INTRODUCTORY ECONOMETRICS

A Modern Approach

7e

Jeffrey M. Wooldridge

Introductory Econometrics

A MODERN APPROACH

SEVENTH EDITION

Jeffrey M. Wooldridge
Michigan State University

CENGAGE

Australia • Brazil • Mexico • Singapore • United Kingdom • United States

CENGAGE

***Introductory Econometrics: A Modern Approach,* Seventh Edition**
Jeffrey M. Wooldridge

Senior Vice President, Higher Education Product Management: Erin Joyner

Product Director: Jason Fremder

Sr. Product Manager: Michael Parthenakis

Sr. Learning Designer: Sarah Keeling

Sr. Content Manager: Anita Verma

In-House Subject Matter Expert (s): Eugenia Belova, Ethan Crist and Kasie Jean

Digital Delivery Lead: Timothy Christy

Product Assistant: Matt Schiesl

Manufacturing Planner: Kevin Kluck

Production Service: SPi-Global

Intellectual Property

Analyst: Jennifer Bowes

Project Manager: Julie Geagan-Chevez

Marketing Manager: John Carey

Sr. Designer: Bethany Bourgeois

Cover Designer: Tin Box Studio

Library of Congress Control Number: 2018956380

ISBN: 978-1-337-55886-0

Cengage
20 Channel Center Street
Boston, MA 02210
USA

Cengage is a leading provider of customized learning solutions with employees residing in nearly 40 different countries and sales in more than 125 countries around the world. Find your local representative at **www.cengage.com**.

Cengage products are represented in Canada by Nelson Education, Ltd.

To learn more about Cengage platforms and services, register or access your online learning solution, or purchase materials for your course, visit **www.cengage.com**.

Printed in the United States of America
Print Number: 01 Print Year: 2018

Brief Contents

Contents

CHAPTER 1 The Nature of Econometrics and Economic Data 1

PART 1

Regression Analysis with Cross-Sectional Data 19

CHAPTER 2 The Simple Regression Model 20

CHAPTER 3 Multiple Regression Analysis: Estimation 66

CHAPTER 7 Multiple Regression Analysis with Qualitative Information 220

CHAPTER 8 Heteroskedasticity 262

CHAPTER 9 More on Specification and Data Issues 294

PART 2

Regression Analysis with Time Series Data 333

PART 3

Advanced Topics 425

CHAPTER 13 Pooling Cross Sections across Time: Simple Panel Data Methods 426

CHAPTER 14 Advanced Panel Data Methods 462

CHAPTER 15 Instrumental Variables Estimation and Two-Stage Least Squares 495

MATH REFRESHER A Basic Mathematical Tools 666

MATH REFRESHER B Fundamentals of Probability 684

MATH REFRESHER C Fundamentals of Mathematical Statistics 714

ADVANCED TREATMENT D Summary of Matrix Algebra 749

ADVANCED TREATMENT E The Linear Regression Model in Matrix Form 760

Preface

In ALL content, please indent the first paragraph as well, like the following ones. My motivation for writing the first edition of *Introductory Econometrics: A Modern Approach* was that I saw a fairly wide gap between how econometrics is taught to undergraduates and how empirical researchers think about and apply econometric methods. I became convinced that teaching introductory econometrics from the perspective of professional users of econometrics would actually simplify the presentation, in addition to making the subject much more interesting.

Based on the positive reactions to the several earlier editions, it appears that my hunch was correct. Many instructors, having a variety of backgrounds and interests and teaching students with different levels of preparation, have embraced the modern approach to econometrics espoused in this text. The emphasis in this edition is still on applying econometrics to real-world problems. Each econometric method is motivated by a particular issue facing researchers analyzing nonexperimental data. The focus in the main text is on understanding and interpreting the assumptions in light of actual empirical applications: the mathematics required is no more than college algebra and basic probability and statistics.

Designed for Today's Econometrics Course

The seventh edition preserves the overall organization of the sixth. The most noticeable feature that distinguishes this text from most others is the separation of topics by the kind of data being analyzed. This is a clear departure from the traditional approach, which presents a linear model, lists all assumptions that may be needed at some future point in the analysis, and then proves or asserts results without clearly connecting them to the assumptions. My approach is first to treat, in Part 1, multiple regression analysis with cross-sectional data, under the assumption of random sampling. This setting is natural to students because they are familiar with random sampling from a population in their introductory statistics courses. Importantly, it allows us to distinguish assumptions made about the underlying population regression model—assumptions that can be given economic or behavioral content—from assumptions about how the data were sampled. Discussions about the consequences of nonrandom sampling can be treated in an intuitive fashion after the students have a good grasp of the multiple regression model estimated using random samples.

An important feature of a modern approach is that the explanatory variables—along with the dependent variable—are treated as outcomes of random variables. For the social sciences, allowing random explanatory variables is much more realistic than the traditional assumption of nonrandom explanatory variables. As a nontrivial benefit, the population model/random sampling approach reduces the number of assumptions that students must absorb and understand. Ironically, the classical approach to regression analysis, which treats the explanatory variables as fixed in repeated samples and is still pervasive in introductory texts, literally applies to data collected in an experimental setting. In addition, the contortions required to state and explain assumptions can be confusing to students.

My focus on the population model emphasizes that the fundamental assumptions underlying regression analysis, such as the zero mean assumption on the unobservable error term, are properly stated conditional on the explanatory variables. This leads to a clear understanding of the kinds of problems, such as heteroskedasticity (nonconstant variance), that can invalidate standard inference procedures. By focusing on the population, I am also able to dispel several misconceptions that arise in econometrics texts at all levels. For example, I explain why the usual *R*-squared is still valid as a goodness-of-fit measure in the presence of heteroskedasticity (Chapter 8) or serially correlated errors (Chapter 12); I provide a simple demonstration that tests for functional form should not be viewed as general tests of omitted variables (Chapter 9); and I explain why one should always include in a regression model extra control variables that are uncorrelated with the explanatory variable of interest, which is often a key policy variable (Chapter 6).

Because the assumptions for cross-sectional analysis are relatively straightforward yet realistic, students can get involved early with serious cross-sectional applications without having to worry about the thorny issues of trends, seasonality, serial correlation, high persistence, and spurious regression that are ubiquitous in time series regression models. Initially, I figured that my treatment of regression with cross-sectional data followed by regression with time series data would find favor with instructors whose own research interests are in applied microeconomics, and that appears to be the case. It has been gratifying that adopters of the text with an applied time series bent have been equally enthusiastic about the structure of the text. By postponing the econometric analysis of time series data, I am able to put proper focus on the potential pitfalls in analyzing time series data that do not arise with cross-sectional data. In effect, time series econometrics finally gets the serious treatment it deserves in an introductory text.

As in the earlier editions, I have consciously chosen topics that are important for reading journal articles and for conducting basic empirical research. Within each topic, I have deliberately omitted many tests and estimation procedures that, while traditionally included in textbooks, have not withstood the empirical test of time. Likewise, I have emphasized more recent topics that have clearly demonstrated their usefulness, such as obtaining test statistics that are robust to heteroskedasticity (or serial correlation) of unknown form, using multiple years of data for policy analysis, or solving the omitted variable problem by instrumental variables methods. I appear to have made fairly good choices, as I have received only a handful of suggestions for adding or deleting material.

I take a systematic approach throughout the text, by which I mean that each topic is presented by building on the previous material in a logical fashion, and assumptions are introduced only as they are needed to obtain a conclusion. For example, empirical researchers who use econometrics in their research understand that not all of the Gauss-Markov assumptions are needed to show that the ordinary least squares (OLS) estimators are unbiased. Yet the vast majority of econometrics texts introduce a complete set of assumptions (many of which are redundant or in some cases even logically conflicting) before proving the unbiasedness of OLS. Similarly, the normality assumption is often included among the assumptions that are needed for the Gauss-Markov Theorem, even though it is fairly well known that normality plays no role in showing that the OLS estimators are the best linear unbiased estimators.

My systematic approach is illustrated by the order of assumptions that I use for multiple regression in Part 1. This structure results in a natural progression for briefly summarizing the role of each assumption:

MLR.1: Introduce the population model and interpret the population parameters (which we hope to estimate).

MLR.2: Introduce random sampling from the population and describe the data that we use to estimate the population parameters.

MLR.3: Add the assumption on the explanatory variables that allows us to compute the estimates from our sample; this is the so-called no perfect collinearity assumption.

MLR.4: Assume that, in the population, the mean of the unobservable error does not depend on the values of the explanatory variables; this is the "mean independence" assumption combined with a zero population mean for the error, and it is the key assumption that delivers unbiasedness of OLS.

After introducing Assumptions MLR.1 to MLR.3, one can discuss the algebraic properties of ordinary least squares—that is, the properties of OLS for a particular set of data. By adding Assumption MLR.4, we can show that OLS is unbiased (and consistent). Assumption MLR.5 (homoskedasticity) is added for the Gauss-Markov Theorem and for the usual OLS variance formulas to be valid. Assumption MLR.6 (normality), which is not introduced until Chapter 4, is added to round out the classical linear model assumptions. The six assumptions are used to obtain exact statistical inference and to conclude that the OLS estimators have the smallest variances among all unbiased estimators.

I use parallel approaches when I turn to the study of large-sample properties and when I treat regression for time series data in Part 2. The careful presentation and discussion of assumptions makes it relatively easy to transition to Part 3, which covers advanced topics that include using pooled cross-sectional data, exploiting panel data structures, and applying instrumental variables methods. Generally, I have strived to provide a unified view of econometrics, where all estimators and test statistics are obtained using just a few intuitively reasonable principles of estimation and testing (which, of course, also have rigorous justification). For example, regression-based tests for heteroskedasticity and serial correlation are easy for students to grasp because they already have a solid understanding of regression. This is in contrast to treatments that give a set of disjointed recipes for outdated econometric testing procedures.

Throughout the text, I emphasize ceteris paribus relationships, which is why, after one chapter on the simple regression model, I move to multiple regression analysis. The multiple regression setting motivates students to think about serious applications early. I also give prominence to policy analysis with all kinds of data structures. Practical topics, such as using proxy variables to obtain ceteris paribus effects and interpreting partial effects in models with interaction terms, are covered in a simple fashion.

Designed at Undergraduates, Applicable to Master's Students

The text is designed for undergraduate economics majors who have taken college algebra and one-semester of introductory probability and statistics. (Math Refresher A, B, and C contain the requisite background material.) A one-semester or one-quarter econometrics course would not be expected to cover all, or even any, of the more advanced material in Part 3. A typical introductory course includes Chapters 1 through 8, which cover the basics of simple and multiple regression for cross-sectional data. Provided the emphasis is on intuition and interpreting the empirical examples, the material from the first eight chapters should be accessible to undergraduates in most economics departments. Most instructors will also want to cover at least parts of the chapters on regression analysis with time series data, Chapters 10 and 12, in varying degrees of depth. In the one-semester course that I teach at Michigan State, I cover Chapter 10 fairly carefully, give an overview of the material in Chapter 11, and cover the material on serial correlation in Chapter 12. I find that this basic one-semester course puts students on a solid footing to write empirical papers, such as a term paper, a senior seminar paper, or a senior thesis. Chapter 9 contains more specialized topics that arise in analyzing cross-sectional data, including data problems such as outliers and nonrandom sampling; for a one-semester course, it can be skipped without loss of continuity.

The structure of the text makes it ideal for a course with a cross-sectional or policy analysis focus: the time series chapters can be skipped in lieu of topics from Chapters 9 or 15. The new material on potential outcomes added to the first nine chapters should help the instructor craft a course that provides an introduction to modern policy analysis. Chapter 13 is advanced only in the sense that it treats two new data structures: independently pooled cross sections and two-period panel data analysis. Such data structures are especially useful for policy analysis, and the chapter provides several examples. Students with a good grasp of Chapters 1 through 8 will have little difficulty with Chapter 13. Chapter 14 covers more advanced panel data methods and would probably be covered only in a second course. A good way to end a course on cross-sectional methods is to cover the rudiments of instrumental variables estimation in Chapter 15.

I have used selected material in Part 3, including Chapters 13 and 17, in a senior seminar geared to producing a serious research paper. Along with the basic one-semester course, students who have been exposed to basic panel data analysis, instrumental variables estimation, and limited dependent variable models are in a position to read large segments of the applied social sciences literature. Chapter 17 provides an introduction to the most common limited dependent variable models.

The text is also well suited for an introductory master's level course, where the emphasis is on applications rather than on derivations using matrix algebra. Several instructors have used the text to teach policy analysis at the master's level. For instructors wanting to present the material in matrix form, Appendices D and E are self-contained treatments of the matrix algebra and the multiple regression model in matrix form.

At Michigan State, PhD students in many fields that require data analysis—including accounting, agricultural economics, development economics, economics of education, finance, international economics, labor economics, macroeconomics, political science, and public finance—have found the text to be a useful bridge between the empirical work that they read and the more theoretical econometrics they learn at the PhD level.

Suggestions for Designing Your Course Beyond the Basic

I have already commented on the contents of most of the chapters as well as possible outlines for courses. Here I provide more specific comments about material in chapters that might be covered or skipped:

Chapter 9 has some interesting examples (such as a wage regression that includes IQ score as an explanatory variable). The rubric of proxy variables does not have to be formally introduced to present these kinds of examples, and I typically do so when finishing up cross-sectional analysis. In Chapter 12, for a one-semester course, I skip the material on serial correlation robust inference for ordinary least squares as well as dynamic models of heteroskedasticity.

Even in a second course I tend to spend only a little time on Chapter 16, which covers simultaneous equations analysis. I have found that instructors differ widely in their opinions on the importance of teaching simultaneous equations models to undergraduates. Some think this material is fundamental; others think it is rarely applicable. My own view is that simultaneous equations models are overused (see Chapter 16 for a discussion). If one reads applications carefully, omitted variables and measurement error are much more likely to be the reason one adopts instrumental variables estimation, and this is why I use omitted variables to motivate instrumental variables estimation in Chapter 15. Still, simultaneous equations models are indispensable for estimating demand and supply functions, and they apply in some other important cases as well.

Chapter 17 is the only chapter that considers models inherently nonlinear in their parameters, and this puts an extra burden on the student. The first material one should cover in this chapter is on probit and logit models for binary response. My presentation of Tobit models and censored regression still appears to be novel in introductory texts. I explicitly recognize that the Tobit model is applied to corner solution outcomes on random samples, while censored regression is applied when the data collection process censors the dependent variable at essentially arbitrary thresholds.

Chapter 18 covers some recent important topics from time series econometrics, including testing for unit roots and cointegration. I cover this material only in a second-semester course at either the undergraduate or master's level. A fairly detailed introduction to forecasting is also included in Chapter 18.

Chapter 19, which would be added to the syllabus for a course that requires a term paper, is much more extensive than similar chapters in other texts. It summarizes some of the methods appropriate for various kinds of problems and data structures, points out potential pitfalls, explains in some detail how to write a term paper in empirical economics, and includes suggestions for possible projects.

What's Changed?

I have added new exercises to many chapters, including to the Math Refresher and Advanced Treatment appendices. Some of the new computer exercises use new data sets, including a data set on performance of men's college basketball teams. I have also added more challenging problems that require derivations.

There are several notable changes to the text. An important organizational change, which should facilitate a wider variety of teaching tastes, is that the notion of binary, or dummy, explanatory variables is introduced in Chapter 2. There, it is shown that ordinary least squares estimation leads to a staple in basic statistics: the difference in means between two subgroups in a population. By introducing qualitative factors into regression early on, the instructor is able to use a wider variety of empirical examples from the very beginning.

The early discussion of binary explanatory variables allows for a formal introduction of potential, or counterfactual, outcomes, which is indispensable in the modern literature on estimating causal effects. The counterfactual approach to studying causality appears in previous editions, but Chapters 2, 3, 4, and 7 now explicitly include new sections on the modern approach to causal inference. Because basic policy analysis involves the binary decision to participate in a program or not, a leading example of using dummy independent variables in simple and multiple regression is to evaluate policy interventions. At the same time, the new material is incorporated into the text so that instructors not wishing to cover the potential outcomes framework may easily skip the material. Several end-of-chapter problems concern extensions of the basic potential outcomes framework, which should be valuable for instructors wishing to cover that material.

Chapter 3 includes a new section on different ways that one can apply multiple regression, including problems of pure prediction, testing efficient markets, and culminating with a discussion of estimating treatment or causal effects. I think this section provides a nice way to organize students' thinking about the scope of multiple regression after they have seen the mechanics of ordinary least squares (OS) and several examples. As with other new material that touches on causal effects, this material can be skipped without loss of continuity. A new section in Chapter 7 continues the discussion of potential outcomes, allowing for nonconstant treatment effects. The material is a nice illustration of estimating different regression functions for two subgroups from a population. New problems in this chapter that allow the student more experience in using full regression adjustment to estimate causal effects.

One notable change to Chapter 9 is a more detailed discussion of using missing data indicators when data are missing on one or more of the explanatory variables. The assumptions underlying the method are discussed in more detail than in the previous edition.

Chapter 12 has been reorganized to reflect a more modern treatment of the problem of serial correlation in the errors of time series regression models. The new structure first covers adjusting the OLS standard errors to allow general forms of serial correlation. Thus, the chapter outline now parallels that in Chapter 8, with the emphasis in both cases on OLS estimation but making inference robust to violation of standard assumptions. Correcting for serial correlation using generalized least squares now comes after OLS and the treatment of testing for serial correlation.

The advanced chapters also include several improvements. Chapter 13 now discusses, at an accessible level, extensions of the standard difference-in-differences setup, allowing for multiple control groups, multiple time periods, and even group-specific trends. In addition, the chapter includes a more detailed discussion of computing standard errors robust to serial correlation when using first-differencing estimation with panel data.

Chapter 14 now provides more detailed discussions of several important issues in estimating panel data models by fixed effects, random effects, and correlated random effects (CRE). The CRE approach with missing data is discussed in more detail, as is how one accounts for general functional forms, such as squares and interactions, which are covered in the cross-sectional setting in Chapter 6. An expanded section on general policy analysis with panel data should be useful for courses with an emphasis on program interventions and policy evaluation.

Chapter 16, which still covers simultaneous equations models, now provides an explicit link between the potential outcomes framework and specification of simultaneous equations models.

Chapter 17 now includes a discussion of using regression adjustment for estimating causal (treatment) effects when the outcome variable has special features, such as when the outcome itself is a binary variable. Then, as the reader is asked to explore in a new problem, logit and probit models can be used to obtain more reliable estimates of average treatment effects by estimating separate models for each treatment group.

Chapter 18 now provides more details about how one can compute a proper standard error for a forecast (as opposed to a prediction) interval. This should help the advanced reader understand in more detail the nature of the uncertainty in the forecast.

About MindTap™

MindTap is an outcome-driven application that propels students from memorization to mastery. It's the only platform that gives you complete ownership of your course. With it, you can challenge every student, build their confidence, and empower them to be unstoppable.

Access Everything You Need In One Place. Cut down on prep with preloaded, organized course materials in MindTap. Teach more efficiently with interactive multimedia, assignments, quizzes and more. And give your students the power to read, listen and study on their phones, so they can learn on their terms.

Empower Your Students To Reach Their Potential. Twelve distinct metrics give you actionable insights into student engagement. Identify topics troubling your entire class and instantly communicate with struggling students. And students can track their scores to stay motivated toward their goals. Together, you can accelerate progress.

Your Course. *Your Content*. Only MindTap gives you complete control over your course. You have the flexibility to reorder textbook chapters, add your own notes and embed a variety of content including OER. Personalize course content to your students' needs. They can even read your notes, add their own and highlight key text to aid their progress.

A Dedicated Team, Whenever You Need Them. MindTap isn't just a tool; it's backed by a personalized team eager to support you. Get help setting up your course and tailoring it to your specific objectives. You'll be ready to make an impact from day one. And, we'll be right here to help you and your students throughout the semester—and beyond.

Design Features

In addition to the didactic material in the chapter, I have included two features to help students better understand and apply what they are learning. Each chapter contains many numbered examples. Several of these are case studies drawn from recently published papers. I have used my judgment to simplify the analysis, hopefully without sacrificing the main point. The "Going Further Questions" in the chapter provide students an opportunity to “go further” in learning the material through analysis or application. Students will find immediate feedback for these questions in the end of the text.

The end-of-chapter problems and computer exercises are heavily oriented toward empirical work, rather than complicated derivations. The students are asked to reason carefully based on what they have learned. The computer exercises often expand on the in-text examples. Several exercises use data sets from published works or similar data sets that are motivated by published research in economics and other fields.

A pioneering feature of this introductory econometrics text is the extensive glossary. The short definitions and descriptions are a helpful refresher for students studying for exams or reading empirical research that uses econometric methods. I have added and updated several entries for the seventh edition.

Instructional Tools

Cengage offers various supplements for instructors and students who use this book. I would like to thank the Subject Matter Expert team who worked on these supplements and made teaching and learning easy.

C. Patrick Scott, Ph.D., Louisiana Tech University (R Videos and Computer exercise reviewer)
Hisham Foad (Aplia Home work reviewer and Glossary)
Kenneth H. Brown, Missouri State University (R Videos creator)
Scott Kostyshak, University of Florida (R Videos reviewer)
Ujwal Kharel (Test Bank and Adaptive Test Prep)

Data Sets—Available in Six Formats

With more than 100 data sets in six different formats, including Stata®, R, EViews®, Minitab®, Microsoft® Excel, and Text, the instructor has many options for problem sets, examples, and term projects. Because most of the data sets come from actual research, some are very large. Except for partial lists of data sets to illustrate the various data structures, the data sets are not reported in the text. This book is geared to a course where computer work plays an integral role.

Updated Data Sets Handbook

An extensive data description manual is also available online. This manual contains a list of data sources along with suggestions for ways to use the data sets that are not described in the text. This unique handbook, created by author Jeffrey M. Wooldridge, lists the source of all data sets for quick reference and how each might be used. Because the data book contains page numbers, it is easy to see how the author used the data in the text. Students may want to view the descriptions of each data set and it can help guide instructors in generating new homework exercises, exam problems, or term projects. The author also provides suggestions on improving the data sets in this detailed resource that is available on the book’s companion website at http://login.cengage.com and students can access it free at www.cengage.com.

Instructor's Manual with Solutions

REVISED INSTRUCTOR'S MANUAL WITH SOLUTIONS SAVES TIME IN PREPARATION AND GRADING. The online Instructor's Manual with solutions contains answers to all exercises in this edition. Teaching tips provide suggestions for presenting each chapter's material. The Instructor's Manual also contains sources for each of the data files with suggestions for using the data to develop problem sets, exams, and term papers. The Instructor's Manual is password-protected and available for download on the book's companion website.

Test Bank

Cengage Testing, powered by Cognero® is a flexible, online system that allows you to import, edit, and manipulate content from the text's test bank or elsewhere, including your own favorite test questions; create multiple test versions in an instant; and deliver tests from your LMS, your classroom, or wherever you want.

PowerPoint Slides

UPDATED POWERPOINT® SLIDES BRING LECTURES TO LIFE WHILE VISUALLY CLARIFYING CONCEPTS. Exceptional PowerPoint® presentation slides, created specifically for this edition, help you create engaging, memorable lectures. The slides are particularly useful for clarifying advanced topics in Part 3. You can modify or customize the slides for your specific course. PowerPoint® slides are available for convenient download on the instructor-only, password-protected section of the book's companion website.

Scientific Word Slides

UPDATED SCIENTIFIC WORD® SLIDES REINFORCE TEXT CONCEPTS AND LECTURE PRESENTATIONS. Created by the text author, this edition's Scientific Word® slides reinforce the book's presentation slides while highlighting the benefits of Scientific Word®, the application created by MacKichan software, Inc. for specifically composing mathematical, scientific and technical documents using LaTeX typesetting. These slides are based on the author's actual lectures and are available for convenient download on the password-protected section of the book's companion website.

Student Supplements

Student Solutions Manual

Now your student's can maximize their study time and further their course success with this dynamic online resource. This helpful Solutions Manual includes detailed steps and solutions to odd-numbered problems as well as computer exercises in the text. This supplement is available as a free resource at www.cengagebrain.com.

Acknowledgments

I would like to thank those who reviewed and provided helpful comments for this and previous editions of the text:

Erica Johnson, *Gonzaga University*

Mary Ellen Benedict, *Bowling Green State University*

Chirok Han, *Korea University*

Yan Li, *Temple University*

Melissa Tartari, *Yale University*

Michael Allgrunn, *University of South Dakota*

Gregory Colman, *Pace University*

Yoo-Mi Chin, *Missouri University of Science and Technology*

Arsen Melkumian, *Western Illinois University*

Kevin J. Murphy, *Oakland University*

Kristine Grimsrud, *University of New Mexico*

Will Melick, *Kenyon College*

Philip H. Brown, *Colby College*

Argun Saatcioglu, *University of Kansas*

Ken Brown, *University of Northern Iowa*

Michael R. Jonas, *University of San Francisco*

Melissa Yeoh, *Berry College*

Nikolaos Papanikolaou, *SUNY at New Paltz*

Konstantin Golyaev, *University of Minnesota*

Soren Hauge, *Ripon College*

Kevin Williams, *University of Minnesota*

Hailong Qian, *Saint Louis University*

Rod Hissong, *University of Texas at Arlington*

Steven Cuellar, *Sonoma State University*

Yanan Di, *Wagner College*

John Fitzgerald, *Bowdoin College*

Philip N. Jefferson, *Swarthmore College*

Yongsheng Wang, *Washington and Jefferson College*

Sheng-Kai Chang, *National Taiwan University*

Damayanti Ghosh, *Binghamton University*

Susan Averett, *Lafayette College*

Kevin J. Mumford, *Purdue University*

Nicolai V. Kuminoff, *Arizona State University*

Subarna K. Samanta, *The College of New Jersey*

Jing Li, *South Dakota State University*

Gary Wagner, *University of Arkansas–Little Rock*

Kelly Cobourn, *Boise State University*

Timothy Dittmer, *Central Washington University*

Daniel Fischmar, *Westminster College*

Subha Mani, *Fordham University*

John Maluccio, *Middlebury College*

James Warner, *College of Wooster*

Christopher Magee, *Bucknell University*

Andrew Ewing, *Eckerd College*

Debra Israel, *Indiana State University*

Jay Goodliffe, *Brigham Young University*

Stanley R. Thompson, *The Ohio State University*

Michael Robinson, *Mount Holyoke College*

Ivan Jeliazkov, *University of California, Irvine*

Heather O'Neill, *Ursinus College*

Leslie Papke, *Michigan State University*

Timothy Vogelsang, *Michigan State University*

Stephen Woodbury, *Michigan State University*

Some of the changes I discussed earlier were driven by comments I received from people on this list, and I continue to mull over other specific suggestions made by one or more reviewers.

Many students and teaching assistants, too numerous to list, have caught mistakes in earlier editions or have suggested rewording some paragraphs. I am grateful to them.

As always, it was a pleasure working with the team at Cengage Learning. Michael Parthenakis, my longtime Product Manager, has learned very well how to guide me with a firm yet gentle hand. Anita Verma and Ethan Crist quickly mastered the difficult challenges of being the content and subject matter expert team of a dense, technical textbook. Their careful reading of the manuscript and fine eye for detail have improved this seventh edition considerably.

This book is dedicated to my family: Leslie, Edmund, and R.G.

Jeffrey M. Wooldridge

About the Author

Jeffrey M. Wooldridge is University Distinguished Professor of Economics at Michigan State University, where he has taught since 1991. From 1986 to 1991, he was an assistant professor of economics at the Massachusetts Institute of Technology. He received his bachelor of arts, with majors in computer science and economics, from the University of California, Berkeley, in 1982, and received his doctorate in economics in 1986 from the University of California, San Diego. He has published more than 60 articles in internationally recognized journals, as well as several book chapters. He is also the author of *Econometric Analysis of Cross Section and Panel Data*, second edition. His awards include an Alfred P. Sloan Research Fellowship, the Plura Scripsit award from *Econometric Theory*, the Sir Richard Stone prize from the *Journal of Applied Econometrics*, and three graduate teacher-of-the-year awards from MIT. He is a fellow of the Econometric Society and of the *Journal of Econometrics*. He is past editor of the *Journal of Business and Economic Statistics*, and past econometrics coeditor of *Economics Letters*. He has served on the editorial boards of *Econometric Theory*, the *Journal of Economic Literature*, the *Journal of Econometrics*, the *Review of Economics and Statistics*, and the *Stata Journal*. He has also acted as an occasional econometrics consultant for Arthur Andersen, Charles River Associates, the Washington State Institute for Public Policy, Stratus Consulting, and Industrial Economics, Incorporated.

CHAPTER 1

The Nature of Econometrics and Economic Data

Chapter 1 discusses the scope of econometrics and raises general issues that arise in the application of econometric methods. Section 1-1 provides a brief discussion about the purpose and scope of econometrics and how it fits into economic analysis. Section 1-2 provides examples of how one can start with an economic theory and build a model that can be estimated using data. Section 1-3 examines the kinds of data sets that are used in business, economics, and other social sciences. Section 1-4 provides an intuitive discussion of the difficulties associated with inferring causality in the social sciences.

1-1 What Is Econometrics?

Imagine that you are hired by your state government to evaluate the effectiveness of a publicly funded job training program. Suppose this program teaches workers various ways to use computers in the manufacturing process. The 20-week program offers courses during nonworking hours. Any hourly manufacturing worker may participate, and enrollment in all or part of the program is voluntary. You are to determine what, if any, effect the training program has on each worker's subsequent hourly wage.

Now, suppose you work for an investment bank. You are to study the returns on different investment strategies involving short-term U.S. treasury bills to decide whether they comply with implied economic theories.

The task of answering such questions may seem daunting at first. At this point, you may only have a vague idea of the kind of data you would need to collect. By the end of this introductory econometrics course, you should know how to use econometric methods to formally evaluate a job training program or to test a simple economic theory.

Econometrics is based upon the development of statistical methods for estimating economic relationships, testing economic theories, and evaluating and implementing government and business policy. A common application of econometrics is the forecasting of such important macroeconomic variables as interest rates, inflation rates, and gross domestic product (GDP). Whereas forecasts of economic indicators are highly visible and often widely published, econometric methods can be used in economic areas that have nothing to do with macroeconomic forecasting. For example, we will study the effects of political campaign expenditures on voting outcomes. We will consider the effect of school spending on student performance in the field of education. In addition, we will learn how to use econometric methods for forecasting economic time series.

Econometrics has evolved as a separate discipline from mathematical statistics because the former focuses on the problems inherent in collecting and analyzing nonexperimental economic data. **Nonexperimental data** are not accumulated through controlled experiments on individuals, firms, or segments of the economy. (Nonexperimental data are sometimes called **observational data**, or **retrospective data**, to emphasize the fact that the researcher is a passive collector of the data.) **Experimental data** are often collected in laboratory environments in the natural sciences, but they are more difficult to obtain in the social sciences. Although some social experiments can be devised, it is often impossible, prohibitively expensive, or morally repugnant to conduct the kinds of controlled experiments that would be needed to address economic issues. We give some specific examples of the differences between experimental and nonexperimental data in Section 1-4.

Naturally, econometricians have borrowed from mathematical statisticians whenever possible. The method of multiple regression analysis is the mainstay in both fields, but its focus and interpretation can differ markedly. In addition, economists have devised new techniques to deal with the complexities of economic data and to test the predictions of economic theories.

1-2 Steps in Empirical Economic Analysis

Econometric methods are relevant in virtually every branch of applied economics. They come into play either when we have an economic theory to test or when we have a relationship in mind that has some importance for business decisions or policy analysis. An **empirical analysis** uses data to test a theory or to estimate a relationship.

How does one go about structuring an empirical economic analysis? It may seem obvious, but it is worth emphasizing that the first step in any empirical analysis is the careful formulation of the question of interest. The question might deal with testing a certain aspect of an economic theory, or it might pertain to testing the effects of a government policy. In principle, econometric methods can be used to answer a wide range of questions.

In some cases, especially those that involve the testing of economic theories, a formal **economic model** is constructed. An economic model consists of mathematical equations that describe various relationships. Economists are well known for their building of models to describe a vast array of behaviors. For example, in intermediate microeconomics, individual consumption decisions, subject to a budget constraint, are described by mathematical models. The basic premise underlying these models is *utility maximization*. The assumption that individuals make choices to maximize their well-being, subject to resource constraints, gives us a very powerful framework for creating tractable economic models and making clear predictions. In the context of consumption decisions, utility maximization leads to a set of *demand equations*. In a demand equation, the quantity demanded of each commodity depends on the price of the goods, the price of substitute and complementary goods, the consumer's income, and the individual's characteristics that affect taste. These equations can form the basis of an econometric analysis of consumer demand.

Economists have used basic economic tools, such as the utility maximization framework, to explain behaviors that at first glance may appear to be noneconomic in nature. A classic example is Becker's (1968) economic model of criminal behavior.

EXAMPLE 1.1 Economic Model of Crime

In a seminal article, Nobel Prize winner Gary Becker postulated a utility maximization framework to describe an individual's participation in crime. Certain crimes have clear economic rewards, but most criminal behaviors have costs. The opportunity costs of crime prevent the criminal from participating in other activities such as legal employment. In addition, there are costs associated with the possibility of being caught and then, if convicted, the costs associated with incarceration. From Becker's perspective, the decision to undertake illegal activity is one of resource allocation, with the benefits and costs of competing activities taken into account.

Under general assumptions, we can derive an equation describing the amount of time spent in criminal activity as a function of various factors. We might represent such a function as

$$y = f(x_1, x_2, x_3, x_4, x_5, x_6, x_7), \quad [1.1]$$

where

y = hours spent in criminal activities,
x_1 = "wage" for an hour spent in criminal activity,
x_2 = hourly wage in legal employment,
x_3 = income other than from crime or employment,
x_4 = probability of getting caught,
x_5 = probability of being convicted if caught,
x_6 = expected sentence if convicted, and
x_7 = age.

Other factors generally affect a person's decision to participate in crime, but the list above is representative of what might result from a formal economic analysis. As is common in economic theory, we have not been specific about the function $f(\cdot)$ in (1.1). This function depends on an underlying utility function, which is rarely known. Nevertheless, we can use economic theory—or introspection—to predict the effect that each variable would have on criminal activity. This is the basis for an econometric analysis of individual criminal activity.

Formal economic modeling is sometimes the starting point for empirical analysis, but it is more common to use economic theory less formally, or even to rely entirely on intuition. You may agree that the determinants of criminal behavior appearing in equation (1.1) are reasonable based on common sense; we might arrive at such an equation directly, without starting from utility maximization. This view has some merit, although there are cases in which formal derivations provide insights that intuition can overlook.

Next is an example of an equation that we can derive through somewhat informal reasoning.

EXAMPLE 1.2 Job Training and Worker Productivity

Consider the problem posed at the beginning of Section 1-1. A labor economist would like to examine the effects of job training on worker productivity. In this case, there is little need for formal economic theory. Basic economic understanding is sufficient for realizing that factors such as education, experience, and training affect worker productivity. Also, economists are well aware that workers are paid commensurate with their productivity. This simple reasoning leads to a model such as

$$wage = f(educ, exper, training), \quad [1.2]$$

where

$wage$ = hourly wage,
$educ$ = years of formal education,
$exper$ = years of workforce experience, and
$training$ = weeks spent in job training.

Again, other factors generally affect the wage rate, but equation (1.2) captures the essence of the problem.

After we specify an economic model, we need to turn it into what we call an **econometric model**. Because we will deal with econometric models throughout this text, it is important to know how an econometric model relates to an economic model. Take equation (1.1) as an example. The form of the function $f(\cdot)$ must be specified before we can undertake an econometric analysis. A second issue concerning (1.1) is how to deal with variables that cannot reasonably be observed. For example, consider the wage that a person can earn in criminal activity. In principle, such a quantity is well defined, but it would be difficult if not impossible to observe this wage for a given individual. Even variables such as the probability of being arrested cannot realistically be obtained for a given individual, but at least we can observe relevant arrest statistics and derive a variable that approximates the probability of arrest. Many other factors affect criminal behavior that we cannot even list, let alone observe, but we must somehow account for them.

The ambiguities inherent in the economic model of crime are resolved by specifying a particular econometric model:

$$\begin{aligned} crime = \beta_0 &+ \beta_1 wage + \beta_2 othinc + \beta_3 freqarr + \beta_4 freqconv \\ &+ \beta_5 avgsen + \beta_6 age + u, \end{aligned} \tag{1.3}$$

where

$crime$ = some measure of the frequency of criminal activity,
$wage$ = the wage that can be earned in legal employment,
$othinc$ = the income from other sources (assets, inheritance, and so on),
$freqarr$ = the frequency of arrests for prior infractions (to approximate the probability of arrest),
$freqconv$ = the frequency of conviction, and
$avgsen$ = the average sentence length after conviction.

The choice of these variables is determined by the economic theory as well as data considerations. The term u contains unobserved factors, such as the wage for criminal activity, moral character, family background, and errors in measuring things like criminal activity and the probability of arrest. We could add family background variables to the model, such as number of siblings, parents' education, and so on, but we can never eliminate u entirely. In fact, dealing with this *error term* or *disturbance term* is perhaps the most important component of any econometric analysis.

The constants $\beta_0, \beta_1, \ldots, \beta_6$ are the *parameters* of the econometric model, and they describe the directions and strengths of the relationship between *crime* and the factors used to determine *crime* in the model.

A complete econometric model for Example 1.2 might be

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 training + u, \tag{1.4}$$

where the term u contains factors such as "innate ability," quality of education, family background, and the myriad other factors that can influence a person's wage. If we are specifically concerned about the effects of job training, then β_3 is the parameter of interest.

For the most part, econometric analysis begins by specifying an econometric model, without consideration of the details of the model's creation. We generally follow this approach, largely because careful derivation of something like the economic model of crime is time consuming and can take us into some specialized and often difficult areas of economic theory. Economic reasoning will play a role in our examples, and we will merge any underlying economic theory into the econometric model specification. In the economic model of crime example, we would start with an econometric model such as (1.3) and use economic reasoning and common sense as guides for choosing the variables. Although this approach loses some of the richness of economic analysis, it is commonly and effectively applied by careful researchers.

Once an econometric model such as (1.3) or (1.4) has been specified, various *hypotheses* of interest can be stated in terms of the unknown parameters. For example, in equation (1.3), we might hypothesize that *wage*, the wage that can be earned in legal employment, has no effect on criminal behavior. In the context of this particular econometric model, the hypothesis is equivalent to $\beta_1 = 0$.

An empirical analysis, by definition, requires data. After data on the relevant variables have been collected, econometric methods are used to estimate the parameters in the econometric model and to formally test hypotheses of interest. In some cases, the econometric model is used to make predictions in either the testing of a theory or the study of a policy's impact.

Because data collection is so important in empirical work, Section 1-3 will describe the kinds of data that we are likely to encounter.

1-3 The Structure of Economic Data

Economic data sets come in a variety of types. Whereas some econometric methods can be applied with little or no modification to many different kinds of data sets, the special features of some data sets must be accounted for or should be exploited. We next describe the most important data structures encountered in applied work.

1-3a Cross-Sectional Data

A **cross-sectional data set** consists of a sample of individuals, households, firms, cities, states, countries, or a variety of other units, taken at a given point in time. Sometimes, the data on all units do not correspond to precisely the same time period. For example, several families may be surveyed during different weeks within a year. In a pure cross-sectional analysis, we would ignore any minor timing differences in collecting the data. If a set of families was surveyed during different weeks of the same year, we would still view this as a cross-sectional data set.

An important feature of cross-sectional data is that we can often assume that they have been obtained by **random sampling** from the underlying population. For example, if we obtain information on wages, education, experience, and other characteristics by randomly drawing 500 people from the working population, then we have a random sample from the population of all working people. Random sampling is the sampling scheme covered in introductory statistics courses, and it simplifies the analysis of cross-sectional data. A review of random sampling is contained in Math Refresher C.

Sometimes, random sampling is not appropriate as an assumption for analyzing cross-sectional data. For example, suppose we are interested in studying factors that influence the accumulation of family wealth. We could survey a random sample of families, but some families might refuse to report their wealth. If, for example, wealthier families are less likely to disclose their wealth, then the resulting sample on wealth is not a random sample from the population of all families. This is an illustration of a sample selection problem, an advanced topic that we will discuss in Chapter 17.

Another violation of random sampling occurs when we sample from units that are large relative to the population, particularly geographical units. The potential problem in such cases is that the population is not large enough to reasonably assume the observations are independent draws. For example, if we want to explain new business activity across states as a function of wage rates, energy prices, corporate and property tax rates, services provided, quality of the workforce, and other state characteristics, it is unlikely that business activities in states near one another are independent. It turns out that the econometric methods that we discuss do work in such situations, but they sometimes need to be refined. For the most part, we will ignore the intricacies that arise in analyzing such situations and treat these problems in a random sampling framework, even when it is not technically correct to do so.

Cross-sectional data are widely used in economics and other social sciences. In economics, the analysis of cross-sectional data is closely aligned with the applied microeconomics fields, such as labor economics, state and local public finance, industrial organization, urban economics, demography, and health economics. Data on individuals, households, firms, and cities at a given point in time are important for testing microeconomic hypotheses and evaluating economic policies.

The cross-sectional data used for econometric analysis can be represented and stored in computers. Table 1.1 contains, in abbreviated form, a cross-sectional data set on 526 working individuals

TABLE 1.1 A Cross-Sectional Data Set on Wages and Other Individual Characteristics

obsno	wage	educ	exper	female	married
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

for the year 1976. (This is a subset of the data in the file WAGE1.) The variables include *wage* (in dollars per hour), *educ* (years of education), *exper* (years of potential labor force experience), *female* (an indicator for gender), and *married* (marital status). These last two variables are binary (zero-one) in nature and serve to indicate qualitative features of the individual (the person is female or not; the person is married or not). We will have much to say about binary variables in Chapter 7 and beyond.

The variable *obsno* in Table 1.1 is the observation number assigned to each person in the sample. Unlike the other variables, it is not a characteristic of the individual. All econometrics and statistics software packages assign an observation number to each data unit. Intuition should tell you that, for data such as that in Table 1.1, it does not matter which person is labeled as observation 1, which person is called observation 2, and so on. The fact that the ordering of the data does not matter for econometric analysis is a key feature of cross-sectional data sets obtained from random sampling.

Different variables sometimes correspond to different time periods in cross-sectional data sets. For example, to determine the effects of government policies on long-term economic growth, economists have studied the relationship between growth in real per capita GDP over a certain period (say, 1960 to 1985) and variables determined in part by government policy in 1960 (government consumption as a percentage of GDP and adult secondary education rates). Such a data set might be represented as in Table 1.2, which constitutes part of the data set used in the study of cross-country growth rates by De Long and Summers (1991).

The variable *gpcrgdp* represents average growth in real per capita GDP over the period 1960 to 1985. The fact that *govcons60* (government consumption as a percentage of GDP) and *second60*

TABLE 1.2 A Data Set on Economic Growth Rates and Country Characteristics

obsno	country	gpcrgdp	govcons60	second60
1	Argentina	0.89	9	32
2	Austria	3.32	16	50
3	Belgium	2.56	13	69
4	Bolivia	1.24	18	12
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
61	Zimbabwe	2.30	17	6

(percentage of adult population with a secondary education) correspond to the year 1960, while *gpcrgdp* is the average growth over the period from 1960 to 1985, does not lead to any special problems in treating this information as a cross-sectional data set. The observations are listed alphabetically by country, but nothing about this ordering affects any subsequent analysis.

1-3b Time Series Data

A **time series data** set consists of observations on a variable or several variables over time. Examples of time series data include stock prices, money supply, consumer price index, GDP, annual homicide rates, and automobile sales figures. Because past events can influence future events and lags in behavior are prevalent in the social sciences, time is an important dimension in a time series data set. Unlike the arrangement of cross-sectional data, the chronological ordering of observations in a time series conveys potentially important information.

A key feature of time series data that makes them more difficult to analyze than cross-sectional data is that economic observations can rarely, if ever, be assumed to be independent across time. Most economic and other time series are related, often strongly related, to their recent histories. For example, knowing something about the GDP from last quarter tells us quite a bit about the likely range of the GDP during this quarter, because GDP tends to remain fairly stable from one quarter to the next. Although most econometric procedures can be used with both cross-sectional and time series data, more needs to be done in specifying econometric models for time series data before standard econometric methods can be justified. In addition, modifications and embellishments to standard econometric techniques have been developed to account for and exploit the dependent nature of economic time series and to address other issues, such as the fact that some economic variables tend to display clear trends over time.

Another feature of time series data that can require special attention is the **data frequency** at which the data are collected. In economics, the most common frequencies are daily, weekly, monthly, quarterly, and annually. Stock prices are recorded at daily intervals (excluding Saturday and Sunday). The money supply in the U.S. economy is reported weekly. Many macroeconomic series are tabulated monthly, including inflation and unemployment rates. Other macro series are recorded less frequently, such as every three months (every quarter). GDP is an important example of a quarterly series. Other time series, such as infant mortality rates for states in the United States, are available only on an annual basis.

Many weekly, monthly, and quarterly economic time series display a strong seasonal pattern, which can be an important factor in a time series analysis. For example, monthly data on housing starts differ across the months simply due to changing weather conditions. We will learn how to deal with seasonal time series in Chapter 10.

Table 1.3 contains a time series data set obtained from an article by Castillo-Freeman and Freeman (1992) on minimum wage effects in Puerto Rico. The earliest year in the data set is the first

TABLE 1.3 Minimum Wage, Unemployment, and Related Data for Puerto Rico

obsno	year	avgmin	avgcov	prunemp	prgnp
1	1950	0.20	20.1	15.4	878.7
2	1951	0.21	20.7	16.0	925.0
3	1952	0.23	22.6	14.8	1015.9
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
37	1986	3.35	58.1	18.9	4281.6
38	1987	3.35	58.2	16.8	4496.7

observation, and the most recent year available is the last observation. When econometric methods are used to analyze time series data, the data should be stored in chronological order.

The variable *avgmin* refers to the average minimum wage for the year, *avgcov* is the average coverage rate (the percentage of workers covered by the minimum wage law), *prunemp* is the unemployment rate, and *prgnp* is the gross national product, in millions of 1954 dollars. We will use these data later in a time series analysis of the effect of the minimum wage on employment.

1-3c Pooled Cross Sections

Some data sets have both cross-sectional and time series features. For example, suppose that two cross-sectional household surveys are taken in the United States, one in 1985 and one in 1990. In 1985, a random sample of households is surveyed for variables such as income, savings, family size, and so on. In 1990, a *new* random sample of households is taken using the same survey questions. To increase our sample size, we can form a **pooled cross section** by combining the two years.

Pooling cross sections from different years is often an effective way of analyzing the effects of a new government policy. The idea is to collect data from the years before and after a key policy change. As an example, consider the following data set on housing prices taken in 1993 and 1995, before and after a reduction in property taxes in 1994. Suppose we have data on 250 houses for 1993 and on 270 houses for 1995. One way to store such a data set is given in Table 1.4.

Observations 1 through 250 correspond to the houses sold in 1993, and observations 251 through 520 correspond to the 270 houses sold in 1995. Although the order in which we store the data turns out not to be crucial, keeping track of the year for each observation is usually very important. This is why we enter *year* as a separate variable.

A pooled cross section is analyzed much like a standard cross section, except that we often need to account for secular differences in the variables across the time. In fact, in addition to increasing the sample size, the point of a pooled cross-sectional analysis is often to see how a key relationship has changed over time.

TABLE 1.4 Pooled Cross Sections: Two Years of Housing Prices

obsno	year	hprice	proptax	sqrft	bdrms	bthrms
1	1993	85,500	42	1600	3	2.0
2	1993	67,300	36	1440	3	2.5
3	1993	134,000	38	2000	4	2.5
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
250	1993	243,600	41	2600	4	3.0
251	1995	65,000	16	1250	2	1.0
252	1995	182,400	20	2200	4	2.0
253	1995	97,500	15	1540	3	2.0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
520	1995	57,200	16	1100	2	1.5

1-3d Panel or Longitudinal Data

A **panel data** (or *longitudinal data*) set consists of a time series for *each* cross-sectional member in the data set. As an example, suppose we have wage, education, and employment history for a set of individuals followed over a 10-year period. Or we might collect information, such as investment and financial data, about the same set of firms over a five-year time period. Panel data can also be collected on geographical units. For example, we can collect data for the same set of counties in the United States on immigration flows, tax rates, wage rates, government expenditures, and so on, for the years 1980, 1985, and 1990.

The key feature of panel data that distinguishes them from a pooled cross section is that the *same* cross-sectional units (individuals, firms, or counties in the preceding examples) are followed over a given time period. The data in Table 1.4 are not considered a panel data set because the houses sold are likely to be different in 1993 and 1995; if there are any duplicates, the number is likely to be so small as to be unimportant. In contrast, Table 1.5 contains a two-year panel data set on crime and related statistics for 150 cities in the United States.

There are several interesting features in Table 1.5. First, each city has been given a number from 1 through 150. Which city we decide to call city 1, city 2, and so on, is irrelevant. As with a pure cross section, the ordering in the cross section of a panel data set does not matter. We could use the city name in place of a number, but it is often useful to have both.

A second point is that the two years of data for city 1 fill the first two rows or observations, observations 3 and 4 correspond to city 2, and so on. Because each of the 150 cities has two rows of data, any econometrics package will view this as 300 observations. This data set can be treated as a pooled cross section, where the same cities happen to show up in each year. But, as we will see in Chapters 13 and 14, we can also use the panel structure to analyze questions that cannot be answered by simply viewing this as a pooled cross section.

In organizing the observations in Table 1.5, we place the two years of data for each city adjacent to one another, with the first year coming before the second in all cases. For just about every practical purpose, this is the preferred way for ordering panel data sets. Contrast this organization with the way the pooled cross sections are stored in Table 1.4. In short, the reason for ordering panel data as in Table 1.5 is that we will need to perform data transformations for each city across the two years.

Because panel data require replication of the same units over time, panel data sets, especially those on individuals, households, and firms, are more difficult to obtain than pooled cross sections. Not surprisingly, observing the same units over time leads to several advantages over cross-sectional data or even pooled cross-sectional data. The benefit that we will focus on in this text is that having

TABLE 1.5 A Two-Year Panel Data Set on City Crime Statistics

obsno	city	year	murders	population	unem	police
1	1	1986	5	350,000	8.7	440
2	1	1990	8	359,200	7.2	471
3	2	1986	2	64,300	5.4	75
4	2	1990	1	65,100	5.5	75
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
297	149	1986	10	260,700	9.6	286
298	149	1990	6	245,000	9.8	334
299	150	1986	25	543,000	4.3	520
300	150	1990	32	546,200	5.2	493