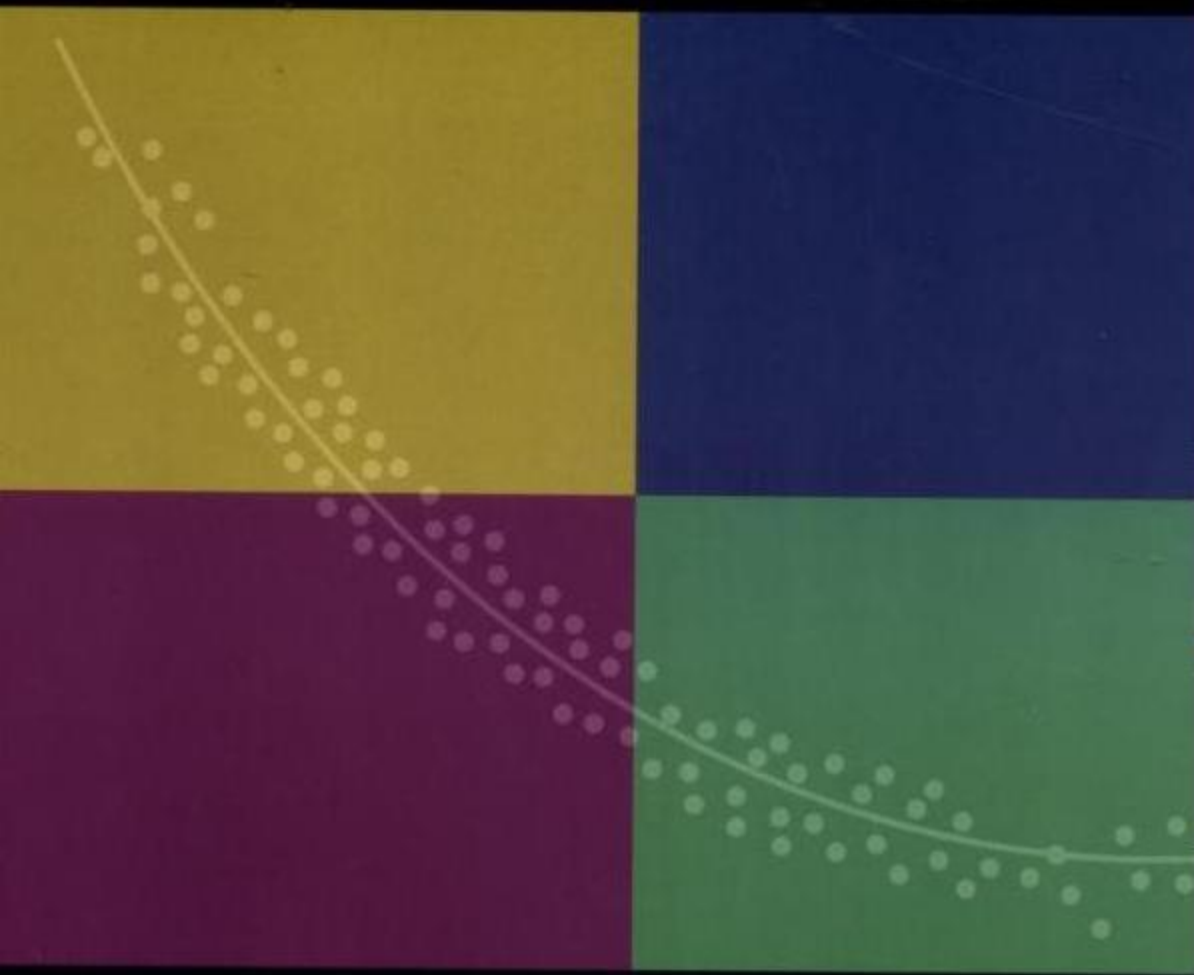
Damodar N. Gujarati



Basic Econometrics

fourth edition

BRIEF CONTENTS

CONTENTS

PREFACE

BACKGROUND AND PURPOSE

As in the previous three editions, the primary objective of the fourth edition of *Basic Econometrics* is to provide an elementary but comprehensive introduction to econometrics without resorting to matrix algebra, calculus, or statistics beyond the elementary level.

In this edition I have attempted to incorporate some of the developments in the theory and practice of econometrics that have taken place since the publication of the third edition in 1995. With the availability of sophisticated and user-friendly statistical packages, such as Eviews, Limdep, Microfit, Minitab, PcGive, SAS, Shazam, and Stata, it is now possible to discuss several econometric techniques that could not be included in the previous editions of the book. I have taken full advantage of these statistical packages in illustrating several examples and exercises in this edition.

I was pleasantly surprised to find that my book is used not only by economics and business students but also by students and researchers in several other disciplines, such as politics, international relations, agriculture, and health sciences. Students in these disciplines will find the expanded discussion of several topics very useful.

THE FOURTH EDITION

The major changes in this edition are as follows:

1. In the introductory chapter, after discussing the steps involved in traditional econometric methodology, I discuss the very important question of how one chooses among competing econometric models.

2. In Chapter 1, I discuss very briefly the measurement scale of economic variables. It is important to know whether the variables are *ratio*

scale, interval scale, ordinal scale, or *nominal scale,* for that will determine the econometric technique that is appropriate in a given situation.

3. The appendices to Chapter 3 now include the large-sample properties of OLS estimators, particularly the property of consistency.

4. The appendix to Chapter 5 now brings into one place the properties and interrelationships among the four important probability distributions that are heavily used in this book, namely, the *normal, t, chi square,* and *F.*

5. Chapter 6, on functional forms of regression models, now includes a discussion of regression on standardized variables.

6. To make the book more accessible to the nonspecialist, I have moved the discussion of the matrix approach to linear regression from old Chapter 9 to Appendix C. Appendix C is slightly expanded to include some advanced material for the benefit of the more mathematically inclined students. The new Chapter 9 now discusses dummy variable regression models.

7. Chapter 10, on multicollinearity, includes an extended discussion of the famous Longley data, which shed considerable light on the nature and scope of multicollinearity.

8. Chapter 11, on heteroscedasticity, now includes in the appendix an intuitive discussion of White's robust standard errors.

9. Chapter 12, on autocorrelation, now includes a discussion of the Newey–West method of correcting the OLS standard errors to take into account likely autocorrelation in the error term. The corrected standard errors are known as HAC standard errors. This chapter also discusses briefly the topic of forecasting with autocorrelated error terms.

10. Chapter 13, on econometric modeling, replaces old Chapters 13 and 14. This chapter has several new topics that the applied researcher will find particularly useful. They include a compact discussion of model selection criteria, such as the *Akaike information criterion,* the *Schwarz information criterion, Mallows's* C_p *criterion,* and *forecast chi square.* The chapter also discusses topics such as *outliers, leverage, influence, recursive least squares,* and *Chow's prediction failure test.* This chapter concludes with some cautionary advice to the practitioner about econometric theory and econometric practice.

11. Chapter 14, on nonlinear regression models, is new. Because of the easy availability of statistical software, it is no longer difficult to estimate regression models that are nonlinear in the parameters. Some econometric models are intrinsically nonlinear in the parameters and need to be estimated by iterative methods. This chapter discusses and illustrates some comparatively simple methods of estimating nonlinear-in-parameter regression models.

12. Chapter 15, on qualitative response regression models, which replaces old Chapter 16, on dummy dependent variable regression models, provides a fairly extensive discussion of regression models that involve a dependent variable that is qualitative in nature. The main focus is on logit

and probit models and their variations. The chapter also discusses the *Poisson regression model,* which is used for modeling count data, such as the number of patents received by a firm in a year; the number of telephone calls received in a span of, say, 5 minutes; etc. This chapter has a brief discussion of multinomial logit and probit models and duration models.

13. Chapter 16, on panel data regression models, is new. A panel data combines features of both time series and cross-section data. Because of increasing availability of panel data in the social sciences, panel data regression models are being increasingly used by researchers in many fields. This chapter provides a nontechnical discussion of the *fixed effects* and *random effects* models that are commonly used in estimating regression models based on panel data.

14. Chapter 17, on dynamic econometric models, has now a rather extended discussion of the Granger causality test, which is routinely used (and misused) in applied research. The Granger causality test is sensitive to the number of lagged terms used in the model. It also assumes that the underlying time series is stationary.

15. Except for new problems and minor extensions of the existing estimation techniques, Chapters 18, 19, and 20 on simultaneous equation models are basically unchanged. This reflects the fact that interest in such models has dwindled over the years for a variety of reasons, including their poor forecasting performance after the OPEC oil shocks of the 1970s.

16. Chapter 21 is a substantial revision of old Chapter 21. Several concepts of time series econometrics are developed and illustrated in this chapter. The main thrust of the chapter is on the nature and importance of stationary time series. The chapter discusses several methods of finding out if a given time series is stationary. Stationarity of a time series is crucial for the application of various econometric techniques discussed in this book.

17. Chapter 22 is also a substantial revision of old Chapter 22. It discusses the topic of economic forecasting based on the *Box–Jenkins (ARIMA)* and *vector autoregression* (VAR) methodologies. It also discusses the topic of measuring volatility in financial time series by the techniques of *autoregressive conditional heteroscedasticity* (ARCH) and *generalized autoregressive conditional heteroscedasticity* (GARCH).

18. Appendix A, on statistical concepts, has been slightly expanded. Appendix C discusses the linear regression model using matrix algebra. This is for the benefit of the more advanced students.

As in the previous editions, all the econometric techniques discussed in this book are illustrated by examples, several of which are based on concrete data from various disciplines. The end-of-chapter questions and problems have several new examples and data sets. For the advanced reader, there are several technical appendices to the various chapters that give proofs of the various theorems and or formulas developed in the text.

ORGANIZATION AND OPTIONS

Changes in this edition have considerably expanded the scope of the text. I hope this gives the instructor substantial flexibility in choosing topics that are appropriate to the intended audience. Here are suggestions about how this book may be used.

One-semester course for the nonspecialist: Appendix A, Chapters 1 through 9, an overview of Chapters 10, 11, 12 (omitting all the proofs).

One-semester course for economics majors: Appendix A, Chapters 1 through 13.

Two-semester course for economics majors: Appendices A, B, C, Chapters 1 to 22. Chapters 14 and 16 may be covered on an optional basis. Some of the technical appendices may be omitted.

Graduate and postgraduate students and researchers: This book is a handy reference book on the major themes in econometrics.

SUPPLEMENTS

Data CD

Every text is packaged with a CD that contains the data from the text in ASCII or text format and can be read by most software packages.

Student Solutions Manual

Free to instructors and salable to students is a Student Solutions Manual (ISBN 0072427922) that contains detailed solutions to the 475 questions and problems in the text.

EViews

With this fourth edition we are pleased to provide Eviews Student Version 3.1 on a CD along with all of the data from the text. This software is available from the publisher packaged with the text (ISBN: 0072565705). Eviews Student Version is available separately from QMS. Go to http://www.eviews.com for further information.

Web Site

A comprehensive web site provides additional material to support the study of econometrics. Go to www.mhhe.com/econometrics/gujarati4.

ACKNOWLEDGMENTS

Since the publication of the first edition of this book in 1978, I have received valuable advice, comments, criticism, and suggestions from a variety of people. In particular, I would like to acknowledge the help I have received

from Michael McAleer of the University of Western Australia, Peter Kennedy of Simon Frazer University in Canada, and Kenneth White, of the University of British Columbia, George K. Zestos of Christopher Newport University, Virginia, and Paul Offner, Georgetown University, Washington, D.C.

I am also grateful to several people who have influenced me by their scholarship. I especially want to thank Arthur Goldberger of the University of Wisconsin, William Greene of New York University, and the late G. S. Maddala. For this fourth edition I am especially grateful to these reviewers who provided their invaluable insight, criticism, and suggestions: Michael A. Grove at the University of Oregon, Harumi Ito at Brown University, Han Kim at South Dakota University, Phanindra V. Wunnava at Middlebury College, and George K. Zestos of Christopher Newport University.

Several authors have influenced my writing. In particular, I am grateful to these authors: Chandan Mukherjee, director of the Centre for Development Studies, Trivandrum, India; Howard White and Marc Wuyts, both at the Institute of Social Studies in the Netherlands; Badi H. Baltagi, Texas A&M University; B. Bhaskara Rao, University of New South Wales, Australia; R. Carter Hill, Louisiana University; William E. Griffiths, University of New England; George G. Judge, University of California at Berkeley; Marno Verbeek, Center for Economic Studies, KU Leuven; Jeffrey Wooldridge, Michigan State University; Kerry Patterson, University of Reading, U.K.; Francis X. Diebold, Wharton School, University of Pennsylvania; Wojciech W. Charemza and Derek F. Deadman, both of the University of Leicester, U.K.; Gary Koop, University of Glasgow.

I am very grateful to several of my colleagues at West Point for their support and encouragement over the years. In particular, I am grateful to Brigadier General Daniel Kaufman, Colonel Howard Russ, Lieutenant Colonel Mike Meese, Lieutenant Colonel Casey Wardynski, Major David Trybulla, Major Kevin Foster, Dean Dudley, and Dennis Smallwood.

I would like to thank students and teachers all over the world who have not only used my book but have communicated with me about various aspects of the book.

For their behind the scenes help at McGraw-Hill, I am grateful to Lucille Sutton, Aric Bright, and Catherine R. Schultz.

George F. Watson, the copyeditor, has done a marvellous job in editing a rather lengthy and demanding manuscript. For that, I am much obliged to him.

Finally, but not least important, I would like to thank my wife, Pushpa, and my daughters, Joan and Diane, for their constant support and encouragement in the preparation of this and the previous editions.

Damodar N. Gujarati

INTRODUCTION

I.1 WHAT IS 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.[2]

> 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.[3]

> Econometrics is concerned with the empirical determination of economic laws.[4]

[1]Gerhard Tintner, *Methodology of Mathematical Economics and Econometrics,* The University of Chicago Press, Chicago, 1968, p. 74.
[2]P. A. Samuelson, T. C. Koopmans, and J. R. N. Stone, "Report of the Evaluative Committee for *Econometrica,*" *Econometrica,* vol. 22, no. 2, April 1954, pp. 141–146.
[3]Arthur S. Goldberger, *Econometric Theory,* John Wiley & Sons, New York, 1964, p. 1.
[4]H. Theil, *Principles of Econometrics,* John Wiley & Sons, New York, 1971, p. 1.

> The art of the econometrician consists in finding the set of assumptions that are both sufficiently specific and sufficiently realistic to allow him to take the best possible advantage of the data available to him.[5]
>
> Econometricians . . . are a positive help in trying to dispel the poor public image of economics (quantitative or otherwise) as a subject in which empty boxes are opened by assuming the existence of can-openers to reveal contents which any ten economists will interpret in 11 ways.[6]
>
> The method of econometric research aims, essentially, at a conjunction of economic theory and actual measurements, using the theory and technique of statistical inference as a bridge pier.[7]

I.2 WHY A SEPARATE DISCIPLINE?

As the preceding definitions suggest, econometrics is an amalgam of economic theory, mathematical economics, economic statistics, and mathematical statistics. Yet the subject deserves to be studied in its own right for the following reasons.

Economic theory makes statements or hypotheses that are mostly qualitative in nature. For example, microeconomic theory states that, other things remaining the same, a reduction in the price of a commodity is expected to increase the quantity demanded of that commodity. Thus, economic theory postulates a negative or inverse relationship between the price and quantity demanded of a commodity. But the theory itself does not provide any numerical measure of the relationship between the two; that is, it does not tell by how much the quantity will go up or down as a result of a certain change in the price of the commodity. It is the job of the econometrician to provide such numerical estimates. Stated differently, econometrics gives empirical 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. As we shall see, the econometrician often uses the mathematical equations proposed by the mathematical economist but puts these equations in such a form that they lend themselves to empirical testing. And this conversion of mathematical into econometric equations requires a great deal of ingenuity and practical skill.

Economic statistics is mainly concerned with collecting, processing, and presenting economic data in the form of charts and tables. These are the

[5]E. Malinvaud, *Statistical Methods of Econometrics,* Rand McNally, Chicago, 1966, p. 514.

[6]Adrian C. Darnell and J. Lynne Evans, *The Limits of Econometrics,* Edward Elgar Publishing, Hants, England, 1990, p. 54.

[7]T. Haavelmo, "The Probability Approach in Econometrics," Supplement to *Econometrica,* vol. 12, 1944, preface p. iii.

jobs of the economic statistician. It is he or she who is primarily responsible for collecting data on gross national product (GNP), employment, unemployment, prices, etc. 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. Of course, one who does that becomes an econometrician.

Although mathematical statistics provides many tools used in the trade, 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, like the meteorologist, generally depends on data that cannot be controlled directly. As Spanos correctly observes:

> In econometrics the modeler is often faced with **observational** as opposed to **experimental** data. This has two important implications for empirical modeling in econometrics. First, the modeler is required to master very different skills than those needed for analyzing experimental data. . . . Second, the separation of the data collector and the data analyst requires the modeler to familiarize himself/herself thoroughly with the nature and structure of data in question.[8]

I.3 METHODOLOGY 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.[9]

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.

---

[8]Aris Spanos, *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data,* Cambridge University Press, United Kingdom, 1999, p. 21.

[9]For an enlightening, if advanced, discussion on econometric methodology, see David F. Hendry, *Dynamic Econometrics,* Oxford University Press, New York, 1995. See also Aris Spanos, *op. cit.*

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:

$$Y = \beta_1 + \beta_2 X \qquad 0 < \beta_2 < 1 \qquad \textbf{(I.3.1)}$$

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, Eq. (I.3.1) is as shown in Figure I.1. This equation, which states that consumption is lin-

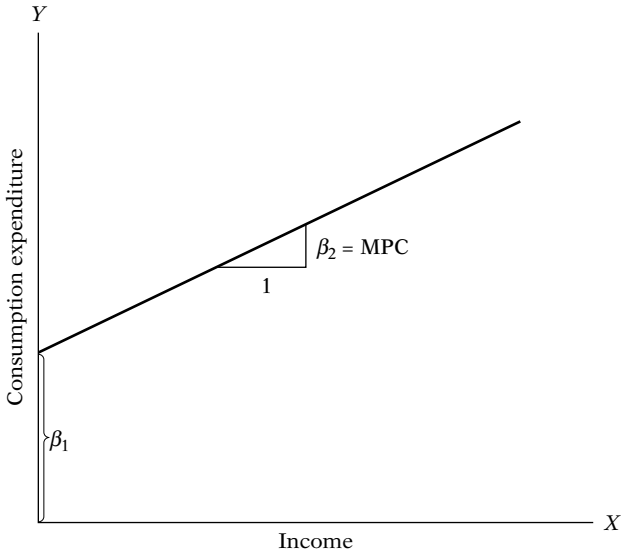


FIGURE I.1 Keynesian consumption function.

[10]John Maynard Keynes, *The General Theory of Employment, Interest and Money,* Harcourt Brace Jovanovich, New York, 1936, p. 96.

early 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 mathematical 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 are called the **independent,** or **explanatory,** variable(s). Thus, in the Keynesian consumption function, Eq. (I.3.1), consumption (expenditure) is the dependent variable and income is the explanatory variable.

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 variables 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 income 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 consumption 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 (I.3.1) as follows:

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

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.

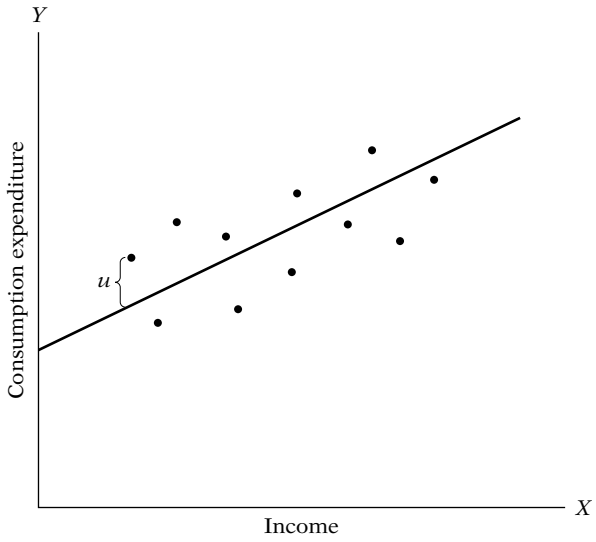


FIGURE I.2 Econometric model of the Keynesian consumption function.

4. Obtaining Data

To estimate the econometric model given in (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

TABLE I.1 DATA ON *Y* (PERSONAL CONSUMPTION EXPENDITURE) AND *X* (GROSS DOMESTIC PRODUCT, 1982–1996), BOTH IN 1992 BILLIONS OF DOLLARS

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.

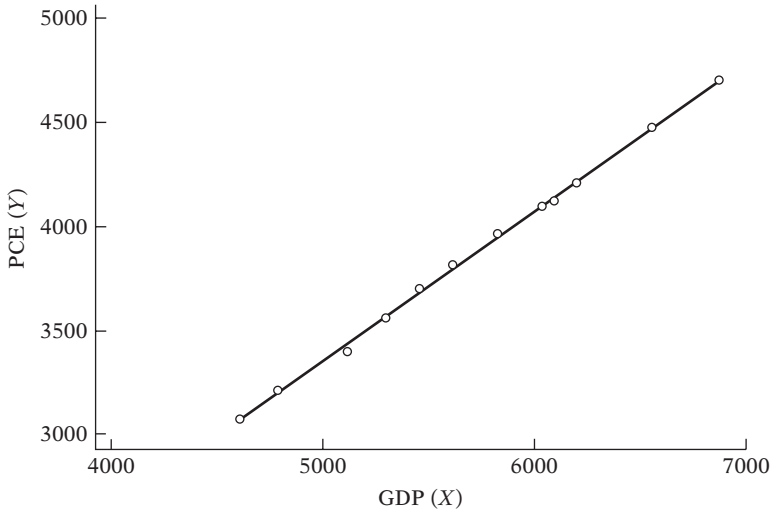


FIGURE I.3 Personal consumption expenditure (*Y*) in relation to GDP (*X*), 1982–1996, both in billions of 1992 dollars.

the U.S. economy for the period 1981–1996. 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 1992 dollars. Therefore, the data are in "real" terms; that is, they are measured in constant (1992) 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 consumption 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, −184.08 and 0.7064. Thus, the estimated consumption function is:

$$\hat{Y} = -184.08 + 0.7064X_i \qquad \textbf{(I.3.3)}$$

The hat on the *Y* indicates that it is an estimate.[11] The estimated consumption function (i.e., regression line) is shown in Figure I.3.

[11]As a matter of convention, a hat over a variable or parameter indicates that it is an estimated value.

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 1982–1996 the slope coefficient (i.e., the **MPC**) was about 0.70, suggesting that for the sample period an increase in real income of 1 dollar led, *on average*, to an increase of about 70 cents in real consumption expenditure.[12] We say *on average* because the relationship between consumption 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 70 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, Eq. (I.3.3) are in accord with the expectations of the theory that is being tested. According to "positive" 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 example we found the MPC to be about 0.70. 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.

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 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.[14] Putting

[12]Do not worry now about how these values were obtained. As we show in Chap. 3, the statistical method of **least squares** has produced these estimates. Also, for now do not worry about the negative value of the intercept.

[13]See Milton Friedman, "The Methodology of Positive Economics," *Essays in Positive Economics,* University of Chicago Press, Chicago, 1953.

[14]Data on PCE and GDP were available for 1997 but we purposely left them out to illustrate the topic discussed in this section. As we will discuss in subsequent chapters, it is a good idea to save a portion of the data to find out how well the fitted model predicts the out-of-sample observations.

this GDP figure on the right-hand side of (I.3.3), we obtain:

$$\begin{aligned}\hat{Y}_{1997} &= -184.0779 + 0.7064\,(7269.8) \\ &= 4951.3167\end{aligned} \tag{I.3.4}$$

or about 4951 billion dollars. Thus, given the value of the GDP, the mean, or average, forecast consumption expenditure is about 4951 billion dollars. The actual value of the consumption expenditure reported in 1997 was 4913.5 billion dollars. The estimated model (I.3.3) thus **overpredicted** the actual consumption expenditure by about 37.82 billion dollars. We could say the **forecast error** is about 37.82 billion dollars, which is about 0.76 percent of the actual GDP value for 1997. 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.

There is another use of the estimated model (I.3.3). Suppose the President decides to propose a reduction in the income tax. What will be the effect of such a policy on income and thereby on consumption expenditure and ultimately on employment?

Suppose that, as a result of the proposed policy change, investment expenditure increases. What will be the effect on the economy? As macroeconomic theory shows, the change in income following, say, a dollar's worth of change in investment expenditure is given by the **income multiplier *M*,** which is defined as

$$M = \frac{1}{1 - \text{MPC}} \tag{I.3.5}$$

If we use the MPC of 0.70 obtained in (I.3.3), this multiplier becomes about $M = 3.33$. That is, an increase (decrease) of a dollar in investment will *eventually* lead to more than a threefold increase (decrease) in income; note that it takes time for the multiplier to work.

The critical value in this computation is MPC, for the multiplier depends on it. And this estimate of the MPC can be obtained from regression models such as (I.3.3). Thus, a quantitative estimate of MPC provides valuable information for policy purposes. Knowing MPC, one can predict the future course of income, consumption expenditure, and employment following a change in the government's fiscal policies.

8. Use of the Model for Control or Policy Purposes

Suppose we have the estimated consumption function given in (I.3.3). Suppose further the government believes that consumer expenditure of about 4900 (billions of 1992 dollars) will keep the unemployment rate at its

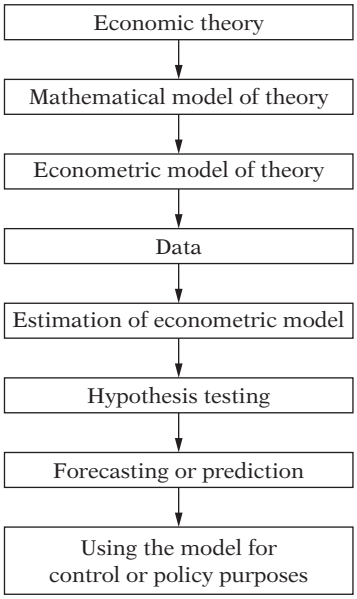


FIGURE I.4 Anatomy of econometric modeling.

current level of about 4.2 percent (early 2000). What level of income will guarantee the target amount of consumption expenditure?

If the regression results given in (I.3.3) seem reasonable, simple arithmetic will show that

$$4900 = -184.0779 + 0.7064X \qquad \textbf{(I.3.6)}$$

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.

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*.**

Figure I.4 summarizes the anatomy of classical econometric modeling.

Choosing among Competing Models

When a governmental agency (e.g., the U.S. Department of Commerce) collects economic data, such as that shown in Table I.1, it does not necessarily have any economic theory in mind. How then does one know that the data really support the Keynesian theory of consumption? Is it because the Keynesian consumption function (i.e., the regression line) shown in Figure I.3 is extremely close to the actual data points? Is it possible that an-

other consumption model (theory) might equally fit the data as well? For example, Milton Friedman has developed a model of consumption, called the *permanent income hypothesis*.[15] Robert Hall has also developed a model of consumption, called the *life-cycle permanent income hypothesis*.[16] Could one or both of these models also fit the data in Table I.1?

In short, the question facing a researcher in practice is how to choose among competing hypotheses or models of a given phenomenon, such as the consumption–income relationship. As Miller contends:

> No encounter with data is step towards genuine confirmation unless the hypothesis does a better job of coping with the data than some natural rival. . . . What strengthens a hypothesis, here, is a victory that is, at the same time, a defeat for a plausible rival.[17]

How then does one choose among competing models or hypotheses? Here the advice given by Clive Granger is worth keeping in mind:[18]

> I would like to suggest that in the future, when you are presented with a new piece of theory or empirical model, you ask these questions:
>
> **(i)** What purpose does it have? What economic decisions does it help with? and;
> **(ii)** Is there any evidence being presented that allows me to evaluate its quality compared to alternative theories or models?
>
> I think attention to such questions will strengthen economic research and discussion.

As we progress through this book, we will come across several competing hypotheses trying to explain various economic phenomena. For example, students of economics are familiar with the concept of the production function, which is basically a relationship between output and inputs (say, capital and labor). In the literature, two of the best known are the *Cobb–Douglas* and the *constant elasticity of substitution* production functions. Given the data on output and inputs, we will have to find out which of the two production functions, if any, fits the data well.

The eight-step classical econometric methodology discussed above is neutral in the sense that it can be used to test any of these rival hypotheses.

Is it possible to develop a methodology that is comprehensive enough to include competing hypotheses? This is an involved and controversial topic.

---

[15]Milton Friedman, *A Theory of Consumption Function,* Princeton University Press, Princeton, N.J., 1957.

[16]R. Hall, "Stochastic Implications of the Life Cycle Permanent Income Hypothesis: Theory and Evidence," *Journal of Political Economy,* 1978, vol. 86, pp. 971–987.

[17]R. W. Miller, *Fact and Method: Explanation, Confirmation, and Reality in the Natural and Social Sciences,* Princeton University Press, Princeton, N.J., 1978, p. 176.

[18]Clive W. J. Granger, *Empirical Modeling in Economics,* Cambridge University Press, U.K., 1999, p. 58.

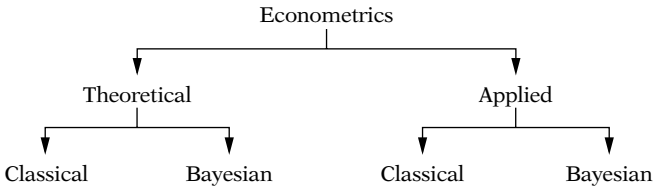


FIGURE I.5 Categories of econometrics.

We will discuss it in Chapter 13, after we have acquired the necessary econometric theory.

I.4 TYPES OF ECONOMETRICS

As the classificatory scheme in Figure I.5 suggests, econometrics may be divided into two broad categories: **theoretical econometrics** and **applied econometrics.** In each category, one can approach the subject in the **classical** or **Bayesian** tradition. In this book the emphasis is on the classical approach. For the Bayesian approach, the reader may consult the references given at the end of the chapter.

Theoretical econometrics is concerned with the development of appropriate methods for measuring economic relationships specified by econometric models. In this aspect, econometrics leans heavily on mathematical statistics. For example, one of the methods used extensively in this book is **least squares.** Theoretical econometrics 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.

This book is concerned largely with the development of econometric methods, their assumptions, their uses, their limitations. These methods are illustrated with examples from various areas of economics and business. But this is *not* a book of applied econometrics in the sense that it delves deeply into any particular field of economic application. That job is best left to books written specifically for this purpose. References to some of these books are provided at the end of this book.

I.5 MATHEMATICAL AND STATISTICAL PREREQUISITES

Although this book is written at an elementary level, the author assumes that the reader is familiar with the basic concepts of statistical estimation and hypothesis testing. However, a broad but nontechnical overview of the basic statistical concepts used in this book is provided in **Appendix A** for

the benefit of those who want to refresh their knowledge. Insofar as mathematics is concerned, a nodding acquaintance with the notions of differential calculus is desirable, although not essential. Although most graduate level books in econometrics make heavy use of matrix algebra, I want to make it clear that it is not needed to study this book. It is my strong belief that the fundamental ideas of econometrics can be conveyed without the use of matrix algebra. However, for the benefit of the mathematically inclined student, **Appendix C** gives the summary of basic regression theory in matrix notation. For these students, **Appendix B** provides a succinct summary of the main results from matrix algebra.

I.6 THE ROLE OF THE COMPUTER

Regression analysis, the bread-and-butter tool of econometrics, these days is unthinkable without the computer and some access to statistical software. (Believe me, I grew up in the generation of the slide rule!) Fortunately, several excellent regression packages are commercially available, both for the mainframe and the microcomputer, and the list is growing by the day. Regression software packages, such as **ET, LIMDEP, SHAZAM, MICRO TSP, MINITAB, EVIEWS, SAS, SPSS, STATA, Microfit, PcGive,** and **BMD** have most of the econometric techniques and tests discussed in this book.

In this book, from time to time, the reader will be asked to conduct **Monte Carlo** experiments using one or more of the statistical packages. Monte Carlo experiments are "fun" exercises that will enable the reader to appreciate the properties of several statistical methods discussed in this book. The details of the Monte Carlo experiments will be discussed at appropriate places.

I.7 SUGGESTIONS FOR FURTHER READING

The topic of econometric methodology is vast and controversial. For those interested in this topic, I suggest the following books:

Neil de Marchi and Christopher Gilbert, eds., *History and Methodology of Econometrics*, Oxford University Press, New York, 1989. This collection of readings discusses some early work on econometric methodology and has an extended discussion of the British approach to econometrics relating to time series data, that is, data collected over a period of time.

Wojciech W. Charemza and Derek F. Deadman, *New Directions in Econometric Practice: General to Specific Modelling, Cointegration and Vector Autogression*, 2d ed., Edward Elgar Publishing Ltd., Hants, England, 1997. The authors of this book critique the traditional approach to econometrics and give a detailed exposition of new approaches to econometric methodology.

Adrian C. Darnell and J. Lynne Evans, *The Limits of Econometrics,* Edward Elgar Publishers Ltd., Hants, England, 1990. The book provides a somewhat

balanced discussion of the various methodological approaches to econometrics, with renewed allegiance to traditional econometric methodology.

Mary S. Morgan, *The History of Econometric Ideas,* Cambridge University Press, New York, 1990. The author provides an excellent historical perspective on the theory and practice of econometrics, with an in-depth discussion of the early contributions of Haavelmo (1990 Nobel Laureate in Economics) to econometrics. In the same spirit, David F. Hendry and Mary S. Morgan, *The Foundation of Econometric Analysis*, Cambridge University Press, U.K., 1995, have collected seminal writings in econometrics to show the evolution of econometric ideas over time.

David Colander and Reuven Brenner, eds., *Educating Economists*, University of Michigan Press, Ann Arbor, Michigan, 1992, present a critical, at times agnostic, view of economic teaching and practice.

For Bayesian statistics and econometrics, the following books are very useful: John H. Dey, *Data in Doubt*, Basic Blackwell Ltd., Oxford University Press, England, 1985. Peter M. Lee, *Bayesian Statistics: An Introduction*, Oxford University Press, England, 1989. Dale J. Porier, *Intermediate Statistics and Econometrics: A Comparative Approach*, MIT Press, Cambridge, Massachusetts, 1995. Arnold Zeller, *An Introduction to Bayesian Inference in Econometrics*, John Wiley & Sons, New York, 1971, is an advanced reference book.

PART ONE

SINGLE-EQUATION REGRESSION MODELS

Part I of this text introduces single-equation regression models. In these models, one variable, called the *dependent variable,* is expressed as a linear function of one or more other variables, called the *explanatory variables*. In such models it is assumed implicitly that causal relationships, if any, between the dependent and explanatory variables flow in one direction only, namely, from the explanatory variables to the dependent variable.

In Chapter 1, we discuss the historical as well as the modern interpretation of the term *regression* and illustrate the difference between the two interpretations with several examples drawn from economics and other fields.

In Chapter 2, we introduce some fundamental concepts of regression analysis with the aid of the two-variable linear regression model, a model in which the dependent variable is expressed as a linear function of only a single explanatory variable.

In Chapter 3, we continue to deal with the two-variable model and introduce what is known as the *classical linear regression model,* a model that makes several simplifying assumptions. With these assumptions, we introduce the method of *ordinary least squares* (OLS) to estimate the parameters of the two-variable regression model. The method of OLS is simple to apply, yet it has some very desirable statistical properties.

In Chapter 4, we introduce the (two-variable) classical *normal* linear regression model, a model that assumes that the random dependent variable follows the normal probability distribution. With this assumption, the OLS estimators obtained in Chapter 3 possess some stronger statistical properties than the nonnormal classical linear regression model—properties that enable us to engage in statistical inference, namely, hypothesis testing.

Chapter 5 is devoted to the topic of hypothesis testing. In this chapter, we try to find out whether the estimated regression coefficients are compatible with the hypothesized values of such coefficients, the hypothesized values being suggested by theory and/or prior empirical work.

Chapter 6 considers some extensions of the two-variable regression model. In particular, it discusses topics such as (1) regression through the origin, (2) scaling and units of measurement, and (3) functional forms of regression models such as double-log, semilog, and reciprocal models.

In Chapter 7, we consider the multiple regression model, a model in which there is more than one explanatory variable, and show how the method of OLS can be extended to estimate the parameters of such models.

In Chapter 8, we extend the concepts introduced in Chapter 5 to the multiple regression model and point out some of the complications arising from the introduction of several explanatory variables.

Chapter 9 on dummy, or qualitative, explanatory variables concludes Part I of the text. This chapter emphasizes that not all explanatory variables need to be quantitative (i.e., ratio scale). Variables, such as gender, race, religion, nationality, and region of residence, cannot be readily quantified, yet they play a valuable role in explaining many an economic phenomenon.

1

THE NATURE OF REGRESSION ANALYSIS

As mentioned in the Introduction, regression is a main tool of econometrics, and in this chapter we consider very briefly the nature of this tool.

1.1 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.[1] 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.[2] 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."

[1]Francis Galton, "Family Likeness in Stature," *Proceedings of Royal Society, London*, vol. 40, 1886, pp. 42–72.

[2]K. Pearson and A. Lee, "On the Laws of Inheritance," *Biometrika*, vol. 2, Nov. 1903, pp. 357–462.

1.2 THE MODERN INTERPRETATION OF REGRESSION

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.

The full import of this view of regression analysis will become clearer as we progress, but a few simple examples will make the basic concept quite clear.

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 **scatter-**

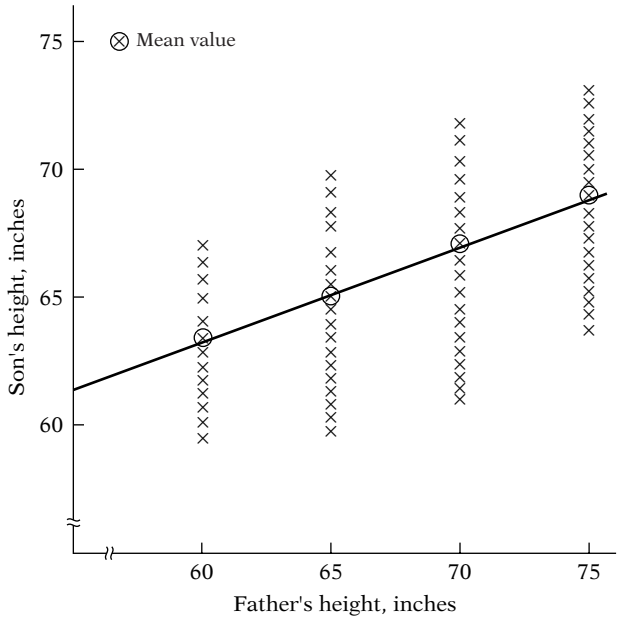


FIGURE 1.1 Hypothetical distribution of sons' heights corresponding to given heights of fathers.

gram. This figure shows the distribution of heights of sons in a hypothetical population corresponding to the given or *fixed* values of the father's height. Notice that corresponding to any given height of a father is a *range* or distribution of the heights of the sons. However, notice that despite the variability of the height of sons for a given value of father's height, the average height of sons generally increases as the height of the father increases. To show this clearly, the circled crosses in the figure indicate the *average* height of sons corresponding to a given height of the father. Connecting these averages, we obtain the line shown in the figure. This line, as we shall see, is known as the **regression line.** It shows how the *average* height of sons increases with the father's height.[3]

2. Consider the scattergram in Figure 1.2, which gives the distribution in a hypothetical population of heights of boys measured at *fixed* ages. Corresponding to any given age, we have a range, or distribution, of heights. Obviously, not all boys of a given age are likely to have identical heights. But height *on the average* increases with age (of course, up to a certain age), which can be seen clearly if we draw a line (the regression line) through the

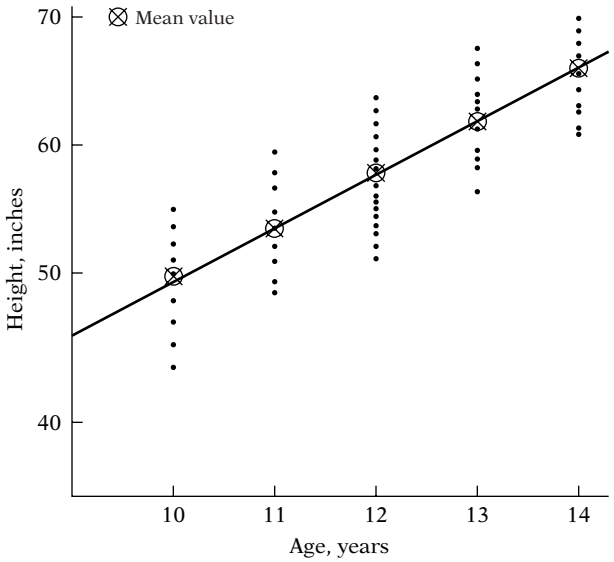


FIGURE 1.2 Hypothetical distribution of heights corresponding to selected ages.

[3]At this stage of the development of the subject matter, we shall call this regression line simply the *line connecting the mean, or average, value of the dependent variable (son's height) corresponding to the given value of the explanatory variable (father's height)*. Note that this line has a positive slope but the slope is less than 1, which is in conformity with Galton's regression to mediocrity. (Why?)

circled points that represent the average height at the given ages. Thus, knowing the age, we may be able to predict from the regression line the average height corresponding to that age.

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), that is, average change in consumption expenditure for, say, a dollar's worth of change in real income (see Figure I.3).

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. Such a scattergram may enable the labor economist to predict the average change in money wages given a certain unemployment rate. Such knowledge may be helpful in stating something about the inflationary process in an economy, for increases in money wages are likely to be reflected in increased prices.

FIGURE 1.3 Hypothetical Phillips curve.

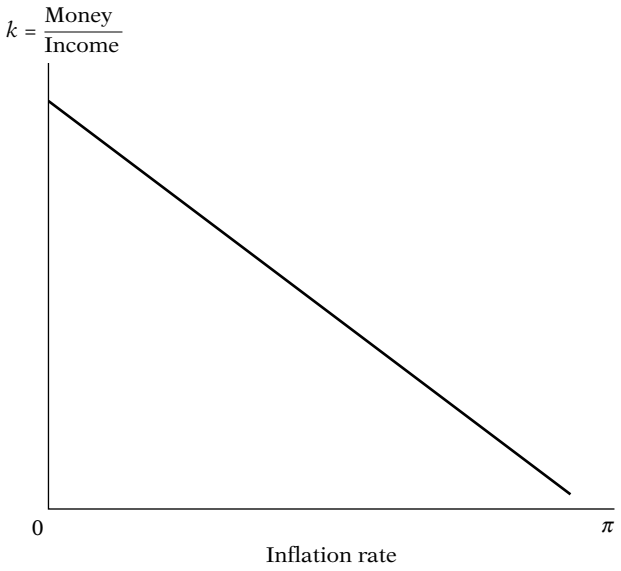


FIGURE 1.4 Money holding in relation to the inflation rate π.

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. A quantitative analysis of this relationship will enable the monetary economist to predict the amount of money, as a proportion of their income, that people would want to hold at various rates of inflation.

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. This knowledge may be helpful in determining the "optimum" advertising budget.

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 reader can supply scores of such examples of the dependence of one variable on one or more other variables. The techniques of regression analysis discussed in this text are specially designed to study such dependence among variables.

1.3 STATISTICAL VERSUS DETERMINISTIC RELATIONSHIPS

From the examples cited in Section 1.2, the reader will notice that in regression analysis we are concerned with what is known as the *statistical*, not *functional* or *deterministic*, dependence among variables, such as those of classical physics. In statistical relationships among variables we essentially deal with **random** or **stochastic**[4] 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 the sense that the explanatory variables, although certainly important, will not enable the agronomist to predict crop yield exactly because of errors involved in measuring these variables as well as a host of other factors (variables) that collectively affect the yield but may be difficult to identify individually. Thus, there is bound to be some "intrinsic" or random variability in the dependent-variable crop yield that cannot be fully explained no matter how many explanatory variables we consider.

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, $F = k(m_1 m_2 / r^2)$, where $F =$ force, m_1 and m_2 are the masses of the two particles, $r =$ distance, and $k =$ constant of proportionality. Another example is Ohm's law, which states: For metallic conductors over a limited range of temperature the current C is proportional to the voltage V; that is, $C = (\frac{1}{k})V$ where $\frac{1}{k}$ is the constant of proportionality. Other examples of such deterministic relationships are Boyle's gas law, Kirchhoff's law of electricity, and Newton's law of motion.

In this text we are not concerned with such deterministic relationships. Of course, if there are errors of measurement, say, in the k of Newton's law of gravity, the otherwise deterministic relationship becomes a statistical relationship. In this situation, force can be predicted only approximately from the given value of k (and m_1, m_2, and r), which contains errors. The variable F in this case becomes a random variable.

1.4 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

[4]The word *stochastic* comes from the Greek word *stokhos* meaning "a bull's eye." The outcome of throwing darts on a dart board is a stochastic process, that is, a process fraught with misses.

suggestive, can never establish causal connection: our ideas of causation must come from outside statistics, ultimately from some theory or other."[5]

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 nonstatistical considerations: Common sense suggests that the relationship cannot be reversed, for we cannot control rainfall by varying crop yield.

In all the examples cited in Section 1.2 the point to note is that **a statistical relationship in itself cannot logically imply causation.** To ascribe causality, one must appeal to a priori or theoretical considerations. Thus, in the third example cited, one can invoke economic theory in saying that consumption expenditure depends on real income.[6]

1.5 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. For example, we may be interested in finding the correlation (coefficient) between smoking and lung cancer, between scores on statistics and mathematics examinations, between high school grades and college grades, and so on. 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. Thus, we may want to know whether we can predict the average score on a statistics examination by knowing a student's score on a mathematics examination.

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),[7] which was made explicit in the definition of regression given in Section 1.2. Thus, in Figure 1.2 we assumed that the variable age was fixed at given levels and height measurements were obtained at these levels. In correlation analysis, on the

[5]M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, Charles Griffin Publishers, New York, 1961, vol. 2, chap. 26, p. 279.

[6]But as we shall see in Chap. 3, classical regression analysis is based on the assumption that the model used in the analysis is the correct model. Therefore, the direction of causality may be implicit in the model postulated.

[7]It is crucial to note that the explanatory variables may be intrinsically stochastic, but for the purpose of regression analysis we assume that their values are fixed in repeated sampling (that is, X assumes the same values in various samples), thus rendering them in effect nonrandom or nonstochastic. But more on this in Chap. 3, Sec. 3.2.

other hand, we treat any (two) variables symmetrically; there is no distinction between the dependent and explanatory variables. After all, 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. As we shall see, most of the correlation theory is based on the assumption of randomness of variables, whereas most of the regression theory to be expounded in this book is conditional upon the assumption that the dependent variable is stochastic but the explanatory variables are fixed or nonstochastic.[8]

1.6 TERMINOLOGY AND NOTATION

Before we proceed to a formal analysis of regression theory, let us dwell briefly on the matter of 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

Although it is a matter of personal taste and tradition, in this text 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 as that of consumption expenditure on real income, 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

[8]In advanced treatment of econometrics, one can relax the assumption that the explanatory variables are nonstochastic (see introduction to Part II).

one explanatory variable, as in the crop-yield, rainfall, temperature, sunshine, and fertilizer examples, it is known as **multiple regression analysis.** In other words, in two-variable regression there is only one explanatory variable, whereas in multiple regression there is more than one explanatory variable.

The term **random** is a synonym for the term **stochastic.** As noted earlier, a random or stochastic variable is a variable that can take on any set of values, positive or negative, with a given probability.[9]

Unless stated otherwise, the letter Y will denote the dependent variable and the X's ($X_1, X_2, \ldots, X_k$) will denote the explanatory variables, X_k being the kth explanatory variable. The subscript i or t will denote the ith or the tth observation or value. X_{ki} (or X_{kt}) will denote the ith (or tth) observation on variable X_k. N (or T) will denote the total number of observations or values in the population, and n (or t) the total number of observations in a sample. As a matter of convention, the observation subscript i will be used for **cross-sectional data** (i.e., data collected at one point in time) and the subscript t will be used for **time series data** (i.e., data collected over a period of time). The nature of cross-sectional and time series data, as well as the important topic of the nature and sources of data for empirical analysis, is discussed in the following section.

1.7 THE NATURE AND SOURCES OF DATA FOR ECONOMIC ANALYSIS[10]

The success of any econometric analysis ultimately depends on the availability of the appropriate data. It is therefore essential that we spend some time discussing the nature, sources, and limitations of the data that one may encounter in empirical analysis.

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) data.

Time Series Data The data shown in Table I.1 of the Introduction are an example of 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.,

[9]See **App. A** for formal definition and further details.

[10]For an informative account, see Michael D. Intriligator, *Econometric Models, Techniques, and Applications,* Prentice Hall, Englewood Cliffs, N.J., 1978, chap. 3.

government budgets), **quinquennially,** that is, every 5 years (e.g., the census of manufactures), or **decennially** (e.g., the census of population). Sometime data are available both quarterly as well as annually, as in the case of the data on GDP and consumer expenditure. With the advent of high-speed computers, data can now be collected over an extremely short interval of time, such as the data on stock prices, which can be obtained literally continuously (the so-called *real-time quote*).

Although time series data are used heavily in econometric studies, they present special problems for econometricians. As we will show in chapters on **time series econometrics** later on, most empirical work based on time series data assumes that the underlying time series is **stationary.** Although it is too early to introduce the precise technical meaning of stationarity at this juncture, *loosely speaking a time series is stationary if its mean and variance do not vary systematically over time.* To see what this means, consider Figure 1.5, which depicts the behavior of the M1 money supply in the United States from January 1, 1959, to July 31, 1999. (The actual data are given in exercise 1.4.) As you can see from this figure, the M1 money supply shows a steady upward **trend** as well as variability over the years, suggesting that the M1 time series is not stationary.[11] We will explore this topic fully in Chapter 21.

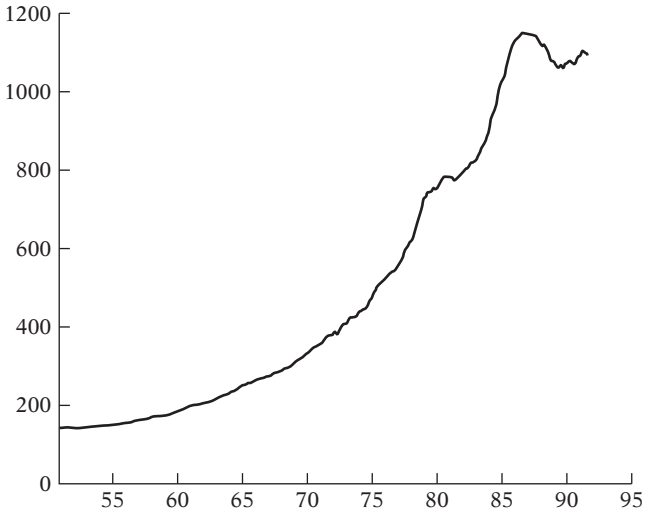


FIGURE 1.5 M1 money supply: United States, 1951:01–1999:09.

[11]To see this more clearly, we divided the data into four time periods: 1951:01 to 1962:12; 1963:01 to 1974:12; 1975:01 to 1986:12, and 1987:01 to 1999:09: For these subperiods the mean values of the money supply (with corresponding standard deviations in parentheses) were, respectively, 165.88 (23.27), 323.20 (72.66), 788.12 (195.43), and 1099 (27.84), all figures in billions of dollars. This is a rough indication of the fact that the money supply over the entire period was not stationary.

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, of course, the opinion polls by Gallup and umpteen other organizations. A concrete example of cross-sectional data is given in Table 1.1 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 1.1 we have two cross-sectional samples.

Just as time series data create their own special problems (because of the stationarity issue), cross-sectional data too have their own problems, specifically the problem of *heterogeneity*. From the data given in 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

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.

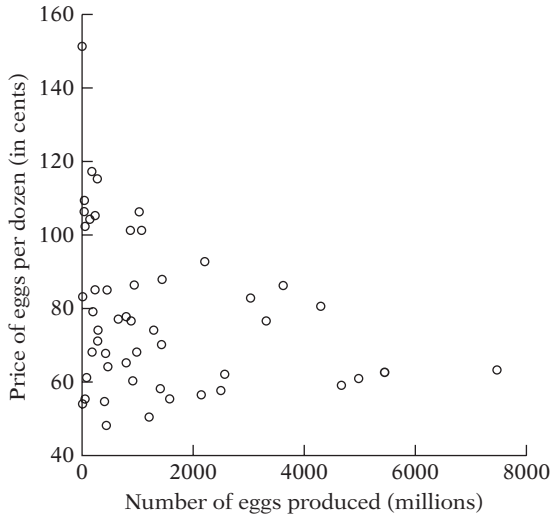


FIGURE 1.6 Relationship between eggs produced and prices, 1990.

include such heterogeneous units in a statistical analysis, the **size** or **scale effect** must be taken into account so as not to mix apples with oranges. 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. In Chapter 11 we will see how the scale effect can be an important factor in assessing relationships among economic variables.

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. Likewise, the data given in exercise 1.1 are pooled data in that the Consumer Price Index (CPI) for each country for 1973–1997 is time series data, whereas the data on the CPI for the seven countries for a single year are cross-sectional data. In the pooled data we have 175 observations—25 annual observations for each of the seven countries.

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. For example, the U.S. Department of Commerce carries out a census of housing at periodic intervals. At each periodic survey the same household (or the people living at the same address) is interviewed to find out if there has been any change in the housing and financial conditions of that household since the last survey. By interviewing the same household periodically, the panel data provides very useful information on the dynamics of household behavior, as we shall see in Chapter 16.

The Sources of Data[12]

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 Internet The Internet has literally revolutionized data gathering. If you just "surf the net" with a keyword (e.g., exchange rates), you will be swamped with all kinds of data sources. In **Appendix E** we provide some of the frequently visited web sites that provide economic and financial data of all sorts. Most of the data can be downloaded without much cost. You may want to bookmark the various web sites that might provide you with useful economic data.

The data collected by various agencies may be **experimental** or **nonexperimental.** In experimental data, often collected in the natural sciences, the investigator may want to collect data while holding certain factors constant in order to assess the impact of some factors on a given phenomenon. For instance, in assessing the impact of obesity on blood pressure, the researcher would want to collect data while holding constant the eating, smoking, and drinking habits of the people in order to minimize the influence of these variables on blood pressure.

In the social sciences, the data that one generally encounters are nonexperimental in nature, that is, not subject to the control of the researcher.[13] For example, the data on GNP, unemployment, stock prices, etc., are not directly under the control of the investigator. As we shall see, this lack of control often creates special problems for the researcher in pinning down the exact cause or causes affecting a particular situation. For example, is it the money supply that determines the (nominal) GDP or is it the other way round?

The Accuracy of Data[14]

Although plenty of data are available for economic research, the quality of the data is often not that good. There are several reasons for that. 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

[12]For an illuminating account, see Albert T. Somers, *The U.S. Economy Demystified: What the Major Economic Statistics Mean and their Significance for Business,* D.C. Heath, Lexington, Mass., 1985.

[13]In the social sciences too sometimes one can have a controlled experiment. An example is given in exercise 1.6.

[14]For a critical review, see O. Morgenstern, *The Accuracy of Economic Observations*, 2d ed., Princeton University Press, Princeton, N.J., 1963.

get a 40 percent response to a questionnaire. Analysis based on such partial response may not truly reflect the behavior of the 60 percent who did not respond, thereby leading to what is known as (sample) **selectivity bias.** Then there is the further problem that those who respond to the questionnaire may not answer all the questions, especially questions of financially sensitive nature, thus leading to additional selectivity bias. 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. Fifth, economic data are generally available at a highly aggregate level. For example, most macrodata (e.g., GNP, employment, inflation, unemployment) are available for the economy as a whole or at the most for some broad geographical regions. 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 IRS, for example, is not allowed by law to disclose data on individual tax returns; it can only release some broad summary data. Therefore, if one wants to find out how much individuals with a certain level of income spent on health care, one cannot do that analysis except at a very highly aggregate level. But such macroanalysis often fails to reveal the dynamics of the behavior of the microunits. Similarly, 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.

Because of all these and many other problems, **the researcher should always keep in mind that the results of research are only as good as the quality of the data.** Therefore, if in given situations researchers find that the results of the research are "unsatisfactory," the cause may be not that they used the wrong model but that the quality of the data was poor. Unfortunately, because of the nonexperimental nature of the data used in most social science studies, researchers very often have no choice but to depend on the available data. But they should always keep in mind that the data used may not be the best and should try not to be too dogmatic about the results obtained from a given study, especially when the quality of the data is suspect.

A Note on the Measurement Scales of Variables[15]

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, X_1 and X_2, the ratio X_1/X_2 and the distance $(X_2 - X_1)$ are meaningful quantities. Also, there is a

[15]The following discussion relies heavily on Aris Spanos, *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data,* Cambridge University Press, New York, 1999, p. 24.

natural ordering (ascending or descending) of the values along the scale. Therefore, comparisons such as $X_2 \leq X_1$ or $X_2 \geq X_1$ are meaningful. Most economic variables belong to this category. Thus, it is meaningful to ask how big is this year's GDP compared with the previous year's GDP.

Interval Scale An interval scale variable satisfies the last two properties of the ratio scale variable but not the first. Thus, the distance between two time periods, say (2000–1995) is meaningful, but not the ratio of two time periods (2000/1995).

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) or income class (upper, middle, lower). For these variables the ordering exists but the distances between the categories cannot be quantified. Students of economics will recall the *indifference curves* between two goods, each higher indifference curve indicating higher level of utility, but one cannot quantify by how much one indifference curve is higher than the others.

Nominal Scale Variables in this category have none of the features of the ratio scale variables. Variables such as gender (male, female) and marital status (married, unmarried, divorced, separated) simply denote categories. *Question:* What is the reason why such variables cannot be expressed on the ratio, interval, or ordinal scales?

As we shall see, 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 of measurement scales discussed above.

1.8 SUMMARY AND CONCLUSIONS

1. The key idea behind regression analysis is the statistical dependence of one variable, the dependent variable, on one or more other variables, the explanatory variables.

2. The objective of such analysis is to estimate and/or predict the mean or average value of the dependent variable on the basis of the known or fixed values of the explanatory variables.

3. In practice the success of regression analysis depends on the availability of the appropriate data. This chapter discussed the nature, sources, and limitations of the data that are generally available for research, especially in the social sciences.

4. In any research, the researcher should clearly state the sources of the data used in the analysis, their definitions, their methods of collection, and any gaps or omissions in the data as well as any revisions in the data. Keep in mind that the macroeconomic data published by the government are often revised.

5. Since the reader may not have the time, energy, or resources to track down the data, the reader has the right to presume that the data used by the researcher are properly gathered and that the computations and analysis are correct.

EXERCISES

1.1. Table 1.2 gives data on the Consumer Price Index (CPI) for seven industrialized countries with 1982–1984 = 100 as the base of the index.
- **a.** From the given data, compute the inflation rate for each country.[16]
- **b.** Plot the inflation rate for each country against time (i.e., use the horizontal axis for time and the vertical axis for the inflation rate.)
- **c.** What broad conclusions can you draw about the inflation experience in the seven countries?
- **d.** Which country's inflation rate seems to be most variable? Can you offer any explanation?

TABLE 1.2 CPI IN SEVEN INDUSTRIAL COUNTRIES, 1973–1997 (1982−1984 = 100)

Year	Canada	France	Germany	Italy	Japan	U.K.	U.S.
1973	40.80000	34.60000	62.80000	20.60000	47.90000	27.90000	44.40000
1974	45.20000	39.30000	67.10000	24.60000	59.00000	32.30000	49.30000
1975	50.10000	43.90000	71.10000	28.80000	65.90000	40.20000	53.80000
1976	53.90000	48.10000	74.20000	33.60000	72.20000	46.80000	56.90000
1977	58.10000	52.70000	76.90000	40.10000	78.10000	54.20000	60.60000
1978	63.30000	57.50000	79.00000	45.10000	81.40000	58.70000	65.20000
1979	69.20000	63.60000	82.20000	52.10000	84.40000	66.60000	72.60000
1980	76.10000	72.30000	86.70000	63.20000	90.90000	78.50000	82.40000
1981	85.60000	81.90000	92.20000	75.40000	95.30000	87.90000	90.90000
1982	94.90000	91.70000	97.10000	87.70000	98.10000	95.40000	96.50000
1983	100.4000	100.4000	100.3000	100.8000	99.80000	99.80000	99.60000
1984	104.7000	108.1000	102.7000	111.5000	102.1000	104.8000	103.9000
1985	109.0000	114.4000	104.8000	121.1000	104.1000	111.1000	107.6000
1986	113.5000	117.3000	104.7000	128.5000	104.8000	114.9000	109.6000
1987	118.4000	121.1000	104.9000	134.4000	104.8000	119.7000	113.6000
1988	123.2000	124.4000	106.3000	141.1000	105.6000	125.6000	118.3000
1989	129.3000	128.7000	109.2000	150.4000	108.1000	135.3000	124.0000
1990	135.5000	133.0000	112.2000	159.6000	111.4000	148.2000	130.7000
1991	143.1000	137.2000	116.3000	169.8000	115.0000	156.9000	136.2000
1992	145.3000	140.5000	122.1000	178.8000	116.9000	162.7000	140.3000
1993	147.9000	143.5000	127.6000	186.4000	118.4000	165.3000	144.5000
1994	148.2000	145.8000	131.1000	193.7000	119.3000	169.4000	148.2000
1995	151.4000	148.4000	133.5000	204.1000	119.1000	175.1000	152.4000
1996	153.8000	151.4000	135.5000	212.0000	119.3000	179.4000	156.9000
1997	156.3000	153.2000	137.8000	215.7000	121.3000	185.0000	160.5000

[16]Subtract from the current year's CPI the CPI from the previous year, divide the difference by the previous year's CPI, and multiply the result by 100. Thus, the inflation rate for Canada for 1974 is $[(45.2 - 40.8)/40.8] \times 100 = 10.78\%$ (approx.).

1.2. **a.** Plot the inflation rate of Canada, France, Germany, Italy, Japan, and the United Kingdom against the United States inflation rate.
b. Comment generally about the behavior of the inflation rate in the six countries vis-à-vis the U.S. inflation rate.
c. If you find that the six countries' inflation rates move in the same direction as the U.S. inflation rate, would that suggest that U.S. inflation "causes" inflation in the other countries? Why or why not?

1.3. Table 1.3 gives the foreign exchange rates for seven industrialized countries for years 1977–1998. Except for the United Kingdom, the exchange rate is defined as the units of foreign currency for one U.S. dollar; for the United Kingdom, it is defined as the number of U.S. dollars for one U.K. pound.
a. Plot these exchange rates against time and comment on the general behavior of the exchange rates over the given time period.
b. The dollar is said to *appreciate* if it can buy more units of a foreign currency. Contrarily, it is said to depreciate if it buys fewer units of a foreign currency. Over the time period 1977–1998, what has been the general behavior of the U.S. dollar? Incidentally, look up any textbook on macroeconomics or international economics to find out what factors determine the appreciation or depreciation of a currency.

1.4. The data behind the M1 money supply in Figure 1.5 are given in Table 1.4. Can you give reasons why the money supply has been increasing over the time period shown in the table?

TABLE 1.3 EXCHANGE RATES FOR SEVEN COUNTRIES: 1977–1998

Year	Canada	France	Germany	Japan	Sweden	Switzerland	U.K.
1977	1.063300	4.916100	2.323600	268.6200	4.480200	2.406500	1.744900
1978	1.140500	4.509100	2.009700	210.3900	4.520700	1.790700	1.918400
1979	1.171300	4.256700	1.834300	219.0200	4.289300	1.664400	2.122400
1980	1.169300	4.225100	1.817500	226.6300	4.231000	1.677200	2.324600
1981	1.199000	5.439700	2.263200	220.6300	5.066000	1.967500	2.024300
1982	1.234400	6.579400	2.428100	249.0600	6.283900	2.032700	1.748000
1983	1.232500	7.620400	2.553900	237.5500	7.671800	2.100700	1.515900
1984	1.295200	8.735600	2.845500	237.4600	8.270800	2.350000	1.336800
1985	1.365900	8.980000	2.942000	238.4700	8.603200	2.455200	1.297400
1986	1.389600	6.925700	2.170500	168.3500	7.127300	1.797900	1.467700
1987	1.325900	6.012200	1.798100	144.6000	6.346900	1.491800	1.639800
1988	1.230600	5.959500	1.757000	128.1700	6.137000	1.464300	1.781300
1989	1.184200	6.380200	1.880800	138.0700	6.455900	1.636900	1.638200
1990	1.166800	5.446700	1.616600	145.0000	5.923100	1.390100	1.784100
1991	1.146000	5.646800	1.661000	134.5900	6.052100	1.435600	1.767400
1992	1.208500	5.293500	1.561800	126.7800	5.825800	1.406400	1.766300
1993	1.290200	5.666900	1.654500	111.0800	7.795600	1.478100	1.501600
1994	1.366400	5.545900	1.621600	102.1800	7.716100	1.366700	1.531900
1995	1.372500	4.986400	1.432100	93.96000	7.140600	1.181200	1.578500
1996	1.363800	5.115800	1.504900	108.7800	6.708200	1.236100	1.560700
1997	1.384900	5.839300	1.734800	121.0600	7.644600	1.451400	1.637600
1998	1.483600	5.899500	1.759700	130.9900	7.952200	1.450600	1.657300

Source: Economic Report of the President, January 2000 and January 2001.

TABLE 1.4 SEASONALLY ADJUSTED M1 SUPPLY: 1959:01–1999:09 (BILLIONS OF DOLLARS)

1959:01	138.8900	139.3900	139.7400	139.6900	140.6800	141.1700
1959:07	141.7000	141.9000	141.0100	140.4700	140.3800	139.9500
1960:01	139.9800	139.8700	139.7500	139.5600	139.6100	139.5800
1960:07	140.1800	141.3100	141.1800	140.9200	140.8600	140.6900
1961:01	141.0600	141.6000	141.8700	142.1300	142.6600	142.8800
1961:07	142.9200	143.4900	143.7800	144.1400	144.7600	145.2000
1962:01	145.2400	145.6600	145.9600	146.4000	146.8400	146.5800
1962:07	146.4600	146.5700	146.3000	146.7100	147.2900	147.8200
1963:01	148.2600	148.9000	149.1700	149.7000	150.3900	150.4300
1963:07	151.3400	151.7800	151.9800	152.5500	153.6500	153.2900
1964:01	153.7400	154.3100	154.4800	154.7700	155.3300	155.6200
1964:07	156.8000	157.8200	158.7500	159.2400	159.9600	160.3000
1965:01	160.7100	160.9400	161.4700	162.0300	161.7000	162.1900
1965:07	163.0500	163.6800	164.8500	165.9700	166.7100	167.8500
1966:01	169.0800	169.6200	170.5100	171.8100	171.3300	171.5700
1966:07	170.3100	170.8100	171.9700	171.1600	171.3800	172.0300
1967:01	171.8600	172.9900	174.8100	174.1700	175.6800	177.0200
1967:07	178.1300	179.7100	180.6800	181.6400	182.3800	183.2600
1968:01	184.3300	184.7100	185.4700	186.6000	187.9900	189.4200
1968:07	190.4900	191.8400	192.7400	194.0200	196.0200	197.4100
1969:01	198.6900	199.3500	200.0200	200.7100	200.8100	201.2700
1969:07	201.6600	201.7300	202.1000	202.9000	203.5700	203.8800
1970:01	206.2200	205.0000	205.7500	206.7200	207.2200	207.5400
1970:07	207.9800	209.9300	211.8000	212.8800	213.6600	214.4100
1971:01	215.5400	217.4200	218.7700	220.0000	222.0200	223.4500
1971:07	224.8500	225.5800	226.4700	227.1600	227.7600	228.3200
1972:01	230.0900	232.3200	234.3000	235.5800	235.8900	236.6200
1972:07	238.7900	240.9300	243.1800	245.0200	246.4100	249.2500
1973:01	251.4700	252.1500	251.6700	252.7400	254.8900	256.6900
1973:07	257.5400	257.7600	257.8600	259.0400	260.9800	262.8800
1974:01	263.7600	265.3100	266.6800	267.2000	267.5600	268.4400
1974:07	269.2700	270.1200	271.0500	272.3500	273.7100	274.2000
1975:01	273.9000	275.0000	276.4200	276.1700	279.2000	282.4300
1975:07	283.6800	284.1500	285.6900	285.3900	286.8300	287.0700
1976:01	288.4200	290.7600	292.7000	294.6600	295.9300	296.1600
1976:07	297.2000	299.0500	299.6700	302.0400	303.5900	306.2500
1977:01	308.2600	311.5400	313.9400	316.0200	317.1900	318.7100
1977:07	320.1900	322.2700	324.4800	326.4000	328.6400	330.8700
1978:01	334.4000	335.3000	336.9600	339.9200	344.8600	346.8000
1978:07	347.6300	349.6600	352.2600	353.3500	355.4100	357.2800
1979:01	358.6000	359.9100	362.4500	368.0500	369.5900	373.3400
1979:07	377.2100	378.8200	379.2800	380.8700	380.8100	381.7700
1980:01	385.8500	389.7000	388.1300	383.4400	384.6000	389.4600
1980:07	394.9100	400.0600	405.3600	409.0600	410.3700	408.0600
1981:01	410.8300	414.3800	418.6900	427.0600	424.4300	425.5000
1981:07	427.9000	427.8500	427.4600	428.4500	430.8800	436.1700
1982:01	442.1300	441.4900	442.3700	446.7800	446.5300	447.8900
1982:07	449.0900	452.4900	457.5000	464.5700	471.1200	474.3000
1983:01	476.6800	483.8500	490.1800	492.7700	499.7800	504.3500
1983:07	508.9600	511.6000	513.4100	517.2100	518.5300	520.7900

(*Continued*)

TABLE 1.4 (*Continued*)

1984:01	524.4000	526.9900	530.7800	534.0300	536.5900	540.5400
1984:07	542.1300	542.3900	543.8600	543.8700	547.3200	551.1900
1985:01	555.6600	562.4800	565.7400	569.5500	575.0700	583.1700
1985:07	590.8200	598.0600	604.4700	607.9100	611.8300	619.3600
1986:01	620.4000	624.1400	632.8100	640.3500	652.0100	661.5200
1986:07	672.2000	680.7700	688.5100	695.2600	705.2400	724.2800
1987:01	729.3400	729.8400	733.0100	743.3900	746.0000	743.7200
1987:07	744.9600	746.9600	748.6600	756.5000	752.8300	749.6800
1988:01	755.5500	757.0700	761.1800	767.5700	771.6800	779.1000
1988:07	783.4000	785.0800	784.8200	783.6300	784.4600	786.2600
1989:01	784.9200	783.4000	782.7400	778.8200	774.7900	774.2200
1989:07	779.7100	781.1400	782.2000	787.0500	787.9500	792.5700
1990:01	794.9300	797.6500	801.2500	806.2400	804.3600	810.3300
1990:07	811.8000	817.8500	821.8300	820.3000	822.0600	824.5600
1991:01	826.7300	832.4000	838.6200	842.7300	848.9600	858.3300
1991:07	862.9500	868.6500	871.5600	878.4000	887.9500	896.7000
1992:01	910.4900	925.1300	936.0000	943.8900	950.7800	954.7100
1992:07	964.6000	975.7100	988.8400	1004.340	1016.040	1024.450
1993:01	1030.900	1033.150	1037.990	1047.470	1066.220	1075.610
1993:07	1085.880	1095.560	1105.430	1113.800	1123.900	1129.310
1994:01	1132.200	1136.130	1139.910	1141.420	1142.850	1145.650
1994:07	1151.490	1151.390	1152.440	1150.410	1150.440	1149.750
1995:01	1150.640	1146.740	1146.520	1149.480	1144.650	1144.240
1995:07	1146.500	1146.100	1142.270	1136.430	1133.550	1126.730
1996:01	1122.580	1117.530	1122.590	1124.520	1116.300	1115.470
1996:07	1112.340	1102.180	1095.610	1082.560	1080.490	1081.340
1997:01	1080.520	1076.200	1072.420	1067.450	1063.370	1065.990
1997:07	1067.570	1072.080	1064.820	1062.060	1067.530	1074.870
1998:01	1073.810	1076.020	1080.650	1082.090	1078.170	1077.780
1998:07	1075.370	1072.210	1074.650	1080.400	1088.960	1093.350
1999:01	1091.000	1092.650	1102.010	1108.400	1104.750	1101.110
1999:07	1099.530	1102.400	1093.460			

Source: Board of Governors, Federal Reserve Bank, USA.

1.5. Suppose you were to develop an economic model of criminal activities, say, the hours spent in criminal activities (e.g., selling illegal drugs). What variables would you consider in developing such a model? See if your model matches the one developed by the Nobel laureate economist Gary Becker.[17]

1.6. *Controlled experiments in economics:* On April 7, 2000, President Clinton signed into law a bill passed by both Houses of the U.S. Congress that lifted earnings limitations on Social Security recipients. Until then, recipients between the ages of 65 and 69 who earned more than $17,000 a year would lose 1 dollar's worth of Social Security benefit for every 3 dollars of income earned in excess of $17,000. How would you devise a study to assess the impact of this change in the law? *Note:* There was no income limitation for recipients over the age of 70 under the old law.

[17]G. S. Becker, "Crime and Punishment: An Economic Approach," *Journal of Political Economy,* vol. 76, 1968, pp. 169–217.

TABLE 1.5 IMPACT OF ADVERTISING EXPENDITURE

Firm	Impressions, millions	Expenditure, millions of 1983 dollars
1. Miller Lite	32.1	50.1
2. Pepsi	99.6	74.1
3. Stroh's	11.7	19.3
4. Fed'l Express	21.9	22.9
5. Burger King	60.8	82.4
6. Coca Cola	78.6	40.1
7. McDonald's	92.4	185.9
8. MCl	50.7	26.9
9. Diet Cola	21.4	20.4
10. Ford	40.1	166.2
11. Levi's	40.8	27.0
12. Bud Lite	10.4	45.6
13. ATT/Bell	88.9	154.9
14. Calvin Klein	12.0	5.0
15. Wendy's	29.2	49.7
16. Polaroid	38.0	26.9
17. Shasta	10.0	5.7
18. Meow Mix	12.3	7.6
19. Oscar Meyer	23.4	9.2
20. Crest	71.1	32.4
21. Kibbles 'N Bits	4.4	6.1

Source: http://lib.stat.cmu.edu/DASL/Datafiles/tvadsdat.html

1.7. The data presented in Table 1.5 was published in the March 1, 1984 issue of the *Wall Street Journal*. It relates to the advertising budget (in millions of dollars) of 21 firms for 1983 and millions of impressions retained per week by the viewers of the products of these firms. The data are based on a survey of 4000 adults in which users of the products were asked to cite a commercial they had seen for the product category in the past week.

a. Plot impressions on the vertical axis and advertising expenditure on the horizontal axis.

b. What can you say about the nature of the relationship between the two variables?

c. Looking at your graph, do you think it pays to advertise? Think about all those commercials shown on Super Bowl Sunday or during the World Series.

Note: We will explore further the data given in Table 1.5 in subsequent chapters.

2

TWO-VARIABLE REGRESSION ANALYSIS: SOME BASIC IDEAS

In Chapter 1 we discussed the concept of regression in broad terms. In this chapter we approach the subject somewhat formally. Specifically, this and the following two chapters introduce the reader to the theory underlying the simplest possible regression analysis, namely, the **bivariate,** or **two-variable,** regression in which the dependent variable (the regressand) is related to a single explanatory variable (the regressor). This case is considered first, not because of its practical adequacy, but because it presents the fundamental ideas of regression analysis as simply as possible and some of these ideas can be illustrated with the aid of two-dimensional graphs. Moreover, as we shall see, the more general **multiple** regression analysis in which the regressand is related to one or more regressors is in many ways a logical extension of the two-variable case.

2.1 A HYPOTHETICAL EXAMPLE[1]

As noted in Section 1.2, regression analysis is largely concerned with estimating and/or predicting the (population) mean value of the dependent variable on the basis of the known or fixed values of the explanatory variable(s).[2] To understand this, consider the data given in Table 2.1. The data

[1]The reader whose statistical knowledge has become somewhat rusty may want to freshen it up by reading the statistical appendix, **App. A,** before reading this chapter.

[2]The *expected value, or expectation, or population mean of a random variable Y* is denoted by the symbol $E(Y)$. On the other hand, the mean value computed from a sample of values from the Y population is denoted as $\bar{Y}$, read as Y bar.

TABLE 2.1 WEEKLY FAMILY INCOME X, $

$Y\downarrow$ / $X\rightarrow$	80	100	120	140	160	180	200	220	240	260
Weekly family consumption expenditure Y, $	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
	–	88	–	113	125	140	–	160	189	185
	–	–	–	115	–	–	–	162	–	191
Total	325	462	445	707	678	750	685	1043	966	1211
Conditional means of Y, $E(Y \mid X)$	65	77	89	101	113	125	137	149	161	173

in the table refer to a total **population** of 60 families in a hypothetical community and their weekly income (X) and weekly consumption expenditure (Y), both in dollars. The 60 families are divided into 10 income groups (from $80 to $260) and the weekly expenditures of each family in the various groups are as shown in the table. Therefore, we have 10 *fixed* values of X and the corresponding Y values against each of the X values; so to speak, there are 10 Y subpopulations.

There is considerable variation in weekly consumption expenditure in each income group, which can be seen clearly from Figure 2.1. But the general picture that one gets is that, despite the variability of weekly consump-

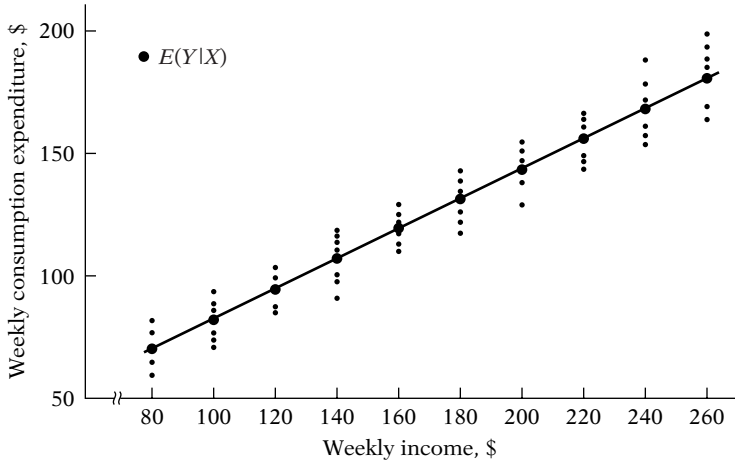


FIGURE 2.1 Conditional distribution of expenditure for various levels of income (data of Table 2.1).

tion expenditure within each income bracket, *on the average,* weekly consumption expenditure increases as income increases. To see this clearly, in Table 2.1 we have given the mean, or average, weekly consumption expenditure corresponding to each of the 10 levels of income. Thus, corresponding to the weekly income level of \$80, the mean consumption expenditure is \$65, while corresponding to the income level of \$200, it is \$137. In all we have 10 mean values for the 10 subpopulations of Y. We call these mean values **conditional expected values,** as they depend on the given values of the (conditioning) variable X. Symbolically, we denote them as $E(Y \mid X)$, which is read as the expected value of Y given the value of X (see also Table 2.2).

It is important to distinguish these conditional expected values from the **unconditional expected value of** weekly consumption expenditure, $E(Y)$. If we add the weekly consumption expenditures for all the 60 families in the *population* and divide this number by 60, we get the number \$121.20 (\$7272/60), which is the unconditional mean, or expected, value of weekly consumption expenditure, $E(Y)$; it is unconditional in the sense that in arriving at this number we have disregarded the income levels of the various families.[3] Obviously, the various conditional expected values of Y given in Table 2.1 are different from the unconditional expected value of Y of \$121.20. When we ask the question, "What is the *expected value* of weekly consumption expenditure of a family," we get the answer \$121.20 (the unconditional mean). But if we ask the question, "What is the *expected value* of weekly consumption expenditure of a family whose monthly income is,

TABLE 2.2 CONDITIONAL PROBABILITIES $p(Y \mid X_i)$ FOR THE DATA OF TABLE 2.1

$p(Y \mid X_i)$ ↓ / X→	80	100	120	140	160	180	200	220	240	260
Conditional probabilities $p(Y \mid X_i)$	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	–	$\frac{1}{6}$	–	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	–	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	–	–	–	$\frac{1}{7}$	–	–	–	$\frac{1}{7}$	–	$\frac{1}{7}$
Conditional means of Y	65	77	89	101	113	125	137	149	161	173

[3]As shown in **App. A**, in general the conditional and unconditional mean values are different.

say, $140," we get the answer $101 (the conditional mean). To put it differently, if we ask the question, "What is the best (mean) prediction of weekly expenditure of families with a weekly income of $140," the answer would be $101. Thus the knowledge of the income level may enable us to better predict the mean value of consumption expenditure than if we do not have that knowledge.[4] This probably is the essence of regression analysis, as we shall discover throughout this text.

The dark circled points in Figure 2.1 show the conditional mean values of Y against the various X values. If we join these conditional mean values, we obtain what is known as the **population regression line (PRL),** or more generally, the **population regression curve.**[5] More simply, it is the **regression of *Y* on *X*.** The adjective "population" comes from the fact that we are dealing in this example with the entire population of 60 families. Of course, in reality a population may have many families.

Geometrically, then, a population regression curve is simply the locus of the conditional means of the dependent variable for the fixed values of the explanatory variable(s). More simply, it is the curve connecting the means of the subpopulations of Y corresponding to the given values of the regressor X. It can be depicted as in Figure 2.2.

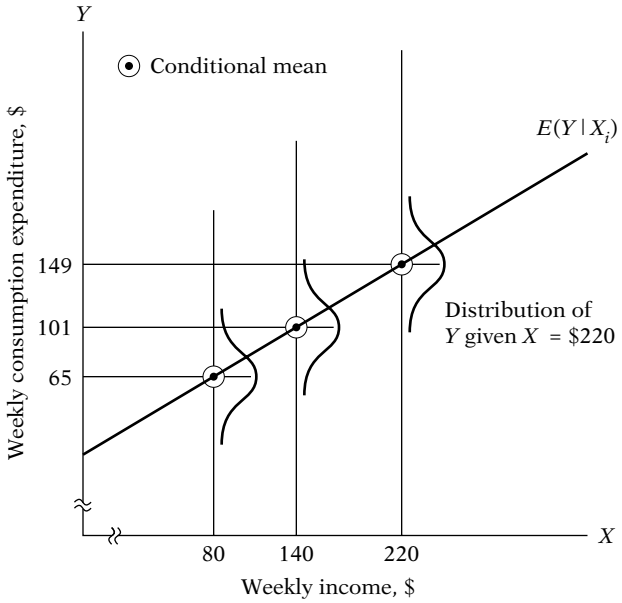


FIGURE 2.2 Population regression line (data of Table 2.1).

[4]I am indebted to James Davidson on this perspective. See James Davidson, *Econometric Theory,* Blackwell Publishers, Oxford, U.K., 2000, p. 11.

[5]In the present example the PRL is a straight line, but it could be a curve (see Figure 2.3).

This figure shows that for each X (i.e., income level) there is a population of Y values (weekly consumption expenditures) that are spread around the (conditional) mean of those Y values. For simplicity, we are assuming that these Y values are distributed symmetrically around their respective (conditional) mean values. And the regression line (or curve) passes through these (conditional) mean values.

With this background, the reader may find it instructive to reread the definition of regression given in Section 1.2.

2.2 THE CONCEPT OF POPULATION REGRESSION FUNCTION (PRF)

From the preceding discussion and Figures. 2.1 and 2.2, it is clear that each conditional mean $E(Y \mid X_i)$ is a function of X_i, where X_i is a given value of X. Symbolically,

$$E(Y \mid X_i) = f(X_i) \tag{2.2.1}$$

where $f(X_i)$ denotes some function of the explanatory variable X. In our example, $E(Y \mid X_i)$ is a linear function of X_i. Equation (2.2.1) is known as the **conditional expectation function (CEF)** or **population regression function (PRF)** or **population regression (PR)** for short. It states merely that the *expected value* of the distribution of Y given X_i is functionally related to X_i. In simple terms, it tells how the mean or average response of Y varies with X.

What form does the function $f(X_i)$ assume? This is an important question because in real situations we do not have the entire population available for examination. The functional form of the PRF is therefore an empirical question, although in specific cases theory may have something to say. For example, an economist might posit that consumption expenditure is linearly related to income. Therefore, as a first approximation or a working hypothesis, we may assume that the PRF $E(Y \mid X_i)$ is a linear function of X_i, say, of the type

$$E(Y \mid X_i) = \beta_1 + \beta_2 X_i \tag{2.2.2}$$

where β_1 and β_2 are unknown but fixed parameters known as the **regression coefficients;** β_1 and β_2 are also known as **intercept** and **slope coefficients,** respectively. Equation (2.2.1) itself is known as the **linear population regression function.** Some alternative expressions used in the literature are *linear population regression model* or simply *linear population regression*. In the sequel, the terms **regression, regression equation,** and **regression model** will be used synonymously.

In regression analysis our interest is in estimating the PRFs like (2.2.2), that is, estimating the values of the unknowns β_1 and β_2 on the basis of observations on Y and X. This topic will be studied in detail in Chapter 3.

2.3 THE MEANING OF THE TERM *LINEAR*

Since this text is concerned primarily with linear models like (2.2.2), it is essential to know what the term *linear* really means, for it can be interpreted in two different ways.

Linearity in the Variables

The first and perhaps more "natural" meaning of linearity is that the conditional expectation of Y is a linear function of X_i, such as, for example, (2.2.2).[6] Geometrically, the regression curve in this case is a straight line. In this interpretation, a regression function such as $E(Y \mid X_i) = \beta_1 + \beta_2 X_i^2$ is not a linear function because the variable X appears with a power or index of 2.

Linearity in the Parameters

The second interpretation of linearity is that the conditional expectation of Y, $E(Y \mid X_i)$, is a linear function of the parameters, the β's; it may or may not be linear in the variable X.[7] In this interpretation $E(Y \mid X_i) = \beta_1 + \beta_2 X_i^2$ is a linear (in the parameter) regression model. To see this, let us suppose X takes the value 3. Therefore, $E(Y \mid X = 3) = \beta_1 + 9\beta_2$, which is obviously linear in β_1 and β_2. All the models shown in Figure 2.3 are thus linear regression models, that is, models linear in the parameters.

Now consider the model $E(Y \mid X_i) = \beta_1 + \beta_2^2 X_i$. Now suppose $X = 3$; then we obtain $E(Y \mid X_i) = \beta_1 + 3\beta_2^2$, which is nonlinear in the parameter β_2. The preceding model is an example of a **nonlinear (in the parameter) regression model.** We will discuss such models in Chapter 14.

Of the two interpretations of linearity, linearity in the parameters is relevant for the development of the regression theory to be presented shortly. Therefore, *from now on the term "linear" regression will always mean a regression that is linear in the parameters; the β's (that is, the parameters are raised to the first power only). It may or may not be linear in the explanatory variables, the X's.* Schematically, we have Table 2.3. Thus, $E(Y \mid X_i) = \beta_1 + \beta_2 X_i$, which is linear both in the parameters and variable, is a LRM, and so is $E(Y \mid X_i) = \beta_1 + \beta_2 X_i^2$, which is linear in the parameters but nonlinear in variable X.

---

[6] A function $Y = f(X)$ is said to be linear in X if X appears with a power or index of 1 only (that is, terms such as X^2, $\sqrt{X}$, and so on, are excluded) and is not multiplied or divided by any other variable (for example, $X \cdot Z$ or X/Z, where Z is another variable). If Y depends on X alone, another way to state that Y is linearly related to X is that the rate of change of Y with respect to X (i.e., the slope, or derivative, of Y with respect to X, dY/dX) is independent of the value of X. Thus, if $Y = 4X$, $dY/dX = 4$, which is independent of the value of X. But if $Y = 4X^2$, $dY/dX = 8X$, which is not independent of the value taken by X. Hence this function is not linear in X.

[7] A function is said to be linear in the parameter, say, β_1, if β_1 appears with a power of 1 only and is not multiplied or divided by any other parameter (for example, $\beta_1\beta_2$, β_2/β_1, and so on).

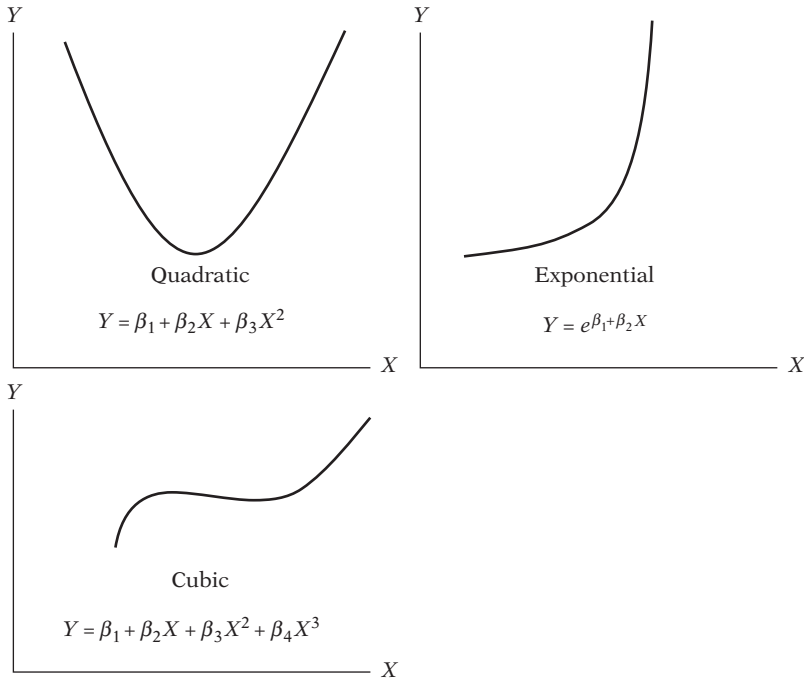


FIGURE 2.3 Linear-in-parameter functions.

TABLE 2.3 LINEAR REGRESSION MODELS

Model linear in parameters?	Model linear in variables?	
	Yes	No
Yes	LRM	LRM
No	NLRM	NLRM

Note: LRM = linear regression model
NLRM = nonlinear regression model

2.4 STOCHASTIC SPECIFICATION OF PRF

It is clear from Figure 2.1 that, as family income increases, family consumption expenditure on the average increases, too. But what about the consumption expenditure of an individual family in relation to its (fixed) level of income? It is obvious from Table 2.1 and Figure 2.1 that an individual family's consumption expenditure does not necessarily increase as the income level increases. For example, from Table 2.1 we observe that corresponding to the income level of $100 there is one family whose consumption expenditure of $65 is less than the consumption expenditures of two families whose weekly income is only $80. But notice that the *average* consumption

expenditure of families with a weekly income of $100 is greater than the average consumption expenditure of families with a weekly income of $80 ($77 versus $65).

What, then, can we say about the relationship between an individual family's consumption expenditure and a given level of income? We see from Figure 2.1 that, given the income level of X_i, an individual family's consumption expenditure is clustered around the average consumption of all families at that X_i, that is, around its conditional expectation. Therefore, we can express the *deviation* of an individual Y_i around its expected value as follows:

$$u_i = Y_i - E(Y \mid X_i)$$

or

$$Y_i = E(Y \mid X_i) + u_i \tag{2.4.1}$$

where the deviation u_i is an unobservable random variable taking positive or negative values. Technically, u_i is known as the **stochastic disturbance** or **stochastic error term.**

How do we interpret (2.4.1)? We can say that the expenditure of an individual family, given its income level, can be expressed as the sum of two components: (1) $E(Y \mid X_i)$, which is simply the mean consumption expenditure of all the families with the same level of income. This component is known as the **systematic,** or **deterministic,** component, and (2) u_i, which is the random, or **nonsystematic,** component. We shall examine shortly the nature of the stochastic disturbance term, but for the moment assume that it is a *surrogate or proxy* for all the omitted or neglected variables that may affect Y but are not (or cannot be) included in the regression model.

If $E(Y \mid X_i)$ is assumed to be linear in X_i, as in (2.2.2), Eq. (2.4.1) may be written as

$$\begin{aligned} Y_i &= E(Y \mid X_i) + u_i \\ &= \beta_1 + \beta_2 X_i + u_i \end{aligned} \tag{2.4.2}$$

Equation (2.4.2) posits that the consumption expenditure of a family is linearly related to its income plus the disturbance term. Thus, the individual consumption expenditures, given $X = \$80$ (see Table 2.1), can be expressed as

$$\begin{aligned} Y_1 &= 55 = \beta_1 + \beta_2(80) + u_1 \\ Y_2 &= 60 = \beta_1 + \beta_2(80) + u_2 \\ Y_3 &= 65 = \beta_1 + \beta_2(80) + u_3 \\ Y_4 &= 70 = \beta_1 + \beta_2(80) + u_4 \\ Y_5 &= 75 = \beta_1 + \beta_2(80) + u_5 \end{aligned} \tag{2.4.3}$$

Now if we take the expected value of (2.4.1) on both sides, we obtain

$$\begin{aligned} E(Y_i \mid X_i) &= E[E(Y \mid X_i)] + E(u_i \mid X_i) \\ &= E(Y \mid X_i) + E(u_i \mid X_i) \end{aligned} \tag{2.4.4}$$

where use is made of the fact that the expected value of a constant is that constant itself.[8] Notice carefully that in (2.4.4) we have taken the conditional expectation, conditional upon the given X's.

Since $E(Y_i \mid X_i)$ is the same thing as $E(Y \mid X_i)$, Eq. (2.4.4) implies that

$$E(u_i \mid X_i) = 0 \tag{2.4.5}$$

Thus, the assumption that the regression line passes through the conditional means of Y (see Figure 2.2) implies that the conditional mean values of u_i (conditional upon the given X's) are zero.

From the previous discussion, it is clear (2.2.2) and (2.4.2) are equivalent forms if $E(u_i \mid X_i) = 0$.[9] But the stochastic specification (2.4.2) has the advantage that it clearly shows that there are other variables besides income that affect consumption expenditure and that an individual family's consumption expenditure cannot be fully explained only by the variable(s) included in the regression model.

2.5 THE SIGNIFICANCE OF THE STOCHASTIC DISTURBANCE TERM

As noted in Section 2.4, the disturbance term u_i is a surrogate for all those variables that are omitted from the model but that collectively affect Y. The obvious question is: Why not introduce these variables into the model explicitly? Stated otherwise, why not develop a multiple regression model with as many variables as possible? The reasons are many.

1. *Vagueness of theory:* The theory, if any, determining the behavior of Y may be, and often is, incomplete. We might know for certain that weekly income X influences weekly consumption expenditure Y, but we might be ignorant or unsure about the other variables affecting Y. Therefore, u_i may be used as a substitute for all the excluded or omitted variables from the model.

2. *Unavailability of data:* Even if we know what some of the excluded variables are and therefore consider a multiple regression rather than a simple regression, we may not have quantitative information about these

[8] See **App. A** for a brief discussion of the properties of the expectation operator E. Note that $E(Y \mid X_i)$, once the value of X_i is fixed, is a constant.

[9] As a matter of fact, in the method of least squares to be developed in Chap. 3, it is assumed explicitly that $E(u_i \mid X_i) = 0$. See Sec. 3.2.

variables. It is a common experience in empirical analysis that the data we would ideally like to have often are not available. For example, in principle we could introduce family wealth as an explanatory variable in addition to the income variable to explain family consumption expenditure. But unfortunately, information on family wealth generally is not available. Therefore, we may be forced to omit the wealth variable from our model despite its great theoretical relevance in explaining consumption expenditure.

3. *Core variables versus peripheral variables:* Assume in our consumption-income example that besides income X_1, the number of children per family X_2, sex X_3, religion X_4, education X_5, and geographical region X_6 also affect consumption expenditure. But it is quite possible that the joint influence of all or some of these variables may be so small and at best nonsystematic or random that as a practical matter and for cost considerations it does not pay to introduce them into the model explicitly. One hopes that their combined effect can be treated as a random variable u_i.[10]

4. *Intrinsic randomness in human behavior:* Even if we succeed in introducing all the relevant variables into the model, there is bound to be some "intrinsic" randomness in individual Y's that cannot be explained no matter how hard we try. The disturbances, the u's, may very well reflect this intrinsic randomness.

5. *Poor proxy variables:* Although the classical regression model (to be developed in Chapter 3) assumes that the variables Y and X are measured accurately, in practice the data may be plagued by errors of measurement. Consider, for example, Milton Friedman's well-known theory of the consumption function.[11] He regards *permanent consumption* (Y^p) as a function of *permanent income* (X^p). But since data on these variables are not directly observable, in practice we use proxy variables, such as current consumption (Y) and current income (X), which can be observable. Since the observed Y and X may not equal Y^p and X^p, there is the problem of errors of measurement. The disturbance term u may in this case then also represent the errors of measurement. As we will see in a later chapter, if there are such errors of measurement, they can have serious implications for estimating the regression coefficients, the β's.

6. *Principle of parsimony:* Following Occam's razor,[12] we would like to keep our regression model as simple as possible. If we can explain the behavior of Y "substantially" with two or three explanatory variables and if

[10]A further difficulty is that variables such as sex, education, and religion are difficult to quantify.

[11]Milton Friedman, *A Theory of the Consumption Function,* Princeton University Press, Princeton, N.J., 1957.

[12]"That descriptions be kept as simple as possible until proved inadequate," *The World of Mathematics,* vol. 2, J. R. Newman (ed.), Simon & Schuster, New York, 1956, p. 1247, or, "Entities should not be multiplied beyond necessity," Donald F. Morrison, *Applied Linear Statistical Methods,* Prentice Hall, Englewood Cliffs, N.J., 1983, p. 58.

our theory is not strong enough to suggest what other variables might be included, why introduce more variables? Let u_i represent all other variables. Of course, we should not exclude relevant and important variables just to keep the regression model simple.

7. *Wrong functional form:* Even if we have theoretically correct variables explaining a phenomenon and even if we can obtain data on these variables, very often we do not know the form of the functional relationship between the regressand and the regressors. Is consumption expenditure a linear (invariable) function of income or a nonlinear (invariable) function? If it is the former, $Y_i = \beta_1 + B_2 X_i + u_i$ is the proper functional relationship between Y and X, but if it is the latter, $Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + u_i$ may be the correct functional form. In two-variable models the functional form of the relationship can often be judged from the scattergram. But in a multiple regression model, it is not easy to determine the appropriate functional form, for graphically we cannot visualize scattergrams in multiple dimensions.

For all these reasons, the stochastic disturbances u_i assume an extremely critical role in regression analysis, which we will see as we progress.

2.6 THE SAMPLE REGRESSION FUNCTION (SRF)

By confining our discussion so far to the population of Y values corresponding to the fixed X's, we have deliberately avoided sampling considerations (note that the data of Table 2.1 represent the population, not a sample). But it is about time to face up to the sampling problems, for in most practical situations what we have is but a sample of Y values corresponding to some fixed X's. Therefore, our task now is to estimate the PRF on the basis of the sample information.

As an illustration, pretend that the population of Table 2.1 was not known to us and the only information we had was a randomly selected sample of Y values for the fixed X's as given in Table 2.4. Unlike Table 2.1, we now have only one Y value corresponding to the given X's; each Y (given X_i) in Table 2.4 is chosen randomly from similar Y's corresponding to the same X_i from the population of Table 2.1.

The question is: From the sample of Table 2.4 can we predict the average weekly consumption expenditure Y in the population as a whole corresponding to the chosen X's? In other words, can we estimate the PRF from the sample data? As the reader surely suspects, we may not be able to estimate the PRF "accurately" because of sampling fluctuations. To see this, suppose we draw another random sample from the population of Table 2.1, as presented in Table 2.5.

Plotting the data of Tables 2.4 and 2.5, we obtain the scattergram given in Figure 2.4. In the scattergram two sample regression lines are drawn so as

TABLE 2.4
A RANDOM SAMPLE FROM THE POPULATION OF TABLE 2.1

Y	X
70	80
65	100
90	120
95	140
110	160
115	180
120	200
140	220
155	240
150	260

TABLE 2.5
ANOTHER RANDOM SAMPLE FROM THE POPULATION OF TABLE 2.1

Y	X
55	80
88	100
90	120
80	140
118	160
120	180
145	200
135	220
145	240
175	260

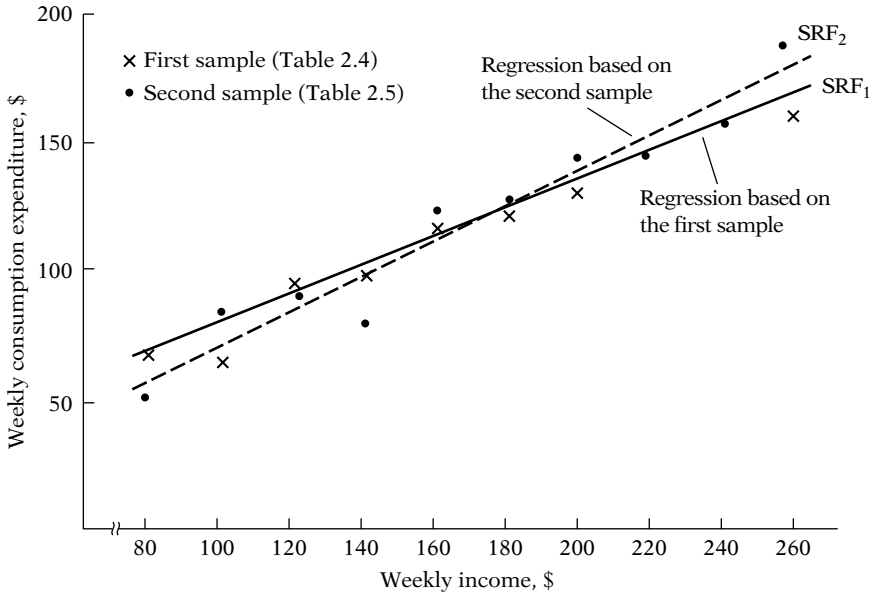


FIGURE 2.4 Regression lines based on two different samples.

to "fit" the scatters reasonably well: SRF_1 is based on the first sample, and SRF_2 is based on the second sample. Which of the two regression lines represents the "true" population regression line? If we avoid the temptation of looking at Figure 2.1, which purportedly represents the PR, there is no way we can be absolutely sure that either of the regression lines shown in Figure 2.4 represents the true population regression line (or curve). The regression lines in Figure 2.4 are known as the **sample regression lines.** Sup-

posedly they represent the population regression line, but because of sampling fluctuations they are at best an approximation of the true PR. In general, we would get N different SRFs for N different samples, and these SRFs are not likely to be the same.

Now, analogously to the PRF that underlies the population regression line, we can develop the concept of the **sample regression function** (SRF) to represent the sample regression line. The sample counterpart of (2.2.2) may be written as

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i \tag{2.6.1}$$

where $\hat{Y}$ is read as "Y-hat" or "Y-cap"
$\hat{Y}_i$ = estimator of $E(Y \mid X_i)$
$\hat{\beta}_1$ = estimator of β_1
$\hat{\beta}_2$ = estimator of β_2

Note that an **estimator,** also known as a (sample) **statistic,** is simply a rule or formula or method that tells how to estimate the population parameter from the information provided by the sample at hand. A particular numerical value obtained by the estimator in an application is known as an **estimate.**[13]

Now just as we expressed the PRF in two equivalent forms, (2.2.2) and (2.4.2), we can express the SRF (2.6.1) in its stochastic form as follows:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i \tag{2.6.2}$$

where, in addition to the symbols already defined, $\hat{u}_i$ denotes the (sample) **residual** term. Conceptually $\hat{u}_i$ is analogous to u_i and can be regarded as an *estimate* of u_i. It is introduced in the SRF for the same reasons as u_i was introduced in the PRF.

To sum up, then, we find our primary objective in regression analysis is to estimate the PRF

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{2.4.2}$$

on the basis of the SRF

$$Y_i = \hat{\beta}_1 + \hat{\beta} x_i = \hat{u}_i \tag{2.6.2}$$

because more often than not our analysis is based upon a single sample from some population. But because of sampling fluctuations our estimate of

[13]As noted in the Introduction, a hat above a variable will signify an estimator of the relevant population value.

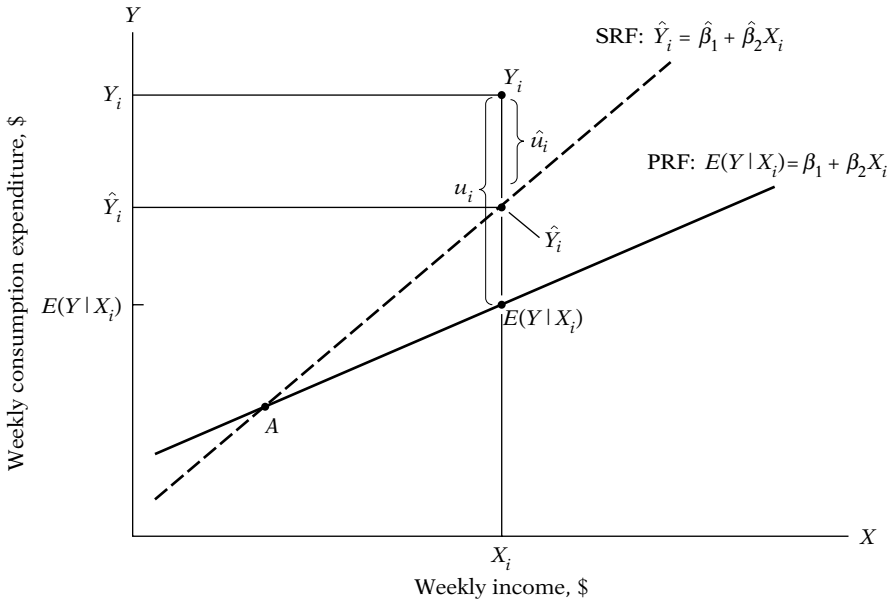


FIGURE 2.5 Sample and population regression lines.

the PRF based on the SRF is at best an approximate one. This approximation is shown diagrammatically in Figure 2.5.

For $X = X_i$, we have one (sample) observation $Y = Y_i$. In terms of the SRF, the observed Y_i can be expressed as

$$Y_i = \hat{Y}_i + \hat{u}_i \tag{2.6.3}$$

and in terms of the PRF, it can be expressed as

$$Y_i = E(Y \mid X_i) + u_i \tag{2.6.4}$$

Now obviously in Figure 2.5 $\hat{Y}_i$ *overestimates* the true $E(Y \mid X_i)$ for the X_i shown therein. By the same token, for any X_i to the left of the point *A*, the SRF will *underestimate* the true PRF. But the reader can readily see that such over- and underestimation is inevitable because of sampling fluctuations.

The critical question now is: Granted that the SRF is but an approximation of the PRF, can we devise a rule or a method that will make this approximation as "close" as possible? In other words, how should the SRF be constructed so that $\hat{\beta}_1$ is as "close" as possible to the true β_1 and $\hat{\beta}_2$ is as "close" as possible to the true β_2 even though we will never know the true β_1 and β_2?

The answer to this question will occupy much of our attention in Chapter 3. We note here that we can develop procedures that tell us how to construct the SRF to mirror the PRF as faithfully as possible. It is fascinating to consider that this can be done even though we never actually determine the PRF itself.

2.7 AN ILLUSTRATIVE EXAMPLE

We conclude this chapter with an example. Table 2.6 gives data on the level of education (measured by the number of years of schooling), the mean hourly wages earned by people at each level of education, and the number of people at the stated level of education. Ernst Berndt originally obtained the data presented in the table, and he derived these data from the current population survey conducted in May 1985.[14] We will explore these data (with additional explanatory variables) in Chapter 3 (Example 3.3, p. 91).

Plotting the (conditional) mean wage against education, we obtain the picture in Figure 2.6. The regression curve in the figure shows how mean wages vary with the level of education; they generally increase with the level of education, a finding one should not find surprising. We will study in a later chapter how variables besides education can also affect the mean wage.

TABLE 2.6
MEAN HOURLY WAGE BY EDUCATION

Years of schooling	Mean wage, $	Number of people
6	4.4567	3
7	5.7700	5
8	5.9787	15
9	7.3317	12
10	7.3182	17
11	6.5844	27
12	7.8182	218
13	7.8351	37
14	11.0223	56
15	10.6738	13
16	10.8361	70
17	13.6150	24
18	13.5310	31
	Total	528

Source: Arthur S. Goldberger, *Introductory Econometrics,* Harvard University Press, Cambridge, Mass., 1998, Table 1.1, p. 5 (adapted).

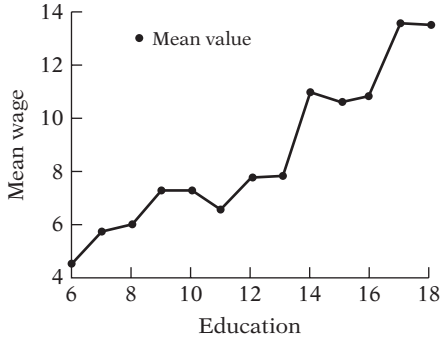


FIGURE 2.6
Relationship between mean wages and education.

[14]Ernst R. Berndt, *The Practice of Econometrics: Classic and Contemporary,* Addison Wesley, Reading, Mass., 1991. Incidentally, this is an excellent book that the reader may want to read to find out how econometricians go about doing research.

2.8 SUMMARY AND CONCLUSIONS

1. The key concept underlying regression analysis is the concept of the **conditional expectation function (CEF), or population regression function (PRF).** Our objective in regression analysis is to find out how the average value of the dependent variable (or regressand) varies with the given value of the explanatory variable (or regressor).

2. This book largely deals with **linear PRFs,** that is, regressions that are linear in the parameters. They may or may not be linear in the regressand or the regressors.

3. For empirical purposes, it is the **stochastic PRF** that matters. The **stochastic disturbance term** u_i plays a critical role in estimating the PRF.

4. The PRF is an idealized concept, since in practice one rarely has access to the entire population of interest. Usually, one has a sample of observations from the population. Therefore, one uses the **stochastic sample regression function (SRF)** to estimate the PRF. How this is actually accomplished is discussed in Chapter 3.

EXERCISES

Questions

2.1. What is the conditional expectation function or the population regression function?

2.2. What is the difference between the population and sample regression functions? Is this a distinction without difference?

2.3. What is the role of the stochastic error term u_i in regression analysis? What is the difference between the stochastic error term and the residual, $\hat{u}_i$?

2.4. Why do we need regression analysis? Why not simply use the mean value of the regressand as its best value?

2.5. What do we mean by a *linear* regression model?

2.6. Determine whether the following models are linear in the parameters, or the variables, or both. Which of these models are linear regression models?

Model	Descriptive title
a. $Y_i = \beta_1 + \beta_2 \left(\frac{1}{X_i}\right) + u_i$	Reciprocal
b. $Y_i = \beta_1 + \beta_2 \ln X_i + u_i$	Semilogarithmic
c. $\ln Y_i = \beta_1 + \beta_2 X_i + u_i$	Inverse semilogarithmic
d. $\ln Y_i = \ln \beta_1 + \beta_2 \ln X_i + u_i$	Logarithmic or double logarithmic
e. $\ln Y_i = \beta_1 - \beta_2 \left(\frac{1}{X_i}\right) + u_i$	Logarithmic reciprocal

Note: ln = natural log (i.e., log to the base e); u_i is the stochastic disturbance term. We will study these models in Chapter 6.

2.7. Are the following models linear regression models? Why or why not?

a. $Y_i = e^{\beta_1 + \beta_2 X_i + u_i}$

b. $Y_i = \dfrac{1}{1 + e^{\beta_1 + \beta_2 X_i + u_i}}$

c. $\ln Y_i = \beta_1 + \beta_2 \left(\frac{1}{X_i}\right) + u_i$

d. $Y_i = \beta_1 + (0.75 - \beta_1)e^{-\beta_2(X_i - 2)} + u_i$

e. $Y_i = \beta_1 + \beta_2^3 X_i + u_i$

2.8. What is meant by an *intrinsically linear* regression model? If β_2 in exercise 2.7**d** were 0.8, would it be a linear or nonlinear regression model?

***2.9.** Consider the following nonstochastic models (i.e., models without the stochastic error term). Are they linear regression models? If not, is it possible, by suitable algebraic manipulations, to convert them into linear models?

a. $Y_i = \dfrac{1}{\beta_1 + \beta_2 X_i}$

b. $Y_i = \dfrac{X_i}{\beta_1 + \beta_2 X_i}$

c. $Y_i = \dfrac{1}{1 + \exp(-\beta_1 - \beta_2 X_i)}$

2.10. You are given the scattergram in Figure 2.7 along with the regression line. What general conclusion do you draw from this diagram? Is the regression line sketched in the diagram a population regression line or the sample regression line?

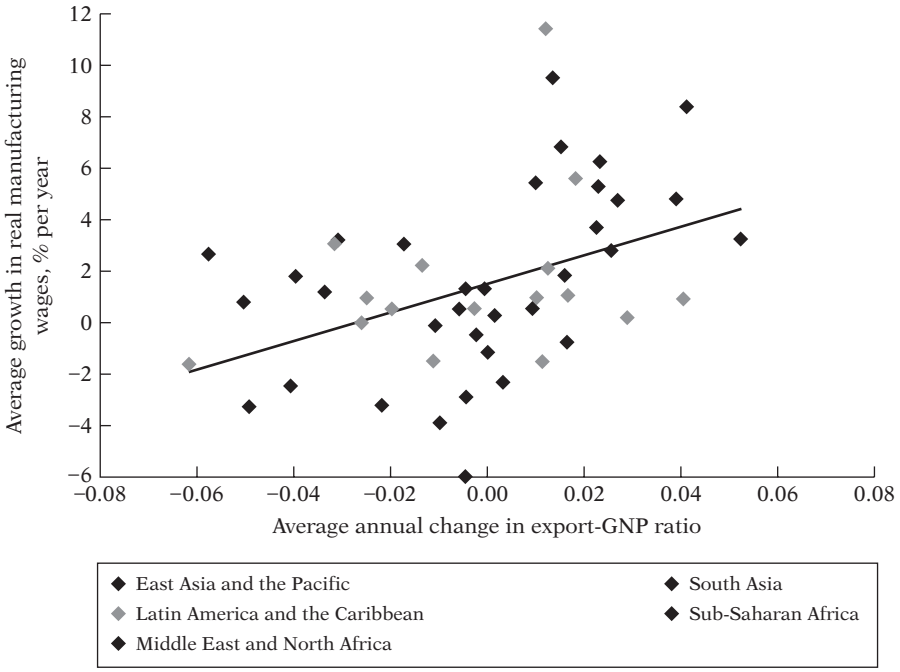


FIGURE 2.7 Growth rates of real manufacturing wages and exports. Data are for 50 developing countries during 1970–90.

Source: The World Bank, *World Development Report 1995,* p. 55. The original source is UNIDO data, World Bank data.

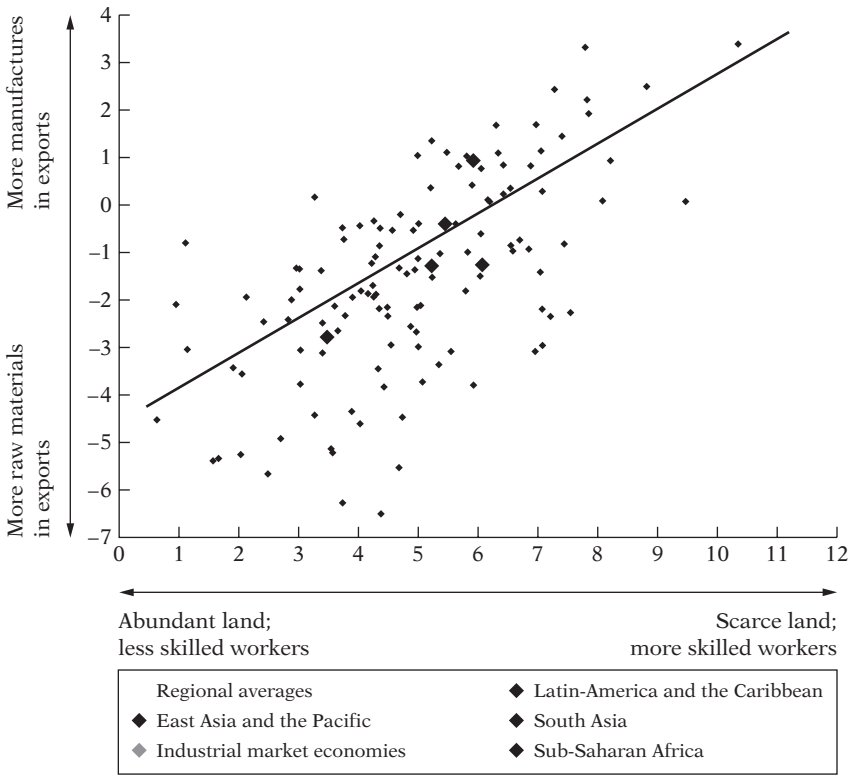


FIGURE 2.8 Skill intensity of exports and human capital endowment. Data are for 126 industrial and developing countries in 1985. Values along the horizontal axis are logarithms of the ratio of the country's average educational attainment to its land area: vertical axis values are logarithms of the ratio of manufactured to primary-products exports.

Source: World Bank, *World Development Report 1995,* p. 59. Original sources: Export data from United Nations Statistical Office COMTRADE data base; education data from UNDP 1990; land data from the World Bank.

2.11. From the scattergram given in Figure 2.8, what general conclusions do you draw? What is the economic theory that underlies this scattergram? (*Hint:* Look up any international economics textbook and read up on the Heckscher–Ohlin model of trade.)

2.12. What does the scattergram in Figure 2.9 reveal? On the basis of this diagram, would you argue that minimum wage laws are good for economic well-being?

2.13. Is the regression line shown in Figure I.3 of the Introduction the PRF or the SRF? Why? How would you interpret the scatterpoints around the regression line? Besides GDP, what other factors, or variables, might determine personal consumption expenditure?

Problems

2.14. You are given the data in Table 2.7 for the United States for years 1980–1996.

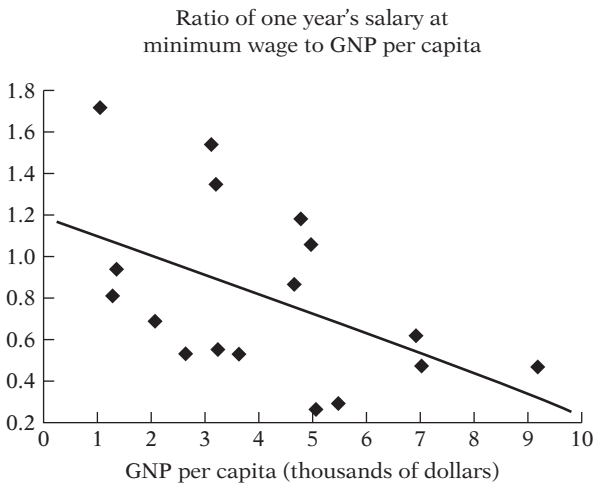


FIGURE 2.9 The minimum wage and GNP per capita. The sample consists of 17 developing countries. Years vary by country from 1988 to 1992. Data are in international prices.

Source: World Bank, *World Development Report 1995,* p. 75.

TABLE 2.7 LABOR FORCE PARTICIPATION DATA

Year	CLFPRM[1]	CLFPRF[2]	UNRM[3]	UNRF[4]	AHE82[5]	AHE[6]
1980	77.4	51.5	6.9	7.4	7.78	6.66
1981	77.0	52.1	7.4	7.9	7.69	7.25
1982	76.6	52.6	9.9	9.4	7.68	7.68
1983	76.4	53.9	9.9	9.2	7.79	8.02
1984	76.4	53.6	7.4	7.6	7.80	8.32
1985	76.3	54.5	7.0	7.4	7.77	8.57
1986	76.3	55.3	6.9	7.1	7.81	8.76
1987	76.2	56.0	6.2	6.2	7.73	8.98
1988	76.2	56.6	5.5	5.6	7.69	9.28
1989	76.4	57.4	5.2	5.4	7.64	9.66
1990	76.4	57.5	5.7	5.5	7.52	10.01
1991	75.8	57.4	7.2	6.4	7.45	10.32
1992	75.8	57.8	7.9	7.0	7.41	10.57
1993	75.4	57.9	7.2	6.6	7.39	10.83
1994	75.1	58.8	6.2	6.0	7.40	11.12
1995	75.0	58.9	5.6	5.6	7.40	11.44
1996	74.9	59.3	5.4	5.4	7.43	11.82

Source: Economic Report of the President, 1997. Table citations below refer to the source document.
[1]*CLFPRM,* Civilian labor force participation rate, male (%), Table B-37, p. 343.
[2]*CLFPRF,* Civilian labor force participation rate, female (%), Table B-37, p. 343.
[3]*UNRM,* Civilian unemployment rate, male (%) Table B-40, p. 346.
[4]*UNRF,* Civilian unemployment rate, female (%) Table B-40, p. 346.
[5]*AHE82,* Average hourly earnings (1982 dollars), Table B-45, p. 352.
[6]*AHE,* Average hourly earnings (current dollars), Table B-45, p. 352.

a. Plot the male civilian labor force participation rate against male civilian unemployment rate. Eyeball a regression line through the scatter points. A priori, what is the expected relationship between the two and what is the underlying economic theory? Does the scattergram support the theory?

b. Repeat part **a** for females.

c. Now plot both the male and female labor participation rates against average hourly earnings (in 1982 dollars). (You may use separate diagrams.) Now what do you find? And how would you rationalize your finding?

d. Can you plot the labor force participation rate against the unemployment rate and the average hourly earnings simultaneously? If not, how would you verbalize the relationship among the three variables?

2.15. Table 2.8 gives data on expenditure on food and total expenditure, measured in rupees, for a sample of 55 rural households from India. (In early 2000, a U.S. dollar was about 40 Indian rupees.)

TABLE 2.8 FOOD AND TOTAL EXPENDITURE (RUPEES)

Observation	Food expenditure	Total expenditure	Observation	Food expenditure	Total expenditure
1	217.0000	382.0000	29	390.0000	655.0000
2	196.0000	388.0000	30	385.0000	662.0000
3	303.0000	391.0000	31	470.0000	663.0000
4	270.0000	415.0000	32	322.0000	677.0000
5	325.0000	456.0000	33	540.0000	680.0000
6	260.0000	460.0000	34	433.0000	690.0000
7	300.0000	472.0000	35	295.0000	695.0000
8	325.0000	478.0000	36	340.0000	695.0000
9	336.0000	494.0000	37	500.0000	695.0000
10	345.0000	516.0000	38	450.0000	720.0000
11	325.0000	525.0000	39	415.0000	721.0000
12	362.0000	554.0000	40	540.0000	730.0000
13	315.0000	575.0000	41	360.0000	731.0000
14	355.0000	579.0000	42	450.0000	733.0000
15	325.0000	585.0000	43	395.0000	745.0000
16	370.0000	586.0000	44	430.0000	751.0000
17	390.0000	590.0000	45	332.0000	752.0000
18	420.0000	608.0000	46	397.0000	752.0000
19	410.0000	610.0000	47	446.0000	769.0000
20	383.0000	616.0000	48	480.0000	773.0000
21	315.0000	618.0000	49	352.0000	773.0000
22	267.0000	623.0000	50	410.0000	775.0000
23	420.0000	627.0000	51	380.0000	785.0000
24	300.0000	630.0000	52	610.0000	788.0000
25	410.0000	635.0000	53	530.0000	790.0000
26	220.0000	640.0000	54	360.0000	795.0000
27	403.0000	648.0000	55	305.0000	801.0000
28	350.0000	650.0000			

Source: Chandan Mukherjee, Howard White, and Marc Wuyts, *Econometrics and Data Analysis for Developing Countries,* Routledge, New York, 1998, p. 457.

a. Plot the data, using the vertical axis for expenditure on food and the horizontal axis for total expenditure, and sketch a regression line through the scatterpoints.
b. What broad conclusions can you draw from this example?
c. A priori, would you expect expenditure on food to increase linearly as total expenditure increases regardless of the level of total expenditure? Why or why not? You can use total expenditure as a proxy for total income.

2.16. Table 2.9 gives data on mean Scholastic Aptitude Test (SAT) scores for college-bound seniors for 1967–1990.
a. Use the horizontal axis for years and the vertical axis for SAT scores to plot the verbal and math scores for males and females separately.
b. What general conclusions can you draw?
c. Knowing the verbal scores of males and females, how would you go about predicting their math scores?
d. Plot the female verbal SAT score against the male verbal SAT score. Sketch a regression line through the scatterpoints. What do you observe?

TABLE 2.9 MEAN SCHOLASTIC APTITUDE TEST SCORES FOR COLLEGE-BOUND SENIORS, 1967–1990*

	Verbal			Math		
Year	Males	Females	Total	Males	Females	Total
1967	463	468	466	514	467	492
1968	464	466	466	512	470	492
1969	459	466	463	513	470	493
1970	459	461	460	509	465	488
1971	454	457	455	507	466	488
1972	454	452	453	505	461	484
1973	446	443	445	502	460	481
1974	447	442	444	501	459	480
1975	437	431	434	495	449	472
1976	433	430	431	497	446	472
1977	431	427	429	497	445	470
1978	433	425	429	494	444	468
1979	431	423	427	493	443	467
1980	428	420	424	491	443	466
1981	430	418	424	492	443	466
1982	431	421	426	493	443	467
1983	430	420	425	493	445	468
1984	433	420	426	495	449	471
1985	437	425	431	499	452	475
1986	437	426	431	501	451	475
1987	435	425	430	500	453	476
1988	435	422	428	498	455	476
1989	434	421	427	500	454	476
1990	429	419	424	499	455	476

*Data for 1967–1971 are estimates.
Source: The College Board. The *New York Times,* Aug. 28, 1990, p. B-5.

3

TWO-VARIABLE REGRESSION MODEL: THE PROBLEM OF ESTIMATION

As noted in Chapter 2, our first task is to estimate the population regression function (PRF) on the basis of the sample regression function (SRF) as accurately as possible. In **Appendix A** we have discussed two generally used methods of estimation: (1) **ordinary least squares (OLS)** and (2) **maximum likelihood (ML).** By and large, it is the method of OLS that is used extensively in regression analysis primarily because it is intuitively appealing and mathematically much simpler than the method of maximum likelihood. Besides, as we will show later, in the linear regression context the two methods generally give similar results.

3.1 THE METHOD OF ORDINARY LEAST SQUARES

The method of ordinary least squares is attributed to Carl Friedrich Gauss, a German mathematician. Under certain assumptions (discussed in Section 3.2), the method of least squares has some very attractive statistical properties that have made it one of the most powerful and popular methods of regression analysis. To understand this method, we first explain the least-squares principle.

Recall the two-variable PRF:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad \textbf{(2.4.2)}$$

However, as we noted in Chapter 2, the PRF is not directly observable. We

estimate it from the SRF:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i \tag{2.6.2}$$

$$= \hat{Y}_i + \hat{u}_i \tag{2.6.3}$$

where $\hat{Y}_i$ is the estimated (conditional mean) value of Y_i.

But how is the SRF itself determined? To see this, let us proceed as follows. First, express (2.6.3) as

$$\begin{aligned} \hat{u}_i &= Y_i - \hat{Y}_i \\ &= Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i \end{aligned} \tag{3.1.1}$$

which shows that the $\hat{u}_i$ (the residuals) are simply the differences between the actual and estimated Y values.

Now given n pairs of observations on Y and X, we would like to determine the SRF in such a manner that it is as close as possible to the actual Y. To this end, we may adopt the following criterion: Choose the SRF in such a way that the sum of the residuals $\sum \hat{u}_i = \sum (Y_i - \hat{Y}_i)$ is as small as possible. Although intuitively appealing, this is not a very good criterion, as can be seen in the hypothetical scattergram shown in Figure 3.1.

If we adopt the criterion of minimizing $\sum \hat{u}_i$, Figure 3.1 shows that the residuals $\hat{u}_2$ and $\hat{u}_3$ as well as the residuals $\hat{u}_1$ and $\hat{u}_4$ receive the same weight in the sum $(\hat{u}_1 + \hat{u}_2 + \hat{u}_3 + \hat{u}_4)$, although the first two residuals are much closer to the SRF than the latter two. In other words, all the residuals receive

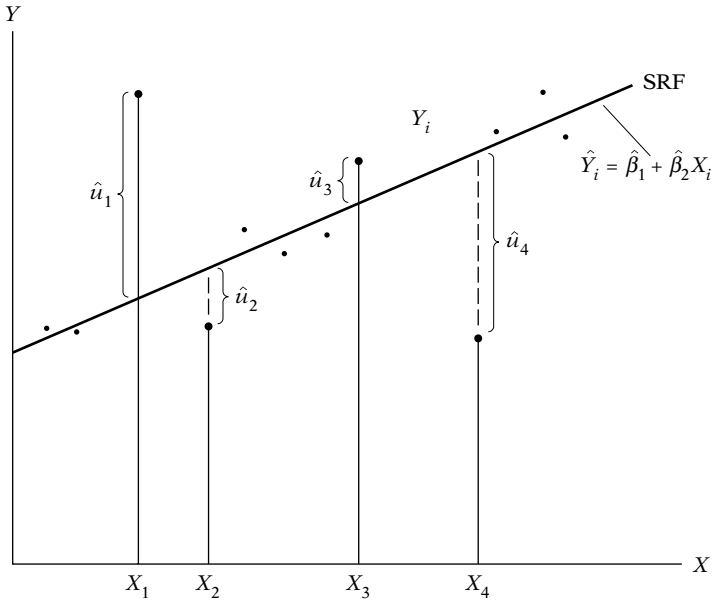


FIGURE 3.1 Least-squares criterion.

equal importance no matter how close or how widely scattered the individual observations are from the SRF. A consequence of this is that it is quite possible that the algebraic sum of the $\hat{u}_i$ is small (even zero) although the $\hat{u}_i$ are widely scattered about the SRF. To see this, let $\hat{u}_1, \hat{u}_2, \hat{u}_3$, and $\hat{u}_4$ in Figure 3.1 assume the values of 10, −2, +2, and −10, respectively. The algebraic sum of these residuals is zero although $\hat{u}_1$ and $\hat{u}_4$ are scattered more widely around the SRF than $\hat{u}_2$ and $\hat{u}_3$. We can avoid this problem if we adopt the *least-squares criterion,* which states that the SRF can be fixed in such a way that

$$\begin{aligned} \sum \hat{u}_i^2 &= \sum (Y_i - \hat{Y}_i)^2 \\ &= \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 \end{aligned} \tag{3.1.2}$$

is as small as possible, where $\hat{u}_i^2$ are the squared residuals. By squaring $\hat{u}_i$, this method gives more weight to residuals such as $\hat{u}_1$ and $\hat{u}_4$ in Figure 3.1 than the residuals $\hat{u}_2$ and $\hat{u}_3$. As noted previously, under the minimum $\sum \hat{u}_i$ criterion, the sum can be small even though the $\hat{u}_i$ are widely spread about the SRF. But this is not possible under the least-squares procedure, for the larger the $\hat{u}_i$ (in absolute value), the larger the $\sum \hat{u}_i^2$. A further justification for the least-squares method lies in the fact that the estimators obtained by it have some very desirable statistical properties, as we shall see shortly.

It is obvious from (3.1.2) that

$$\sum \hat{u}_i^2 = f(\hat{\beta}_1, \hat{\beta}_2) \tag{3.1.3}$$

that is, the sum of the squared residuals is some function of the estimators $\hat{\beta}_1$ and $\hat{\beta}_2$. For any given set of data, choosing different values for $\hat{\beta}_1$ and $\hat{\beta}_2$ will give different $\hat{u}$'s and hence different values of $\sum \hat{u}_i^2$. To see this clearly, consider the hypothetical data on Y and X given in the first two columns of Table 3.1. Let us now conduct two experiments. In experiment 1,

TABLE 3.1 EXPERIMENTAL DETERMINATION OF THE SRF

Y_i (1)	X_t (2)	$\hat{Y}_{1i}$ (3)	$\hat{u}_{1i}$ (4)	$\hat{u}_{1i}^2$ (5)	$\hat{Y}_{2i}$ (6)	$\hat{u}_{2i}$ (7)	$\hat{u}_{2i}^2$ (8)
4	1	2.929	1.071	1.147	4	0	0
5	4	7.000	−2.000	4.000	7	−2	4
7	5	8.357	−1.357	1.841	8	−1	1
12	6	9.714	2.286	5.226	9	3	9
Sum: 28	16		0.0	12.214		0	14

Notes: $\hat{Y}_{1i} = 1.572 + 1.357X_i$ (i.e., $\hat{\beta}_1 = 1.572$ and $\hat{\beta}_2 = 1.357$)
$\hat{Y}_{2i} = 3.0 + 1.0X_i$ (i.e., $\hat{\beta}_1 = 3$ and $\hat{\beta}_2 = 1.0$)
$\hat{u}_{1i} = (Y_i - \hat{Y}_{1i})$
$\hat{u}_{2i} = (Y_i - \hat{Y}_{2i})$

let $\hat{\beta}_1 = 1.572$ and $\hat{\beta}_2 = 1.357$ (let us not worry right now about how we got these values; say, it is just a guess).[1] Using these $\hat{\beta}$ values and the X values given in column (2) of Table 3.1, we can easily compute the estimated Y_i given in column (3) of the table as $\hat{Y}_{1i}$ (the subscript 1 is to denote the first experiment). Now let us conduct another experiment, but this time using the values of $\hat{\beta}_1 = 3$ and $\hat{\beta}_2 = 1$. The estimated values of Y_i from this experiment are given as $\hat{Y}_{2i}$ in column (6) of Table 3.1. Since the $\hat{\beta}$ values in the two experiments are different, we get different values for the estimated residuals, as shown in the table; $\hat{u}_{1i}$ are the residuals from the first experiment and $\hat{u}_{2i}$ from the second experiment. The squares of these residuals are given in columns (5) and (8). Obviously, as expected from (3.1.3), these residual sums of squares are different since they are based on different sets of $\hat{\beta}$ values.

Now which sets of $\hat{\beta}$ values should we choose? Since the $\hat{\beta}$ values of the first experiment give us a lower $\sum \hat{u}_i^2 (= 12.214)$ than that obtained from the $\hat{\beta}$ values of the second experiment $(= 14)$, we might say that the $\hat{\beta}$'s of the first experiment are the "best" values. But how do we know? For, if we had infinite time and infinite patience, we could have conducted many more such experiments, choosing different sets of $\hat{\beta}$'s each time and comparing the resulting $\sum \hat{u}_i^2$ and then choosing that set of $\hat{\beta}$ values that gives us the least possible value of $\sum \hat{u}_i^2$ assuming of course that we have considered all the conceivable values of β_1 and β_2. But since time, and certainly patience, are generally in short supply, we need to consider some shortcuts to this trial-and-error process. Fortunately, the method of least squares provides us such a shortcut. The principle or the method of least squares chooses $\hat{\beta}_1$ and $\hat{\beta}_2$ in such a manner that, for a given sample or set of data, $\sum \hat{u}_i^2$ is as small as possible. In other words, for a given sample, the method of least squares provides us with unique estimates of β_1 and β_2 that give the smallest possible value of $\sum \hat{u}_i^2$. How is this accomplished? This is a straight-forward exercise in differential calculus. As shown in Appendix 3A, Section 3A.1, the process of differentiation yields the following equations for estimating β_1 and β_2:

$$\sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i \tag{3.1.4}$$

$$\sum Y_i X_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2 \tag{3.1.5}$$

where n is the sample size. These simultaneous equations are known as the **normal equations.**

[1] For the curious, these values are obtained by the method of least squares, discussed shortly. See Eqs. (3.1.6) and (3.1.7).

Solving the normal equations simultaneously, we obtain

$$\begin{aligned}\hat{\beta}_2 &= \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2} \\ &= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \\ &= \frac{\sum x_i y_i}{\sum x_i^2}\end{aligned} \tag{3.1.6}$$

where $\bar{X}$ and $\bar{Y}$ are the sample means of X and Y and where we define $x_i = (X_i - \bar{X})$ and $y_i = (Y_i - \bar{Y})$. *Henceforth we adopt the convention of letting the lowercase letters denote deviations from mean values.*

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2} \\ &= \bar{Y} - \hat{\beta}_2 \bar{X}\end{aligned} \tag{3.1.7}$$

The last step in (3.1.7) can be obtained directly from (3.1.4) by simple algebraic manipulations.

Incidentally, note that, by making use of simple algebraic identities, formula (3.1.6) for estimating β_2 can be alternatively expressed as

$$\begin{aligned}\hat{\beta}_2 &= \frac{\sum x_i y_i}{\sum x_i^2} \\ &= \frac{\sum x_i Y_i}{\sum X_i^2 - n\bar{X}^2} \\ &= \frac{\sum X_i y_i}{\sum X_i^2 - n\bar{X}^2}\end{aligned} \tag{3.1.8}$$[2]

The estimators obtained previously are known as the **least-squares estimators,** for they are derived from the least-squares principle. Note the following **numerical properties** of estimators obtained by the method of OLS: "Numerical properties are those that hold as a consequence of the use

[2] *Note 1:* $\sum x_i^2 = \sum (X_i - \bar{X})^2 = \sum X_i^2 - 2\sum X_i \bar{X} + \sum \bar{X}^2 = \sum X_i^2 - 2\bar{X}\sum X_i + \sum \bar{X}^2$, since $\bar{X}$ is a constant. Further noting that $\sum X_i = n\bar{X}$ and $\sum \bar{X}^2 = n\bar{X}^2$ since $\bar{X}$ is a constant, we finally get $\sum x_i^2 = \sum X_i^2 - n\bar{X}^2$.
Note 2: $\sum x_i y_i = \sum x_i (Y_i - \bar{Y}) = \sum x_i Y_i - \bar{Y}\sum x_i = \sum x_i Y_i - \bar{Y}\sum (X_i - \bar{X}) = \sum x_i Y_i$, since $\bar{Y}$ is a constant and since the sum of deviations of a variable from its mean value [e.g., $\sum (X_i - \bar{X})$] is always zero. Likewise, $\sum y_i = \sum (Y_i - \bar{Y}) = 0$.

of ordinary least squares, regardless of how the data were generated."[3] Shortly, we will also consider the **statistical properties** of OLS estimators, that is, properties "that hold only under certain assumptions about the way the data were generated."[4] (See the classical linear regression model in Section 3.2.)

I. The OLS estimators are expressed solely in terms of the observable (i.e., sample) quantities (i.e., X and Y). Therefore, they can be easily computed.

II. They are **point estimators;** that is, given the sample, each estimator will provide only a single (point) value of the relevant population parameter. (In Chapter 5 we will consider the so-called **interval estimators,** which provide a range of possible values for the unknown population parameters.)

III. Once the OLS estimates are obtained from the sample data, the sample regression line (Figure 3.1) can be easily obtained. The regression line thus obtained has the following properties:

1. It passes through the sample means of Y and X. This fact is obvious from (3.1.7), for the latter can be written as $\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}$, which is shown diagrammatically in Figure 3.2.

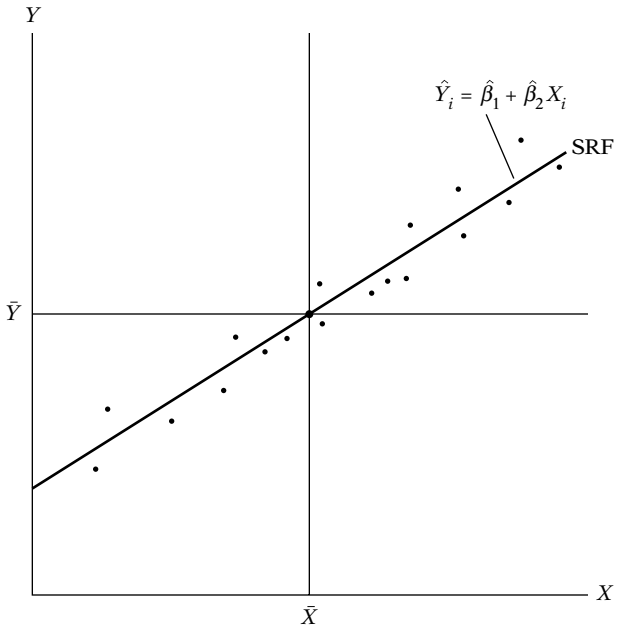


FIGURE 3.2 Diagram showing that the sample regression line passes through the sample mean values of Y and X.

[3]Russell Davidson and James G. MacKinnon, *Estimation and Inference in Econometrics,* Oxford University Press, New York, 1993, p. 3.

[4]*Ibid.*

2. The mean value of the estimated $Y = \hat{Y}_i$ is equal to the mean value of the actual Y for

$$\begin{aligned} \hat{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i \\ &= (\bar{Y} - \hat{\beta}_2 \bar{X}) + \hat{\beta}_2 X_i \\ &= \bar{Y} + \hat{\beta}_2 (X_i - \bar{X}) \end{aligned} \tag{3.1.9}$$

Summing both sides of this last equality over the sample values and dividing through by the sample size n gives

$$\bar{\hat{Y}} = \bar{Y} \tag{3.1.10}$$[5]

where use is made of the fact that $\sum(X_i - \bar{X}) = 0$. (Why?)

3. The mean value of the residuals $\hat{u}_i$ is zero. From Appendix 3A, Section 3A.1, the first equation is

$$-2 \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$$

But since $\hat{u}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$, the preceding equation reduces to $-2 \sum \hat{u}_i = 0$, whence $\bar{\hat{u}} = 0$.[6]

As a result of the preceding property, the sample regression

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i \tag{2.6.2}$$

can be expressed in an alternative form where both Y and X are expressed as deviations from their mean values. To see this, sum (2.6.2) on both sides to give

$$\begin{aligned} \sum Y_i &= n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i + \sum \hat{u}_i \\ &= n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i \qquad \text{since } \sum \hat{u}_i = 0 \end{aligned} \tag{3.1.11}$$

Dividing Eq. (3.1.11) through by n, we obtain

$$\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X} \tag{3.1.12}$$

which is the same as (3.1.7). Subtracting Eq. (3.1.12) from (2.6.2), we obtain

$$Y_i - \bar{Y} = \hat{\beta}_2 (X_i - \bar{X}) + \hat{u}_i$$

[5]Note that this result is true only when the regression model has the intercept term β_1 in it. As **App. 6A**, **Sec. 6A.1** shows, this result need not hold when β_1 is absent from the model.

[6]This result also requires that the intercept term β_1 be present in the model (see **App. 6A**, **Sec. 6A.1**).

or

$$y_i = \hat{\beta}_2 x_i + \hat{u}_i \tag{3.1.13}$$

where y_i and x_i, following our convention, are deviations from their respective (sample) mean values.

Equation (3.1.13) is known as the **deviation form.** Notice that the intercept term $\hat{\beta}_1$ is no longer present in it. But the intercept term can always be estimated by (3.1.7), that is, from the fact that the sample regression line passes through the sample means of Y and X. An advantage of the deviation form is that it often simplifies computing formulas.

In passing, note that in the deviation form, the SRF can be written as

$$\hat{y}_i = \hat{\beta}_2 x_i \tag{3.1.14}$$

whereas in the original units of measurement it was $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$, as shown in (2.6.1).

4. The residuals $\hat{u}_i$ are uncorrelated with the predicted Y_i. This statement can be verified as follows: using the deviation form, we can write

$$\begin{aligned} \sum \hat{y}_i \hat{u}_i &= \hat{\beta}_2 \sum x_i \hat{u}_i \\ &= \hat{\beta}_2 \sum x_i (y_i - \hat{\beta}_2 x_i) \\ &= \hat{\beta}_2 \sum x_i y_i - \hat{\beta}_2^2 \sum x_i^2 \\ &= \hat{\beta}_2^2 \sum x_i^2 - \hat{\beta}_2^2 \sum x_i^2 \\ &= 0 \end{aligned} \tag{3.1.15}$$

where use is made of the fact that $\hat{\beta}_2 = \sum x_i y_i / \sum x_i^2$.

5. The residuals $\hat{u}_i$ are uncorrelated with X_i; that is, $\sum \hat{u}_i X_i = 0$. This fact follows from Eq. (2) in Appendix 3A, Section 3A.1.

3.2 THE CLASSICAL LINEAR REGRESSION MODEL: THE ASSUMPTIONS UNDERLYING THE METHOD OF LEAST SQUARES

If our objective is to estimate β_1 and β_2 only, the method of OLS discussed in the preceding section will suffice. But recall from Chapter 2 that in regression analysis our objective is not only to obtain $\hat{\beta}_1$ and $\hat{\beta}_2$ but also to draw inferences about the true β_1 and β_2. For example, we would like to know how close $\hat{\beta}_1$ and $\hat{\beta}_2$ are to their counterparts in the population or how close $\hat{Y}_i$ is to the true $E(Y \mid X_i)$. To that end, we must not only specify the functional form of the model, as in (2.4.2), but also make certain assumptions about

the manner in which Y_i are generated. To see why this requirement is needed, look at the PRF: $Y_i = \beta_1 + \beta_2 X_i + u_i$. It shows that Y_i depends on both X_i and u_i. Therefore, unless we are specific about how X_i and u_i are created or generated, there is no way we can make any statistical inference about the Y_i and also, as we shall see, about β_1 and β_2. Thus, the assumptions made about the X_i variable(s) and the error term are extremely critical to the valid interpretation of the regression estimates.

The Gaussian, standard, or classical linear regression model (CLRM), which is the cornerstone of most econometric theory, makes 10 assumptions.[7] We first discuss these assumptions in the context of the two-variable regression model; and in Chapter 7 we extend them to multiple regression models, that is, models in which there is more than one regressor.

Assumption 1: Linear regression model. The regression model is **linear in the parameters,** as shown in (2.4.2)

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad \textbf{(2.4.2)}$$

We already discussed model (2.4.2) in Chapter 2. Since linear-in-parameter regression models are the starting point of the CLRM, we will maintain this assumption throughout this book. Keep in mind that the regressand Y and the regressor X themselves may be nonlinear, as discussed in Chapter 2.[8]

Assumption 2: *X* values are fixed in repeated sampling. Values taken by the regressor X are considered fixed in repeated samples. More technically, X is assumed to be *nonstochastic.*

This assumption is implicit in our discussion of the PRF in Chapter 2. But it is very important to understand the concept of "fixed values in repeated sampling," which can be explained in terms of our example given in Table 2.1. Consider the various Y populations corresponding to the levels of income shown in that table. Keeping the value of income X fixed, say, at level \$80, we draw at random a family and observe its weekly family consumption expenditure Y as, say, \$60. Still keeping X at \$80, we draw at random another family and observe its Y value as \$75. In each of these drawings (i.e., repeated sampling), the value of X is fixed at \$80. We can repeat this process for all the X values shown in Table 2.1. As a matter of fact, the sample data shown in Tables 2.4 and 2.5 were drawn in this fashion.

What all this means is that our regression analysis is **conditional regression analysis,** that is, conditional on the given values of the regressor(s) X.

[7]It is classical in the sense that it was developed first by Gauss in 1821 and since then has served as a norm or a standard against which may be compared the regression models that do not satisfy the Gaussian assumptions.

[8]However, a brief discussion of nonlinear-in-the-parameter regression models is given in Chap. 14.

Assumption 3: Zero mean value of disturbance u_i. Given the value of X, the mean, or expected, value of the random disturbance term u_i is zero. Technically, the conditional mean value of u_i is zero. Symbolically, we have

$$E(u_i \mid X_i) = 0 \qquad \textbf{(3.2.1)}$$

Assumption 3 states that the mean value of u_i, conditional upon the given X_i, is zero. Geometrically, this assumption can be pictured as in Figure 3.3, which shows a few values of the variable X and the Y populations associated with each of them. As shown, each Y population corresponding to a given X is distributed around its mean value (shown by the circled points on the PRF) with some Y values above the mean and some below it. The distances above and below the mean values are nothing but the u_i, and what (3.2.1) requires is that the average or mean value of these deviations corresponding to any given X should be zero.[9]

This assumption should not be difficult to comprehend in view of the discussion in Section 2.4 [see Eq. (2.4.5)]. All that this assumption says is that the factors not explicitly included in the model, and therefore subsumed in u_i, do not systematically affect the mean value of Y; so to speak, the positive u_i

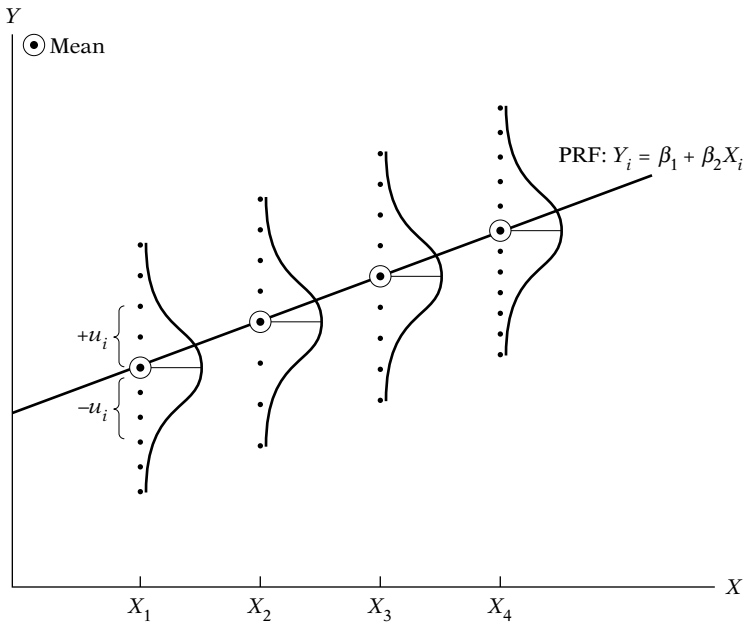


FIGURE 3.3 Conditional distribution of the disturbances u_i.

[9]For illustration, we are assuming merely that the u's are distributed symmetrically as shown in Figure 3.3. But shortly we will assume that the u's are distributed normally.

values cancel out the negative u_i values so that their average or mean effect on Y is zero.[10]

In passing, note that the assumption $E(u_i \mid X_i) = 0$ implies that $E(Y_i \mid X_i) = \beta_i + \beta_2 X_i$. (Why?) Therefore, the two assumptions are equivalent.

Assumption 4: Homoscedasticity or equal variance of u_i. Given the value of X, the variance of u_i is the same for all observations. That is, the conditional variances of u_i are identical. Symbolically, we have

$$\begin{aligned} \textbf{var}\,(u_i \mid X_i) &= E[u_i - E(u_i \mid X_i)]^2 \\ &= E\left(u_i^2 \mid X_i\right) \text{ because of Assumption 3} \\ &= \sigma^2 \end{aligned} \qquad \textbf{(3.2.2)}$$

where **var** stands for variance.

Eq. (3.2.2) states that the variance of u_i for each X_i (i.e., the conditional variance of u_i) is some positive constant number equal to σ^2. Technically, (3.2.2) represents the assumption of **homoscedasticity,** or *equal* (homo) *spread* (*scedasticity*) or *equal variance*. The word comes from the Greek verb *skedanime,* which means to disperse or scatter. Stated differently, (3.2.2) means that the Y populations corresponding to various X values have the same variance. Put simply, the variation around the regression line (which is the line of average relationship between Y and X) is the same across the X values; it neither increases or decreases as X varies. Diagrammatically, the situation is as depicted in Figure 3.4.

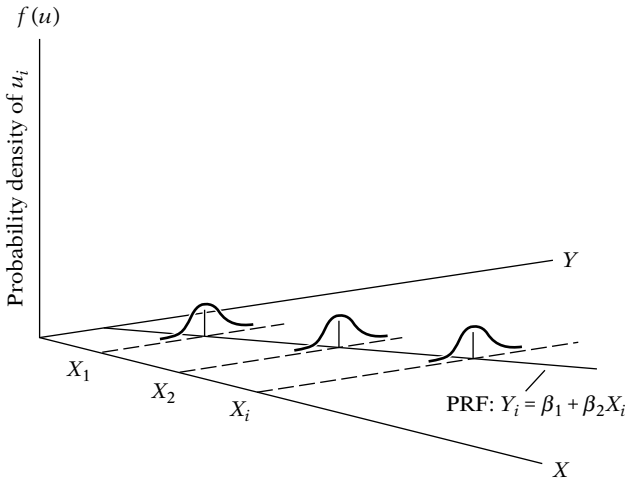


FIGURE 3.4 Homoscedasticity.

[10]For a more technical reason why Assumption 3 is necessary see E. Malinvaud, *Statistical Methods of Econometrics,* Rand McNally, Chicago, 1966, p. 75. See also exercise 3.3.

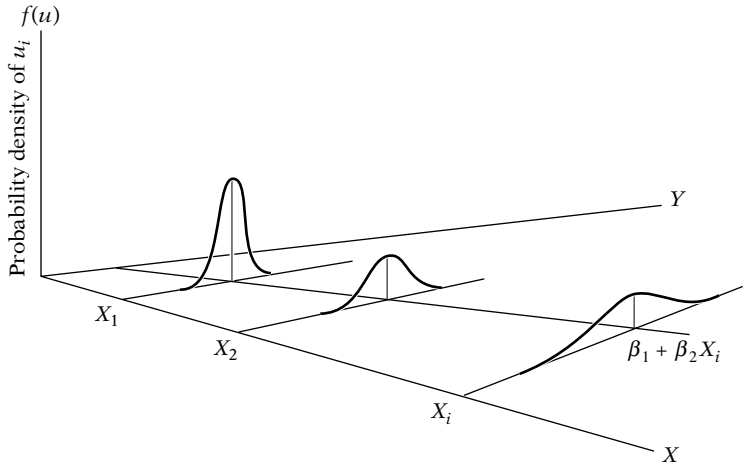


FIGURE 3.5 Heteroscedasticity.

In contrast, consider Figure 3.5, where the conditional variance of the Y population varies with X. This situation is known appropriately as **heteroscedasticity,** or *unequal spread,* or *variance.* Symbolically, in this situation (3.2.2) can be written as

$$\text{var}\,(u_i \mid X_i) = \sigma_i^2 \qquad \textbf{(3.2.3)}$$

Notice the subscript on σ^2 in Eq. (3.2.3), which indicates that the variance of the Y population is no longer constant.

To make the difference between the two situations clear, let Y represent weekly consumption expenditure and X weekly income. Figures 3.4 and 3.5 show that as income increases the average consumption expenditure also increases. But in Figure 3.4 the variance of consumption expenditure remains the same at all levels of income, whereas in Figure 3.5 it increases with increase in income. In other words, richer families on the average consume more than poorer families, but there is also more variability in the consumption expenditure of the former.

To understand the rationale behind this assumption, refer to Figure 3.5. As this figure shows, $\text{var}\,(u \mid X_1) < \text{var}\,(u \mid X_2), \ldots, < \text{var}\,(u \mid X_i)$. Therefore, the likelihood is that the Y observations coming from the population with $X = X_1$ would be closer to the PRF than those coming from populations corresponding to $X = X_2$, $X = X_3$, and so on. In short, not all Y values corresponding to the various X's will be equally reliable, reliability being judged by how closely or distantly the Y values are distributed around their means, that is, the points on the PRF. If this is in fact the case, would we not prefer to sample from those Y populations that are closer to their mean than those that are widely spread? But doing so might restrict the variation we obtain across X values.

By invoking Assumption 4, we are saying that at this stage all Y values corresponding to the various X's are equally important. In Chapter 11 we shall see what happens if this is not the case, that is, where there is heteroscedasticity.

In passing, note that Assumption 4 implies that the conditional variances of Y_i are also homoscedastic. That is,

$$\text{var}\,(Y_i \mid X_i) = \sigma^2 \tag{3.2.4}$$

Of course, the *unconditional variance* of Y is σ_Y^2. Later we will see the importance of distinguishing between conditional and unconditional variances of Y (see Appendix A for details of conditional and unconditional variances).

Assumption 5: No autocorrelation between the disturbances. Given any two X values, X_i and X_j $(i \neq j)$, the correlation between any two u_i and u_j $(i \neq j)$ is zero. Symbolically,

$$\begin{aligned} \textbf{cov}\,(u_i, u_j \mid X_i, X_j) &= E\{[u_i - E(u_i)] \mid X_i\}\{[u_j - E(u_j)] \mid X_j\} \\ &= E(u_i \mid X_i)(u_j \mid X_j) \qquad \text{(why?)} \\ &= 0 \end{aligned} \tag{3.2.5}$$

where i and j are two different observations and where **cov** means **covariance.**

In words, (3.2.5) postulates that the disturbances u_i and u_j are uncorrelated. Technically, this is the assumption of **no serial correlation,** or **no autocorrelation.** This means that, given X_i, the deviations of any two Y values from their mean value do not exhibit patterns such as those shown in Figure 3.6*a* and *b*. In Figure 3.6*a*, we see that the u's are **positively correlated,** a positive u followed by a positive u or a negative u followed by a negative u. In Figure 3.6*b*, the u's are **negatively correlated,** a positive u followed by a negative u and vice versa.

If the disturbances (deviations) follow systematic patterns, such as those shown in Figure 3.6*a* and *b,* there is auto- or serial correlation, and what Assumption 5 requires is that such correlations be absent. Figure 3.6*c* shows that there is no systematic pattern to the u's, thus indicating zero correlation.

The full import of this assumption will be explained thoroughly in Chapter 12. But intuitively one can explain this assumption as follows. Suppose in our PRF $(Y_t = \beta_1 + \beta_2 X_t + u_t)$ that u_t and u_{t-1} are positively correlated. Then Y_t depends not only on X_t but also on u_{t-1} for u_{t-1} to some extent determines u_t. At this stage of the development of the subject matter, by invoking Assumption 5, we are saying that we will consider the systematic effect, if any, of X_t on Y_t and not worry about the other influences that might act on Y as a result of the possible intercorrelations among the u's. But, as noted in Chapter 12, we will see how intercorrelations among the disturbances can be brought into the analysis and with what consequences.

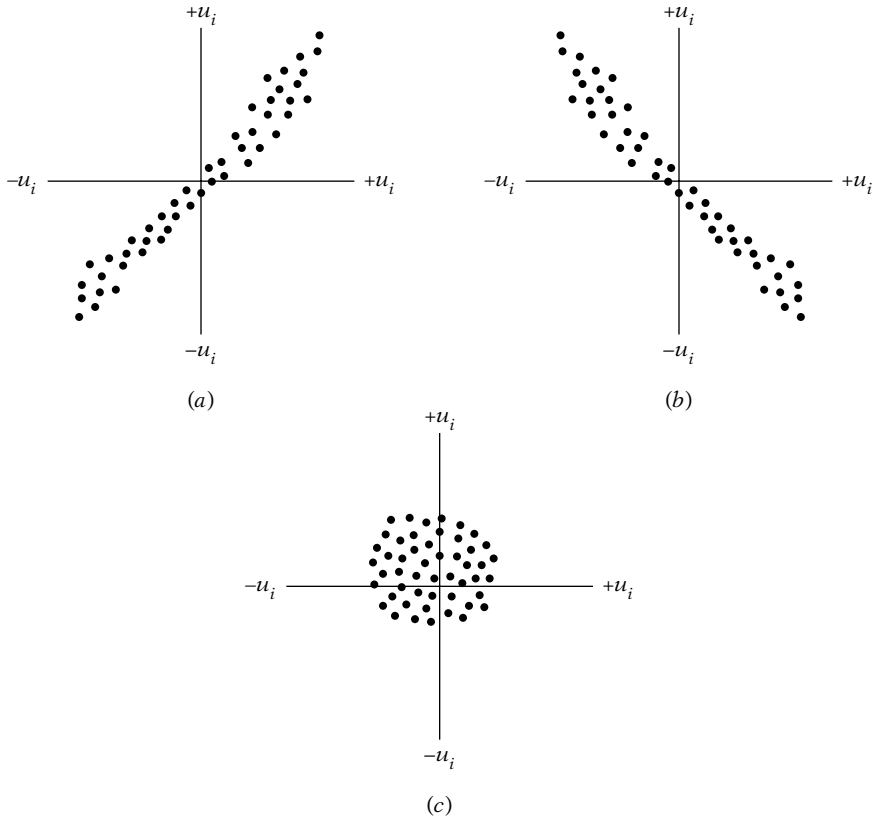


FIGURE 3.6 Patterns of correlation among the disturbances. (*a*) positive serial correlation; (*b*) negative serial correlation; (*c*) zero correlation.

Assumption 6: Zero covariance between u_i and X_i, or $E(u_i X_i) = 0$. Formally,

$$\begin{aligned}
\textbf{cov}\ (u_i, X_i) &= E[u_i - E(u_i)][X_i - E(X_i)] \\
&= E[u_i(X_i - E(X_i))] \qquad \text{since } E(u_i) = 0 \\
&= E(u_i X_i) - E(X_i)E(u_i) \qquad \text{since } E(X_i) \text{ is nonstochastic} \\
&= E(u_i X_i) \qquad \text{since } E(u_i) = 0 \\
&= 0 \qquad \text{by assumption}
\end{aligned} \tag{3.2.6}$$

Assumption 6 states that the disturbance u and explanatory variable X are uncorrelated. The rationale for this assumption is as follows: When we expressed the PRF as in (2.4.2), we assumed that X and u (which may represent the influence of all the omitted variables) have separate (and additive) influence on Y. But if X and u are correlated, it is not possible to assess their individual effects on Y. Thus, if X and u are positively correlated, X increases

when u increases and it decreases when u decreases. Similarly, if X and u are negatively correlated, X increases when u decreases and it decreases when u increases. In either case, it is difficult to isolate the influence of X and u on Y.

Assumption 6 is automatically fulfilled if X variable is nonrandom or nonstochastic and Assumption 3 holds, for in that case, $\text{cov}\,(u_i, X_i) = [X_i - E(X_i)]E[u_i - E(u_i)] = 0$. (Why?) But since we have assumed that our X variable not only is nonstochastic but also assumes fixed values in repeated samples,[11] Assumption 6 is not very critical for us; it is stated here merely to point out that the regression theory presented in the sequel holds true even if the X's are stochastic or random, provided they are independent or at least uncorrelated with the disturbances u_i.[12] (We shall examine the consequences of relaxing Assumption 6 in Part II.)

Assumption 7: The number of observations *n* must be greater than the number of parameters to be estimated. Alternatively, the number of observations *n* must be greater than the number of explanatory variables.

This assumption is not so innocuous as it seems. In the hypothetical example of Table 3.1, imagine that we had only the first pair of observations on Y and X (4 and 1). From this single observation there is no way to estimate the two unknowns, β_1 and β_2. We need at least two pairs of observations to estimate the two unknowns. In a later chapter we will see the critical importance of this assumption.

Assumption 8: Variability in *X* values. The *X* values in a given sample must not all be the same. Technically, var (*X*) must be a finite positive number.[13]

This assumption too is not so innocuous as it looks. Look at Eq. (3.1.6). If all the X values are identical, then $X_i = \bar{X}$ (Why?) and the denominator of that equation will be zero, making it impossible to estimate β_2 and therefore β_1. Intuitively, we readily see why this assumption is important. Looking at

[11] Recall that in obtaining the samples shown in Tables 2.4 and 2.5, we kept the same X values.

[12] As we will discuss in Part II, if the X's are stochastic but distributed independently of u_i, the properties of least estimators discussed shortly continue to hold, but if the stochastic X's are merely uncorrelated with u_i, the properties of OLS estimators hold true only if the sample size is very large. At this stage, however, there is no need to get bogged down with this theoretical point.

[13] The sample variance of X is

$$\text{var}\,(X) = \frac{\sum(X_i - \bar{X})^2}{n-1}$$

where n is sample size.

our family consumption expenditure example in Chapter 2, if there is very little variation in family income, we will not be able to explain much of the variation in the consumption expenditure. The reader should keep in mind that variation in both Y and X is essential to use regression analysis as a research tool. In short, the variables must vary!

Assumption 9: The regression model is correctly specified. Alternatively, there is no **specification bias or error** in the model used in empirical analysis.

As we discussed in the Introduction, the classical econometric methodology assumes implicitly, if not explicitly, that the model used to test an economic theory is "correctly specified." This assumption can be explained informally as follows. An econometric investigation begins with the specification of the econometric model underlying the phenomenon of interest. Some important questions that arise in the specification of the model include the following: (1) What variables should be included in the model? (2) What is the functional form of the model? Is it linear in the parameters, the variables, or both? (3) What are the probabilistic assumptions made about the Y_i, the X_i, and the u_i entering the model?

These are extremely important questions, for, as we will show in Chapter 13, by omitting important variables from the model, or by choosing the wrong functional form, or by making wrong stochastic assumptions about the variables of the model, the validity of interpreting the estimated regression will be highly questionable. To get an intuitive feeling about this, refer to the Phillips curve shown in Figure 1.3. Suppose we choose the following two models to depict the underlying relationship between the rate of change of money wages and the unemployment rate:

$$Y_i = \alpha_1 + \alpha_2 X_i + u_i \tag{3.2.7}$$

$$Y_i = \beta_1 + \beta_2 \left(\frac{1}{X_i}\right) + u_i \tag{3.2.8}$$

where Y_i = the rate of change of money wages, and X_i = the unemployment rate.

The regression model (3.2.7) is linear both in the parameters and the variables, whereas (3.2.8) is linear in the parameters (hence a linear regression model by our definition) but nonlinear in the variable X. Now consider Figure 3.7.

If model (3.2.8) is the "correct" or the "true" model, fitting the model (3.2.7) to the scatterpoints shown in Figure 3.7 will give us wrong predictions: Between points A and B, for any given X_i the model (3.2.7) is going to overestimate the true mean value of Y, whereas to the left of A (or to the right of B) it is going to underestimate (or overestimate, in absolute terms) the true mean value of Y.

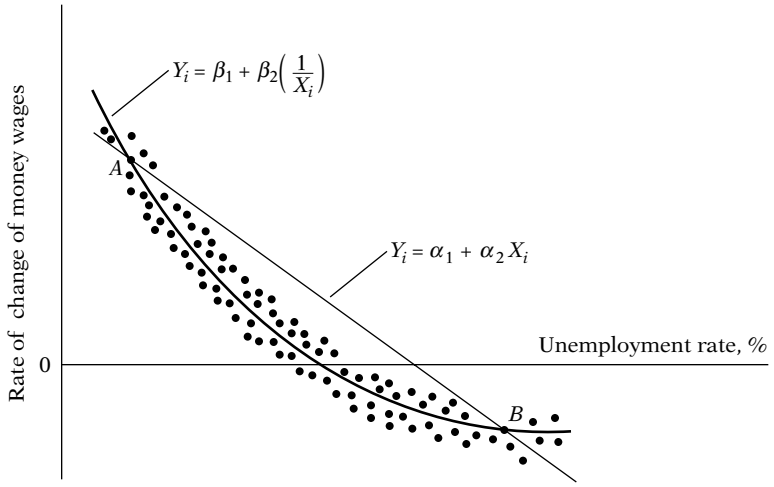


FIGURE 3.7 Linear and nonlinear Phillips curves.

The preceding example is an instance of what is called a **specification bias** or a **specification error;** here the bias consists in choosing the wrong functional form. We will see other types of specification errors in Chapter 13.

Unfortunately, in practice one rarely knows the correct variables to include in the model or the correct functional form of the model or the correct probabilistic assumptions about the variables entering the model for the theory underlying the particular investigation (e.g., the Phillips-type money wage change–unemployment rate tradeoff) may not be strong or robust enough to answer all these questions. Therefore, in practice, the econometrician has to use some judgment in choosing the number of variables entering the model and the functional form of the model and has to make some assumptions about the stochastic nature of the variables included in the model. To some extent, there is some trial and error involved in choosing the "right" model for empirical analysis.[14]

If judgment is required in selecting a model, what is the need for Assumption 9? Without going into details here (see Chapter 13), this assumption is there to remind us that our regression analysis and therefore the results based on that analysis are conditional upon the chosen model and to warn us that we should give very careful thought in formulating econometric

[14]But one should avoid what is known as **"data mining,"** that is, trying every possible model with the hope that at least one will fit the data well. That is why it is essential that there be some economic reasoning underlying the chosen model and that any modifications in the model should have some economic justification. A purely ad hoc model may be difficult to justify on theoretical or a priori grounds. In short, theory should be the basis of estimation. But we will have more to say about data mining in Chap. 13, for there are some who argue that in some situations data mining can serve a useful purpose.

models, especially when there may be several competing theories trying to explain an economic phenomenon, such as the inflation rate, or the demand for money, or the determination of the appropriate or equilibrium value of a stock or a bond. *Thus, econometric model-building, as we shall discover, is more often an art rather than a science.*

Our discussion of the assumptions underlying the classical linear regression model is now completed. It is important to note that all these assumptions pertain to the PRF only and not the SRF. But it is interesting to observe that the method of least squares discussed previously has some properties that are similar to the assumptions we have made about the PRF. For example, the finding that $\sum \hat{u}_i = 0$, and, therefore, $\bar{\hat{u}} = 0$, is akin to the assumption that $E(u_i \mid X_i) = 0$. Likewise, the finding that $\sum \hat{u}_i X_i = 0$ is similar to the assumption that $\text{cov}(u_i, X_i) = 0$. It is comforting to note that the method of least squares thus tries to "duplicate" some of the assumptions we have imposed on the PRF.

Of course, the SRF does not duplicate all the assumptions of the CLRM. As we will show later, although $\text{cov}(u_i, u_j) = 0 \, (i \neq j)$ by assumption, it is *not* true that the *sample* $\text{cov}(\hat{u}_i, \hat{u}_j) = 0 \, (i \neq j)$. As a matter of fact, we will show later that the residuals not only are autocorrelated but also are heteroscedastic (see Chapter 12).

When we go beyond the two-variable model and consider multiple regression models, that is, models containing several regressors, we add the following assumption.

> **Assumption 10: There is no perfect multicollinearity.** That is, there are *no perfect linear relationships among the explanatory variables.*

We will discuss this assumption in Chapter 7, where we discuss multiple regression models.

A Word about These Assumptions

The million-dollar question is: How realistic are all these assumptions? The "reality of assumptions" is an age-old question in the philosophy of science. Some argue that it does not matter whether the assumptions are realistic. What matters are the predictions based on those assumptions. Notable among the "irrelevance-of-assumptions thesis" is Milton Friedman. To him, unreality of assumptions is a positive advantage: "to be important . . . a hypothesis must be descriptively false in its assumptions."[15]

One may not subscribe to this viewpoint fully, but recall that in any scientific study we make certain assumptions because they facilitate the

[15]Milton Friedman, *Essays in Positive Economics,* University of Chicago Press, Chicago, 1953, p. 14.

development of the subject matter in gradual steps, not because they are necessarily realistic in the sense that they replicate reality exactly. As one author notes, ". . . if simplicity is a desirable criterion of good theory, all good theories idealize and oversimplify outrageously."[16]

What we plan to do is first study the properties of the CLRM thoroughly, and then in later chapters examine in depth what happens if one or more of the assumptions of CLRM are not fulfilled. At the end of this chapter, we provide in Table 3.4 a guide to where one can find out what happens to the CLRM if a particular assumption is not satisfied.

As a colleague pointed out to me, when we review research done by others, we need to consider whether the assumptions made by the researcher are appropriate to the data and problem. All too often, published research is based on implicit assumptions about problem and data that are likely not correct and that produce estimates based on these assumptions. Clearly, the knowledgeable reader should, realizing these problems, adopt a skeptical attitude toward the research. The assumptions listed in Table 3.4 therefore provide a checklist for guiding our research and for evaluating the research of others.

With this backdrop, we are now ready to study the CLRM. In particular, we want to find out the **statistical properties** of OLS compared with the purely **numerical properties** discussed earlier. The statistical properties of OLS are based on the assumptions of CLRM already discussed and are enshrined in the famous **Gauss–Markov theorem.** But before we turn to this theorem, which provides the theoretical justification for the popularity of OLS, we first need to consider the **precision** or **standard errors** of the least-squares estimates.

3.3 PRECISION OR STANDARD ERRORS OF LEAST-SQUARES ESTIMATES

From Eqs. (3.1.6) and (3.1.7), it is evident that least-squares estimates are a function of the sample data. But since the data are likely to change from sample to sample, the estimates will change ipso facto. Therefore, what is needed is some measure of "reliability" or **precision** of the estimators $\hat{\beta}_1$ and $\hat{\beta}_2$. In statistics the precision of an estimate is measured by its standard error (se).[17] Given the Gaussian assumptions, it is shown in Appendix 3A, Section 3A.3 that the standard errors of the OLS estimates can be obtained

[16]Mark Blaug, *The Methodology of Economics: Or How Economists Explain,* 2d ed., Cambridge University Press, New York, 1992, p. 92.

[17]The **standard error** is nothing but the standard deviation of the sampling distribution of the estimator, and the sampling distribution of an estimator is simply a probability or frequency distribution of the estimator, that is, a distribution of the set of values of the estimator obtained from all possible samples of the same size from a given population. Sampling distributions are used to draw inferences about the values of the population parameters on the basis of the values of the estimators calculated from one or more samples. (For details, see **App. A**.)

as follows:

$$\text{var}\,(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2} \tag{3.3.1}$$

$$\text{se}\,(\hat{\beta}_2) = \frac{\sigma}{\sqrt{\sum x_i^2}} \tag{3.3.2}$$

$$\text{var}\,(\hat{\beta}_1) = \frac{\sum X_i^2}{n \sum x_i^2}\sigma^2 \tag{3.3.3}$$

$$\text{se}\,(\hat{\beta}_1) = \sqrt{\frac{\sum X_i^2}{n \sum x_i^2}}\sigma \tag{3.3.4}$$

where var = variance and se = standard error and where σ^2 is the constant or homoscedastic variance of u_i of Assumption 4.

All the quantities entering into the preceding equations except σ^2 can be estimated from the data. As shown in Appendix 3A, Section 3A.5, σ^2 itself is estimated by the following formula:

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2} \tag{3.3.5}$$

where $\hat{\sigma}^2$ is the OLS estimator of the true but unknown σ^2 and where the expression $n - 2$ is known as the **number of degrees of freedom (df),** $\sum \hat{u}_i^2$ being the sum of the residuals squared or the **residual sum of squares (RSS).**[18]

Once $\sum \hat{u}_i^2$ is known, $\hat{\sigma}^2$ can be easily computed. $\sum \hat{u}_i^2$ itself can be computed either from (3.1.2) or from the following expression (see Section 3.5 for the proof):

$$\sum \hat{u}_i^2 = \sum y_i^2 - \hat{\beta}_2^2 \sum x_i^2 \tag{3.3.6}$$

Compared with Eq. (3.1.2), Eq. (3.3.6) is easy to use, for it does not require computing $\hat{u}_i$ for each observation although such a computation will be useful in its own right (as we shall see in Chapters 11 and 12).

Since

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2}$$

[18]The term **number of degrees of freedom** means the total number of observations in the sample ($= n$) less the number of independent (linear) constraints or restrictions put on them. In other words, it is the number of independent observations out of a total of n observations. For example, before the RSS (3.1.2) can be computed, $\hat{\beta}_1$ and $\hat{\beta}_2$ must first be obtained. These two estimates therefore put two restrictions on the RSS. Therefore, there are $n - 2$, not n, independent observations to compute the RSS. Following this logic, in the three-variable regression RSS will have $n - 3$ df, and for the k-variable model it will have $n - k$ df. **The general rule is this:** df = (n − number of parameters estimated).

an alternative expression for computing $\sum \hat{u}_i^2$ is

$$\sum \hat{u}_i^2 = \sum y_i^2 - \frac{\left(\sum x_i y_i\right)^2}{\sum x_i^2} \tag{3.3.7}$$

In passing, note that the positive square root of $\hat{\sigma}^2$

$$\hat{\sigma} = \sqrt{\frac{\sum \hat{u}_i^2}{n-2}} \tag{3.3.8}$$

is known as the **standard error of estimate** or **the standard error of the regression (se).** It is simply the standard deviation of the Y values about the estimated regression line and is often used as a summary measure of the "goodness of fit" of the estimated regression line, a topic discussed in Section 3.5.

Earlier we noted that, given X_i, σ^2 represents the (conditional) variance of both u_i and Y_i. Therefore, the standard error of the estimate can also be called the (conditional) standard deviation of u_i and Y_i. Of course, as usual, σ_Y^2 and σ_Y represent, respectively, the unconditional variance and unconditional standard deviation of Y.

Note the following features of the variances (and therefore the standard errors) of $\hat{\beta}_1$ and $\hat{\beta}_2$.

1. The variance of $\hat{\beta}_2$ is directly proportional to σ^2 but inversely proportional to $\sum x_i^2$. That is, given σ^2, the larger the variation in the X values, the smaller the variance of $\hat{\beta}_2$ and hence the greater the precision with which β_2 can be estimated. In short, given σ^2, if there is substantial variation in the X values (recall Assumption 8), β_2 can be measured more accurately than when the X_i do not vary substantially. Also, given $\sum x_i^2$, the larger the variance of σ^2, the larger the variance of β_2. Note that as the sample size n increases, the number of terms in the sum, $\sum x_i^2$, will increase. As n increases, the precision with which β_2 can be estimated also increases. (Why?)

2. The variance of $\hat{\beta}_1$ is directly proportional to σ^2 and $\sum X_i^2$ but inversely proportional to $\sum x_i^2$ and the sample size n.

3. Since $\hat{\beta}_1$ and $\hat{\beta}_2$ are estimators, they will not only vary from sample to sample but in a given sample they are likely to be dependent on each other, this dependence being measured by the covariance between them. It is shown in Appendix 3A, Section 3A.4 that

$$\begin{aligned} \operatorname{cov}(\hat{\beta}_1, \hat{\beta}_2) &= -\bar{X} \operatorname{var}(\hat{\beta}_2) \\ &= -\bar{X}\left(\frac{\sigma^2}{\sum x_i^2}\right) \end{aligned} \tag{3.3.9}$$

Since var $(\hat{\beta}_2)$ is always positive, as is the variance of any variable, the nature of the covariance between $\hat{\beta}_1$ and $\hat{\beta}_2$ depends on the sign of $\bar{X}$. If $\bar{X}$ is positive, then as the formula shows, the covariance will be negative. Thus, if the slope coefficient β_2 is *overestimated* (i.e., the slope is too steep), the intercept coefficient β_1 will be *underestimated* (i.e., the intercept will be too small). Later on (especially in the chapter on multicollinearity, Chapter 10), we will see the utility of studying the covariances between the estimated regression coefficients.

How do the variances and standard errors of the estimated regression coefficients enable one to judge the reliability of these estimates? This is a problem in statistical inference, and it will be pursued in Chapters 4 and 5.

3.4 PROPERTIES OF LEAST-SQUARES ESTIMATORS: THE GAUSS–MARKOV THEOREM[19]

As noted earlier, given the assumptions of the classical linear regression model, the least-squares estimates possess some ideal or optimum properties. These properties are contained in the well-known **Gauss–Markov theorem.** To understand this theorem, we need to consider the **best linear unbiasedness property** of an estimator.[20] As explained in Appendix A, an estimator, say the OLS estimator $\hat{\beta}_2$, is said to be a best linear unbiased estimator (BLUE) of β_2 if the following hold:

1. It is **linear,** that is, a linear function of a random variable, such as the dependent variable Y in the regression model.

2. It is **unbiased,** that is, its average or expected value, $E(\hat{\beta}_2)$, is equal to the true value, β_2.

3. It has minimum variance in the class of all such linear unbiased estimators; an unbiased estimator with the least variance is known as an **efficient estimator.**

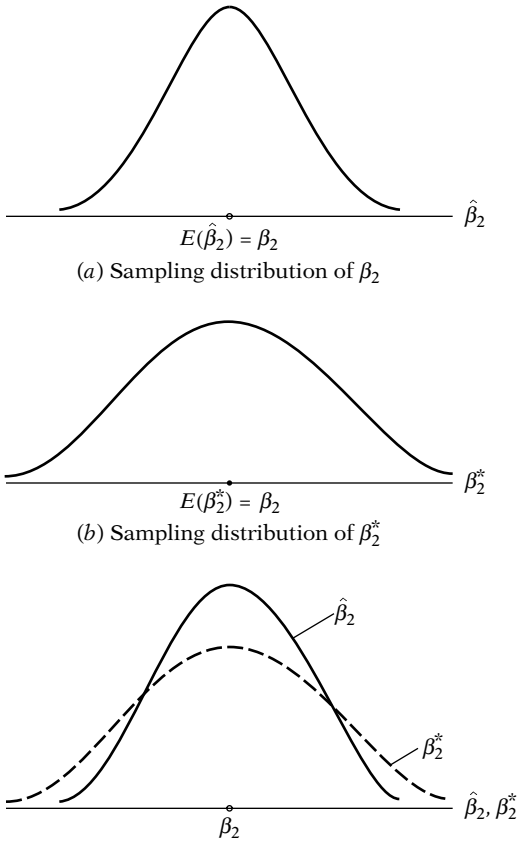
In the regression context it can be proved that the OLS estimators are BLUE. This is the gist of the famous Gauss–Markov theorem, which can be stated as follows:

> **Gauss–Markov Theorem:** Given the assumptions of the classical linear regression model, the least-squares estimators, in the class of unbiased linear estimators, have minimum variance, that is, they are BLUE.

The proof of this theorem is sketched in **Appendix 3A**, **Section 3A.6**. The full import of the Gauss–Markov theorem will become clearer as we move

[19]Although known as the *Gauss–Markov theorem,* the least-squares approach of Gauss antedates (1821) the minimum-variance approach of Markov (1900).

[20]The reader should refer to **App. A** for the importance of linear estimators as well as for a general discussion of the desirable properties of statistical estimators.



(*a*) Sampling distribution of β_2

(*b*) Sampling distribution of β_2^*

(*c*) Sampling distributions of β_2 and β_2^*

FIGURE 3.8 Sampling distribution of OLS estimator $\hat{\beta}_2$ and alternative estimator β_2^*.

along. It is sufficient to note here that the theorem has theoretical as well as practical importance.[21]

What all this means can be explained with the aid of Figure 3.8.

In Figure 3.8(*a*) we have shown the **sampling distribution** of the OLS estimator $\hat{\beta}_2$, that is, the distribution of the values taken by $\hat{\beta}_2$ in repeated sampling experiments (recall Table 3.1). For convenience we have assumed $\hat{\beta}_2$ to be distributed symmetrically (but more on this in Chapter 4). As the figure shows, the mean of the $\hat{\beta}_2$ values, $E(\hat{\beta}_2)$, is equal to the true β_2. In this situation we say that $\hat{\beta}_2$ is an *unbiased estimator* of β_2. In Figure 3.8(*b*) we have shown the sampling distribution of β_2^*, an alternative estimator of β_2

[21]For example, it can be proved that any linear combination of the β's, such as $(\beta_1 - 2\beta_2)$, can be estimated by $(\hat{\beta}_1 - 2\hat{\beta}_2)$, and this estimator is BLUE. For details, see Henri Theil, *Introduction to Econometrics,* Prentice-Hall, Englewood Cliffs, N.J., 1978, pp. 401–402. Note a technical point about the Gauss–Markov theorem: It provides only the sufficient (but not necessary) condition for OLS to be efficient. I am indebted to Michael McAleer of the University of Western Australia for bringing this point to my attention.

obtained by using another (i.e., other than OLS) method. For convenience, assume that β_2^*, like $\hat{\beta}_2$, is unbiased, that is, its average or expected value is equal to β_2. Assume further that both $\hat{\beta}_2$ and β_2^* are linear estimators, that is, they are linear functions of Y. Which estimator, $\hat{\beta}_2$ or β_2^*, would you choose?

To answer this question, superimpose the two figures, as in Figure 3.8(c). It is obvious that although both $\hat{\beta}_2$ and β_2^* are unbiased the distribution of β_2^* is more diffused or widespread around the mean value than the distribution of $\hat{\beta}_2$. In other words, the variance of β_2^* is larger than the variance of $\hat{\beta}_2$. Now given two estimators that are both linear and unbiased, one would choose the estimator with the smaller variance because it is more likely to be close to β_2 than the alternative estimator. In short, one would choose the BLUE estimator.

The Gauss–Markov theorem is remarkable in that it makes no assumptions about the probability distribution of the random variable u_i, and therefore of Y_i (in the next chapter we will take this up). As long as the assumptions of CLRM are satisfied, the theorem holds. As a result, we need not look for another linear unbiased estimator, for we will not find such an estimator whose variance is smaller than the OLS estimator. Of course, if one or more of these assumptions do not hold, the theorem is invalid. For example, if we consider nonlinear-in-the-parameter regression models (which are discussed in Chapter 14), we may be able to obtain estimators that may perform better than the OLS estimators. Also, as we will show in the chapter on heteroscedasticity, if the assumption of homoscedastic variance is not fulfilled, the OLS estimators, although unbiased and consistent, are no longer minimum variance estimators even in the class of linear estimators.

The statistical properties that we have just discussed are known as **finite sample properties:** These properties hold regardless of the sample size on which the estimators are based. Later we will have occasions to consider the **asymptotic properties,** that is, properties that hold only if the sample size is very large (technically, infinite). A general discussion of finite-sample and large-sample properties of estimators is given in **Appendix A.**

3.5 THE COEFFICIENT OF DETERMINATION r^2: A MEASURE OF "GOODNESS OF FIT"

Thus far we were concerned with the problem of estimating regression coefficients, their standard errors, and some of their properties. We now consider the **goodness of fit** of the fitted regression line to a set of data; that is, we shall find out how "well" the sample regression line fits the data. From Figure 3.1 it is clear that if all the observations were to lie on the regression line, we would obtain a "perfect" fit, but this is rarely the case. Generally, there will be some positive $\hat{u}_i$ and some negative $\hat{u}_i$. What we hope for is that these residuals around the regression line are as small as possible. The **coefficient of determination** r^2 (two-variable case) or R^2 (multiple regression) is a summary measure that tells how well the sample regression line fits the data.

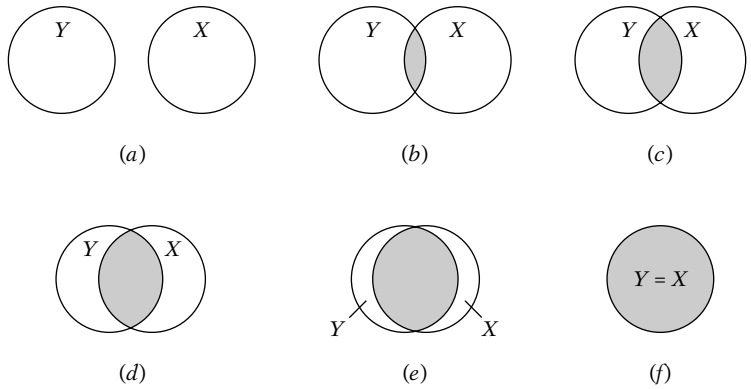


FIGURE 3.9 The Ballentine view of r^2: (*a*) $r^2 = 0$; (*f*) $r^2 = 1$.

Before we show how r^2 is computed, let us consider a heuristic explanation of r^2 in terms of a graphical device, known as the **Venn diagram,** or the **Ballentine,** as shown in Figure 3.9.[22]

In this figure the circle Y represents variation in the dependent variable Y and the circle X represents variation in the explanatory variable X.[23] The overlap of the two circles (the shaded area) indicates the extent to which the variation in Y is explained by the variation in X (say, via an OLS regression). The greater the extent of the overlap, the greater the variation in Y is explained by X. The r^2 is simply a numerical measure of this overlap. In the figure, as we move from left to right, the area of the overlap increases, that is, successively a greater proportion of the variation in Y is explained by X. In short, r^2 increases. When there is no overlap, r^2 is obviously zero, but when the overlap is complete, r^2 is 1, since 100 percent of the variation in Y is explained by X. As we shall show shortly, r^2 lies between 0 and 1.

To compute this r^2, we proceed as follows: Recall that

$$Y_i = \hat{Y}_i + \hat{u}_i \qquad \textbf{(2.6.3)}$$

or in the deviation form

$$y_i = \hat{y}_i + \hat{u}_i \qquad \textbf{(3.5.1)}$$

where use is made of (3.1.13) and (3.1.14). Squaring (3.5.1) on both sides

[22] See Peter Kennedy, "Ballentine: A Graphical Aid for Econometrics," *Australian Economics Papers,* vol. 20, 1981, pp. 414–416. The name Ballentine is derived from the emblem of the well-known Ballantine beer with its circles.

[23] The term *variation* and *variance* are different. Variation means the sum of squares of the deviations of a variable from its mean value. Variance is this sum of squares divided by the appropriate degrees of freedom. In short, variance = variation/df.

and summing over the sample, we obtain

$$\begin{aligned}\sum y_i^2 &= \sum \hat{y}_i^2 + \sum \hat{u}_i^2 + 2\sum \hat{y}_i \hat{u}_i \\ &= \sum \hat{y}_i^2 + \sum \hat{u}_i^2 \\ &= \hat{\beta}_2^2 \sum x_i^2 + \sum \hat{u}_i^2\end{aligned} \tag{3.5.2}$$

since $\sum \hat{y}_i \hat{u}_i = 0$ (why?) and $\hat{y}_i = \hat{\beta}_2 x_i$.

The various sums of squares appearing in (3.5.2) can be described as follows: $\sum y_i^2 = \sum (Y_i - \bar{Y})^2 =$ total variation of the actual Y values about their sample mean, which may be called the **total sum of squares (TSS).** $\sum \hat{y}_i^2 = \sum (\hat{Y}_i - \bar{\hat{Y}})^2 = \sum (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}_2^2 \sum x_i^2 =$ variation of the estimated Y values about their mean ($\bar{\hat{Y}} = \bar{Y}$), which appropriately may be called the sum of squares due to regression [i.e., due to the explanatory variable(s)], or explained by regression, or simply the **explained sum of squares (ESS).** $\sum \hat{u}_i^2 =$ residual or **unexplained** variation of the Y values about the regression line, or simply the **residual sum of squares (RSS).** Thus, (3.5.2) is

$$\text{TSS} = \text{ESS} + \text{RSS} \tag{3.5.3}$$

and shows that the total variation in the observed Y values about their mean value can be partitioned into two parts, one attributable to the regression line and the other to random forces because not all actual Y observations lie on the fitted line. Geometrically, we have Figure 3.10.

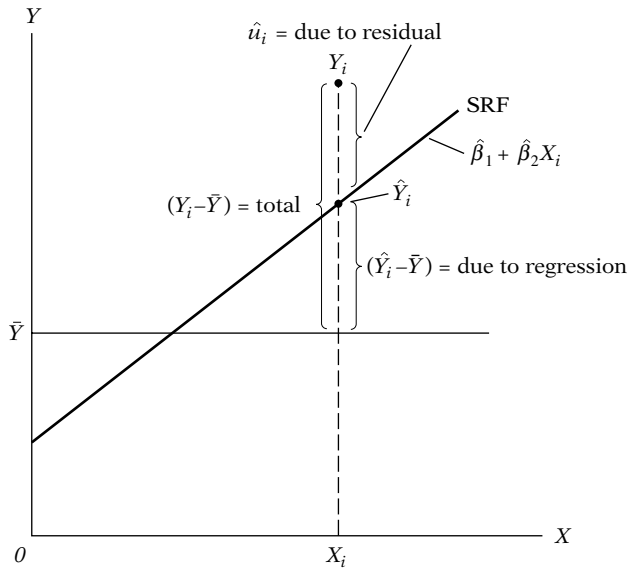


FIGURE 3.10 Breakdown of the variation of Y_i into two components.

Now dividing (3.5.3) by TSS on both sides, we obtain

$$\begin{aligned} 1 &= \frac{\text{ESS}}{\text{TSS}} + \frac{\text{RSS}}{\text{TSS}} \\ &= \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} + \frac{\sum \hat{u}_i^2}{\sum(Y_i - \bar{Y})^2} \end{aligned} \tag{3.5.4}$$

We now define r^2 as

$$r^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{\text{ESS}}{\text{TSS}} \tag{3.5.5}$$

or, alternatively, as

$$\begin{aligned} r^2 &= 1 - \frac{\sum \hat{u}_i^2}{\sum(Y_i - \bar{Y})^2} \\ &= 1 - \frac{\text{RSS}}{\text{TSS}} \end{aligned} \tag{3.5.5a}$$

The quantity r^2 thus defined is known as the (sample) **coefficient of determination** and is the most commonly used measure of the goodness of fit of a regression line. Verbally, r^2 *measures the proportion or percentage of the total variation in Y explained by the regression model.*

Two properties of r^2 may be noted:

1. It is a nonnegative quantity. (Why?)

2. Its limits are $0 \leq r^2 \leq 1$. An r^2 of 1 means a perfect fit, that is, $\hat{Y}_i = Y_i$ for each i. On the other hand, an r^2 of zero means that there is no relationship between the regressand and the regressor whatsoever (i.e., $\hat{\beta}_2 = 0$). In this case, as (3.1.9) shows, $\hat{Y}_i = \hat{\beta}_1 = \bar{Y}$, that is, the best prediction of any Y value is simply its mean value. In this situation therefore the regression line will be horizontal to the X axis.

Although r^2 can be computed directly from its definition given in (3.5.5), it can be obtained more quickly from the following formula:

$$\begin{aligned} r^2 &= \frac{\text{ESS}}{\text{TSS}} \\ &= \frac{\sum \hat{y}_i^2}{\sum y_i^2} \\ &= \frac{\hat{\beta}_2^2 \sum x_i^2}{\sum y_i^2} \\ &= \hat{\beta}_2^2 \left(\frac{\sum x_i^2}{\sum y_i^2} \right) \end{aligned} \tag{3.5.6}$$

If we divide the numerator and the denominator of (3.5.6) by the sample size n (or $n-1$ if the sample size is small), we obtain

$$r^2 = \hat{\beta}_2^2\left(\frac{S_x^2}{S_y^2}\right) \tag{3.5.7}$$

where S_y^2 and S_x^2 are the sample variances of Y and X, respectively.

Since $\hat{\beta}_2 = \sum x_i y_i / \sum x_i^2$, Eq. (3.5.6) can also be expressed as

$$r^2 = \frac{\left(\sum x_i y_i\right)^2}{\sum x_i^2 \sum y_i^2} \tag{3.5.8}$$

an expression that may be computationally easy to obtain.

Given the definition of r^2, we can express ESS and RSS discussed earlier as follows:

$$\begin{aligned} \text{ESS} &= r^2 \cdot \text{TSS} \\ &= r^2 \sum y_i^2 \end{aligned} \tag{3.5.9}$$

$$\begin{aligned} \text{RSS} &= \text{TSS} - \text{ESS} \\ &= \text{TSS}(1 - \text{ESS/TSS}) \\ &= \sum y_i^2 \cdot (1 - r^2) \end{aligned} \tag{3.5.10}$$

Therefore, we can write

$$\begin{aligned} &\text{TSS} = \text{ESS} + \text{RSS} \\ &\sum y_i^2 = r^2 \sum y_i^2 + (1 - r^2) \sum y_i^2 \end{aligned} \tag{3.5.11}$$

an expression that we will find very useful later.

A quantity closely related to but conceptually very much different from r^2 is the **coefficient of correlation,** which, as noted in Chapter 1, is a measure of the degree of association between two variables. It can be computed either from

$$r = \pm\sqrt{r^2} \tag{3.5.12}$$

or from its definition

$$\begin{aligned} r &= \frac{\sum x_i y_i}{\sqrt{\left(\sum x_i^2\right)\left(\sum y_i^2\right)}} \\ &= \frac{n\sum X_i Y_i - \left(\sum X_i\right)\left(\sum Y_i\right)}{\sqrt{\left[n\sum X_i^2 - \left(\sum X_i\right)^2\right]\left[n\sum Y_i^2 - \left(\sum Y_i\right)^2\right]}} \end{aligned} \tag{3.5.13}$$

which is known as the **sample correlation coefficient.**[24]

[24]The population correlation coefficient, denoted by ρ, is defined in **App. A**.

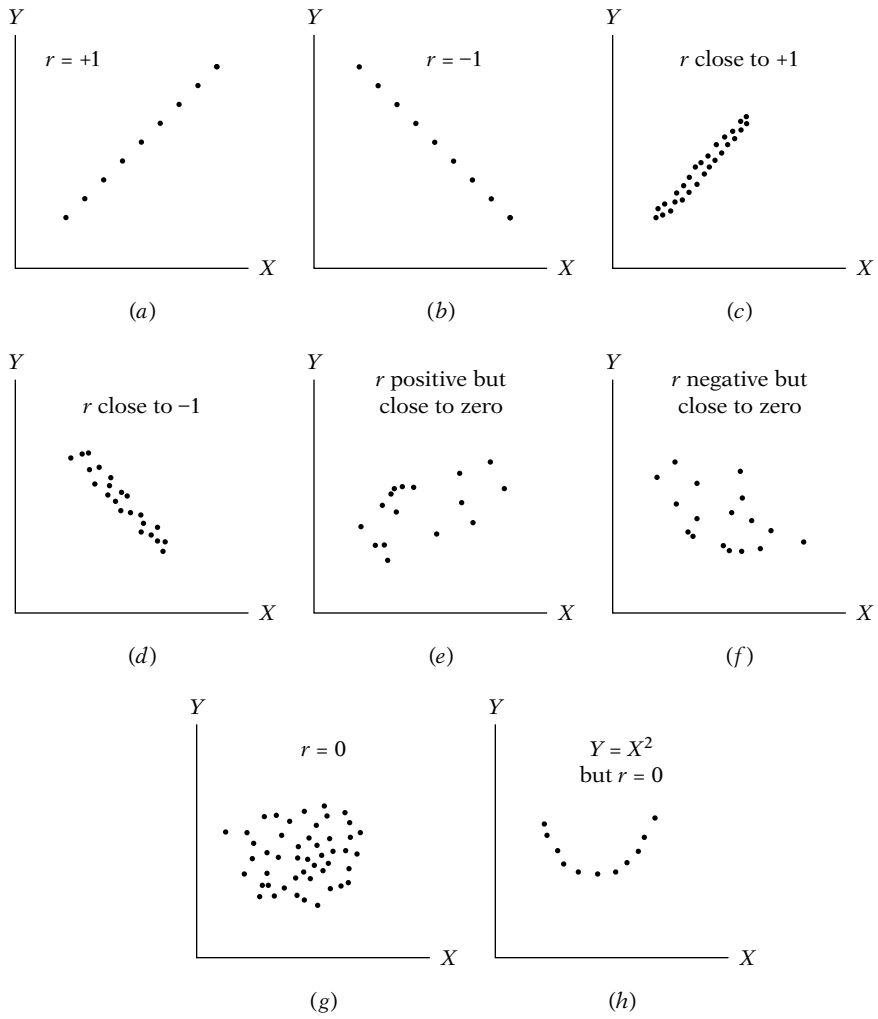


FIGURE 3.11 Correlation patterns (adapted from Henri Theil, *Introduction to Econometrics,* Prentice-Hall, Englewood Cliffs, N.J., 1978, p. 86).

Some of the properties of r are as follows (see Figure 3.11):

1. It can be positive or negative, the sign depending on the sign of the term in the numerator of (3.5.13), which measures the sample *covariation* of two variables.

2. It lies between the limits of -1 and $+1$; that is, $-1 \leq r \leq 1$.

3. It is symmetrical in nature; that is, the coefficient of correlation between X and $Y(r_{XY})$ is the same as that between Y and $X(r_{YX})$.

4. It is independent of the origin and scale; that is, if we define $X_i^* = aX_i + C$ and $Y_i^* = bY_i + d$, where $a > 0$, $b > 0$, and c and d are constants,

then r between X^* and Y^* is the same as that between the original variables X and Y.

5. If X and Y are statistically independent (see **Appendix A** for the definition), the correlation coefficient between them is zero; but if $r = 0$, it does not mean that two variables are independent. In other words, **zero correlation does not necessarily imply independence.** [See Figure 3.11(h).]

6. It is a measure of *linear association* or *linear dependence* only; it has no meaning for describing nonlinear relations. Thus in Figure 3.11(h), $Y = X^2$ is an exact relationship yet r is zero. (Why?)

7. Although it is a measure of linear association between two variables, it does not necessarily imply any cause-and-effect relationship, as noted in Chapter 1.

In the regression context, r^2 is a more meaningful measure than r, for the former tells us the proportion of variation in the dependent variable explained by the explanatory variable(s) and therefore provides an overall measure of the extent to which the variation in one variable determines the variation in the other. The latter does not have such value.[25] Moreover, as we shall see, the interpretation of r ($= R$) in a multiple regression model is of dubious value. However, we will have more to say about r^2 in Chapter 7.

In passing, note that the r^2 defined previously *can also be computed as the squared coefficient of correlation between actual Y_i and the estimated Y_i*, namely, $\hat{Y}_i$. That is, using (3.5.13), we can write

$$r^2 = \frac{\left[\sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})\right]^2}{\sum (Y_i - \bar{Y})^2 \sum (\hat{Y}_i - \bar{Y})^2}$$

That is,

$$r^2 = \frac{\left(\sum y_i \hat{y}_i\right)^2}{\left(\sum y_i^2\right)\left(\sum \hat{y}_i^2\right)} \tag{3.5.14}$$

where Y_i = actual Y, $\hat{Y}_i$ = estimated Y, and $\bar{Y} = \bar{\hat{Y}}$ = the mean of Y. For proof, see exercise 3.15. Expression (3.5.14) justifies the description of r^2 as a measure of goodness of fit, for it tells how close the estimated Y values are to their actual values.

3.6 A NUMERICAL EXAMPLE

We illustrate the econometric theory developed so far by considering the Keynesian consumption function discussed in the Introduction. Recall that Keynes stated that "The fundamental psychological law . . . is that men

[25] In regression modeling the underlying theory will indicate the direction of causality between Y and X, which, in the context of single-equation models, is generally from X to Y.

TABLE 3.2 HYPOTHETICAL DATA ON WEEKLY FAMILY CONSUMPTION EXPENDITURE *Y* AND WEEKLY FAMILY INCOME *X*

Y, $	*X*, $
70	80
65	100
90	120
95	140
110	160
115	180
120	200
140	220
155	240
150	260

[women] are disposed, as a rule and on average, to increase their consumption as their income increases, but not by as much as the increase in their income," that is, the marginal propensity to consume (MPC) is greater than zero but less than one. Although Keynes did not specify the exact functional form of the relationship between consumption and income, for simplicity assume that the relationship is linear as in (2.4.2). As a test of the Keynesian consumption function, we use the sample data of Table 2.4, which for convenience is reproduced as Table 3.2. The raw data required to obtain the estimates of the regression coefficients, their standard errors, etc., are given in Table 3.3. From these raw data, the following calculations are obtained, and the reader is advised to check them.

$$\begin{aligned} \hat{\beta}_1 &= 24.4545 \quad \text{var}(\hat{\beta}_1) = 41.1370 \quad \text{and} \quad \text{se}(\hat{\beta}_1) = 6.4138 \\ \hat{\beta}_2 &= \ \ 0.5091 \quad \text{var}(\hat{\beta}_2) = \ \ 0.0013 \quad \text{and} \quad \text{se}(\hat{\beta}_2) = 0.0357 \\ &\text{cov}(\hat{\beta}_1, \hat{\beta}_2) = -0.2172 \quad \hat{\sigma}^2 = 42.1591 \\ &r^2 = 0.9621 \quad r = 0.9809 \quad \text{df} = 8 \end{aligned} \tag{3.6.1}$$

The estimated regression line therefore is

$$\hat{Y}_i = 24.4545 + 0.5091 X_i \tag{3.6.2}$$

which is shown geometrically as Figure 3.12.

Following Chapter 2, the SRF [Eq. (3.6.2)] and the associated regression line are interpreted as follows: Each point on the regression line gives an *estimate* of the expected or mean value of Y corresponding to the chosen X value; that is, $\hat{Y}_i$ is an estimate of $E(Y \mid X_i)$. The value of $\hat{\beta}_2 = 0.5091$, which measures the slope of the line, shows that, within the sample range of X between \$80 and \$260 per week, as X increases, say, by \$1, the estimated increase in the mean or average weekly consumption expenditure amounts to about 51 cents. The value of $\hat{\beta}_1 = 24.4545$, which is the intercept of the

TABLE 3.3 RAW DATA BASED ON TABLE 3.2

Y_i (1)	X_i (2)	Y_iX_i (3)	X_i^2 (4)	$x_i = X_i - \bar{X}$ (5)	$y_i = Y_i - \bar{Y}$ (6)	x_i^2 (7)	x_iy_i (8)	$\hat{Y}_i$ (9)	$\hat{u}_i = Yi - \hat{Y}_i$ (10)	$\hat{Y}_i\hat{u}_i$ (11)
70	80	5600	6400	−90	−41	8100	3690	65.1818	4.8181	314.0524
65	100	6500	10000	−70	−46	4900	3220	75.3636	−10.3636	−781.0382
90	120	10800	14400	−50	−21	2500	1050	85.5454	4.4545	381.0620
95	140	13300	19600	−30	−16	900	480	95.7272	−0.7272	−69.6128
110	160	17600	25600	−10	−1	100	10	105.9090	4.0909	433.2631
115	180	20700	32400	10	4	100	40	116.0909	−1.0909	−126.6434
120	200	24000	40000	30	9	900	270	125.2727	−6.2727	−792.0708
140	220	30800	48400	50	29	2500	1450	136.4545	3.5454	483.7858
155	240	37200	57600	70	44	4900	3080	145.6363	8.3636	1226.4073
150	260	39000	67600	90	39	8100	3510	156.8181	−6.8181	−1069.2014
Sum 1110	1700	205500	322000	0	0	33000	16800	1109.9995 ≈ 1110.0	0	0.0040 ≈ 0.0
Mean 111	170	nc	nc	0	0	nc	nc	110	0	0

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} = 16{,}800/33{,}000 = 0.5091$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X} = 111 - 0.5091(170) = 24.4545$$

Notes: ≈ symbolizes "approximately equal to"; nc means "not computed."

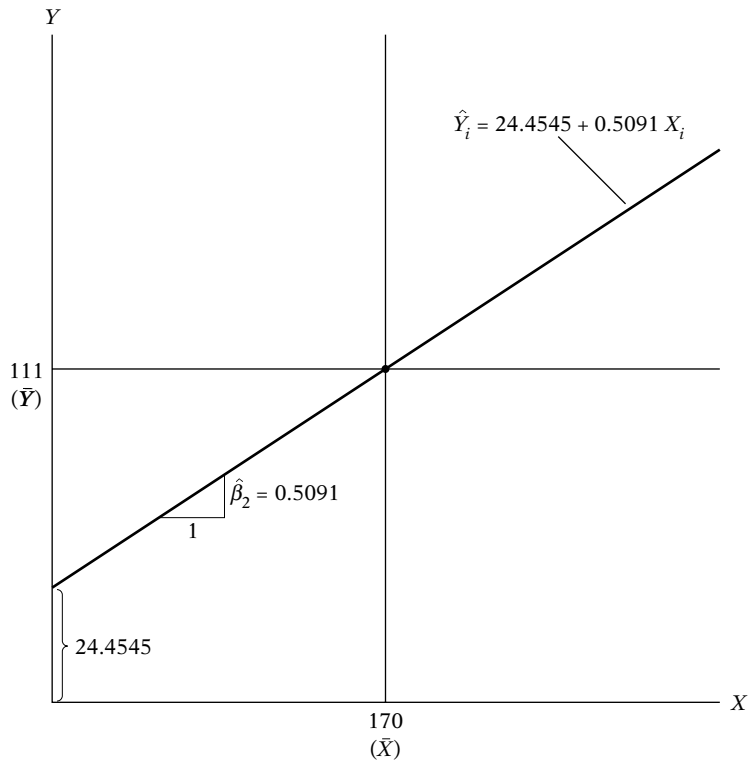


FIGURE 3.12 Sample regression line based on the data of Table 3.2.

line, indicates the average level of weekly consumption expenditure when weekly income is zero. However, this is a mechanical interpretation of the intercept term. In regression analysis such literal interpretation of the intercept term may not be always meaningful, although in the present example it can be argued that a family without any income (because of unemployment, layoff, etc.) might maintain some minimum level of consumption expenditure either by borrowing or dissaving. But in general one has to use common sense in interpreting the intercept term, for very often the sample range of X values may not include zero as one of the observed values.

Perhaps it is best to interpret the intercept term as the mean or average effect on Y of all the variables omitted from the regression model. The value of r^2 of 0.9621 means that about 96 percent of the variation in the weekly consumption expenditure is explained by income. Since r^2 can at most be 1, the observed r^2 suggests that the sample regression line fits the data very well.[26] The coefficient of correlation of 0.9809 shows that the two variables, consumption expenditure and income, are highly positively correlated. The estimated standard errors of the regression coefficients will be interpreted in Chapter 5.

3.7 ILLUSTRATIVE EXAMPLES

EXAMPLE 3.1

CONSUMPTION–INCOME RELATIONSHIP IN THE UNITED STATES, 1982–1996

Let us return to the consumption income data given in Table I.1 of the Introduction. We have already shown the data in Figure I.3 along with the estimated regression line (I.3.3). Now we provide the underlying OLS regression results. (The results were obtained from the statistical package *Eviews 3*.) *Note:* Y = personal consumption expenditure (PCE) and X = gross domestic product (GDP), all measured in 1992 billions of dollars. In this example, our data are *time series* data.

$$\hat{Y}_i = -184.0780 + 0.7064X_i \qquad \textbf{(3.7.1)}$$

$$\text{var}\,(\hat{\beta}_1) = 2140.1707 \qquad \text{se}\,(\hat{\beta}_1) = 46.2619$$

$$\text{var}\,(\hat{\beta}_2) = 0.000061 \qquad \text{se}\,(\hat{\beta}_2) = 0.007827$$

$$r^2 = 0.998406 \qquad \hat{\sigma}^2 = 411.4913$$

Equation (3.7.1) is the aggregate (i.e., for the economy as a whole) Keynesian consumption function. As this equation shows, the **marginal propensity to consume (MPC)** is about 0.71, suggesting that if income goes up by a dollar, the average personal consumption expenditure (PCE) goes up by about 71 cents. From Keynesian theory, the MPC is less than 1. The intercept value of about −184 tells us that if income were zero, the PCE would be about −184 billion dollars. Of course, such a mechanical interpretation of the intercept term does not make economic sense in the present instance because the zero income value is out of the range of values we are working with and does not represent a likely outcome (see Table I.1). As we will see on many an occasion, very often the intercept term may not make much economic sense. Therefore, in practice the intercept term may not be very meaningful, although on occasions it can be very meaningful, as we will see in some illustrative examples. The more meaningful value is the slope coefficient, MPC in the present case.

The r^2 value of 0.9984 means approximately 99 percent of the variation in the PCE is explained by variation in the GDP. Since r^2 at most can be 1, we can say that the regression line in (3.7.1), which is shown in Figure I.3, fits our data extremely well; as you can see from that figure the actual data points are very tightly clustered around the estimated regression line. As we will see throughout this book, in regressions involving time series data one generally obtains high r^2 values. In the chapter on autocorrelation, we will see the reasons behind this phenomenon.

[26] A formal test of the significance of r^2 will be presented in Chap. 8.

EXAMPLE 3.2

FOOD EXPENDITURE IN INDIA

Refer to the data given in Table 2.8 of exercise 2.15. The data relate to a sample of 55 rural households in India. The regressand in this example is expenditure on food and the regressor is total expenditure, a proxy for income, both figures in rupees. The data in this example are thus *cross-sectional* data.

On the basis of the given data, we obtained the following regression:

$$\widehat{\text{FoodExp}}_i = 94.2087 + 0.4368\ \text{TotalExp}_i \qquad \textbf{(3.7.2)}$$

$$\text{var}\,(\hat{\beta}_1) = 2560.9401 \qquad \text{se}\,(\hat{\beta}_1) = 50.8563$$

$$\text{var}\,(\hat{\beta}_2) = 0.0061 \qquad \text{se}\,(\hat{\beta}_2) = 0.0783$$

$$r^2 = 0.3698 \qquad \hat{\sigma}^2 = 4469.6913$$

From (3.7.2) we see that if total expenditure increases by 1 rupee, on average, expenditure on food goes up by about 44 paise (1 rupee = 100 paise). If total expenditure were zero, the average expenditure on food would be about 94 rupees. Again, such a mechanical interpretation of the intercept may not be meaningful. However, in this example one could argue that even if total expenditure is zero (e.g., because of loss of a job), people may still maintain some minimum level of food expenditure by borrowing money or by dissaving.

The r^2 value of about 0.37 means that only 37 percent of the variation in food expenditure is explained by the total expenditure. This might seem a rather low value, but as we will see throughout this text, in cross-sectional data, typically one obtains low r^2 values, possibly because of the diversity of the units in the sample. We will discuss this topic further in the chapter on heteroscedasticity (see Chapter 11).

EXAMPLE 3.3

THE RELATIONSHIP BETWEEN EARNINGS AND EDUCATION

In Table 2.6 we looked at the data relating average hourly earnings and education, as measured by years of schooling. Using that data, if we regress[27] average hourly earnings (Y) on education (X), we obtain the following results.

$$\hat{Y}_i = -0.0144 + 0.7241\ X_i \qquad \textbf{(3.7.3)}$$

$$\text{var}\,(\hat{\beta}_1) = 0.7649 \quad se\,(\hat{\beta}_1) = 0.8746$$

$$\text{var}\,(\hat{\beta}_2) = 0.00483 \quad se\,(\hat{\beta}_2) = 0.0695$$

$$r^2 = 0.9077 \qquad \hat{\sigma}^2 = 0.8816$$

As the regression results show, there is a positive association between education and earnings, an unsurprising finding. For every additional year of schooling, the average hourly earnings go up by about 72 cents an hour. The intercept term is positive but it may have no economic meaning. The r^2 value suggests that about 89 percent of the variation in average hourly earnings is explained by education. For cross-sectional data, such a high r^2 is rather unusual.

3.8 A NOTE ON MONTE CARLO EXPERIMENTS

In this chapter we showed that under the assumptions of CLRM the least-squares estimators have certain desirable statistical features summarized in the BLUE property. In the appendix to this chapter we prove this property

[27] Every field of study has its jargon. The expression "regress Y on X" simply means treat Y as the regressand and X as the regressor.

more formally. But in practice how does one know that the BLUE property holds? For example, how does one find out if the OLS estimators are unbiased? The answer is provided by the so-called **Monte Carlo** experiments, which are essentially computer simulation, or sampling, experiments.

To introduce the basic ideas, consider our two-variable PRF:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{3.8.1}$$

A Monte Carlo experiment proceeds as follows:

1. Suppose the true values of the parameters are as follows: $\beta_1 = 20$ and $\beta_2 = 0.6$.

2. You choose the sample size, say $n = 25$.

3. You fix the values of X for each observation. In all you will have 25 X values.

4. Suppose you go to a random number table, choose 25 values, and call them u_i (these days most statistical packages have built-in random number generators).[28]

5. Since you *know* β_1, β_2, X_i, and u_i, using (3.8.1) you obtain 25 Y_i values.

6. Now using the 25 Y_i values thus generated, you regress these on the 25 X values chosen in step 3, obtaining $\hat{\beta}_1$ and $\hat{\beta}_2$, the least-squares estimators.

7. Suppose you repeat this experiment 99 times, each time using the same β_1, β_2, and X values. Of course, the u_i values will vary from experiment to experiment. Therefore, in all you have 100 experiments, thus generating 100 values each of β_1 and β_2. (In practice, many such experiments are conducted, sometimes 1000 to 2000.)

8. You take the averages of these 100 estimates and call them $\bar{\hat{\beta}}_1$ and $\bar{\hat{\beta}}_2$.

9. If these average values are about the same as the true values of β_1 and β_2 assumed in step 1, this Monte Carlo experiment "establishes" that the least-squares estimators are indeed unbiased. Recall that under CLRM $E(\hat{\beta}_1) = \beta_1$ and $E(\hat{\beta}_2) = \beta_2$.

These steps characterize the general nature of the Monte Carlo experiments. Such experiments are often used to study the statistical properties of various methods of estimating population parameters. They are particularly useful to study the behavior of estimators in small, or finite, samples. These experiments are also an excellent means of driving home the concept of **repeated sampling** that is the basis of most of classical statistical inference, as we shall see in Chapter 5. We shall provide several examples of Monte Carlo experiments by way of exercises for classroom assignment. (See exercise 3.27.)

[28] In practice it is assumed that u_i follows a certain probability distribution, say, normal, with certain parameters (e.g., the mean and variance). Once the values of the parameters are specified, one can easily generate the u_i using statistical packages.

3.9 SUMMARY AND CONCLUSIONS

The important topics and concepts developed in this chapter can be summarized as follows.

1. The basic framework of regression analysis is the **CLRM.**

2. The CLRM is based on a set of assumptions.

3. Based on these assumptions, the least-squares estimators take on certain properties summarized in the Gauss–Markov theorem, which states that in the class of linear unbiased estimators, the least-squares estimators have minimum variance. In short, they are BLUE.

4. The *precision* of OLS estimators is measured by their **standard errors.** In Chapters 4 and 5 we shall see how the standard errors enable one to draw inferences on the population parameters, the β coefficients.

5. The overall goodness of fit of the regression model is measured by the **coefficient of determination,** r^2. It tells what proportion of the variation in the dependent variable, or regressand, is explained by the explanatory variable, or regressor. This r^2 lies between 0 and 1; the closer it is to 1, the better is the fit.

6. A concept related to the coefficient of determination is the **coefficient of correlation,** r. It is a measure of *linear association* between two variables and it lies between -1 and $+1$.

7. The CLRM is a theoretical construct or abstraction because it is based on a set of assumptions that may be stringent or "unrealistic." But such abstraction is often necessary in the initial stages of studying any field of knowledge. Once the CLRM is mastered, one can find out what happens if one or more of its assumptions are not satisfied. The first part of this book is devoted to studying the CLRM. The other parts of the book consider the refinements of the CLRM. Table 3.4 gives the road map ahead.

TABLE 3.4 WHAT HAPPENS IF THE ASSUMPTIONS OF CLRM ARE VIOLATED?

Assumption number	Type of violation	Where to study?
1	Nonlinearity in parameters	Chapter 14
2	Stochastic regressor(s)	Introduction to Part II
3	Nonzero mean of u_i	Introduction to Part II
4	Heteroscedasticity	Chapter 11
5	Autocorrelated disturbances	Chapter 12
6	Nonzero covariance between disturbances and regressor	Introduction to Part II and Part IV
7	Sample observations less than the number of regressors	Chapter 10
8	Insufficient variability in regressors	Chapter 10
9	Specification bias	Chapters 13, 14
10	Multicollinearity	Chapter 10
11*	Nonnormality of disturbances	Introduction to Part II

**Note:* The assumption that the disturbances u_i are normally distributed is not a part of the CLRM. But more on this in Chapter 4.

EXERCISES

Questions

3.1. Given the assumptions in column 1 of the table, show that the assumptions in column 2 are equivalent to them.

ASSUMPTIONS OF THE CLASSICAL MODEL

(1)	(2)
$E(u_i \mid X_i) = 0$	$E(Y_i \mid X_i) = \beta_2 + \beta_2 X$
cov $(u_i, u_j) = 0$ $i \neq j$	cov $(Y_i, Y_j) = 0$ $i \neq j$
var $(u_i \mid X_i) = \sigma^2$	var $(Y_i \mid X_i) = \sigma^2$

3.2. Show that the estimates $\hat{\beta}_1 = 1.572$ and $\hat{\beta}_2 = 1.357$ used in the first experiment of Table 3.1 are in fact the OLS estimators.

3.3. According to Malinvaud (see footnote 10), the assumption that $E(u_i \mid X_i) = 0$ is quite important. To see this, consider the PRF: $Y = \beta_1 + \beta_2 X_i + u_i$. Now consider two situations: (i) $\beta_1 = 0$, $\beta_2 = 1$, and $E(u_i) = 0$; and (ii) $\beta_1 = 1$, $\beta_2 = 0$, and $E(u_i) = (X_i - 1)$. Now take the expectation of the PRF conditional upon X in the two preceding cases and see if you agree with Malinvaud about the significance of the assumption $E(u_i \mid X_i) = 0$.

3.4. Consider the sample regression

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$$

Imposing the restrictions (i) $\sum \hat{u}_i = 0$ and (ii) $\sum \hat{u}_i X_i = 0$, obtain the estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ and show that they are identical with the least-squares estimators given in (3.1.6) and (3.1.7). This method of obtaining estimators is called the **analogy principle.** Give an intuitive justification for imposing restrictions (i) and (ii). (*Hint:* Recall the CLRM assumptions about u_i.) In passing, note that the analogy principle of estimating unknown parameters is also known as the **method of moments** in which sample moments (e.g., sample mean) are used to estimate population moments (e.g., the population mean). As noted in **Appendix A**, a **moment** is a summary statistic of a probability distribution, such as the expected value and variance.

3.5. Show that r^2 defined in (3.5.5) ranges between 0 and 1. You may use the Cauchy–Schwarz inequality, which states that for any random variables X and Y the following relationship holds true:

$$[E(XY)]^2 \leq E(X^2)E(Y^2)$$

3.6. Let $\hat{\beta}_{YX}$ and $\hat{\beta}_{XY}$ represent the slopes in the regression of Y on X and X on Y, respectively. Show that

$$\hat{\beta}_{YX}\hat{\beta}_{XY} = r^2$$

where r is the coefficient of correlation between X and Y.

3.7. Suppose in exercise 3.6 that $\hat{\beta}_{YX}\hat{\beta}_{XY} = 1$. Does it matter then if we regress Y on X or X on Y? Explain carefully.

3.8. Spearman's rank correlation coefficient r_s is defined as follows:

$$r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

where d = difference in the ranks assigned to the same individual or phenomenon and n = number of individuals or phenomena ranked. Derive r_s from r defined in (3.5.13). *Hint:* Rank the X and Y values from 1 to n. Note that the sum of X and Y ranks is $n(n + 1)/2$ each and therefore their means are $(n + 1)/2$.

3.9. Consider the following formulations of the two-variable PRF:

$$\text{Model I:} \quad Y_i = \beta_1 + \beta_2 X_i + u_i$$

$$\text{Model II:} \quad Y_i = \alpha_1 + \alpha_2 (X_i - \bar{X}) + u_i$$

a. Find the estimators of β_1 and α_1. Are they identical? Are their variances identical?

b. Find the estimators of β_2 and α_2. Are they identical? Are their variances identical?

c. What is the advantage, if any, of model II over model I?

3.10. Suppose you run the following regression:

$$y_i = \hat{\beta}_1 + \hat{\beta}_2 x_i + \hat{u}_i$$

where, as usual, y_i and x_i are deviations from their respective mean values. What will be the value of $\hat{\beta}_1$? Why? Will $\hat{\beta}_2$ be the same as that obtained from Eq. (3.1.6)? Why?

3.11. Let r_1 = coefficient of correlation between n pairs of values (Y_i, X_i) and r_2 = coefficient of correlation between n pairs of values $(aX_i + b,\ cY_i + d)$, where a, b, c, and d are constants. Show that $r_1 = r_2$ and hence *establish the principle that the coefficient of correlation is invariant with respect to the change of scale and the change of origin.*

Hint: Apply the definition of r given in (3.5.13).

Note: The operations aX_i, $X_i + b$, and $aX_i + b$ are known, respectively, as the *change of scale, change of origin,* and *change of both scale and origin.*

3.12. If r, the coefficient of correlation between n pairs of values (X_i, Y_i), is positive, then determine whether each of the following statements is true or false:

a. r between $(-X_i, -Y_i)$ is also positive.

b. r between $(-X_i, Y_i)$ and that between $(X_i, -Y_i)$ can be either positive or negative.

c. Both the slope coefficients β_{yx} and β_{xy} are positive, where β_{yx} = slope coefficient in the regression of Y on X and β_{xy} = slope coefficient in the regression of X on Y.

3.13. If X_1, X_2, and X_3 are uncorrelated variables each having the same standard deviation, show that the coefficient of correlation between $X_1 + X_2$ and $X_2 + X_3$ is equal to $\frac{1}{2}$. Why is the correlation coefficient not zero?

3.14. In the regression $Y_i = \beta_1 + \beta_2 X_i + u_i$ suppose we *multiply* each X value by a constant, say, 2. Will it change the residuals and fitted values of Y? Explain. What if we *add* a constant value, say, 2, to each X value?

3.15. Show that (3.5.14) in fact measures the coefficient of determination. *Hint:* Apply the definition of r given in (3.5.13) and recall that $\sum y_i \hat{y}_i = \sum(\hat{y}_i + \hat{u}_i)\hat{y}_i = \sum \hat{y}_i^2$, and remember (3.5.6).

3.16. Explain *with reason* whether the following statements are true, false, or uncertain:

- **a.** Since the correlation between two variables, Y and X, can range from -1 to $+1$, this also means that cov (Y, X) also lies between these limits.
- **b.** If the correlation between two variables is zero, it means that there is no relationship between the two variables whatsoever.
- **c.** If you regress Y_i on $\hat{Y}_i$ (i.e., actual Y on estimated Y), the intercept and slope values will be 0 and 1, respectively.

3.17. *Regression without any regressor.* Suppose you are given the model: $Y_i = \beta_1 + u_i$. Use OLS to find the estimator of β_1. What is its variance and the RSS? Does the estimated β_1 make intuitive sense? Now consider the two-variable model $Y_i = \beta_1 + \beta_2 X_i + u_i$. Is it worth adding X_i to the model? If not, why bother with regression analysis?

Problems

3.18. In Table 3.5, you are given the ranks of 10 students in midterm and final examinations in statistics. Compute Spearman's coefficient of rank correlation and interpret it.

3.19. *The relationship between nominal exchange rate and relative prices.* From the annual observations from 1980 to 1994, the following regression results were obtained, where Y = exchange rate of the German mark to the U.S. dollar (GM/$) and X = ratio of the U.S. consumer price index to the German consumer price index; that is, X represents the relative prices in the two countries:

$$\hat{Y}_t = 6.682 - 4.318 X_t \qquad r^2 = 0.528$$
$$\text{se} = (1.22)(1.333)$$

- **a.** Interpret this regression. How would you interpret r^2?
- **b.** Does the negative value of X_t make economic sense? What is the underlying economic theory?
- **c.** Suppose we were to redefine X as the ratio of German CPI to the U.S. CPI. Would that change the sign of X? And why?

3.20. Table 3.6 gives data on indexes of output per hour (X) and real compensation per hour (Y) for the business and nonfarm business sectors of the U.S. economy for 1959–1997. The base year of the indexes is 1982 = 100 and the indexes are seasonally adjusted.

TABLE 3.5

	Student									
Rank	A	B	C	D	E	F	G	H	I	J
Midterm	1	3	7	10	9	5	4	8	2	6
Final	3	2	8	7	9	6	5	10	1	4

TABLE 3.6 PRODUCTIVITY AND RELATED DATA, BUSINESS SECTOR, 1959–98
[Index numbers, 1992 = 100; quarterly data seasonally adjusted]

	Output per hour of all persons[1]		Compensation per hour[2]	
Year or quarter	Business sector	Nonfarm business sector	Business sector	Nonfarm business sector
1959	50.5	54.2	13.1	13.7
1960	51.4	54.8	13.7	14.3
1961	53.2	56.6	14.2	14.8
1962	55.7	59.2	14.8	15.4
1963	57.9	61.2	15.4	15.9
1964	60.6	63.8	16.2	16.7
1965	62.7	65.8	16.8	17.2
1966	65.2	68.0	17.9	18.2
1967	66.6	69.2	18.9	19.3
1968	68.9	71.6	20.5	20.8
1969	69.2	71.7	21.9	22.2
1970	70.6	72.7	23.6	23.8
1971	73.6	75.7	25.1	25.4
1972	76.0	78.3	26.7	27.0
1973	78.4	80.7	29.0	29.2
1974	77.1	79.4	31.8	32.1
1975	79.8	81.6	35.1	35.3
1976	82.5	84.5	38.2	38.4
1977	84.0	85.8	41.2	41.5
1978	84.9	87.0	44.9	45.2
1979	84.5	86.3	49.2	49.5
1980	84.2	86.0	54.5	54.8
1981	85.8	87.0	59.6	60.2
1982	85.3	88.3	64.1	64.6
1983	88.0	89.9	66.8	67.3
1984	90.2	91.4	69.7	70.2
1985	91.7	92.3	73.1	73.4
1986	94.1	94.7	76.8	77.2
1987	94.0	94.5	79.8	80.1
1988	94.7	95.3	83.6	83.7
1989	95.5	95.8	85.9	86.0
1990	96.1	96.3	90.8	90.7
1991	96.7	97.0	95.1	95.1
1992	100.0	100.0	100.0	100.0
1993	100.1	100.1	102.5	102.2
1994	100.7	100.6	104.4	104.2
1995	101.0	101.2	106.8	106.7
1996	103.7	103.7	110.7	110.4
1997	105.4	105.1	114.9	114.5

[1]Output refers to real gross domestic product in the sector.
[2]Wages and salaries of employees plus employers' contributions for social insurance and private benefit plans. Also includes an estimate of wages, salaries, and supplemental payments for the self-employed.
Source: Economic Report of the President, 1999, Table B-49, p. 384.

a. Plot Y against X for the two sectors separately.
b. What is the economic theory behind the relationship between the two variables? Does the scattergram support the theory?
c. Estimate the OLS regression of Y on X. Save the results for a further look after we study Chapter 5.

3.21. From a sample of 10 observations, the following results were obtained:

$$\sum Y_i = 1110 \qquad \sum X_i = 1700 \qquad \sum X_i Y_i = 205{,}500$$

$$\sum X_i^2 = 322{,}000 \qquad \sum Y_i^2 = 132{,}100$$

with coefficient of correlation $r = 0.9758$. But on rechecking these calculations it was found that two pairs of observations were recorded:

Y	X		Y	X
90	120	instead of	80	110
140	220		150	210

What will be the effect of this error on r? Obtain the correct r.

3.22. Table 3.7 gives data on gold prices, the Consumer Price Index (CPI), and the New York Stock Exchange (NYSE) Index for the United States for the period 1977–1991. The NYSE Index includes most of the stocks listed on the NYSE, some 1500 plus.

TABLE 3.7

Year	Price of gold at New York, $ per troy ounce	Consumer Price Index (CPI), 1982–84 = 100	New York Stock Exchange (NYSE) Index, Dec. 31, 1965 = 100
1977	147.98	60.6	53.69
1978	193.44	65.2	53.70
1979	307.62	72.6	58.32
1980	612.51	82.4	68.10
1981	459.61	90.9	74.02
1982	376.01	96.5	68.93
1983	423.83	99.6	92.63
1984	360.29	103.9	92.46
1985	317.30	107.6	108.90
1986	367.87	109.6	136.00
1987	446.50	113.6	161.70
1988	436.93	118.3	149.91
1989	381.28	124.0	180.02
1990	384.08	130.7	183.46
1991	362.04	136.2	206.33

Source: Data on CPI and NYSE Index are from the *Economic Report of the President,* January 1993, Tables B-59 and B-91, respectively. Data on gold prices are from U.S. Department of Commerce, Bureau of Economic Analysis, *Business Statistics, 1963–1991,* p. 68.

a. Plot in the same scattergram gold prices, CPI, and the NYSE Index.
b. An investment is supposed to be a hedge against inflation if its price and/or rate of return at least keeps pace with inflation. To test this hypothesis, suppose you decide to fit the following model, assuming the scatterplot in **a** suggests that this is appropriate:

$$\text{Gold price}_t = \beta_1 + \beta_2\, \text{CPI}_t + u_t$$
$$\text{NYSE index}_t = \beta_1 + \beta_2\, \text{CPI}_t + u_t$$

3.23. Table 3.8 gives data on gross domestic product (GDP) for the United States for the years 1959–1997.
a. Plot the GDP data in current and constant (i.e., 1992) dollars against time.
b. Letting Y denote GDP and X time (measured chronologically starting with 1 for 1959, 2 for 1960, through 39 for 1997), see if the following model fits the GDP data:

$$Y_t = \beta_1 + \beta_2\, X_t + u_t$$

Estimate this model for both current and constant-dollar GDP.
c. How would you interpret β_2?
d. If there is a difference between β_2 estimated for current-dollar GDP and that estimated for constant-dollar GDP, what explains the difference?

TABLE 3.8 NOMINAL AND REAL GDP, UNITED STATES, 1959–1997

Year	NGDP	RGDP	Year	NGDP	RGDP
1959	507.2000	2210.200	1979	2557.500	4630.600
1960	526.6000	2262.900	1980	2784.200	4615.000
1961	544.8000	2314.300	1981	3115.900	4720.700
1962	585.2000	2454.800	1982	3242.100	4620.300
1963	617.4000	2559.400	1983	3514.500	4803.700
1964	663.0000	2708.400	1984	3902.400	5140.100
1965	719.1000	2881.100	1985	4180.700	5323.500
1966	787.7000	3069.200	1986	4422.200	5487.700
1967	833.6000	3147.200	1987	4692.300	5649.500
1968	910.6000	3293.900	1988	5049.600	5865.200
1969	982.2000	3393.600	1989	5438.700	6062.000
1970	1035.600	3397.600	1990	5743.800	6136.300
1971	1125.400	3510.000	1991	5916.700	6079.400
1972	1237.300	3702.300	1992	6244.400	6244.400
1973	1382.600	3916.300	1993	6558.100	6389.600
1974	1496.900	3891.200	1994	6947.000	6610.700
1975	1630.600	3873.900	1995	7269.600	6761.700
1976	1819.000	4082.900	1996	7661.600	6994.800
1977	2026.900	4273.600	1997	8110.900	7269.800
1978	2291.400	4503.000			

Note: NGDP = nominal GDP (current dollars in billions).
RGDP = real GDP (1992 billions of dollars).
Source: Economic Report of the President, 1999, Tables B-1 and B-2, pp. 326–328.

e. From your results what can you say about the nature of inflation in the United States over the sample period?

3.24. Using the data given in Table I.1 of the Introduction, verify Eq. (3.7.1).

3.25. For the S.A.T. example given in exercise 2.16 do the following:

a. Plot the female verbal score against the male verbal score.

b. If the scatterplot suggests that a linear relationship between the two seems appropriate, obtain the regression of female verbal score on male verbal score.

c. If there is a relationship between the two verbal scores, is the relationship *causal?*

3.26. Repeat exercise 3.24, replacing math scores for verbal scores.

3.27. Monte Carlo study *classroom assignment:* Refer to the 10 X values given in Table 3.2. Let $\beta_1 = 25$ and $\beta_2 = 0.5$. Assume $u_i \approx N(0, 9)$, that is, u_i are normally distributed with mean 0 and variance 9. Generate 100 samples using these values, obtaining 100 estimates of β_1 and β_2. Graph these estimates. What conclusions can you draw from the Monte Carlo study? *Note:* Most statistical packages now can generate random variables from most well-known probability distributions. Ask your instructor for help, in case you have difficulty generating such variables.

APPENDIX 3A

3A.1 DERIVATION OF LEAST-SQUARES ESTIMATES

Differentiating (3.1.2) partially with respect to $\hat{\beta}_1$ and $\hat{\beta}_2$, we obtain

$$\frac{\partial(\sum \hat{u}_i^2)}{\partial \hat{\beta}_1} = -2\sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = -2\sum \hat{u}_i \tag{1}$$

$$\frac{\partial(\sum \hat{u}_i^2)}{\partial \hat{\beta}_2} = -2\sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)X_i = -2\sum \hat{u}_i X_i \tag{2}$$

Setting these equations to zero, after algebraic simplification and manipulation, gives the estimators given in Eqs. (3.1.6) and (3.1.7).

3A.2 LINEARITY AND UNBIASEDNESS PROPERTIES OF LEAST-SQUARES ESTIMATORS

From (3.1.8) we have

$$\hat{\beta}_2 = \frac{\sum x_i Y_i}{\sum x_i^2} = \sum k_i Y_i \tag{3}$$

where

$$k_i = \frac{x_i}{(\sum x_i^2)}$$

which shows that $\hat{\beta}_2$ is a **linear estimator** because it is a linear function of Y; actually it is a weighted average of Y_i with k_i serving as the weights. It can similarly be shown that $\hat{\beta}_1$ too is a linear estimator.

Incidentally, note these properties of the weights k_i:

1. Since the X_i are assumed to be nonstochastic, the k_i are nonstochastic too.

2. $\sum k_i = 0$.

3. $\sum k_i^2 = 1/\sum x_i^2$.

4. $\sum k_i x_i = \sum k_i X_i = 1$. These properties can be directly verified from the definition of k_i.

For example,

$$\begin{aligned}\sum k_i = \sum\left(\frac{x_i}{\sum x_i^2}\right) &= \frac{1}{\sum x_i^2}\sum x_i, && \text{since for a given sample } \textstyle\sum x_i^2 \text{ is known}\\ &= 0, && \text{since } \textstyle\sum x_i, \text{ the sum of deviations from the mean value, is always zero}\end{aligned}$$

Now substitute the PRF $Y_i = \beta_1 + \beta_2 X_i + u_i$ into (3) to obtain

$$\begin{aligned}\hat{\beta}_2 &= \sum k_i(\beta_1 + \beta_2 X_i + u_i)\\ &= \beta_1\sum k_i + \beta_2\sum k_i X_i + \sum k_i u_i\\ &= \beta_2 + \sum k_i u_i\end{aligned} \tag{4}$$

where use is made of the properties of k_i noted earlier.

Now taking expectation of (4) on both sides and noting that k_i, being nonstochastic, can be treated as constants, we obtain

$$\begin{aligned}E(\hat{\beta}_2) &= \beta_2 + \sum k_i E(u_i)\\ &= \beta_2\end{aligned} \tag{5}$$

since $E(u_i) = 0$ by assumption. Therefore, $\hat{\beta}_2$ is an unbiased estimator of β_2. Likewise, it can be proved that $\hat{\beta}_1$ is also an unbiased estimator of β_1.

3A.3 VARIANCES AND STANDARD ERRORS OF LEAST-SQUARES ESTIMATORS

Now by the definition of variance, we can write

$$\begin{aligned}\text{var}\,(\hat{\beta}_2) &= E[\hat{\beta}_2 - E(\hat{\beta}_2)]^2\\ &= E(\hat{\beta}_2 - \beta_2)^2 && \text{since } E(\hat{\beta}_2) = \beta_2\\ &= E\left(\sum k_i u_i\right)^2 && \text{using Eq. (4) above}\\ &= E\left(k_1^2u_1^2 + k_2^2u_2^2 + \cdots + k_n^2u_n^2 + 2k_1k_2u_1u_2 + \cdots + 2k_{n-1}k_nu_{n-1}u_n\right)\end{aligned} \tag{6}$$

Since by assumption, $E(u_i^2) = \sigma^2$ for each i and $E(u_i u_j) = 0$, $i \neq j$, it follows that

$$\begin{aligned} \operatorname{var}(\hat{\beta}_2) &= \sigma^2 \sum k_i^2 \\ &= \frac{\sigma^2}{\sum x_i^2} \qquad \text{(using the definition of } k_i^2\text{)} \\ &= \text{Eq. (3.3.1)} \end{aligned} \tag{7}$$

The variance of $\hat{\beta}_1$ can be obtained following the same line of reasoning already given. Once the variances of $\hat{\beta}_1$ and $\hat{\beta}_2$ are obtained, their positive square roots give the corresponding standard errors.

3A.4 COVARIANCE BETWEEN $\hat{\beta}_1$ AND $\hat{\beta}_2$

By definition,

$$\begin{aligned} \operatorname{cov}(\hat{\beta}_1, \hat{\beta}_2) &= E\{[\hat{\beta}_1 - E(\hat{\beta}_1)][\hat{\beta}_2 - E(\hat{\beta}_2)]\} \\ &= E(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2) \qquad \text{(Why?)} \\ &= -\bar{X} E(\hat{\beta}_2 - \beta_2)^2 \\ &= -\bar{X} \operatorname{var}(\hat{\beta}_2) \\ &= \text{Eq. (3.3.9)} \end{aligned} \tag{8}$$

where use is made of the fact that $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$ and $E(\hat{\beta}_1) = \bar{Y} - \beta_2 \bar{X}$, giving $\hat{\beta}_1 - E(\hat{\beta}_1) = -\bar{X}(\hat{\beta}_2 - \beta_2)$. *Note:* $\operatorname{var}(\hat{\beta}_2)$ is given in (3.3.1).

3A.5 THE LEAST-SQUARES ESTIMATOR OF σ^2

Recall that

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{9}$$

Therefore,

$$\bar{Y} = \beta_1 + \beta_2 \bar{X} + \bar{u} \tag{10}$$

Subtracting (10) from (9) gives

$$y_i = \beta_2 x_i + (u_i - \bar{u}) \tag{11}$$

Also recall that

$$\hat{u}_i = y_i - \hat{\beta}_2 x_i \tag{12}$$

Therefore, substituting (11) into (12) yields

$$\hat{u}_i = \beta_2 x_i + (u_i - \bar{u}) - \hat{\beta}_2 x_i \tag{13}$$

Collecting terms, squaring, and summing on both sides, we obtain

$$\sum \hat{u}_i^2 = (\hat{\beta}_2 - \beta_2)^2 \sum x_i^2 + \sum (u_i - \bar{u})^2 - 2(\hat{\beta}_2 - \beta_2) \sum x_i (u_i - \bar{u}) \quad \textbf{(14)}$$

Taking expectations on both sides gives

$$\begin{aligned} E\left(\sum \hat{u}_i^2\right) &= \sum x_i^2 E(\hat{\beta}_2 - \beta_2)^2 + E\left[\sum (u_i - \bar{u})^2\right] - 2E\left[(\hat{\beta}_2 - \beta_2) \sum x_i (u_i - \bar{u})\right] \\ &= \sum x_i^2 \operatorname{var}(\hat{\beta}_2) + (n-1) \operatorname{var}(u_i) - 2E\left[\sum k_i u_i (x_i u_i)\right] \\ &= \sigma^2 + (n-1)\sigma^2 - 2E\left[\sum k_i x_i u_i^2\right] \\ &= \sigma^2 + (n-1)\sigma^2 - 2\sigma^2 \\ &= (n-2)\sigma^2 \end{aligned} \quad \textbf{(15)}$$

where, in the last but one step, use is made of the definition of k_i given in Eq. (3) and the relation given in Eq. (4). Also note that

$$\begin{aligned} E\sum (u_i - \bar{u})^2 &= E\left[\sum u_i^2 - n\bar{u}^2\right] \\ &= E\left[\sum u_i^2 - n\left(\frac{\sum u_i}{n}\right)^2\right] \\ &= E\left[\sum u_i^2 - \frac{1}{n}\sum (u_i^2)\right] \\ &= n\sigma^2 - \frac{n}{n}\sigma^2 = (n-1)\sigma^2 \end{aligned}$$

where use is made of the fact that the u_i are uncorrelated and the variance of each u_i is σ^2.

Thus, we obtain

$$E\left(\sum \hat{u}_i^2\right) = (n-2)\sigma^2 \quad \textbf{(16)}$$

Therefore, if we define

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2} \quad \textbf{(17)}$$

its expected value is

$$E(\hat{\sigma}^2) = \frac{1}{n-2} E\left(\sum \hat{u}_i^2\right) = \sigma^2 \qquad \text{using (16)} \quad \textbf{(18)}$$

which shows that $\hat{\sigma}^2$ is an unbiased estimator of true σ^2.

3A.6 MINIMUM-VARIANCE PROPERTY OF LEAST-SQUARES ESTIMATORS

It was shown in Appendix 3A, Section 3A.2, that the least-squares estimator $\hat{\beta}_2$ is linear as well as unbiased (this holds true of $\hat{\beta}_1$ too). To show that these estimators are also minimum variance in the class of all linear unbiased estimators, consider the least-squares estimator $\hat{\beta}_2$:

$$\hat{\beta}_2 = \sum k_i Y_i$$

where

$$k_i = \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} = \frac{x_i}{\sum x_i^2} \qquad \text{(see Appendix 3A.2)} \tag{19}$$

which shows that $\hat{\beta}_2$ is a weighted average of the Y's, with k_i serving as the weights.

Let us define an alternative linear estimator of β_2 as follows:

$$\beta_2^* = \sum w_i Y_i \tag{20}$$

where w_i are also weights, not necessarily equal to k_i. Now

$$\begin{aligned} E(\beta_2^*) &= \sum w_i E(Y_i) \\ &= \sum w_i (\beta_1 + \beta_2 X_i) \\ &= \beta_1 \sum w_i + \beta_2 \sum w_i X_i \end{aligned} \tag{21}$$

Therefore, for β_2^* to be unbiased, we must have

$$\sum w_i = 0 \tag{22}$$

and

$$\sum w_i X_i = 1 \tag{23}$$

Also, we may write

$$\begin{aligned} \operatorname{var}(\beta_2^*) &= \operatorname{var} \sum w_i Y_i \\ &= \sum w_i^2 \operatorname{var} Y_i \qquad [\textit{Note:} \operatorname{var} Y_i = \operatorname{var} u_i = \sigma^2] \\ &= \sigma^2 \sum w_i^2 \qquad [\textit{Note:} \operatorname{cov}(Y_i, Y_j) = 0 \, (i \neq j)] \\ &= \sigma^2 \sum \left(w_i - \frac{x_i}{\sum x_i^2} + \frac{x_i}{\sum x_i^2} \right)^2 \qquad \text{(Note the mathematical trick)} \\ &= \sigma^2 \sum \left(w_i - \frac{x_i}{\sum x_i^2} \right)^2 + \sigma^2 \frac{\sum x_i^2}{\left(\sum x_i^2 \right)^2} + 2\sigma^2 \sum \left(w_i - \frac{x_i}{\sum x_i^2} \right) \left(\frac{x_i}{\sum x_i^2} \right) \\ &= \sigma^2 \sum \left(w_i - \frac{x_i}{\sum x_i^2} \right)^2 + \sigma^2 \left(\frac{1}{\sum x_i^2} \right) \end{aligned} \tag{24}$$

because the last term in the next to the last step drops out. (Why?)

Since the last term in (24) is constant, the variance of (β_2^*) can be minimized only by manipulating the first term. If we let

$$w_i = \frac{x_i}{\sum x_i^2}$$

Eq. (24) reduces to

$$\begin{aligned} \text{var}\,(\beta_2^*) &= \frac{\sigma^2}{\sum x_i^2} \\ &= \text{var}\,(\hat{\beta}_2) \end{aligned} \tag{25}$$

In words, with weights $w_i = k_i$, which are the least-squares weights, the variance of the linear estimator β_2^* is equal to the variance of the least-squares estimator $\hat{\beta}_2$; otherwise $\text{var}\,(\beta_2^*) > \text{var}\,(\hat{\beta}_2)$. To put it differently, if there is a minimum-variance linear unbiased estimator of β_2, it must be the least-squares estimator. Similarly it can be shown that $\hat{\beta}_1$ is a minimum-variance linear unbiased estimator of β_1.

3A.7 CONSISTENCY OF LEAST-SQUARES ESTIMATORS

We have shown that, in the framework of the classical linear regression model, the least-squares estimators are unbiased (and efficient) in any sample size, small or large. But sometimes, as discussed in **Appendix A,** an estimator may not satisfy one or more desirable statistical properties in small samples. But as the sample size increases indefinitely, the estimators possess several desirable statistical properties. These properties are known as the **large sample,** or **asymptotic, properties.** In this appendix, we will discuss one large sample property, namely, the property of **consistency,** which is discussed more fully in **Appendix A.** For the two-variable model we have already shown that the OLS estimator $\hat{\beta}_2$ is an unbiased estimator of the true β_2. Now we show that $\hat{\beta}_2$ is also a consistent estimator of β_2. As shown in **Appendix A,** a sufficient condition for consistency is that $\hat{\beta}_2$ is unbiased and that its variance tends to zero as the sample size n tends to infinity.

Since we have already proved the unbiasedness property, we need only show that the variance of $\hat{\beta}_2$ tends to zero as n increases indefinitely. We know that

$$\text{var}\,(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2} = \frac{\sigma^2/n}{\sum x_i^2/n} \tag{26}$$

By dividing the numerator and denominator by n, we do not change the equality.

Now

$$\underbrace{\lim \text{var}\,(\hat{\beta}_2)}_{n \to \infty} = \underbrace{\lim \left(\frac{\sigma^2/n}{\sum x_i^2/n} \right)}_{n \to \infty} = 0 \tag{27}$$

where use is made of the facts that (1) the limit of a ratio quantity is the limit of the quantity in the numerator to the limit of the quantity in the denominator (refer to any calculus book); (2) as n tends to infinity, σ^2/n tends to zero because σ^2 is a finite number; and $[(\sum x_i^2)/n] \neq 0$ because the variance of X has a finite limit because of Assumption 8 of CLRM.

The upshot of the preceding discussion is that the OLS estimator $\hat{\beta}_2$ is a consistent estimator of true β_2. In like fashion, we can establish that $\hat{\beta}_1$ is also a consistent estimator. Thus, in repeated (small) samples, the OLS estimators are unbiased and as the sample size increases indefinitely the OLS estimators are consistent. As we shall see later, even if some of the assumptions of CLRM are not satisfied, we may be able to obtain consistent estimators of the regression coefficients in several situations.