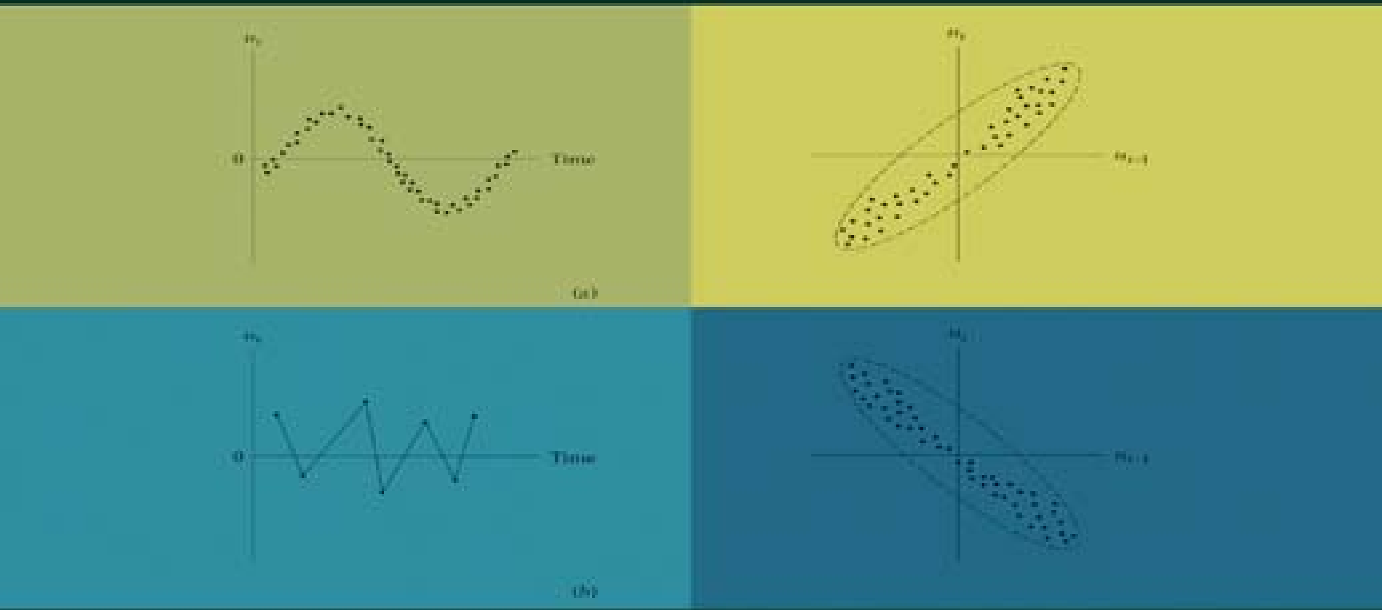
BASIC ECONOMETRICS



Fifth Edition

Damodar N. Gujarati
Dawn C. Porter

The McGraw-Hill Series Economics

ESSENTIALS OF ECONOMICS

Brue, McConnell, and Flynn
Essentials of Economics
Second Edition

Mandel
Economics: The Basics
First Edition

Schiller
Essentials of Economics
Seventh Edition

PRINCIPLES OF ECONOMICS

Colander
Economics, Microeconomics, and Macroeconomics
Seventh Edition

Frank and Bernanke
Principles of Economics, Principles of Microeconomics, Principles of Macroeconomics
Fourth Edition

Frank and Bernanke
Brief Editions: Principles of Economics, Principles of Microeconomics, Principles of Macroeconomics
First Edition

McConnell, Brue, and Flynn
Economics, Microeconomics, and Macroeconomics
Eighteenth Edition

McConnell, Brue, and Flynn
Brief Editions: Economics, Microeconomics, Macroeconomics
First Edition

Miller
Principles of Microeconomics
First Edition

Samuelson and Nordhaus
Economics, Microeconomics, and Macroeconomics
Eighteenth Edition

Schiller
The Economy Today, The Micro Economy Today, and The Macro Economy Today
Eleventh Edition

Slavin
Economics, Microeconomics, and Macroeconomics
Ninth Edition

ECONOMICS OF SOCIAL ISSUES

Guell
Issues in Economics Today
Fourth Edition

Sharp, Register, and Grimes
Economics of Social Issues
Eighteenth Edition

ECONOMETRICS

Gujarati and Porter
Basic Econometrics
Fifth Edition

Gujarati and Porter
Essentials of Econometrics
Fourth Edition

MANAGERIAL ECONOMICS

Baye
Managerial Economics and Business Strategy
Sixth Edition

Brickley, Smith, and Zimmerman
Managerial Economics and Organizational Architecture
Fifth Edition

Thomas and Maurice
Managerial Economics
Ninth Edition

INTERMEDIATE ECONOMICS

Bernheim and Whinston
Microeconomics
First Edition

Dornbusch, Fischer, and Startz
Macroeconomics
Tenth Edition

Frank
Microeconomics and Behavior
Seventh Edition

ADVANCED ECONOMICS

Romer
Advanced Macroeconomics
Third Edition

MONEY AND BANKING

Cecchetti
Money, Banking, and Financial Markets
Second Edition

URBAN ECONOMICS

O'Sullivan
Urban Economics
Seventh Edition

LABOR ECONOMICS

Borjas
Labor Economics
Fourth Edition

McConnell, Brue, and Macpherson
Contemporary Labor Economics
Eighth Edition

PUBLIC FINANCE

Rosen and Gayer
Public Finance
Eighth Edition

Seidman
Public Finance
First Edition

ENVIRONMENTAL ECONOMICS

Field and Field
Environmental Economics: An Introduction
Fifth Edition

INTERNATIONAL ECONOMICS

Appleyard, Field, and Cobb
International Economics
Sixth Edition

King and King
International Economics, Globalization, and Policy: A Reader
Fifth Edition

Pugel
International Economics
Fourteenth Edition

Basic Econometrics

Fifth Edition

Damodar N. Gujarati

Professor Emeritus of Economics,
United States Military Academy, West Point

Dawn C. Porter

University of Southern California

McGraw-Hill Irwin

Boston Burr Ridge, IL Dubuque, IA New York San Francisco St. Louis
Bangkok Bogotá Caracas Kuala Lumpur Lisbon London Madrid Mexico City
Milan Montreal New Delhi Santiago Seoul Singapore Sydney Taipei Toronto

The McGraw-Hill Companies

McGraw-Hill Irwin

BASIC ECONOMETRICS

Published by McGraw-Hill/Irwin, a business unit of The McGraw-Hill Companies, Inc., 1221 Avenue of the Americas, New York, NY, 10020.

This book is printed on acid-free paper.

1 2 3 4 5 6 7 8 9 0 VNH/VNH 0 9 8

ISBN 978-0-07-337577-9
MHID 0-07-337577-2

Publisher: *Douglas Reiner*
Developmental editor: *Anne E. Hilbert*
Editorial coordinator: *Noelle Fox*
Associate marketing manager: *Dean Karampelas*
Lead Project manager: *Christine A. Vaughan*
Full-service project manager: *Michael Ryder, ICC Macmillan Inc.*
Lead production supervisor: *Carol A. Bielski*
Design coordinator: *Joanne Mennemeier*
Media project manager: *Srikanth Potluri, Hurix Systems Pvt. Ltd.*
Cover design: *Brittany Skwierczynski*
Typeface: *10/12 TimesNewRomanPS*
Compositor: *ICC Macmillan Inc.*
Printer: *R. R. Donnelley*

Library of Congress Cataloging-in-Publication Data

Gujarati, Damodar N.
Basic econometrics / Damodar N. Gujarati, Dawn C. Porter. — 5th ed.
p. cm.
Includes bibliographical references and index.
ISBN-13: 978-0-07-337577-9 (alk. paper)
ISBN-10: 0-07-337577-2 (alk. paper)
1. Econometrics. I. Porter, Dawn C. II. Title.
HB139.G84 2009
330.01′5195—dc22
2008035934

www.mhhe.com

About the Authors

Damodar N. Gujarati

After teaching for more than 25 years at the City University of New York and 17 years in the Department of Social Sciences, U.S. Military Academy at West Point, New York, Dr. Gujarati is currently Professor Emeritus of economics at the Academy. Dr. Gujarati received his M.Com. degree from the University of Bombay in 1960, his M.B.A. degree from the University of Chicago in 1963, and his Ph.D. degree from the University of Chicago in 1965. Dr. Gujarati has published extensively in recognized national and international journals, such as the *Review of Economics and Statistics,* the *Economic Journal,* the *Journal of Financial and Quantitative Analysis,* and the *Journal of Business.* Dr. Gujarati was a member of the Board of Editors of the *Journal of Quantitative Economics,* the official journal of the Indian Econometric Society. Dr. Gujarati is also the author of *Pensions and the New York City Fiscal Crisis* (the American Enterprise Institute, 1978), *Government and Business* (McGraw-Hill, 1984), and *Essentials of Econometrics* (McGraw-Hill, 3d ed., 2006). Dr. Gujarati's books on econometrics have been translated into several languages.

Dr. Gujarati was a Visiting Professor at the University of Sheffield, U.K. (1970–1971), a Visiting Fulbright Professor to India (1981–1982), a Visiting Professor in the School of Management of the National University of Singapore (1985–1986), and a Visiting Professor of Econometrics, University of New South Wales, Australia (summer of 1988). Dr. Gujarati has lectured extensively on micro- and macroeconomic topics in countries such as Australia, China, Bangladesh, Germany, India, Israel, Mauritius, and the Republic of South Korea.

Dawn C. Porter

Dawn Porter has been an assistant professor in the Information and Operations Management Department at the Marshall School of Business of the University of Southern California since the fall of 2006. She currently teaches both introductory undergraduate and MBA statistics in the business school. Prior to joining the faculty at USC, from 2001–2006, Dawn was an assistant professor at the McDonough School of Business at Georgetown University, and before that was a visiting professor in the psychology department at the Graduate School of Arts and Sciences at NYU. At NYU she taught a number of advanced statistical methods courses and was also an instructor at the Stern School of Business. Her Ph.D. is from the Stern School in Statistics.

Dawn's areas of research interest include categorical analysis, agreement measures, multivariate modeling, and applications to the field of psychology. Her current research examines online auction models from a statistical perspective. She has presented her research at the Joint Statistical Meetings, the Decision Sciences Institute meetings, the International Conference on Information Systems, several universities including the London School of Economics and NYU, and various e-commerce and statistics seminar series. Dawn is also a co-author on *Essentials of Business Statistics,* 2nd edition, McGraw-Hill Irwin, 2008. Outside of academics, Dawn has been employed as a statistical consultant for KPMG, Inc. She has also worked as a statistical consultant for many other major companies, including Ginnie Mae, Inc., Toys R Us Corporation, IBM, Cosmaire, Inc., and New York University (NYU) Medical Center.

For Joan Gujarati, Diane Gujarati-Chesnut,
Charles Chesnut, and my grandchildren,
"Tommy" and Laura Chesnut.

—DNG

For Judy, Lee, Brett, Bryan, Amy, and Autumn Porter.
But especially for my adoring father, Terry.

—DCP

Brief Contents

Contents

CHAPTER 7
Multiple Regression Analysis: The Problem of Estimation 188

CHAPTER 8
Multiple Regression Analysis: The Problem of Inference 233

CHAPTER 12
Autocorrelation: What Happens If the Error Terms Are Correlated? 412

CHAPTER 13
Econometric Modeling: Model Specification and Diagnostic Testing 467

APPENDIX A
A Review of Some Statistical Concepts 801

APPENDIX B
Rudiments of Matrix Algebra 838

Preface

Objective of the Book

The first edition of *Basic Econometrics* was published thirty years ago. Over the years, there have been important developments in the theory and practice of econometrics. In each of the subsequent editions, I have tried to incorporate the major developments in the field. The fifth edition continues that tradition.

What has not changed, however, over all these years is my firm belief that econometrics can be taught to the beginner in an intuitive and informative way without resorting to matrix algebra, calculus, or statistics beyond the introductory level. Some subject material is inherently technical. In that case I have put the material in the appropriate appendix or refer the reader to the appropriate sources. Even then, I have tried to simplify the technical material so that the reader can get an intuitive understanding of this material.

I am pleasantly surprised not only by the longevity of this book but also by the fact that the book is widely used not only by students of economics and finance but also by students and researchers in the fields of politics, international relations, agriculture, and health sciences. All these students will find the new edition with its expanded topics and concrete applications very useful. In this edition I have paid even more attention to the relevance and timeliness of the real data used in the text. In fact, I have added about fifteen new illustrative examples and more than thirty new end-of-chapter exercises. Also, I have updated the data for about two dozen of the previous edition's examples and more than twenty exercises.

Although I am in the eighth decade of my life, I have not lost my love for econometrics, and I strive to keep up with the major developments in the field. To assist me in this endeavor, I am now happy to have Dr. Dawn Porter, Assistant Professor of Statistics at the Marshall School of Business at the University of Southern California in Los Angeles, as my co-author. Both of us have been deeply involved in bringing the fifth edition of *Basic Econometrics* to fruition.

Major Features of the Fifth Edition

Before discussing the specific changes in the various chapters, the following features of the new edition are worth noting:

1. Practically all of the data used in the illustrative examples have been updated.
2. Several new examples have been added.
3. In several chapters, we have included extended concluding examples that illustrate the various points made in the text.
4. Concrete computer printouts of several examples are included in the book. Most of these results are based on **EViews** (version 6) and **STATA** (version 10), as well as **MINITAB** (version 15).
5. Several new diagrams and graphs are included in various chapters.
6. Several new data-based exercises are included in the various chapters.
7. Small-sized data are included in the book, but large sample data are posted on the book's website, thereby minimizing the size of the text. The website will also publish all of the data used in the book and will be periodically updated.

8. In a few chapters, we have included class exercises in which students are encouraged to obtain their own data and implement the various techniques discussed in the book. Some Monte Carlo simulations are also included in the book.

Specific Changes to the Fifth Edition

Some chapter-specific changes are as follows:

1. The assumptions underlying the classical linear regression model (CLRM) introduced in Chapter 3 now make a careful distinction between fixed regressors (explanatory variables) and random regressors. We discuss the importance of the distinction.
2. The appendix to Chapter 6 discusses the properties of logarithms, the Box-Cox transformations, and various growth formulas.
3. Chapter 7 now discusses not only the marginal impact of a single regressor on the dependent variable but also the impacts of simultaneous changes of all the explanatory variables on the dependent variable. This chapter has also been reorganized in the same structure as the assumptions from Chapter 3.
4. A comparison of the various tests of heteroscedasticity is given in Chapter 11.
5. There is a new discussion of the impact of *structural breaks* on autocorrelation in Chapter 12.
6. New topics included in Chapter 13 are *missing data, non-normal error term,* and *stochastic,* or *random,* regressors.
7. A non-linear regression model discussed in Chapter 14 has a concrete application of the Box-Cox transformation.
8. Chapter 15 contains several new examples that illustrate the use of logit and probit models in various fields.
9. Chapter 16 on *panel data regression models* has been thoroughly revised and illustrated with several applications.
10. An extended discussion of Sims and Granger causality tests is now included in Chapter 17.
11. Stationary and non-stationary time series, as well as some of the problems associated with various tests of stationarity, are now thoroughly discussed in Chapter 21.
12. Chapter 22 includes a discussion on why taking the first differences of a time series for the purpose of making it stationary may not be the appropriate strategy in some situations.

Besides these specific changes, errors and misprints in the previous editions have been corrected and the discussions of several topics in the various chapters have been streamlined.

Organization and Options

The extensive coverage in this edition 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

A comprehensive website contains the following supplementary material:

–Data from the text, as well as additional large set data referenced in the book; the data will be periodically updated by the authors.

–A Solutions Manual, written by Dawn Porter, providing answers to all of the questions and problems throughout the text.

–A digital image library containing all of the graphs and figures from the text.

For more information, please go to www.mhhe.com/gujarati5e

Acknowledgments

Since the publication of the first edition of this book in 1978, we have received valuable advice, comments, criticism, and suggestions from a variety of people. In particular, we would like to acknowledge the help we have received from Michael McAleer of the University of Western Australia, Peter Kennedy of Simon Frazer University in Canada, Kenneth White, of the University of British Columbia, George K. Zestos, of Christopher Newport University, Virginia, and Paul Offner, of Georgetown University, Washington, D.C.

We are also grateful to several people who have influenced us by their scholarship. We especially want to thank Arthur Goldberger of the University of Wisconsin, William Greene of New York University, and the late G. S. Maddala. We continue to be grateful to the following reviewers who provided valuable insight, criticism, and suggestions for previous editions of this text: 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 Andrew Paizis of the City University of New York.

Several authors have influenced the writing of this text. In particular, we are 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.; and Gary Koop, University of Glasgow.

A number of very valuable comments and suggestions given by reviewers of the fourth edition have greatly improved this edition. We would like to thank the following:

Valerie Bencivenga
University of Texas–Austin

Andrew Economopoulos
Ursinus College

Eric Eide
Brigham Young University

Gary Ferrier
University of Arkansas–Fayetteville

David Garman
Tufts University

David Harris
Benedictine College

Don Holley
Boise State University

George Jakubson
Cornell University

Bruce Johnson
Centre College of Kentucky

Duke Kao
Syracuse University

Gary Krueger
Macalester College

Subal Kumbhakar
Binghamton University

Tae-Hwy Lee
University of California–Riverside

Solaiman Miah
West Virginia State University

Fabio Milani
University of California–Irvine

Helen Naughton
University of Oregon

Solomon Smith
Langston University

Kay Strong
Bowling Green State University

Derek Tittle
Georgia Institute of Technology

Tiemen Woutersen
Johns Hopkins University

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

For their behind-the-scenes help at McGraw-Hill, we are grateful to Douglas Reiner, Noelle Fox, and Anne Hilbert.

Finally, but not least important, Dr. Gujarati would like to thank his daughters, Joan and Diane, for their constant support and encouragement in the preparation of this and the previous editions.

Damodar N. Gujarati
Dawn C. Porter

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]

> 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]

[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.

[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.

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 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, and so on. 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]

[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.*

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

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

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

1. Statement of Theory or Hypothesis

Keynes stated:

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

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

2. Specification of the Mathematical Model of Consumption

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

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

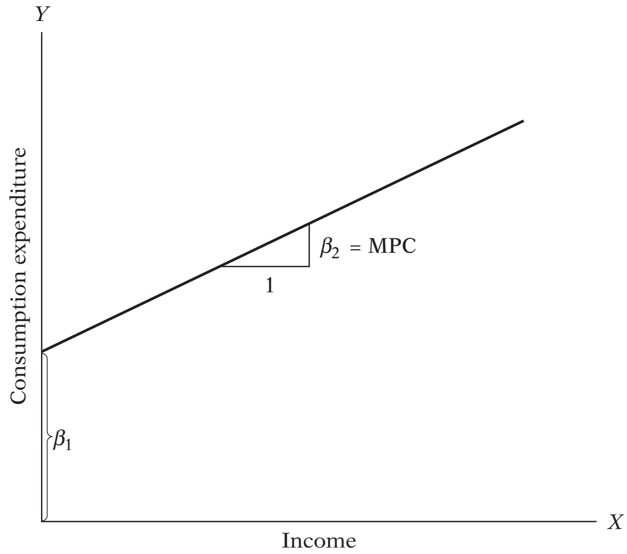
where $Y =$ consumption expenditure and $X =$ income, and where β_1 and β_2, known as the **parameters** of the model, are, respectively, the **intercept** and **slope** coefficients.

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

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

FIGURE I.1
Keynesian consumption function.



3. Specification of the Econometric Model of Consumption

The purely mathematical model of the consumption function given in Eq. (I.3.1) is of limited interest to the econometrician, for it assumes that there is an *exact* or *deterministic* relationship between consumption and income. But relationships between economic 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 in Eq. (I.3.1) as follows:

$$Y = \beta_1 + \beta_2 X + u \tag{I.3.2}$$

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.

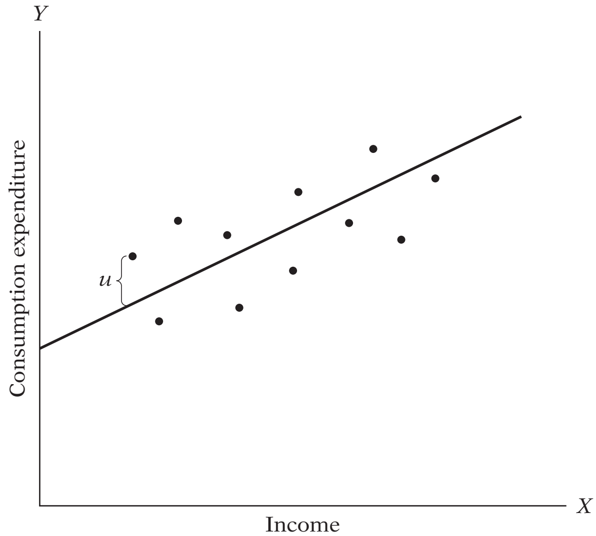


FIGURE I.2
Econometric model of the Keynesian consumption function.

4. Obtaining Data

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

5. Estimation of the Econometric Model

Now that we have the data, our next task is to estimate the parameters of the consumption function. The numerical estimates of the parameters give empirical content to the 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, -299.5913 and 0.7218. Thus, the estimated consumption function is:

$$\hat{Y}_t = -299.5913 + 0.7218X_t \qquad \text{(I.3.3)}$$

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.

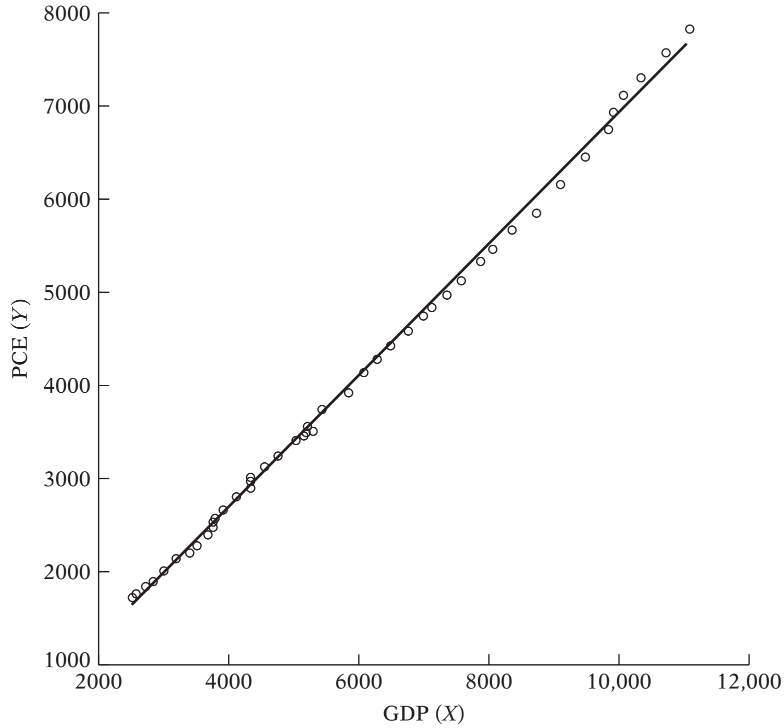
[11]As a matter of convention, a hat over a variable or parameter indicates that it is an estimated value.

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

Source: *Economic Report of the President,* 2007, Table B–2, p. 230.

Year	PCE(Y)	GDP(X)
1960	1597.4	2501.8
1961	1630.3	2560.0
1962	1711.1	2715.2
1963	1781.6	2834.0
1964	1888.4	2998.6
1965	2007.7	3191.1
1966	2121.8	3399.1
1967	2185.0	3484.6
1968	2310.5	3652.7
1969	2396.4	3765.4
1970	2451.9	3771.9
1971	2545.5	3898.6
1972	2701.3	4105.0
1973	2833.8	4341.5
1974	2812.3	4319.6
1975	2876.9	4311.2
1976	3035.5	4540.9
1977	3164.1	4750.5
1978	3303.1	5015.0
1979	3383.4	5173.4
1980	3374.1	5161.7
1981	3422.2	5291.7
1982	3470.3	5189.3
1983	3668.6	5423.8
1984	3863.3	5813.6
1985	4064.0	6053.7
1986	4228.9	6263.6
1987	4369.8	6475.1
1988	4546.9	6742.7
1989	4675.0	6981.4
1990	4770.3	7112.5
1991	4778.4	7100.5
1992	4934.8	7336.6
1993	5099.8	7532.7
1994	5290.7	7835.5
1995	5433.5	8031.7
1996	5619.4	8328.9
1997	5831.8	8703.5
1998	6125.8	9066.9
1999	6438.6	9470.3
2000	6739.4	9817.0
2001	6910.4	9890.7
2002	7099.3	10048.8
2003	7295.3	10301.0
2004	7577.1	10703.5
2005	7841.2	11048.6

FIGURE I.3
Personal consumption expenditure (Y) in relation to GDP (X), 1960–2005, in billions of 2000 dollars.



As Figure I.3 shows, the regression line fits the data quite well in that the data points are very close to the regression line. From this figure we see that for the period 1960–2005 the slope coefficient (i.e., the **MPC**) was about 0.72, suggesting that for the sample period an increase in real income of one dollar led, *on average,* to an increase of about 72 cents in real consumption expenditure.[12] We say *on average* because the relationship between 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 72 cents for a dollar's increase in real income.

6. Hypothesis Testing

Assuming that the fitted model is a reasonably good approximation of reality, we have to develop suitable criteria to find out whether the estimates obtained in, say, Equation I.3.3 are in accord with the expectations of the theory that is being tested. According to "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 0.72. But before we accept this finding as confirmation of Keynesian consumption theory, we must enquire whether this estimate is sufficiently

[12]Do not worry now about how these values were obtained. As we show in Chapter 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.

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 0.72 *statistically less than 1?* If it is, it may support Keynes's theory.

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

7. Forecasting or Prediction

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

To illustrate, suppose we want to predict the mean consumption expenditure for 2006. The GDP value for 2006 was 11319.4 billion dollars.[14] Putting this GDP figure on the right-hand side of Eq. (I.3.3), we obtain:

$$\begin{aligned}\hat{Y}_{2006} &= -299.5913 + 0.7218\,(11319.4)\\ &= 7870.7516\end{aligned} \tag{I.3.4}$$

or about 7870 billion dollars. Thus, given the value of the GDP, the mean, or average, forecast consumption expenditure is about 7870 billion dollars. The actual value of the consumption expenditure reported in 2006 was 8044 billion dollars. The estimated model Eq. (I.3.3) thus **underpredicted** the actual consumption expenditure by about 174 billion dollars. We could say the **forecast error** is about 174 billion dollars, which is about 1.5 percent of the actual GDP value for 2006. When we fully discuss the linear regression model in subsequent chapters, we will try to find out if such an error is "small" or "large." But what is important for now is to note that such forecast errors are inevitable given the statistical nature of our analysis.

There is another use of the estimated model Eq. (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 0.72 obtained in Eq. (I.3.3), this multiplier becomes about $M = 3.57$. 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 Eq. (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.

[14] Data on PCE and GDP were available for 2006 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.

8. Use of the Model for Control or Policy Purposes

Suppose we have the estimated consumption function given in Eq. (I.3.3). Suppose further the government believes that consumer expenditure of about 8750 (billions of 2000 dollars) will keep the unemployment rate at its current level of about 4.2 percent (early 2006). What level of income will guarantee the target amount of consumption expenditure?

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

$$8750 = -299.5913 + 0.7218(GDP_{2006}) \tag{I.3.6}$$

which gives $X = 12537$, approximately. That is, an income level of about 12537 (billion) dollars, given an MPC of about 0.72, will produce an expenditure of about 8750 billion dollars.

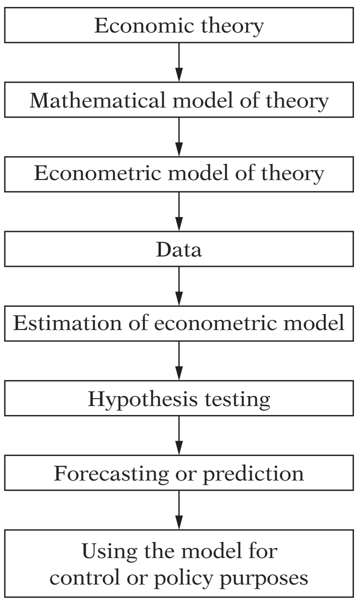
As these calculations suggest, an estimated model may be used for control, or policy, purposes. By appropriate fiscal and monetary policy mix, the government can manipulate the **control variable *X*** to produce the desired level of the **target variable *Y*.**

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 another consumption model (theory) might equally fit the data as well? For example, Milton Friedman has developed a model of consumption, called the *permanent income*

FIGURE I.4
Anatomy of econometric modeling.



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 [a] 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?
>
> (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. 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.

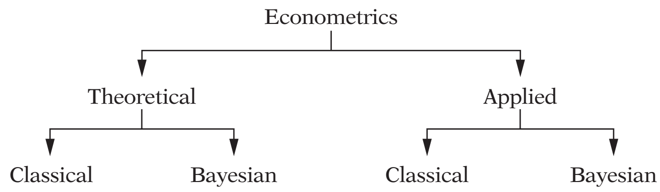
[15]Milton Friedman, *A Theory of Consumption Function,* Princeton University Press, Princeton, N.J., 1957.

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



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, and 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 Publishing 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. This text presents 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,* Basil Blackwell Ltd., Oxford University Press, England, 1985; Peter M. Lee, *Bayesian Statistics: An Introduction,* Oxford University Press, England, 1989; and 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. Another advanced reference book is the *Palgrave Handbook of Econometrics*: Volume 1: *Econometric Theory,* edited by Terence C. Mills and Kerry Patterson, Palgrave Macmillan, New York, 2007.

Part 1

Single-Equation Regression Models

Part 1 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 1 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.

Chapter 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."

1.2 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.

[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.

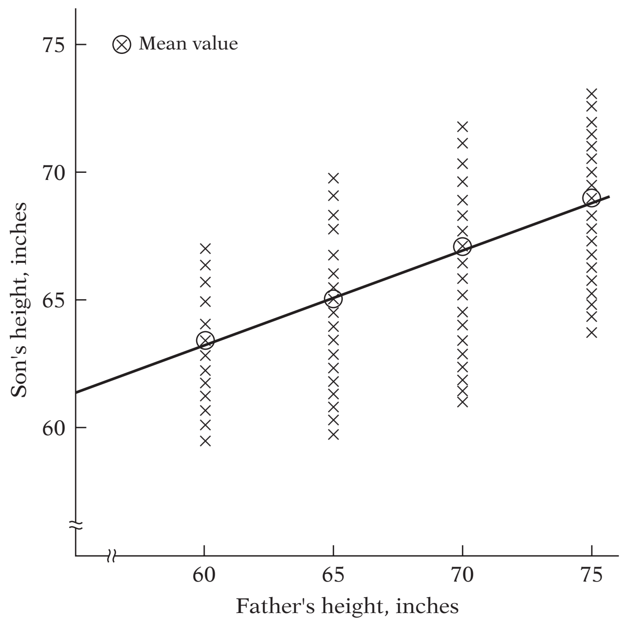
The full import of this view of regression analysis will become clearer as we progress, but a few simple examples will make the basic concept quite clear.

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 **scattergram.** 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

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



[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?)

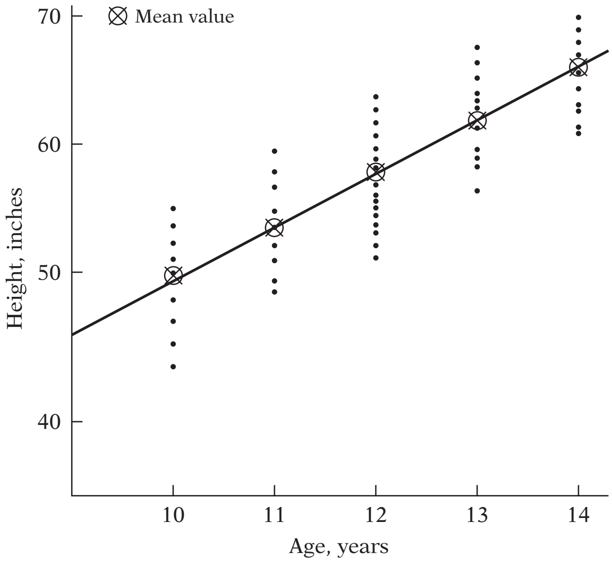


FIGURE 1.2 Hypothetical distribution of heights corresponding to selected ages.

certain age), which can be seen clearly if we draw a line (the regression line) through the 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 aftertax 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 1.3).

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.

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. The slope of this line represents the change in k given a change in the inflation rate. 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.

FIGURE 1.3
Hypothetical Phillips curve.

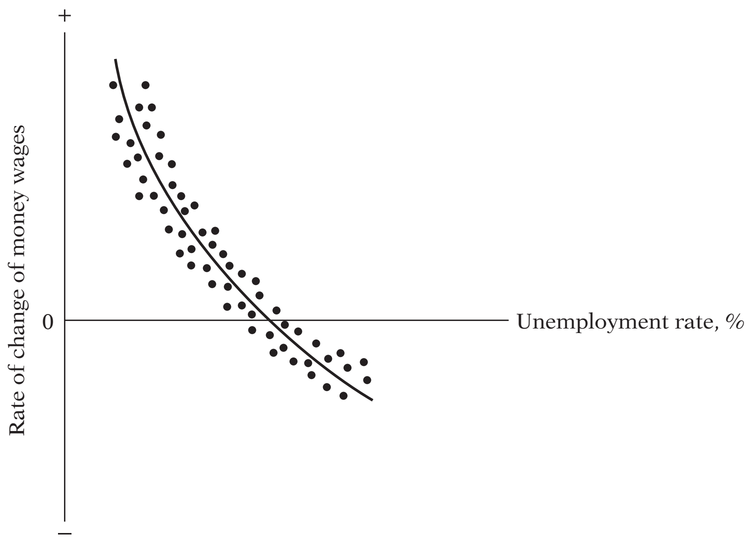
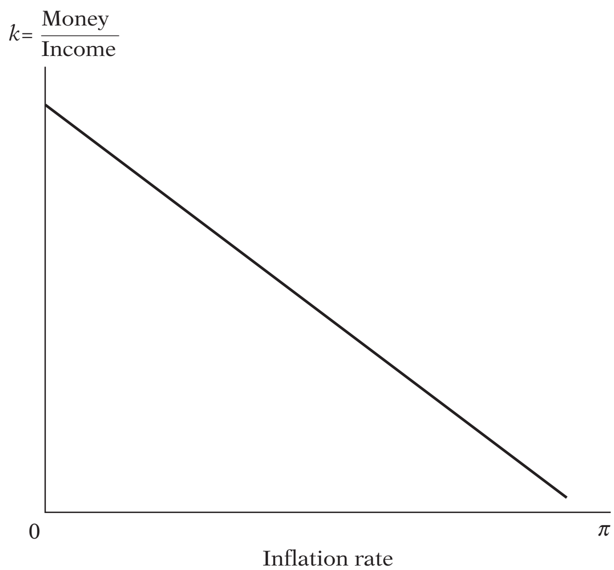


FIGURE 1.4
Money holding in relation to the inflation rate π.



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

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, $F = k(m_1m_2/r^2)$, 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 suggestive, can never establish causal connection: our ideas of causation must come from outside statistics, ultimately from some theory or other."[5]

[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.

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

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 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]

[6]But as we shall see in Chapter 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 Chapter 3, Sec. 3.2.

[8]In advanced treatment of econometrics, one can relax the assumption that the explanatory variables are nonstochastic (see introduction to Part 2).

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 one explanatory variable, as in the crop-yield, rainfall, temperature, sunshine, and fertilizer example, 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.

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

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 1.1 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., government budgets), **quinquennially,** that is, every 5 years (e.g., the census of manufactures), or **decennially,** that is, every 10 years (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 September, 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.

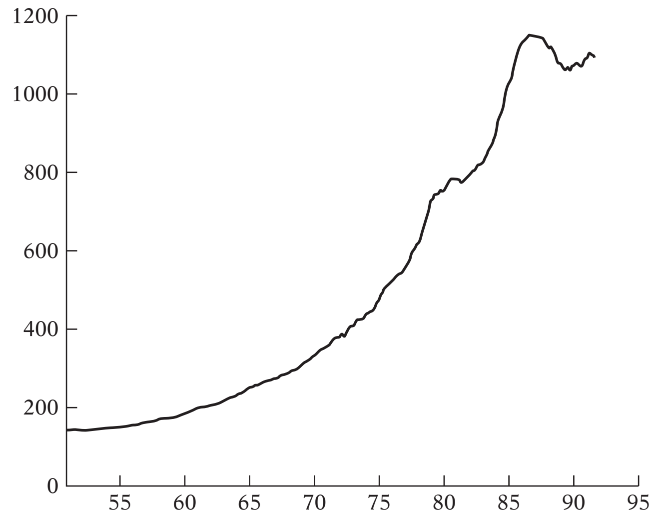
Cross-Section Data

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

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

[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.

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



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 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 1980–2005 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 182 observations—26 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 provide very useful information on the dynamics of household behavior, as we shall see in Chapter 16.

FIGURE 1.6
Relationship between eggs produced and prices, 1990.

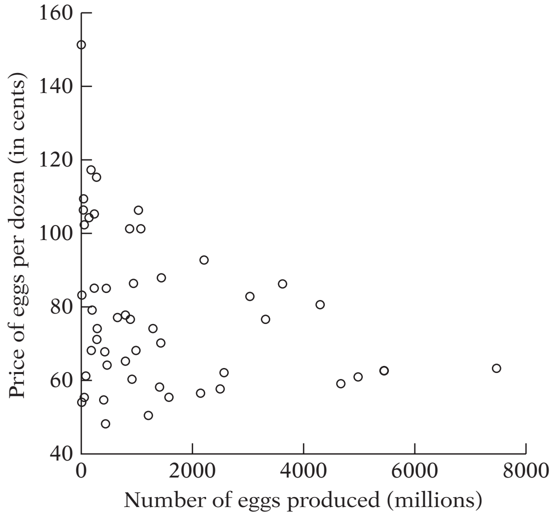


TABLE 1.1 **U.S. Egg Production**

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.

As a concrete example, consider the data given in Table 1.2. The data in the table, originally collected by Y. Grunfeld, refer to the real investment, the real value of the firm, and the real capital stock of four U.S. companies, namely, General Electric (GM), U.S. Steel (US), General Motors (GM), and Westinghouse (WEST), for the period 1935–1954.[12] Since the data are for several companies collected over a number of years, this is a classic example of panel data. In this table, the number of observations for each company is the same, but this is not always the case. If all the companies have the same number of observations, we have what is called a **balanced panel.** If the number of observations is not the same for each company, it is called an **unbalanced panel.** In Chapter 16, Panel Data Regression Models, we will examine such data and show how to estimate such models.

Grunfeld's purpose in collecting these data was to find out how real gross investment (I) depends on the real value of the firm (F) a year earlier and real capital stock (C) a year earlier. Since the companies included in the sample operate in the same capital market, by studying them together, Grunfeld wanted to find out if they had similar investment functions.

The Sources of Data[13]

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 websites 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 websites 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.[14] 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 around?

[12] Y. Grunfeld, "The Determinants of Corporate Investment," unpublished PhD thesis, Department of Economics, University of Chicago, 1958. These data have become a workhorse for illustrating panel data regression models.

[13] 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.

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

TABLE 1.2 Investment Data for Four Companies, 1935–1954

Observation	I	F_{-1}	C_{-1}	Observation	I	F_{-1}	C_{-1}
GE				US			
1935	33.1	1170.6	97.8	1935	209.9	1362.4	53.8
1936	45.0	2015.8	104.4	1936	355.3	1807.1	50.5
1937	77.2	2803.3	118.0	1937	469.9	2673.3	118.1
1938	44.6	2039.7	156.2	1938	262.3	1801.9	260.2
1939	48.1	2256.2	172.6	1939	230.4	1957.3	312.7
1940	74.4	2132.2	186.6	1940	361.6	2202.9	254.2
1941	113.0	1834.1	220.9	1941	472.8	2380.5	261.4
1942	91.9	1588.0	287.8	1942	445.6	2168.6	298.7
1943	61.3	1749.4	319.9	1943	361.6	1985.1	301.8
1944	56.8	1687.2	321.3	1944	288.2	1813.9	279.1
1945	93.6	2007.7	319.6	1945	258.7	1850.2	213.8
1946	159.9	2208.3	346.0	1946	420.3	2067.7	232.6
1947	147.2	1656.7	456.4	1947	420.5	1796.7	264.8
1948	146.3	1604.4	543.4	1948	494.5	1625.8	306.9
1949	98.3	1431.8	618.3	1949	405.1	1667.0	351.1
1950	93.5	1610.5	647.4	1950	418.8	1677.4	357.8
1951	135.2	1819.4	671.3	1951	588.2	2289.5	341.1
1952	157.3	2079.7	726.1	1952	645.2	2159.4	444.2
1953	179.5	2371.6	800.3	1953	641.0	2031.3	623.6
1954	189.6	2759.9	888.9	1954	459.3	2115.5	669.7
GM				WEST			
1935	317.6	3078.5	2.8	1935	12.93	191.5	1.8
1936	391.8	4661.7	52.6	1936	25.90	516.0	0.8
1937	410.6	5387.1	156.9	1937	35.05	729.0	7.4
1938	257.7	2792.2	209.2	1938	22.89	560.4	18.1
1939	330.8	4313.2	203.4	1939	18.84	519.9	23.5
1940	461.2	4643.9	207.2	1940	28.57	628.5	26.5
1941	512.0	4551.2	255.2	1941	48.51	537.1	36.2
1942	448.0	3244.1	303.7	1942	43.34	561.2	60.8
1943	499.6	4053.7	264.1	1943	37.02	617.2	84.4
1944	547.5	4379.3	201.6	1944	37.81	626.7	91.2
1945	561.2	4840.9	265.0	1945	39.27	737.2	92.4
1946	688.1	4900.0	402.2	1946	53.46	760.5	86.0
1947	568.9	3526.5	761.5	1947	55.56	581.4	111.1
1948	529.2	3245.7	922.4	1948	49.56	662.3	130.6
1949	555.1	3700.2	1020.1	1949	32.04	583.8	141.8
1950	642.9	3755.6	1099.0	1950	32.24	635.2	136.7
1951	755.9	4833.0	1207.7	1951	54.38	732.8	129.7
1952	891.2	4924.9	1430.5	1952	71.78	864.1	145.5
1953	1304.4	6241.7	1777.3	1953	90.08	1193.5	174.8
1954	1486.7	5593.6	2226.3	1954	68.60	1188.9	213.5

Notes: $Y = I$ = gross investment = additions to plant and equipment plus maintenance and repairs, in millions of dollars deflated by P_1.
$X_2 = F$ = value of the firm = price of common and preferred shares at Dec. 31 (or average price of Dec. 31 and Jan. 31 of the following year) times number of common and preferred shares outstanding plus total book value of debt at Dec. 31, in millions of dollars deflated by P_2.
$X_3 = C$ = stock of plant and equipment = accumulated sum of net additions to plant and equipment deflated by P_1 minus depreciation allowance deflated by P_3 in these definitions.
P_1 = implicit price deflator of producers' durable equipment (1947 = 100).
P_2 = implicit price deflator of GNP (1947 = 100).
P_3 = depreciation expense deflator = 10-year moving average of wholesale price index of metals and metal products (1947 = 100).

Source: Reproduced from H. D. Vinod and Aman Ullah, *Recent Advances in Regression Methods,* Marcel Dekker, New York, 1981, pp. 259–261.

The Accuracy of Data[15]

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.

1. As noted, most social science data are nonexperimental in nature. Therefore, there is the possibility of observational errors, either of omission or commission.
2. Even in experimentally collected data, errors of measurement arise from approximations and roundoffs.
3. In questionnaire-type surveys, the problem of nonresponse can be serious; a researcher is lucky to get a 40 percent response rate to a questionnaire. Analysis based on such a partial response rate 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 do respond to the questionnaire may not answer all the questions, especially questions of a financially sensitive nature, thus leading to additional selectivity bias.
4. 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.
5. 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 individuals or microunits that may be the ultimate object of study.
6. 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 so except at a very highly aggregate level. 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 of 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[16]

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.

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

[16]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.

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 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 this year's GDP is compared with the previous year's GDP. Personal income, measured in dollars, is a ratio variable; someone earning $100,000 is making twice as much as another person earning $50,000 (before taxes are assessed, of course!).

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). At 11:00 a.m. PST on August 11, 2007, Portland, Oregon, reported a temperature of 60 degrees Fahrenheit while Tallahassee, Florida, reached 90 degrees. Temperature is not measured on a ratio scale since it does not make sense to claim that Tallahassee was 50 percent warmer than Portland. This is mainly due to the fact that the Fahrenheit scale does not use 0 degrees as a natural base.

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 indicates a 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.

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 have been properly gathered and that the computations and analysis are correct.

EXERCISES

1.1. Table 1.3 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.[17]

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?

1.2. *a.* Using Table 1.3, 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?

TABLE 1.3
CPI in Seven Industrial Countries, 1980–2005 (1982–1984 = 100)

Source: *Economic Report of the President,* 2007, Table 108, p. 354.

Year	U.S.	Canada	Japan	France	Germany	Italy	U.K.
1980	82.4	76.1	91.0	72.2	86.7	63.9	78.5
1981	90.9	85.6	95.3	81.8	92.2	75.5	87.9
1982	96.5	94.9	98.1	91.7	97.0	87.8	95.4
1983	99.6	100.4	99.8	100.3	100.3	100.8	99.8
1984	103.9	104.7	102.1	108.0	102.7	111.4	104.8
1985	107.6	109.0	104.2	114.3	104.8	121.7	111.1
1986	109.6	113.5	104.9	117.2	104.6	128.9	114.9
1987	113.6	118.4	104.9	121.1	104.9	135.1	119.7
1988	118.3	123.2	105.6	124.3	106.3	141.9	125.6
1989	124.0	129.3	108.0	128.7	109.2	150.7	135.4
1990	130.7	135.5	111.4	132.9	112.2	160.4	148.2
1991	136.2	143.1	115.0	137.2	116.3	170.5	156.9
1992	140.3	145.3	117.0	140.4	122.2	179.5	162.7
1993	144.5	147.9	118.5	143.4	127.6	187.7	165.3
1994	148.2	148.2	119.3	145.8	131.1	195.3	169.3
1995	152.4	151.4	119.2	148.4	133.3	205.6	175.2
1996	156.9	153.8	119.3	151.4	135.3	213.8	179.4
1997	160.5	156.3	121.5	153.2	137.8	218.2	185.1
1998	163.0	157.8	122.2	154.2	139.1	222.5	191.4
1999	166.6	160.5	121.8	155.0	140.0	226.2	194.3
2000	172.2	164.9	121.0	157.6	142.0	231.9	200.1
2001	177.1	169.1	120.1	160.2	144.8	238.3	203.6
2002	179.9	172.9	119.0	163.3	146.7	244.3	207.0
2003	184.0	177.7	118.7	166.7	148.3	250.8	213.0
2004	188.9	181.0	118.7	170.3	150.8	256.3	219.4
2005	195.3	184.9	118.3	173.2	153.7	261.3	225.6

[17]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 1981 is $[(85.6 - 76.1)/76.1] \times 100 = 12.48\%$ (approx.).

1.3. Table 1.4 gives the foreign exchange rates for nine industrialized countries for the years 1985–2006. 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 1985–2006, 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 1.5. Can you give reasons why the money supply has been increasing over the time period shown in the table?

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.[18]

TABLE 1.4 **Exchange Rates for Nine Countries: 1985–2006**

Year	Australia	Canada	China P. R.	Japan	Mexico	South Korea	Sweden	Switzerland	United Kingdom
1985	0.7003	1.3659	2.9434	238.47	0.257	872.45	8.6032	2.4552	1.2974
1986	0.6709	1.3896	3.4616	168.35	0.612	884.60	7.1273	1.7979	1.4677
1987	0.7014	1.3259	3.7314	144.60	1.378	826.16	6.3469	1.4918	1.6398
1988	0.7841	1.2306	3.7314	128.17	2.273	734.52	6.1370	1.4643	1.7813
1989	0.7919	1.1842	3.7673	138.07	2.461	674.13	6.4559	1.6369	1.6382
1990	0.7807	1.1668	4.7921	145.00	2.813	710.64	5.9231	1.3901	1.7841
1991	0.7787	1.1460	5.3337	134.59	3.018	736.73	6.0521	1.4356	1.7674
1992	0.7352	1.2085	5.5206	126.78	3.095	784.66	5.8258	1.4064	1.7663
1993	0.6799	1.2902	5.7795	111.08	3.116	805.75	7.7956	1.4781	1.5016
1994	0.7316	1.3664	8.6397	102.18	3.385	806.93	7.7161	1.3667	1.5319
1995	0.7407	1.3725	8.3700	93.96	6.447	772.69	7.1406	1.1812	1.5785
1996	0.7828	1.3638	8.3389	108.78	7.600	805.00	6.7082	1.2361	1.5607
1997	0.7437	1.3849	8.3193	121.06	7.918	953.19	7.6446	1.4514	1.6376
1998	0.6291	1.4836	8.3008	130.99	9.152	1,400.40	7.9522	1.4506	1.6573
1999	0.6454	1.4858	8.2783	113.73	9.553	1,189.84	8.2740	1.5045	1.6172
2000	0.5815	1.4855	8.2784	107.80	9.459	1,130.90	9.1735	1.6904	1.5156
2001	0.5169	1.5487	8.2770	121.57	9.337	1,292.02	10.3425	1.6891	1.4396
2002	0.5437	1.5704	8.2771	125.22	9.663	1,250.31	9.7233	1.5567	1.5025
2003	0.6524	1.4008	8.2772	115.94	10.793	1,192.08	8.0787	1.3450	1.6347
2004	0.7365	1.3017	8.2768	108.15	11.290	1,145.24	7.3480	1.2428	1.8330
2005	0.7627	1.2115	8.1936	110.11	10.894	1,023.75	7.4710	1.2459	1.8204
2006	0.7535	1.1340	7.9723	116.31	10.906	954.32	7.3718	1.2532	1.8434

Source: *Economic Report of the President,* 2007, Table B–110, p. 356.

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

TABLE 1.5
Seasonally Adjusted M1 Supply: 1959:01–1999:07 (billions of dollars)

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

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
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

(*Continued*)

TABLE 1.5
(Continued)

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			

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 worth of Social Security benefit for every $3 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.

1.7. The data presented in Table 1.6 were published in the March 1, 1984, issue of *The Wall Street Journal*. They relate 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 1.6 in subsequent chapters.

TABLE 1.6
Impact of Advertising Expenditure

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

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

Chapter 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 three 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 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

[1]The reader whose statistical knowledge has become somewhat rusty may want to freshen it up by reading the statistical appendix, **Appendix 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

that, despite the variability of weekly consumption 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,

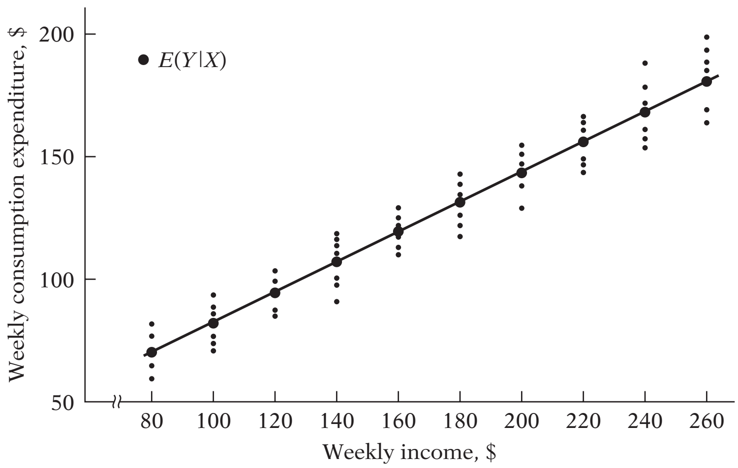


FIGURE 2.1
Conditional distribution of expenditure for various levels of income (data of Table 2.1).

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

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

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, 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.

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.

[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).

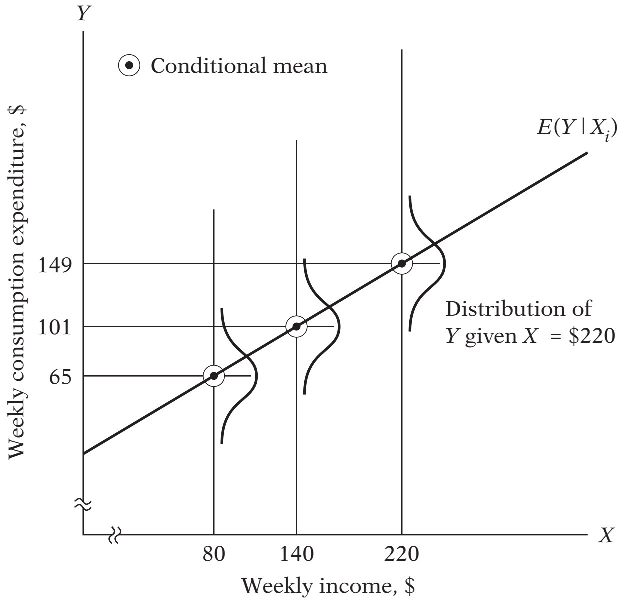


FIGURE 2.2
Population regression line (data of Table 2.1).

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 Equation 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 Eq. (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, Eq. (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.

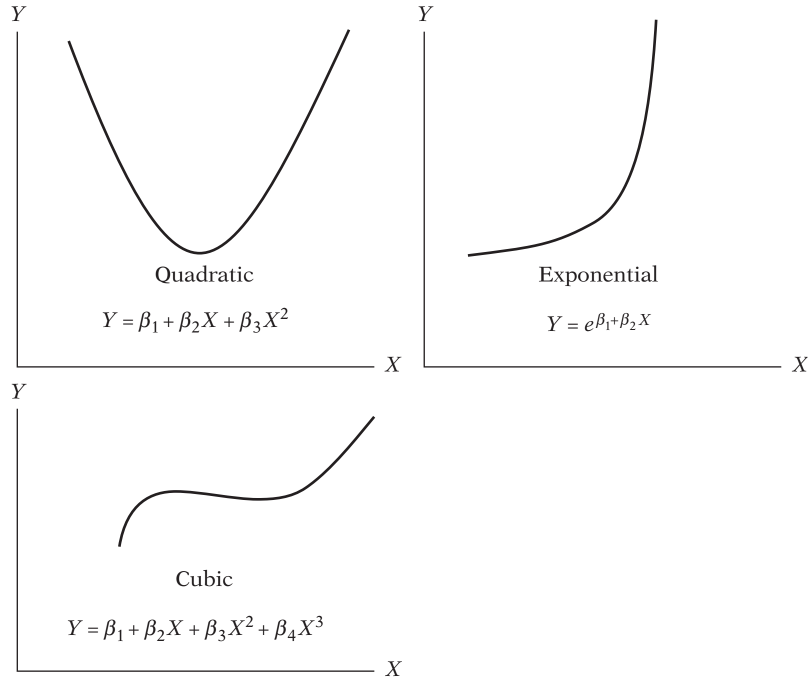


TABLE 2.3
Linear Regression Models

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 Equation 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 Eq. (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 Eq. (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 Equation 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

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

[8]See **Appendix 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.

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 Eq. (2.2.2) and Eq. (2.4.2) are equivalent forms if $E(u_i \mid X_i) = 0$.[9] But the stochastic specification in Eq. (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 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

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

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

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 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, $Y_i = \beta_1 + \beta_2 X_i + u_i$ 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.

[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.

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 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

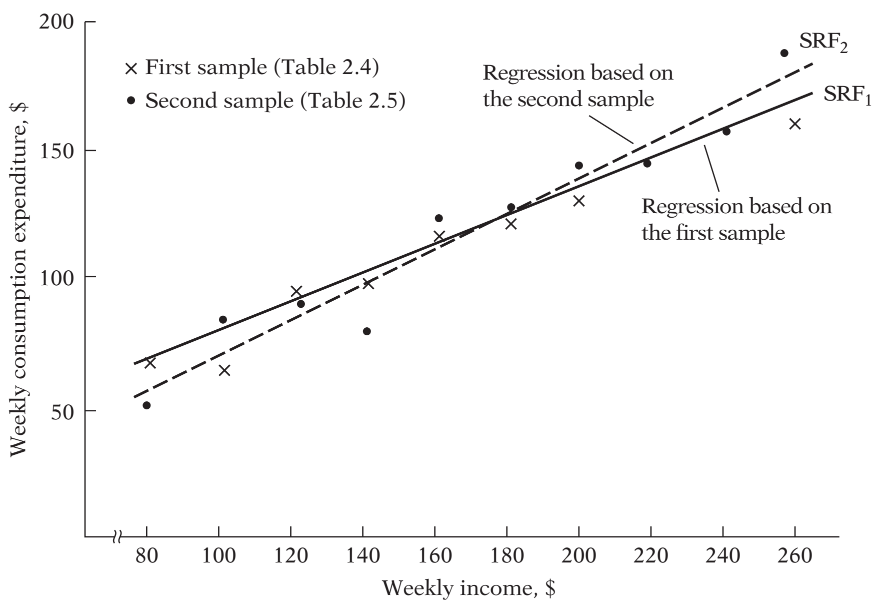
TABLE 2.4
A Random Sample from the Population of Table 2.1

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.



as the **sample regression lines.** Supposedly 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 Eq. (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] It should be noted that an estimator is random, but an estimate is nonrandom. (Why?)

Now just as we expressed the PRF in two equivalent forms, Eq. (2.2.2) and Eq. (2.4.2), we can express the SRF in Equation 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

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

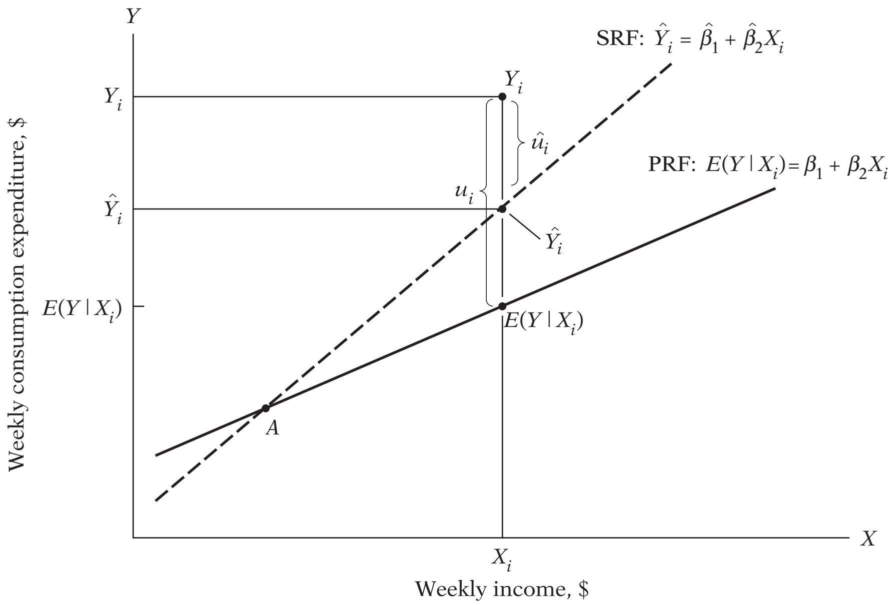
on the basis of the SRF

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

because more often than not our analysis is based upon a single sample from some population. But because of sampling fluctuations, our estimate of the PRF based on the SRF is at best an approximate one. This approximation is shown diagrammatically in Figure 2.5.

[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.



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 Illustrative Examples

We conclude this chapter with two examples.

EXAMPLE 2.1
Mean Hourly Wage by Education

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 population survey conducted in May 1985.[14]

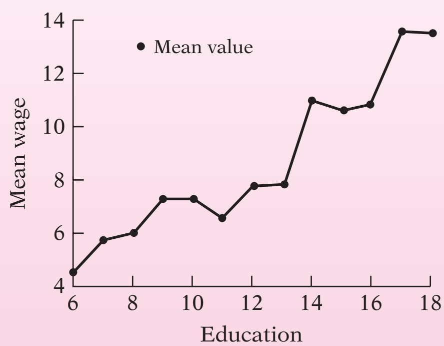
Plotting the (conditional) mean wage against education, we obtain the picture in Figure 2.6. The regression curve in the figure shows how mean wages vary with the level of education; they generally increase with the level of education, a finding one should not find surprising. We will study in a later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

TABLE 2.6
Mean Hourly Wage by Education

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

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

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.

EXAMPLE 2.2

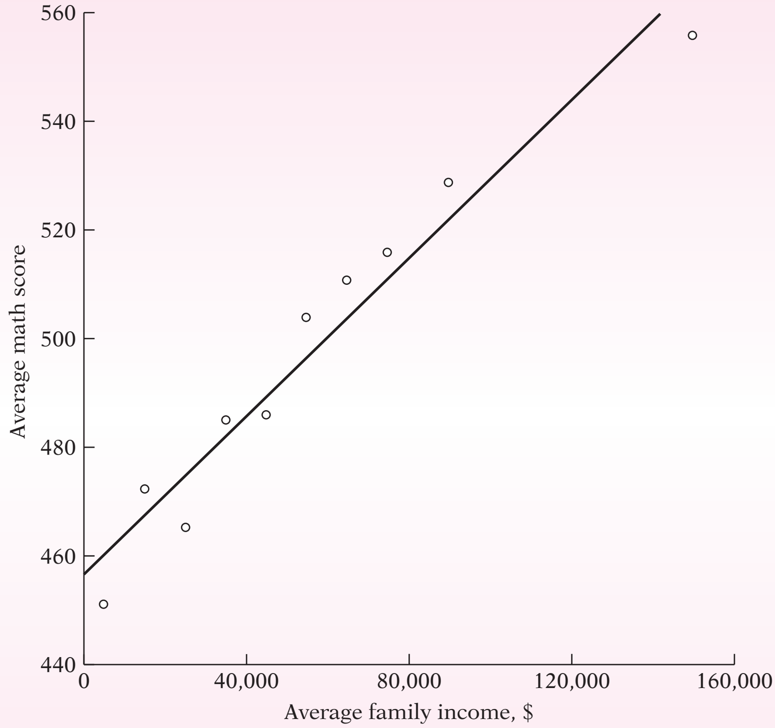
Mathematics SAT Scores by Family Income

Table 2.10 in Exercise 2.17 provides data on mean SAT (Scholastic Aptitude Test) scores on critical reading, mathematics, and writing for college-bound seniors based on 947,347 students taking the SAT examination in 2007. Plotting the mean mathematics scores on mean family income, we obtain the picture in Figure 2.7.

Note: Because of the open-ended income brackets for the first and last income categories shown in Table 2.10, the lowest average family income is assumed to be $5,000 and the highest average family income is assumed to be $150,000.

FIGURE 2.7

Relationship between mean mathematics SAT scores and mean family income.



As Figure 2.7 shows, the average mathematics score increases as average family income increases. Since the number of students taking the SAT examination is quite large, it probably represents the entire population of seniors taking the examination. Therefore, the regression line sketched in Figure 2.7 probably represents the population regression line.

There may be several reasons for the observed positive relationship between the two variables. For example, one might argue that students with higher family income can better afford private tutoring for the SAT examinations. In addition, students with higher family income are more likely to have parents who are highly educated. It is also possible that students with higher mathematics scores come from better schools. The reader can provide other explanations for the observed positive relationship between the two variables.

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

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 2.7d 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 2.8 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?

2.11. From the scattergram given in Figure 2.9, 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 2.10 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?

FIGURE 2.8
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.

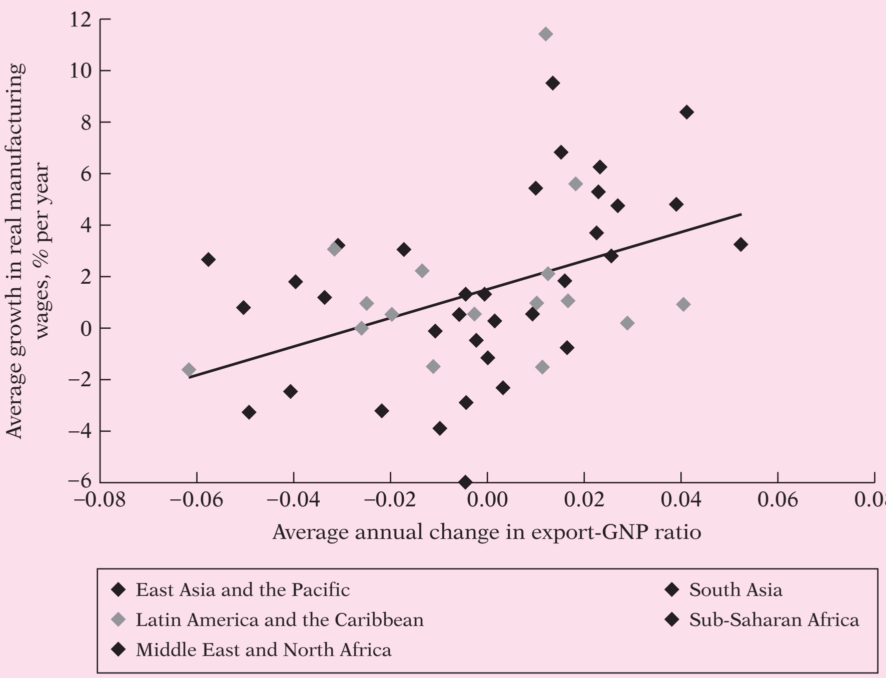


FIGURE 2.9
Skill intensity of exports and human capital endowment. Data are for 126 industrial and developing countries in 1985. Values along the horizontal axis are logarithms of the ratio of the country's average educational attainment to its land area; vertical axis values are logarithms of the ratio 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 database; education data from UNDP 1990; land data from the World Bank.

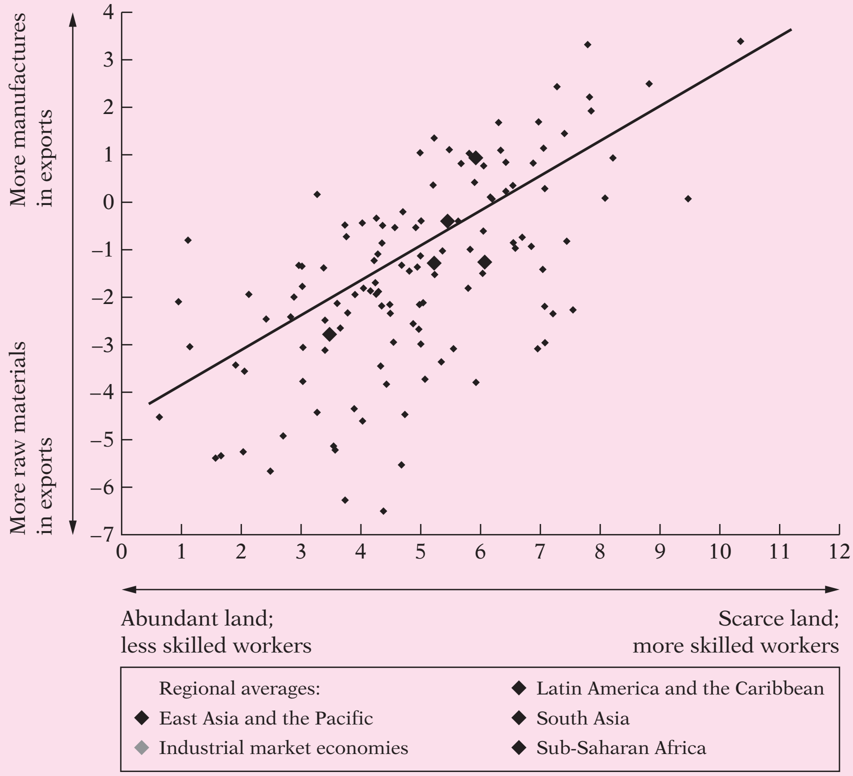
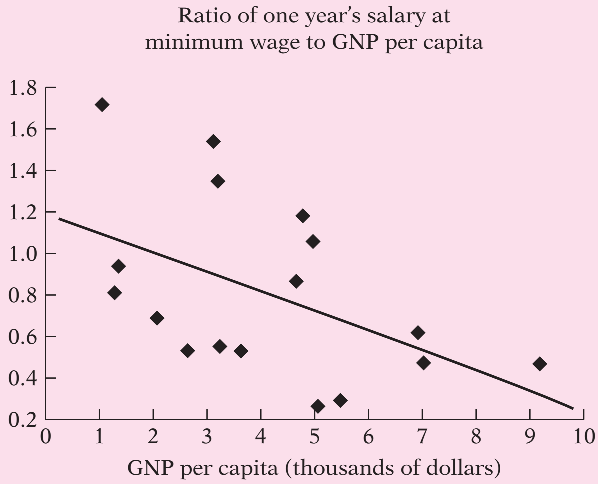


FIGURE 2.10
The minimum wage and GNP per capita. The sample consists of 17 developing countries. Years vary by country from 1988 to 1992. Data are in international prices.

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



Empirical Exercises

2.14. You are given the data in Table 2.7 for the United States for years 1980–2006.

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?

TABLE 2.7
Labor Force Participation Data for U.S. for 1980–2006

Source: *Economic Report of the President, 2007.*

Year	CLFPRM[1]	CLFPRF[2]	UNRM[3]	UNRF[4]	AHE82[5]	AHE[6]
1980	77.40000	51.50000	6.900000	7.400000	7.990000	6.840000
1981	77.00000	52.10000	7.400000	7.900000	7.880000	7.430000
1982	76.60000	52.60000	9.900000	9.400000	7.860000	7.860000
1983	76.40000	52.90000	9.900000	9.200000	7.950000	8.190000
1984	76.40000	53.60000	7.400000	7.600000	7.950000	8.480000
1985	76.30000	54.50000	7.000000	7.400000	7.910000	8.730000
1986	76.30000	55.30000	6.900000	7.100000	7.960000	8.920000
1987	76.20000	56.00000	6.200000	6.200000	7.860000	9.130000
1988	76.20000	56.60000	5.500000	5.600000	7.810000	9.430000
1989	76.40000	57.40000	5.200000	5.400000	7.750000	9.800000
1990	76.40000	57.50000	5.700000	5.500000	7.660000	10.190000
1991	75.80000	57.40000	7.200000	6.400000	7.580000	10.500000
1992	75.80000	57.80000	7.900000	7.000000	7.550000	10.760000
1993	75.40000	57.90000	7.200000	6.600000	7.520000	11.030000
1994	75.10000	58.80000	6.200000	6.000000	7.530000	11.320000
1995	75.00000	58.90000	5.600000	5.600000	7.530000	11.640000
1996	74.90000	59.30000	5.400000	5.400000	7.570000	12.030000
1997	75.00000	59.80000	4.900000	5.000000	7.680000	12.490000
1998	74.90000	59.80000	4.400000	4.600000	7.890000	13.000000
1999	74.70000	60.00000	4.100000	4.300000	8.000000	13.470000
2000	74.80000	59.90000	3.900000	4.100000	8.030000	14.000000
2001	74.40000	59.80000	4.800000	4.700000	8.110000	14.530000
2002	74.10000	59.60000	5.900000	5.600000	8.240000	14.950000
2003	73.50000	59.50000	6.300000	5.700000	8.270000	15.350000
2004	73.30000	59.20000	5.600000	5.400000	8.230000	15.670000
2005	73.30000	59.30000	5.100000	5.100000	8.170000	16.110000
2006	73.50000	59.40000	4.600000	4.600000	8.230000	16.730000

Table citations below refer to the source document.
[1]CLFPRM, Civilian labor force participation rate, male (%), Table B-39, p. 277.
[2]CLFPRF, Civilian labor force participation rate, female (%), Table B-39, p. 277.
[3]UNRM, Civilian unemployment rate, male (%) Table B-42, p. 280.
[4]UNRF, Civilian unemployment rate, female (%) Table B-42, p. 280.
[5]AHE82, Average hourly earnings (1982 dollars), Table B-47, p. 286.
[6]AHE, Average hourly earnings (current dollars), Table B-47, p. 286.

b. Repeat (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.)

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?

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.

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 1972–2007. These data represent the critical reading and mathematics test scores for both male and female students. The writing category was introduced in 2006. Therefore, these data are not included.

a. Use the horizontal axis for years and the vertical axis for SAT scores to plot the critical reading and math scores for males and females separately.

b. What general conclusions do you draw from these graphs?

c. Knowing the critical reading scores of males and females, how would you go about predicting their math scores?

d. Plot the female math scores against the male math scores. What do you observe?

TABLE 2.9
Total Group Mean SAT Reasoning Test Scores: College-Bound Seniors, 1972–2007

Source: College Board, 2007.

	Critical Reading			Mathematics		
Year	Male	Female	Total	Male	Female	Total
1972	531	529	530	527	489	509
1973	523	521	523	525	489	506
1974	524	520	521	524	488	505
1975	515	509	512	518	479	498
1976	511	508	509	520	475	497
1977	509	505	507	520	474	496
1978	511	503	507	517	474	494
1979	509	501	505	516	473	493
1980	506	498	502	515	473	492
1981	508	496	502	516	473	492
1982	509	499	504	516	473	493
1983	508	498	503	516	474	494
1984	511	498	504	518	478	497
1985	514	503	509	522	480	500
1986	515	504	509	523	479	500
1987	512	502	507	523	481	501
1988	512	499	505	521	483	501
1989	510	498	504	523	482	502
1990	505	496	500	521	483	501
1991	503	495	499	520	482	500
1992	504	496	500	521	484	501
1993	504	497	500	524	484	503
1994	501	497	499	523	487	504
1995	505	502	504	525	490	506
1996	507	503	505	527	492	508
1997	507	503	505	530	494	511
1998	509	502	505	531	496	512
1999	509	502	505	531	495	511
2000	507	504	505	533	498	514
2001	509	502	506	533	498	514
2002	507	502	504	534	500	516
2003	512	503	507	537	503	519
2004	512	504	508	537	501	518
2005	513	505	508	538	504	520
2006	505	502	503	536	502	518
2007	504	502	502	533	499	515

Note: For 1972–1986 a formula was applied to the original mean and standard deviation to convert the mean to the recentered scale. For 1987–1995 individual student scores were converted to the recentered scale and then the mean was recomputed. From 1996–1999, nearly all students received scores on the recentered scale. Any score on the original scale was converted to the recentered scale prior to computing the mean. From 2000–2007, all scores are reported on the recentered scale.

2.17. Table 2.10 presents data on mean SAT reasoning test scores classified by income for three kinds of tests: critical reading, mathematics, and writing. In Example 2.2, we presented Figure 2.7, which plotted mean math scores on mean family income.

a. Refer to Figure 2.7 and prepare a similar graph relating average critical reading scores to average family income. Compare your results with those shown in Figure 2.7.

TABLE 2.10
SAT Reasoning Test Classified by Family Income

Source: College Board, 2007 College-Bound Seniors, Table 11.

Family Income ($)	Number of Test Takers	Critical Reading		Mathematics		Writing	
		Mean	SD	Mean	SD	Mean	SD
<10,000	40610	427	107	451	122	423	104
10000–20000	72745	453	106	472	113	446	102
20000–30000	61244	454	102	465	107	444	97
30000–40000	83685	476	103	485	106	466	98
40000–50000	75836	489	103	486	105	477	99
50000–60000	80060	497	102	504	104	486	98
60000–70000	75763	504	102	511	103	493	98
70000–80000	81627	508	101	516	103	498	98
80000–100000	130752	520	102	529	104	510	100
>100000	245025	544	105	556	107	537	103

b. Repeat (a), relating average writing scores to average family income and compare your results with the other two graphs.

c. Looking at the three graphs, what general conclusion can you draw?