ECONOMETRICS
Chapter # 1: THE NATURE OF REGRESSION ANALYSIS
Domodar N. Gujarati

By: Zahir Mohamed Omar

Facebook Page: Zahir Mohamed (@zahirpinhani)

YouTube Channel: Zahir Mohamed

Chapter Outline

- Historical Origin of the Term Regression
- The Modern Interpretation of Regression
- Statistical versus Deterministic Relationship
- Regression versus Causation
- Regression versus Correlation
- Terminology and Notation
- The nature and sources of data for economic analysis
- The accuracy of data
- A Note On The Measurement Scales Of Variables

Historical Origin of the Term 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. In other words, the height of the children of unusually tall or unusually short parents tends to move toward the average height of the population.
- Galton’s *law of universal regression* was confirmed by his friend Karl Pearson, who collected more than a thousand records of heights of members of family groups. He found that the average height of sons of a group of tall fathers was less than their fathers’ height and the average height of sons of a group of short fathers was greater than their fathers’ height, thus “regressing” tall and short sons alike toward the average height of all men.
- In the words of Galton, this was “regression to mediocrity.”

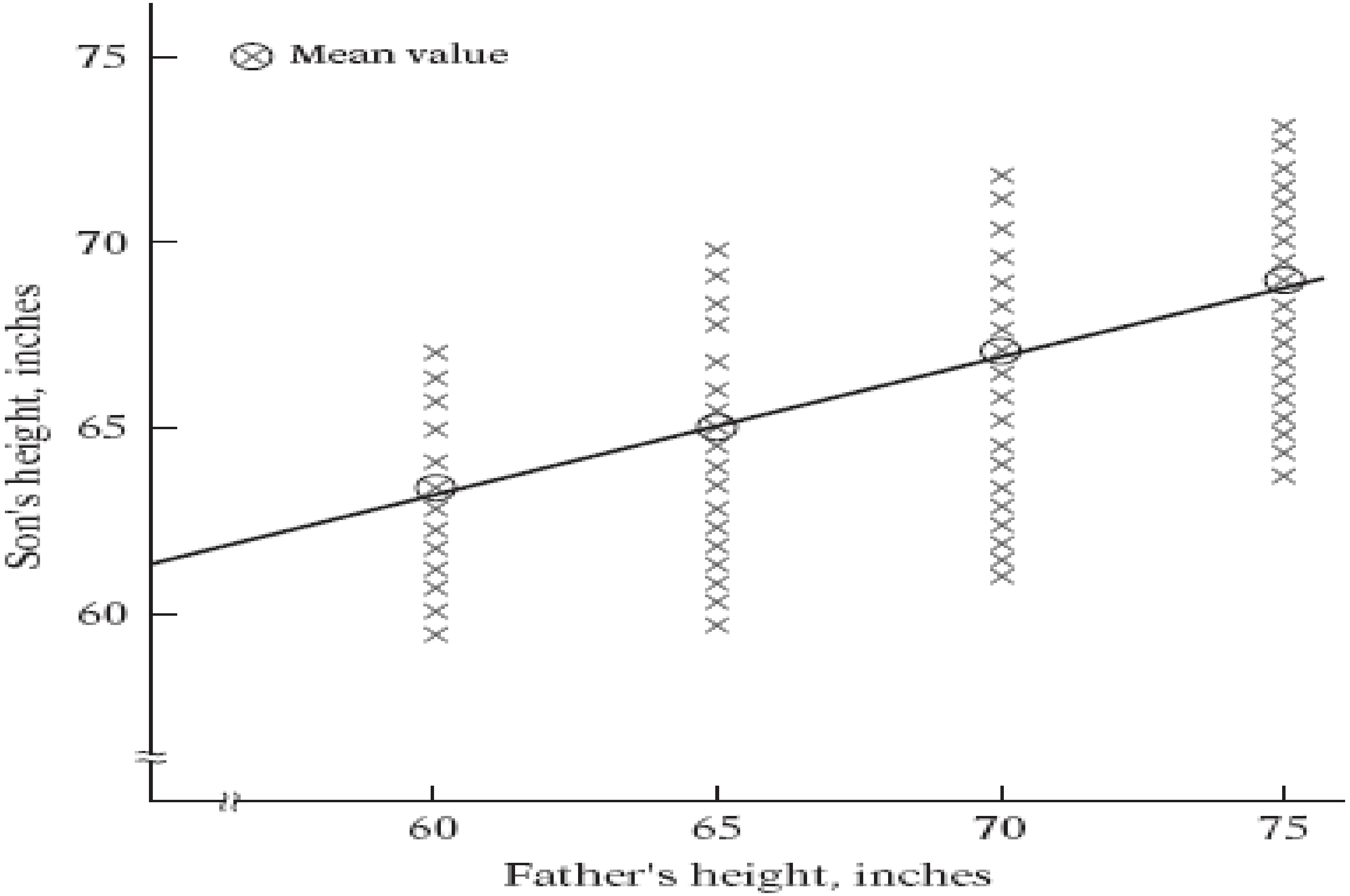
The Modern Interpretation of Regression

- 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. In other words, our concern is with predicting the average height of sons knowing the height of their fathers. To see how this can be done, consider Figure 1.1, which is a **scatter diagram,** or **scattergram.**

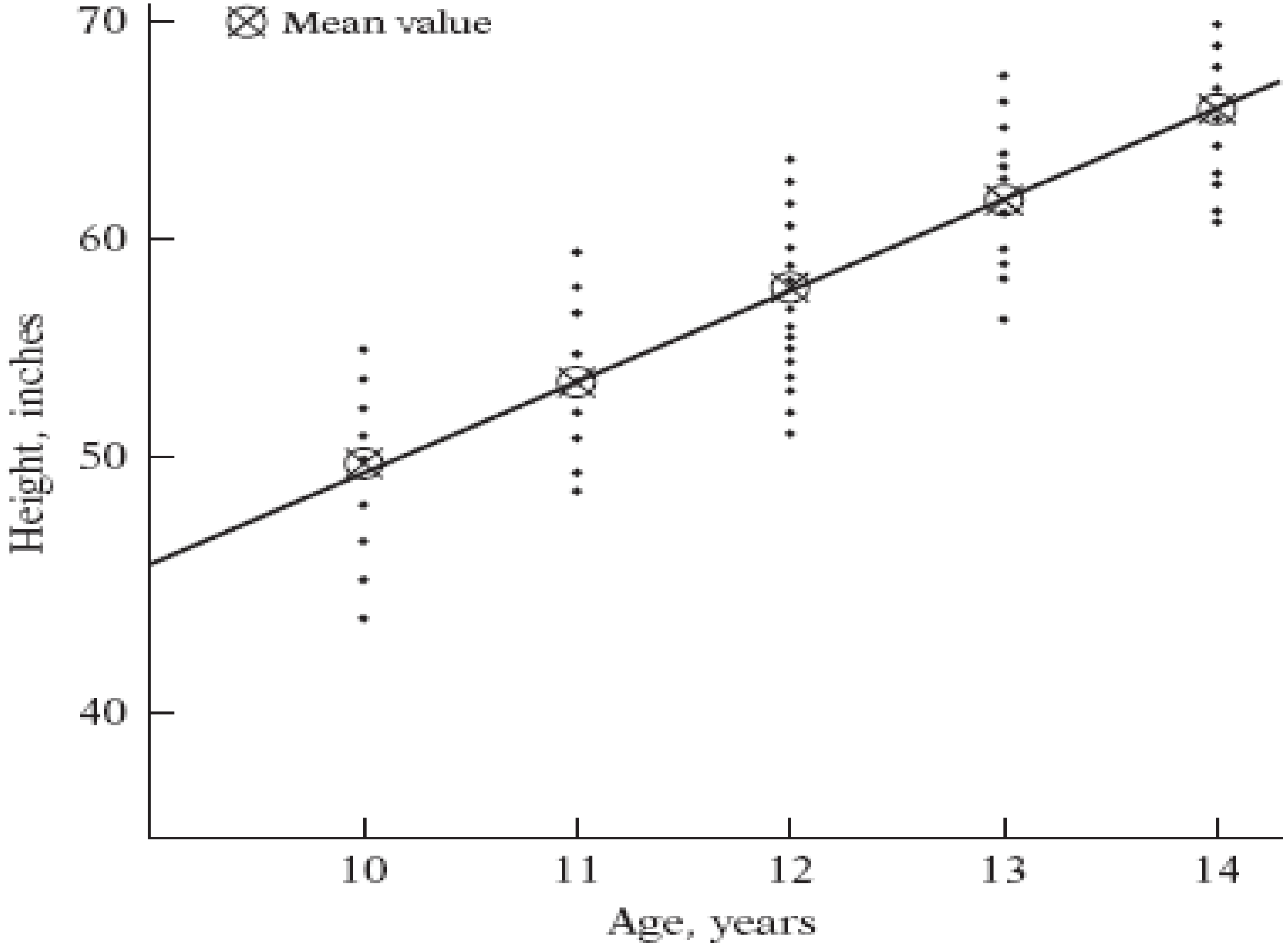
Cont….



⊗ Mean value
75
70
65
60
Son's height, inches
60
65
70
75
Father's height, inches

Cont.....

2. Consider the scattergram in Figure 1.2, which gives the distribution in a hypothetical population of heights of boys measured at *fixed* ages.



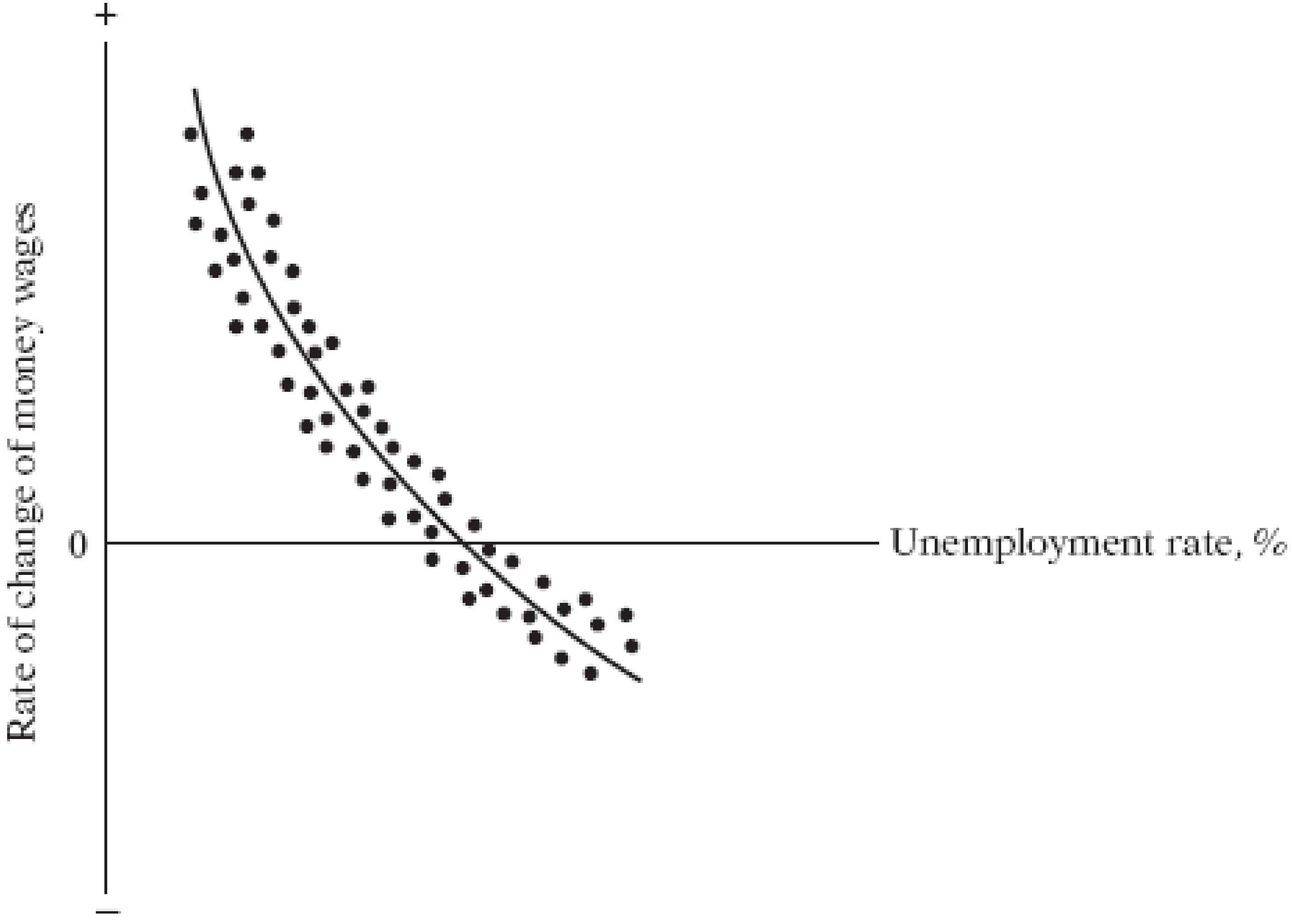
Cont..

3. Turning to economic examples, an economist may be interested in studying the dependence of personal consumption expenditure on after-tax or disposable real personal income. Such an analysis may be helpful in estimating the marginal propensity to consume(MPC).

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.

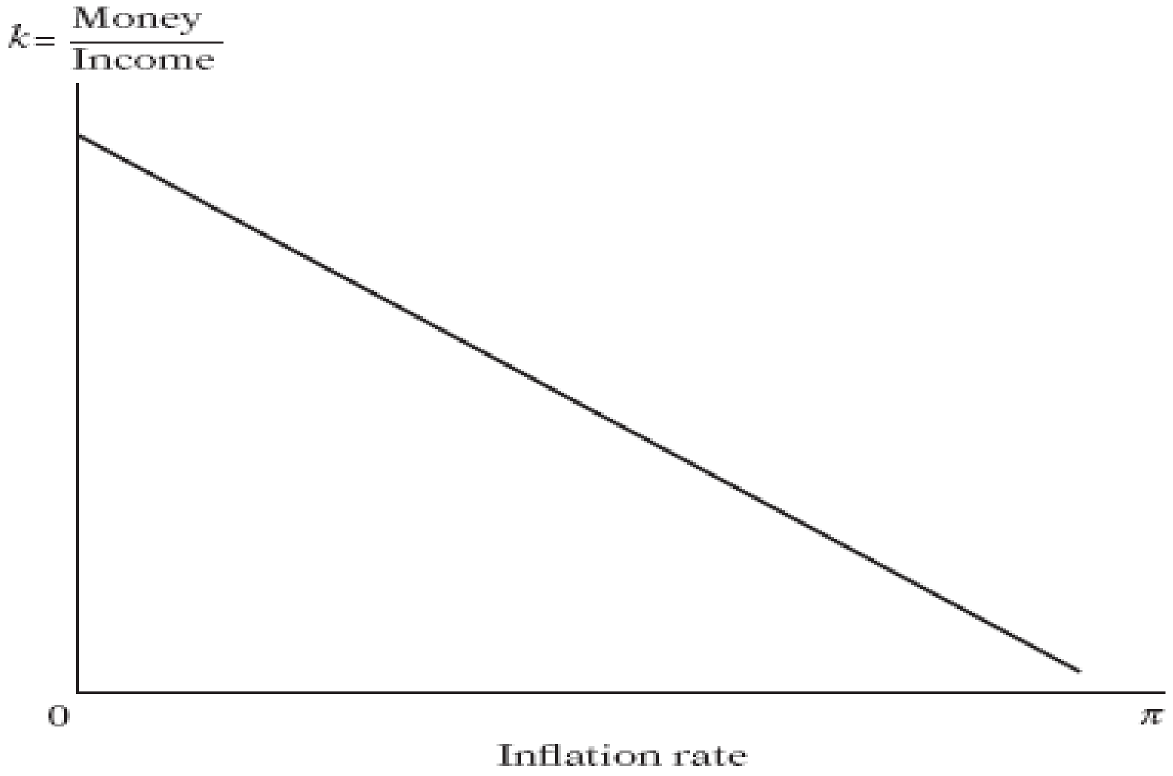
5. A labor economist may want to study the rate of change of money wages in relation to the unemployment rate. The historical data are shown in the scattergram given in Figure 1.3. The curve in Figure 1.3 is an example of the celebrated *Phillips curve* relating changes in the money wages to the unemployment rate.

Figure 1.3



Cont..

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, as depicted in Figure 1.4.



Cont..

- 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. Such a study will be of considerable help in finding out the **elasticity of demand** with respect to advertising expenditure, that is, the percent change in demand in response to, say, a 1 percent change in the advertising budget.
- 8. Finally, an agronomist may be interested in studying the dependence of a particular 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.

Statistical versus Deterministic Relationship

- 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.
- The dependence of crop yield on temperature, rainfall, sunshine, and fertilizer, for example, is statistical in nature.
- In deterministic phenomena, on the other hand, we deal with relationships of the type, say, exhibited by Newton’s law of gravity, which states: Every particle in the universe attracts every other particle with a force directly proportional to the product of their masses and inversely proportional to the square of the distance between them. Symbolically, $F = k(m1m2/r\ 2)$, where F = force, $m1$ and $m2$ are the masses of the two particles, r = distance, and k = constant of proportionality. In this text we are not concerned with such deterministic relationships.

Regression versus Causation

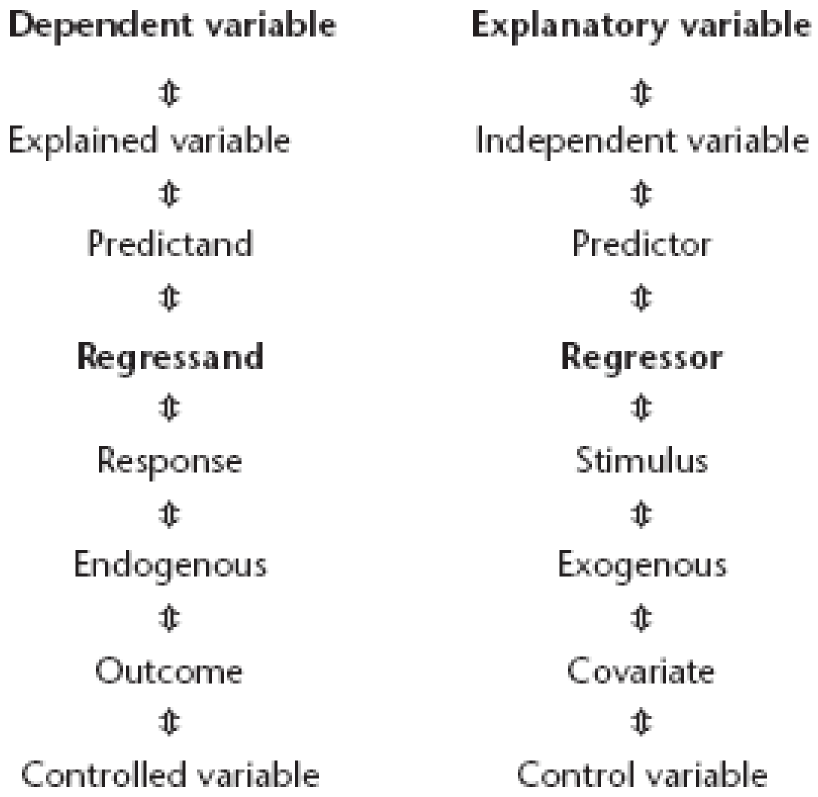
- Although regression analysis deals with the dependence of one variable on other variables, it does not necessarily imply causation.
- In the crop-yield example cited previously, there is no *statistical reason* to assume that rainfall does not depend on crop yield. The fact that we treat crop yield as dependent on rainfall (among other things) is due to non statistical considerations: Common sense suggests that the relationship cannot be reversed, for we cannot control rainfall by varying crop yield.
- **a statistical relationship in itself cannot logically imply causation.** To ascribe causality, one must appeal to a priori or theoretical considerations.

Regression versus Correlation

- In correlation analysis, the primary objective is to measure the strength or degree of linear association between two variables. For example, smoking and lung cancer, scores on statistics and mathematics examinations, and so on. In regression analysis, 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. In regression analysis there is an asymmetry in the way the dependent and explanatory variables are treated.
- In correlation analysis, we treat any (two) variables symmetrically; there is no distinction between the dependent and explanatory variables. The correlation between scores on mathematics and statistics examinations is the same as that between scores on statistics and mathematics examinations. Moreover, both variables are assumed to be random. Whereas most of the regression theory to be dealt with here is conditional upon the assumption that the dependent variable is stochastic but the explanatory variables are fixed or nonstochastic.

Terminology and Notation

In the literature the terms *dependent variable and explanatory variable are described variously. A representative* list is:



Dependent variable	**Explanatory variable**
⇕	⇕
Explained variable	Independent variable
⇕	⇕
Predictand	Predictor
⇕	⇕
Regressand	**Regressor**
⇕	⇕
Response	Stimulus
⇕	⇕
Endogenous	Exogenous
⇕	⇕
Outcome	Covariate
⇕	⇕
Controlled variable	Control variable

Cont..

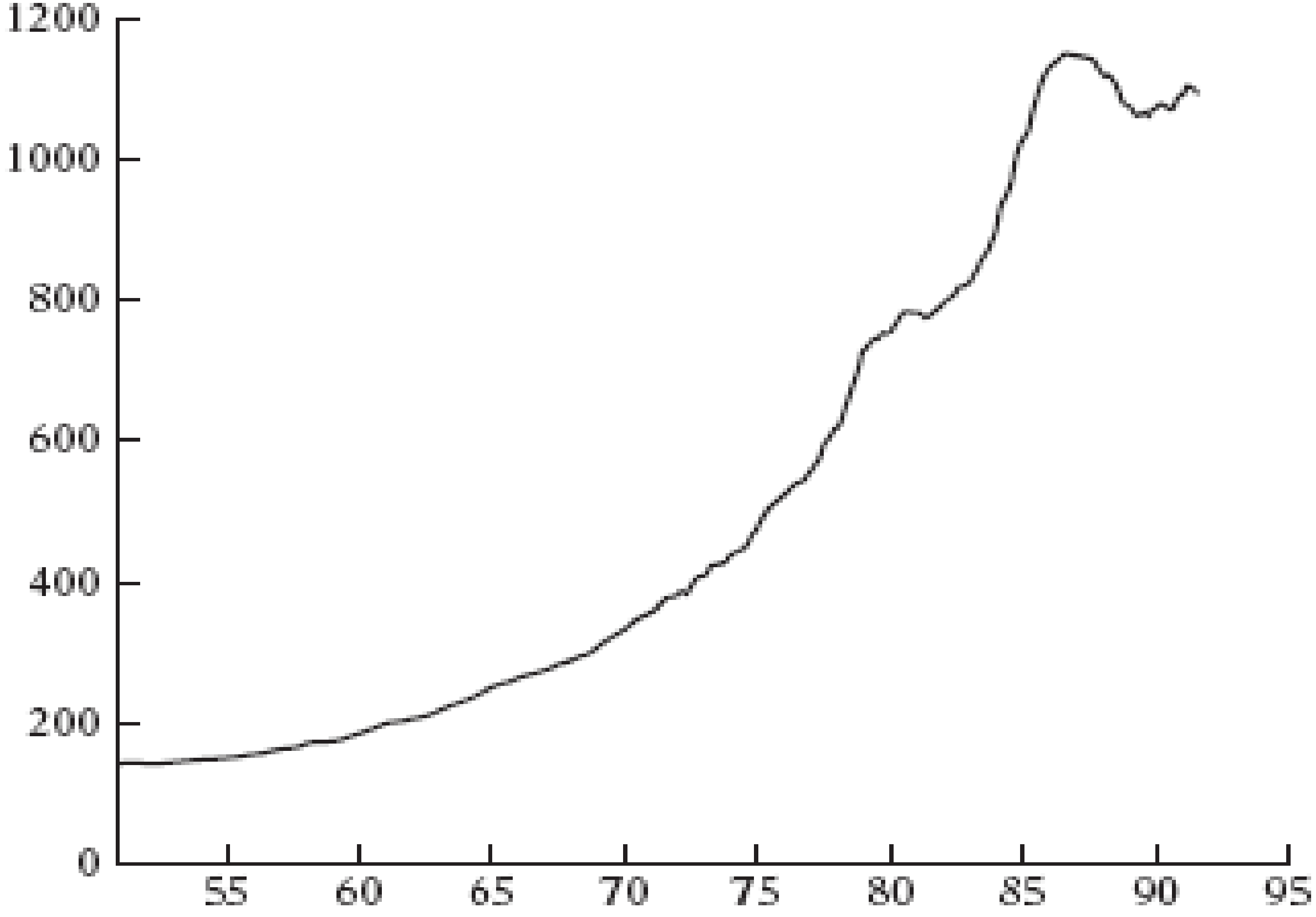
- We will use the dependent variable/explanatory variable or the more neutral, regressand and regressor terminology.
- If we are studying the dependence of a variable on only a single explanatory variable, such a study is known as simple, or two-variable, regression analysis.
- However, if we are studying the dependence of one variable on more than one explanatory variable, it is known as multiple regression analysis.
- The term random is a synonym for the term stochastic. A random or stochastic variable is a variable that can take on any set of values, positive or negative, with a given probability.
- Unless stated otherwise, the letter Y will denote the dependent variable and the X's (X1, X2, . . . , Xk) will denote the explanatory variables, Xk being the kth explanatory variable

The nature and sources of data for economic analysis

Types of Data

- There are three types of data: time series, cross-section, and pooled (i.e., combination of time series and cross-section) data.
- **A time series** is a set of observations on the values that a variable takes at different times. It is collected at regular time intervals, such as daily, weekly, monthly quarterly, annually, quinquennially, that is, every 5 years (e.g., the census of manufactures), or decennially (e.g., the census of population).
- Most empirical work based on time series data assumes that the underlying time series is stationary. Loosely speaking a time series is stationary if its mean and variance do not vary systematically over time.

Figure 1.5 Money Supply m1



Cont..

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. example of cross-sectional data is given in Table 1.1. For each year the data on the 50 states are cross-sectional data. because of the stationarity issue, cross-sectional data too have their own problems, specifically the problem of heterogeneity.

From Table 1.1 we see that we have some states that produce huge amounts of eggs (e.g., Pennsylvania) and some that produce very little (e.g., Alaska). When we include such heterogeneous units in a statistical analysis, the size or scale effect must be taken into account. To see this clearly, we plot in Figure 1.6 the data on eggs produced and their prices in 50 states for the year 1990. This figure shows how widely scattered the observations are

Table 1.1 US egg production

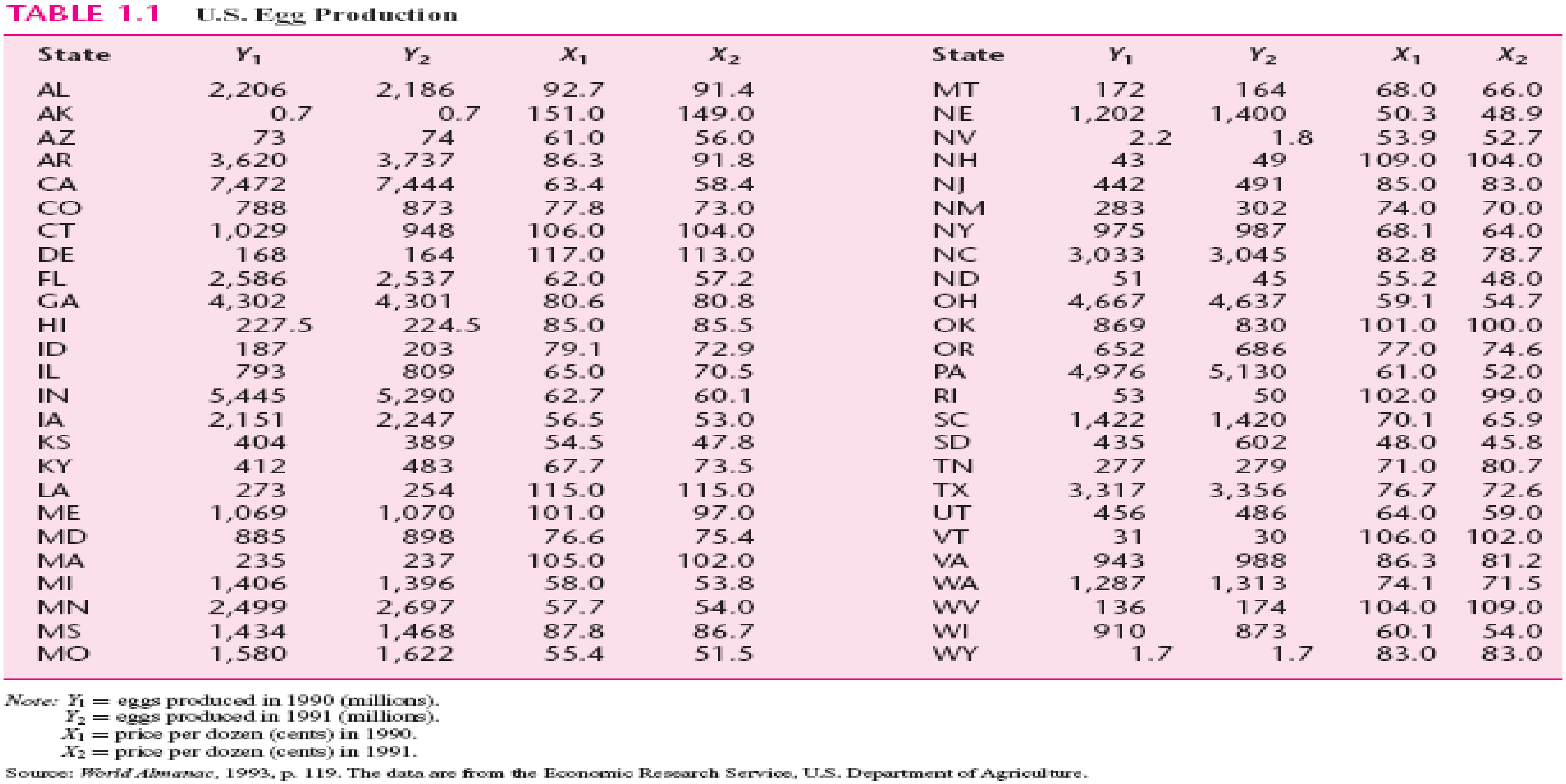


TABLE 1.1 U.S. Egg Production

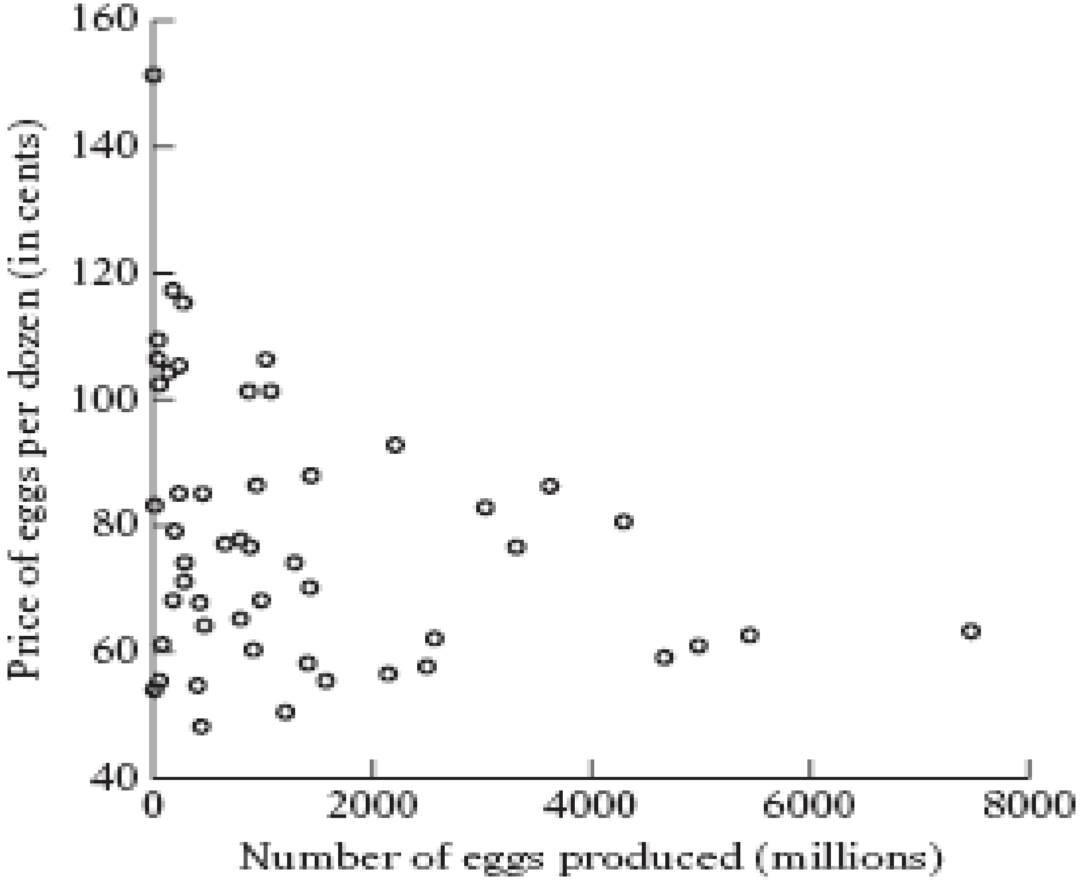
State	Y_1	Y_2	X_1	X_2	State	Y_1	Y_2	X_1	X_2
AL	2,206	2,186	92.7	91.4	MT	172	164	68.0	66.0
AK	0.7	0.7	151.0	149.0	NE	1,202	1,400	50.3	48.9
AZ	73	74	61.0	56.0	NV	2.2	1.8	53.9	52.7
AR	3,620	3,737	86.3	91.8	NH	43	49	109.0	104.0
CA	7,472	7,444	63.4	58.4	NJ	442	491	85.0	83.0
CO	788	873	77.8	73.0	NM	283	302	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.

Figure 1.6

FIGURE 1.6
Relationship between eggs produced and prices, 1990.



Cont..

Pooled Data. In pooled, or combined, data are elements of both time series and cross-section data. The data in Table 1.1 are an example of pooled data. For each year we have 50 cross-sectional observations and for each state we have two time series observations on prices and output of eggs, a total of 100 pooled (or combined) observations.

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.

The Sources of Data

The data used in empirical analysis may be collected by a governmental agency (e.g., the Department of Commerce), an international agency (e.g., the International Monetary Fund (IMF) or the World Bank), a private organization (e.g., the Standard & Poor’s Corporation), or an individual. Literally, there are thousands of such agencies collecting data for one purpose or another.

The accuracy of data

The quality of the data is often not that good. Some reasons for that are:

First, as noted, most social science data are nonexperimental in nature. Therefore, there is the possibility of observational errors, either of omission or commission.

Second, even in experimentally collected data errors of measurement arise from approximations and roundoffs.

Third, in questionnaire-type surveys, the problem of nonresponse can be serious; a researcher is lucky to get a 40% response to a questionnaire.

Fourth, the sampling methods used in obtaining the data may vary so widely that it is often difficult to compare the results obtained from the various samples.

Cont..

Fifth, economic data are generally available at a highly aggregate level. For example, most macrodata (e.g., GNP, inflation, unemployment). Such highly aggregated data may not tell us much about the individual or microunits that may be the ultimate object of study.

Sixth, because of confidentiality, certain data can be published only in highly aggregate form. the Department of Commerce, which conducts the census of business every 5 years, is not allowed to disclose information on production, employment, energy consumption, research and development expenditure, etc., at the firm level. It is therefore difficult to study the interfirm differences on these items.

The researcher should always keep in mind that the results of research are only as good as the quality of the data.

A Note On The Measurement Scales Of Variables

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

Ratio Scale. For a variable X, taking two values, X1 and X2, the ratio X1/X2. Comparisons such as X2 ≤ X1 or X2 ≥ X1 are meaningful.

Interval Scale. The distance between two time periods, say (2000–1995) is meaningful, but not the ratio of two time periods (2000/1995).

Ordinal Scale. Examples are grading systems (A, B, C grades) or income class (upper, middle, lower). For these variables the ordering exists but the distances between the categories cannot be quantified.

Nominal Scale. Variables such as gender and marital status simply denote categories. Such variables cannot be expressed on the ratio, interval, or ordinal scales.

Econometric techniques that may be suitable for ratio scale variables may not be suitable for nominal scale variables. Therefore, it is important to bear in mind the distinctions among the four types