 Concepts

Statistical versus deterministic relationships

In regression analysis we are concerned with what is known as the statistical, not deterministic, dependence among variables. In statistical relationships among variables we essentially deal with random or stochastic variables, that is, variables that have probability distributions.

In functional or deterministic dependency, on the other hand, we also deal with variables, but these variables are not random or stochastic.

Correlation versus causation

Correlation is the pure mathematical relationships between two variables. The judgment of whether or not to have correlation between two variables is only dependent on the data.

Causation indicates the dependence of variables in behavior mechanism; the variable as the result is determined by the variable as the causation.

The variables with causation are bound to have mathematical correlation, but not vice versa.

Regression versus causation

Although regression analysis deals with the dependence of one variable on other variables, it does not necessarily imply causation.

In the words of Kendall and Stuart, “A statistical relationship, however strong and however suggestive, can never establish causal connection: our ideas of causation must come from outside statistics, ultimately from some theory or other”.

Regression versus correlation

Closely related to but conceptually very much different from regression analysis is correlation analysis, where the primary objective is to measure the strength or degree of linear association between two variables. The correlation coefficient, which we shall study in detail in Chapter 3, measures this strength of (linear) association.

In regression analysis, as already noted, we are not primarily interested in such a measure. Instead, we try to estimate or predict the average value of one variable on the basis of the fixed values of other variables.

Regression and correlation have some fundamental differences that are worth mentioning.

In regression analysis there is an asymmetry in the way the dependent and explanatory variables are treated. The dependent variable is assumed to be statistical, random, or stochastic, that is, to have a probability distribution. The explanatory variables, on the other hand, are assumed to have fixed values (in repeated sampling).

In correlation analysis, on the other hand, we treat any (two) variables symmetrically; there is no distinction between the dependent and explanatory variables. Moreover, both variables are assumed to be random.

3.3 Terminology and notation

1) Terminology

Dependent variable	explanatory variable
Explained variable	independent variable
Predictand	predictor
Regressand	regressor
Response	stimulus
Endogenous	exogenous
Outcome	covariate
Target (controlled) variable	control variable
Output	input

2) Notation

$y\,/\,\mathbf{y}$	the dependent variable
$x\,/\,\mathbf{x}\,/\,\mathbf{X}$	the explanatory variable
$\mathbf{x}_1,\cdots\mathbf{x}_k,\cdots\mathbf{x}_K$	$\mathbf{x}_k$ being the *k*th explanatory variable
$y_i\,/\,y_t$	the *i*th (or *t*th) observation on variable $\mathbf{y}$
$x_{ki}\,/\,x_{kt}$	the *i*th (or *t*th) observation on variable $\mathbf{x}_k$
i	used for cross-sectional data
t	used for time series data
$N\,/\,T$	the total number of observations in the population
$n\,/\,t$	the total number of observations in a sample

4. Data and Variables

4.1 Types of Data

Three types of data may be available for empirical analysis: time series, cross-section, and pooled (i.e., combination of time series and cross-section) or panel data.

Time Series Data

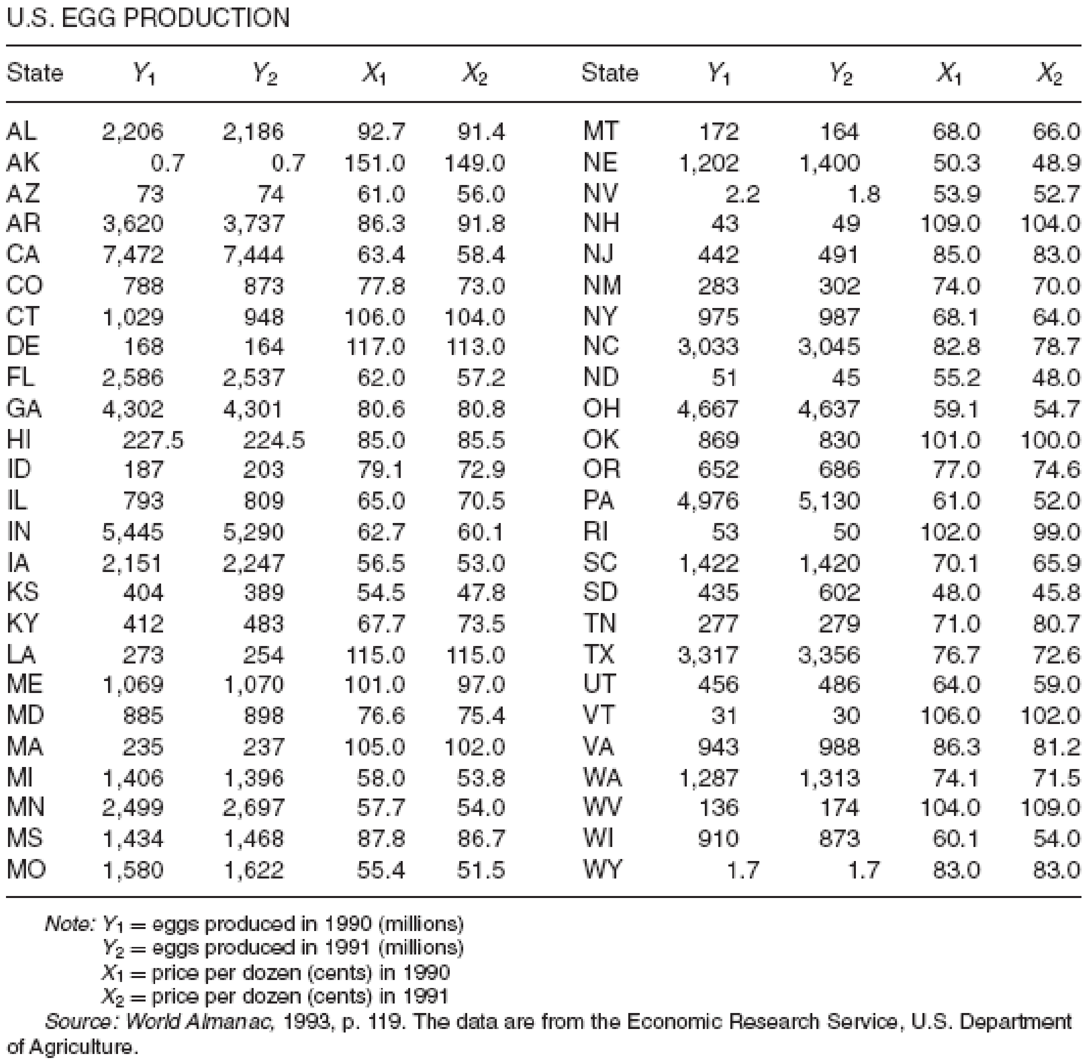
A time series is a set of observations on the values that a variable takes at different times.

Such data may be collected at regular time intervals, such as daily (e.g., stock prices, weather reports), weekly (e.g., money supply figures), monthly [e.g., the unemployment rate, the Consumer Price Index (CPI)], quarterly (e.g., GDP), annually (e.g., government budgets), quinquennially, that is, every 5 years (e.g., the census of manufactures), or decennially (e.g., the census of population).

The data shown in Table 1.2.1 are an example of time series data.

Although time series data are used heavily in econometric studies, they present special problems for econometricians such as autocorrelation, non-stationarity etc.

Table 1.4.1



U.S. EGG PRODUCTION

State	Y_1	Y_2	X_1	X_2	State	Y_1	Y_2	X_1	X_2
AL	2,206	2,186	92.7	91.4	MT	172	164	68.0	66.0
AK	0.7	0.7	151.0	149.0	NE	1,202	1,400	50.3	48.9
AZ	73	74	61.0	56.0	NV	2.2	1.8	53.9	52.7
AR	3,620	3,737	86.3	91.8	NH	43	49	109.0	104.0
CA	7,472	7,444	63.4	58.4	NJ	442	491	85.0	83.0
CO	788	873	77.8	73.0	NM	283	302	74.0	70.0
CT	1,029	948	106.0	104.0	NY	975	987	68.1	64.0
DE	168	164	117.0	113.0	NC	3,033	3,045	82.8	78.7
FL	2,586	2,537	62.0	57.2	ND	51	45	55.2	48.0
GA	4,302	4,301	80.6	80.8	OH	4,667	4,637	59.1	54.7
HI	227.5	224.5	85.0	85.5	OK	869	830	101.0	100.0
ID	187	203	79.1	72.9	OR	652	686	77.0	74.6
IL	793	809	65.0	70.5	PA	4,976	5,130	61.0	52.0
IN	5,445	5,290	62.7	60.1	RI	53	50	102.0	99.0
IA	2,151	2,247	56.5	53.0	SC	1,422	1,420	70.1	65.9
KS	404	389	54.5	47.8	SD	435	602	48.0	45.8
KY	412	483	67.7	73.5	TN	277	279	71.0	80.7
LA	273	254	115.0	115.0	TX	3,317	3,356	76.7	72.6
ME	1,069	1,070	101.0	97.0	UT	456	486	64.0	59.0
MD	885	898	76.6	75.4	VT	31	30	106.0	102.0
MA	235	237	105.0	102.0	VA	943	988	86.3	81.2
MI	1,406	1,396	58.0	53.8	WA	1,287	1,313	74.1	71.5
MN	2,499	2,697	57.7	54.0	WV	136	174	104.0	109.0
MS	1,434	1,468	87.8	86.7	WI	910	873	60.1	54.0
MO	1,580	1,622	55.4	51.5	WY	1.7	1.7	83.0	83.0

Note: Y_1 = eggs produced in 1990 (millions)
Y_2 = eggs produced in 1991 (millions)
X_1 = price per dozen (cents) in 1990
X_2 = price per dozen (cents) in 1991

Source: World Almanac, 1993, p. 119. The data are from the Economic Research Service, U.S. Department of Agriculture.

Cross-Section Data

Cross-section data are data on one or more variables collected at the same point in time, such as the census of population conducted by the Census Bureau every 10 years (the latest being in year 2000), the surveys of consumer expenditures conducted by the University of Michigan, and the opinion polls by Gallup and umpteen other organizations.

A concrete example of cross-sectional data is given in Table 1.4.1 This table gives data on egg production and egg prices for the 50 states in the union for 1990 and 1991. For each year the data on the 50 states are cross-sectional data. Thus, in Table 1.4.1 we have two cross-sectional samples.

Just as time series data create their own special problems, cross-sectional data too have their own problems, specifically the problem of heterogeneity.

Pooled Data

In pooled, or combined, data are elements of both time series and cross-section data.

The data in Table 1.4.1 are an example of pooled data.

Panel, Longitudinal, or Micropanel Data

This is a special type of pooled data in which the same cross-sectional unit (say, a family or a firm) is surveyed over time.

4.2 Categories of Variables

The variables that we will generally encounter fall into four broad categories: ratio scale, interval scale, ordinal scale, and nominal scale.

Ratio Scale

A ratio scale variable satisfies three properties. For a variable X, taking two values, X1 and X2, the ratio X1/X2 and the distance (X2 − X1) are meaningful quantities. Also, there is a natural ordering (ascending or descending) of the values along the scale. Therefore, comparisons such as X2 ≤ X1 or X2 ≥ X1 are meaningful.

Most economic variables such as GDP, income, price, cost etc. belong to this category.

Interval Scale

An interval scale variable satisfies the last two properties of the ratio scale variable but not the first.

Ordinal Scale

A variable belongs to this category only if it satisfies the third property of the ratio scale (i.e., natural ordering).
Examples are grading systems (A, B, C grades), income class (upper, middle, lower) or indifference curves.

Nominal Scale

Variables in this category have none of the features of the ratio scale variables.

Variables such as gender (male, female), marital status (married, unmarried, divorced, separated) , race, nationality, region, party, simply denote categories.

Nominal scale variable can be expressed by dummy variable with two values, 0 and 1.

The former two variables are called the quantitative variable; but the last two are called qualitative variable.