

Introductory Econometrics

Jeffrey M. Wooldridge

THOMSON LEARNING
★
CUSTOM PUBLISHING



Introductory Econometrics

Jeffrey M. Wooldridge
Michigan State University

SOUTH-WESTERN
™
THOMSON LEARNING

COPYRIGHT © 2002 by South-Western College Publishing, a division of Thomson Learning Inc. Thomson Learning™ is a trademark used herein under license.

Printed in the United States of America

South-Western College Publishing
5191 Natorp Boulevard
Mason, OH 45040
USA

For information about our products, contact us:
Thomson Learning Academic Resource Center
1-800-423-0563
<http://www.swcollege.com>

International Headquarters
Thomson Learning
International Division
290 Harbor Drive, 2nd Floor
Stamford, CT 06902-7477
USA

UK/Europe/Middle East/South Africa
Thomson Learning
Berkshire House
168-173 High Holborn
London WC1V 7AA

Asia
Thomson Learning
60 Albert Street, #15-01
Albert Complex
Singapore 189969

Canada
Nelson Thomson Learning
1120 Birchmount Road
Toronto, Ontario M1K 5G4
Canada
United Kingdom

ALL RIGHTS RESERVED. No part of this work covered by the copyright hereon may be reproduced or used in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, Web distribution, or information storage and retrieval systems—without the written permission of the publisher.

ISBN 0-324-29473-5

The Adaptable Courseware Program consists of products and additions to existing South-Western College Publishing products that are produced from camera-ready copy. Peer review, class testing, and accuracy are primarily the responsibility of the author(s).

Custom Contents

Chapter 1	The Nature of Econometrics and Economic Data	1
	What is Econometrics?	1
	Steps in Empirical Economic Analysis	2
	The Structure of Economic Data	5
	Causality and the Notion of Ceteris Paribus in Econometric Analysis	13
	Summary	18
	Key Terms	19
Chapter 2	The Simple Regression Model	21
	Definition of the Simple Regression Model	21
	Deriving the Ordinary Least Squares Estimates	26
	Mechanics of OLS	34
	Units of Measurement and Functional Form	40
	Expected Values and Variances of the OLS Estimators	45
	Regression Through the Origin	57
	Summary	58
	Key Terms	59
	Problems	59
	Computer Exercises	62
	Appendix 2A	64
Chapter 3	Motivation for Multiple Regression	65

Appendix: Fundamentals of Probability 79

Random Variables and Their Probability Distributions	79
Joint Distributions, Conditional Distributions, and Independence	84
Features of Probability Distributions	87
Features of Joint and Conditional Distributions	95
The Normal and Related Distributions	103
Summary	110
Key Terms	111
Problems	112

Appendix: Fundamentals of Mathematical Statistics 115

Populations, Parameters, and Random Sampling	115
Finite Sample Properties of Estimators	116
Asymptotic or Large Sample Properties of Estimators	124
General Approaches to Parameter Estimation	129
Interval Estimation and Confidence Intervals	131
Hypothesis Testing	140
Remarks on Notation	152
Summary	153
Key Terms	153
Problems	154

Chapter 4 Mechanics and Interpretation of Ordinary Least Squares 161

Chapter 5 Multiple Regression Analysis: Inference 197

Sampling Distributions of the OLS Estimators 197

Testing Hypotheses About a Single Population Parameter: The t Test 200

Confidence Intervals 218

Testing Hypotheses About a Single Linear Combination of the
Parameters 220

Testing Multiple Linear Restrictions: The F Test 223

Reporting Regression Results 234

Summary 237

Key Terms 238

Problems 238

Computer Exercises 243

Chapter 6 Multiple Regression Analysis: Further Issues 247

Effects of Data Scaling on OLS Statistics 247

More on Functional Form 252

More on Goodness-of-Fit and Selection of Regressors 261

Prediction and Residual Analysis 266

Summary 274

Key Terms 274

Problems 275

Computer Exercises 277

Chapter 7 Multiple Regression Analysis with Qualitative Information: Binary (or Dummy) Variables 281

Describing Qualitative Information 281

A Single Dummy Independent Variable 283

Using Dummy Variables for Multiple Categories 290

Interactions Involving Dummy Variables 294

A Binary Dependent Variable: The Linear Probability Model 302

More on Policy Analysis And Program Evaluation	308
Summary	310
Key Terms	311
Problems	311
Computer Exercises	314

Chapter 8 Missing Data, Nonrandom Samples, and Outlying Observations 319

Appendix: Basic Mathematical Tools 331

The Summation Operator and Descriptive Statistics	331
Properties of Linear Functions	333
Proportions and Percentages	336
Some Special Functions and Their Properties	338
Differential Calculus	346
Summary	348
Key Terms	349
Problems	349

Appendix: Statistical Tables 353

1

The Nature of Econometrics and Economic Data

Chapter 1 discusses the scope of econometrics and raises general issues that result from the application of econometric methods. 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 the inference of 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 twenty-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. The most common application of econometrics is the forecasting of such important macroeconomic variables as interest rates, inflation rates, and gross domestic product. While forecasts of economic indicators are highly visible and are 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** 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 much more difficult to obtain in the social sciences. While 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.

E X A M P L E 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

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 where formal derivations provide insights that intuition can overlook.

Here is an example of an equation that was derived 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* is hourly wage, *educ* is years of formal education, *exper* is years of workforce experience, and *training* is weeks spent in job training. Again, other factors generally affect the wage rate, but (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**. Since 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:

$$crime = \beta_0 + \beta_1 wage_m + \beta_2 othinc + \beta_3 freqarr + \beta_4 freqconv + \beta_5 avgsen + \beta_6 age + u, \quad (1.3)$$

where *crime* is some measure of the frequency of criminal activity, *wage_m* is the wage that can be earned in legal employment, *othinc* is the income from other sources (assets, inheritance, etc.), *freqarr* is the frequency of arrests for prior infractions (to approximate the probability of arrest), *freqconv* is the frequency of conviction, and *avgsen* is 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 unob-

served 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 $\beta_0, \beta_1, \dots, \beta_6$ 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

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 training + u, \quad (1.4)$$

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. While 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 $wage_m$, 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. While 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.

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 section 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 Appendix 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 for the year 1976. (This is a subset of the data in the file WAGE1.RAW.) 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

Table 1.1

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

<i>obsno</i>	<i>wage</i>	<i>educ</i>	<i>exper</i>	<i>female</i>	<i>married</i>
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

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 one, which person is called Observation Two, 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, in order to determine the effects of government policies on long-term economic growth, economists have studied the relationship between growth in real per capita gross domestic product (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).

Table 1.2

A Data Set on Economic Growth Rates and Country Characteristics

<i>obsno</i>	<i>country</i>	<i>gpcrgdp</i>	<i>govcons60</i>	<i>second60</i>
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

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* (percent 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 order of the observations is listed alphabetically by country, but there is nothing about this ordering that affects any subsequent analysis.

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, gross domestic product, 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 it more difficult to analyze than cross-sectional data is the fact 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 gross domestic product from last quarter tells us quite a bit about the likely range of the GDP during this quarter, since GDP tends to remain fairly stable from one quarter to

the next. While 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 employment rates. Other macro series are recorded less frequently, such as every three months (every quarter). Gross domestic product 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 differs 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 observation, and the most recent year available is the last

Table 1.3

Minimum Wage, Unemployment, and Related Data for Puerto Rico

<i>obsno</i>	<i>year</i>	<i>avgmin</i>	<i>avgcov</i>	<i>unemp</i>	<i>gnp</i>
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. 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), *unemp* is the unemployment rate, and *gnp* is the gross national product. We will use these data later in a time series analysis of the effect of the minimum wage on employment.

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. In order to increase our sample size, we can form a **pooled cross section** by combining the two years. Because random samples are taken in each year, it would be a fluke if the same household appeared in the sample during both years. (The size of the sample is usually very small compared with the number of households in the United States.) This important factor distinguishes a pooled cross section from a panel data set.

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, when there was 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. While 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.

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 ten-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, etc., for the years 1980, 1985, and 1990.

The key feature of panel data that distinguishes it from a pooled cross section is the fact that the *same* cross-sectional units (individuals, firms, or counties in the above